UNIVERSITY OF GOTHENBURG

DEPARTMENT OF CHEMISTRY AND MOLECULAR BIOLOGY

Metabolic pathway analysis of bacteria associated with the marine diatom *Skeletonema marinoi*

Matt Pinder

Degree project for Master of Science (120 HEC) with a major in Molecular Biology with Specialization in Genomics and Systems Biology

2017

60 HEC project

Second Cycle

Images courtesy of Oskar Johansson,
Department of Biological and Environmental Sciences, University of Gothenburg

Contents

1. Abstract
2. Background
    - 2.1. Diatoms
    - 2.2. *Skeletonema marinoi*
    - 2.3. Interactions between diatoms and bacteria
    - 2.4. Identifying diatom-bacteria interactions
3. Materials and methods
    - 3.1. Raw data
    - 3.2. Assembly
        - 3.2.1. HGAP
        - 3.2.2. Falcon
        - 3.2.3. Canu
        - 3.2.4. Circularisation
        - 3.2.5. Assembly selection
    - 3.3. First annotation
    - 3.4. Phylotaxonomic analysis
    - 3.5. 16S rRNA analysis
    - 3.6. Additional classification analyses
    - 3.7. Second annotation
    - 3.8. GenBank consensus
    - 3.9. Pathway analysis
    - 3.10. Additional analyses
    - 3.11. NCBI submission
    - 3.12. Code written for this project
        - 3.12.1. Sequence_Reverse.py
        - 3.12.2. NCBI_Downloader.py
        - 3.12.3. GenBank_Consensus.py (and GenBank_Consensus_Names.py)
4. Results
    - 4.1. Assembly results
    - 4.2. Phylotaxonomic analysis and classification
        - 4.2.1. pb_359_2
        - 4.2.2. pb_359_3
        - 4.2.3. pb_359_4
        - 4.2.4. pb_359_5
        - 4.2.5. pb_359_6
        - 4.2.6. pb_359_7
        - 4.2.7. pb_359_8
    - 4.3. Annotation
    - 4.4. Pathway predictions and interesting genes
    - 4.5. Other observations
5. Discussion
    - 5.1. pb_359_2 - *Roseovarius mucosus* strain SMR3
    - 5.2. pb_359_3 - *Loktanella vestfoldensis* strain SMR4r
    - 5.3. pb_359_4 - *Sphingorhabdus flavimaris* strain SMR4y

5.4. pb_359_5 - *Marinobacter salarius* strain SMR5
5.5. pb_359_6 - *Sulfitobacter pseudonitzschiae* strain SMR1
5.6. pb_359_7 - *Antarctobacter heliothermus* strain SMS3
5.7. pb_359_8 - *Arenibacter algicola* strain SMS7
5.8. Conclusions
6. Acknowledgements
7. References

---

1. Abstract

Diatoms are known to associate with a number of bacterial species, forming an alga-bacteria holobiont - the host and its associated microbiome. The precise relationship between the diatom and an individual bacterial species is not always known, but as axenic growth of the diatom *Skeletonema marinoi* has proven difficult, it can be assumed that the role of bacteria is important. To that end, a collection of bacteria associated with *S. marinoi* strains RO5AC and ST54 were isolated and cultured, and their genomes sequenced using PacBio Single Molecule, Real-Time (SMRT) technology. The genomes were assembled and annotated, and were used to identify the diatom-associated species. Pathway analysis was then used to determine the potential relationships between host and bacteria. The bacteria were found to belong to classes which are often seen in such associations with diatoms, with most belonging to the Alphaproteobacteria. Examination of the annotations, along with pathway analysis, revealed a number of potential features which may explain the roles of these species in the diatom holobiont, primarily focused on nutrient exchange, and resistance to drugs and heavy metals.

---

2. Background

2.1. Diatoms

Diatoms are eukaryotic phytoplankton belonging to the class Bacillariophyceae, noted for their silica-composed cell wall, or frustule [1]. They can be broadly divided into four categories based on their morphology - radial and polar centrics, and raphid and araphid pennates - which have arisen at different times since the emergence of diatoms during the Mesozoic era [2].
Diatoms are important primary producers, performing around 20% of Earth's photosynthesis and sequestering atmospheric CO_2, thereby influencing Earth's climate; this sequestered carbon is locked away when the diatoms then sink to the ocean floor, which contributes to the creation of petroleum over geological time periods [2]. In addition to this natural utility, diatoms - particularly their frustules - are also being investigated for their potential biotechnological applications, such as drug delivery and filtration [1].
Another trait for which some diatom species are known is the production of harmful algal blooms (HABs). HABs occur when the population of certain algae (such as diatoms, dinoflagellates and cyanobacteria) explodes under favourable conditions, causing damage either through the production of toxins, or through problems associated with their sheer weight of numbers, such as preventing other species from obtaining sunlight or oxygen [3].

2.2. *Skeletonema marinoi*

Skeletonema marinoi is a centric, chain-forming diatom species found in locations around the world, particularly in temperate coastal waters [4, 5, 6]. It was originally described as part of a study to classify *Skeletonema* species which did not match the characteristics of the type species, *Skeletonema costatum* [4]. As with other diatoms, it is an important contributor to marine food webs, particularly in the North Atlantic, and is especially prevalent during spring blooms [2, 5].
Skeletonema marinoi is currently being developed into a model organism, on account of features such as a short (24h) generation time, relative ease of culturing, and its habit of entering a benthic resting stage in sediments [5]. These resting auxospores can be revived and cultured after up to a century in this state, allowing them to act as a 'snapshot' of the species' recent evolutionary history for genomic and phenotypic analysis, and comparison with contemporary samples [5]. To that end, an annotated reference genome is being generated (Töpel *et al.*, unpublished work), along with a knockout mutant library and methods for phenotyping the mutants generated (Johansson *et al.*, unpublished work).

2.3. Interactions between diatoms and bacteria

Diatoms and bacteria are known to interact with one another in a variety of ways within an alga-bacteria holobiont - the host and its associated bacteria; these interactions range from symbiotic exchange of nutrients to algicidal parasitism [7]. Multiple studies into the taxonomy of these associated bacteria have shown that the most commonly associated bacterial species come from the various classes of Proteobacteria, as well as the phylum Bacteroidetes [7, 8, 9, 10].
Nutrient exchange is one way in which bacteria and diatoms have been shown to benefit one another. In a direct sense, some bacteria produce cobalamin (vitamin B_{12}), which many species of algae require from external sources; in one study on this subject, it was found that the addition of algal extract to a vitamin B_{12}-producing bacterial culture increased both bacterial growth and vitamin B_{12} production, implying a mutually beneficial interaction [11]. In a more indirect sense, many bacteria also secrete siderophores - small organic molecules which bind iron (present at very low concentrations in marine environments) and make it more soluble, facilitating its uptake; many phytoplankton are able to access the iron bound to some of these siderophores, such as vibrioferrin [12]. In addition, some bacteria have been found to produce auxins - plant growth hormones - which also benefit the host alga [13].
In the opposite direction, some diatoms also produce nutrients useful to bacteria, one well-studied example being dimethylsulfoniopropionate (DMSP), which is known to be produced by *Skeletonema marinoi* [14]. This compound, used by marine algae for processes such as osmotic regulation, can be broken down by some bacteria and used as a sulphur and carbon source, and when degraded to dimethyl sulfide (DMS) is an important contributor to world climate [15].
In contrast to these beneficial interactions, some bacteria exhibit algicidal activity, which can be achieved either through direct contact with the alga, or through the production of algicidal compounds, including proteases and pigments [16, 17]. A specific example involving a member of the *Skeletonema* genus is the bacterium *Kordia algicida*, which is known to produce proteases that trigger cell lysis; there is some specificity in this interaction, as only three of the four diatom species tested in one study were affected by this protease, including *Skeletonema costatum* [17]. The release of these algicidal factors has been suggested to be controlled via quorum sensing, a signalling system whereby bacteria determine their population density using molecules called autoinducers, such as acyl homoserine lactones, and alter their gene expression accordingly [18]. Some diatoms are also able to produce antibacterial compounds in return, including *Skeletonema costatum*, which produces

compounds effective against bacteria of the genus *Vibrio* [19].
An additional bacterial feature suggestive of a relationship with eukaryotes is the presence of biosynthesis genes for phosphatidylcholine; this essential eukaryotic membrane phospholipid is only found in around 10% of bacteria, primarily those associated with eukaryotes, and its absence can have an adverse effect on symbiosis and pathogenicity [20].
Another indicator of how closely diatoms and bacteria interact is the level of horizontal gene transfer between the two groups. Specifically, genes of bacterial origin can be numerous in diatom genomes, including those involved in metabolic processes, cell wall synthesis and DNA-related processes; for example, around 7.5% of genes in the pennate diatom *Phaeodactylum tricornutum* appear to have been transferred from bacteria, a high level among eukaryotes [21].

2.4. Identifying diatom-bacteria interactions

One way of predicting the relationships between diatoms and their associated bacteria is through metabolic pathway analysis - determining which pathways may be present in the bacteria, based on their gene content, and identifying those pathways which might be indicative of diatom-bacteria interactions. One available piece of software for the prediction of pathways in an organism is Pathway Tools [22]. Using an annotated GenBank file, the program's PathoLogic component makes predictions about which pathways may be present in the organism, based on information from the MetaCyc database [23]. The software also gives users the ability to try and improve the program's predictions, such as by manually adding the EC numbers (which denote a particular function) of known enzymes based on experimental evidence, and using Pathway Tools' Pathway Hole Filler component, which uses "homology and pathway-based evidence" to try and fill holes in partially-predicted pathways [24].
Understanding the relationship between diatoms and their microbiota is important, as it could lead to the development of control measures for HABs. In human terms, such events have an impact on both health and the economy; toxic species can cause various forms of poisoning associated with shellfish, and one example of an algal bloom in Texas is thought to have caused losses of over $6 million in 2003 [3]. At least one species of *Skeletonema* - *S. costatum* - has been implicated in HABs [25].
As *Skeletonema marinoi* is currently being developed into a model organism, developing an understanding of its microbiome would help to enhance the model's usefulness, particularly considering the difficulties encountered when trying to grow the diatom axenically (Johansson, O.N., unpublished work). With a view to improving this understanding, this study aimed to assemble and annotate the genomes of a number of bacteria found in association with *S. marinoi*, discern their taxonomic placements, and form a hypothesis regarding their potential roles within the *S. marinoi* holobiome.

3. Materials and methods

3.1. Raw data

Eight bacterial samples were isolated from cultures of *Skeletonema marinoi* strains RO5AC and ST54, and cultured individually (Johansson, O.N., unpublished work). Sequencing was performed by SciLifeLab using the PacBio RSII protocol [26, 27], with each sample being run in a single SMRT (Single Molecule, Real-Time) sequencing cell; the resultant sequence data sets were designated pb_359_1 through pb_359_8. A preliminary assembly was performed by SciLifeLab using the HGAP.3 protocol (protocol explained in section 3.2.1).

3.2. Assembly

3.2.1. HGAP

The HGAP (Hierarchical Genome Assembly Process) assembly pipeline is a method designed for *de novo* genome assembly using only long reads, as opposed to hybrid methods using both long reads and shorter reads (such as those obtained from second-generation sequencing methods, e.g. the Illumina platforms) [28]. The assembler maps reads shorter than a user-defined 'minimum seed read length' to reads which are longer than this value, resulting in more accurate 'preassembled reads'; these are then assembled into contigs using the Celera assembler [29], followed by a correction step using the Quiver algorithm [28].
The SMRT Portal version 2.3.0 interface currently offers two different versions of HGAP - RS_HGAP_Assembly.2 (HGAP.2) and RS_HGAP_Assembly.3 (HGAP.3) [30]. According to Pacific Biosciences, the HGAP.2 protocol is optimised for quality [31], whereas the HGAP.3 protocol is optimised for speed [32]. The difference between the two lies in the main assembly module used; whereas HGAP.2 uses the **Celera Assembler**, HGAP.3 uses PacBio's **AssembleUnitig**, resulting in a ten-fold increase in assembly speed [32]. Despite this, as HGAP.2 is stated to be optimised for quality, this was the protocol chosen for use in this study.
Based on the minimum seed read length and sum of assembled contigs from SciLifeLab's preliminary assembly, new assemblies were performed using the HGAP.2 protocol via SMRT Portal [30, 31]; the minimum seed read length and genome size parameters of the initial HGAP.2 assembly attempt were informed by the preliminary assembly (all other parameters were unchanged from default values), with the parameters of each subsequent assembly attempt being informed by those prior. Several values of minimum seed read length were used in an attempt to minimise the number of polished contigs obtained in the assembly, while maximising overall assembly length.
Some common issues were found with many of the HGAP.2 assemblies, however; in particular, read coverage on at least one of the contigs would often be much higher or lower in places than would be expected of a correctly-assembled sequence, or attempts to circularise the contig would fail (see section 3.2.4). Therefore, two other *de novo* assemblers were also used to try and obtain better assemblies - Falcon and Canu.

3.2.2. Falcon

While it follows the same general method as HGAP, the Falcon assembler uses different components at each step in the pipeline, and is overall easier to experiment with in order to produce a better assembly [33, 34]; it was also recently used to assemble the gorilla genome [35]. One major advantage that Falcon possesses when compared to HGAP is that the ends of each assembled contig don't overlap, therefore no trimming is required prior to attempting circularisation (see section 3.2.4). However, Falcon lacks the Quiver correction step present in HGAP, which is instead achieved during the circularisation step. Although the Falcon assembler has a multitude of settings which can be changed, the main parameters adjusted in this study were length_cutoff and length_cutoff_pr; together, these are roughly analogous to the 'minimum seed read length' parameter in HGAP. As with the HGAP assembly process described above, several values were used for these parameters in order to obtain the best possible assembly. The version of Falcon used in this study was v1.7.5.

3.2.3. Canu

The Canu assembler is described by its creators as "a fork of the Celera Assembler designed for high-noise single-molecule sequencing", including the PacBio RS II sequences used in this study [36]. Canu's advantage over Falcon and Celera Assembler (the assembler component of the HGAP.2 protocol) is its improved ability to resolve repeat regions [36]. However, as with Falcon, Canu lacks the final Quiver correction step present in HGAP; the Canu manual recommends the use of Quiver for correcting PacBio-derived assemblies, which is implemented during the circularisation step (see section 3.2.4) [36]. The only parameter requiring user input in Canu is estimated genome size, which was informed by the results of the previously-attempted assemblies using HGAP and Falcon; several estimated values were used to try and obtain the best possible assembly. The version of Canu used in this study was v1.3.

3.2.4. Circularisation

Bacterial chromosomes and plasmids are generally circular, therefore before an assembly was considered finished, it had to be shown that the replicons could feasibly be circularised.
For sequences assembled using HGAP and Canu, a preparatory step was required to identify and trim overlaps at the ends of each contig. BLASTn was used to identify identical or near-identical regions at both ends of each contig [37]; for ease of interpretation, these results were visualised in BlastViewer version 2.2 [38]. If any such regions were found, one of them was removed before circularisation was attempted. The entire assembly was also checked for large areas of similarity between contigs; this was to rule out the possibility that multiple contigs were part of the same replicon, but had not been assembled together. Once any overlap had been trimmed, each sequence was cut down the middle and rejoined in the opposite order (as shown in Figure 1a), and used as a reference for SMRT Portal's RS_Resequencing.1 (Resequencing) protocol [39]; if the original reads mapped back to this reference without any major dips in average read coverage across the cut point, it was deemed reasonable to assume that the replicon was circular (examples of both unsuccessful and successful results are shown in Figures 1b and 1c, respectively). This protocol also provided an additional Quiver correction step, particularly important for the Falcon and Canu assemblies which lacked Quiver correction during the initial assembly. To automate the process of cutting and rejoining the contigs, a Python script - Sequence_Reverse.py - was written (see section 3.12.1).

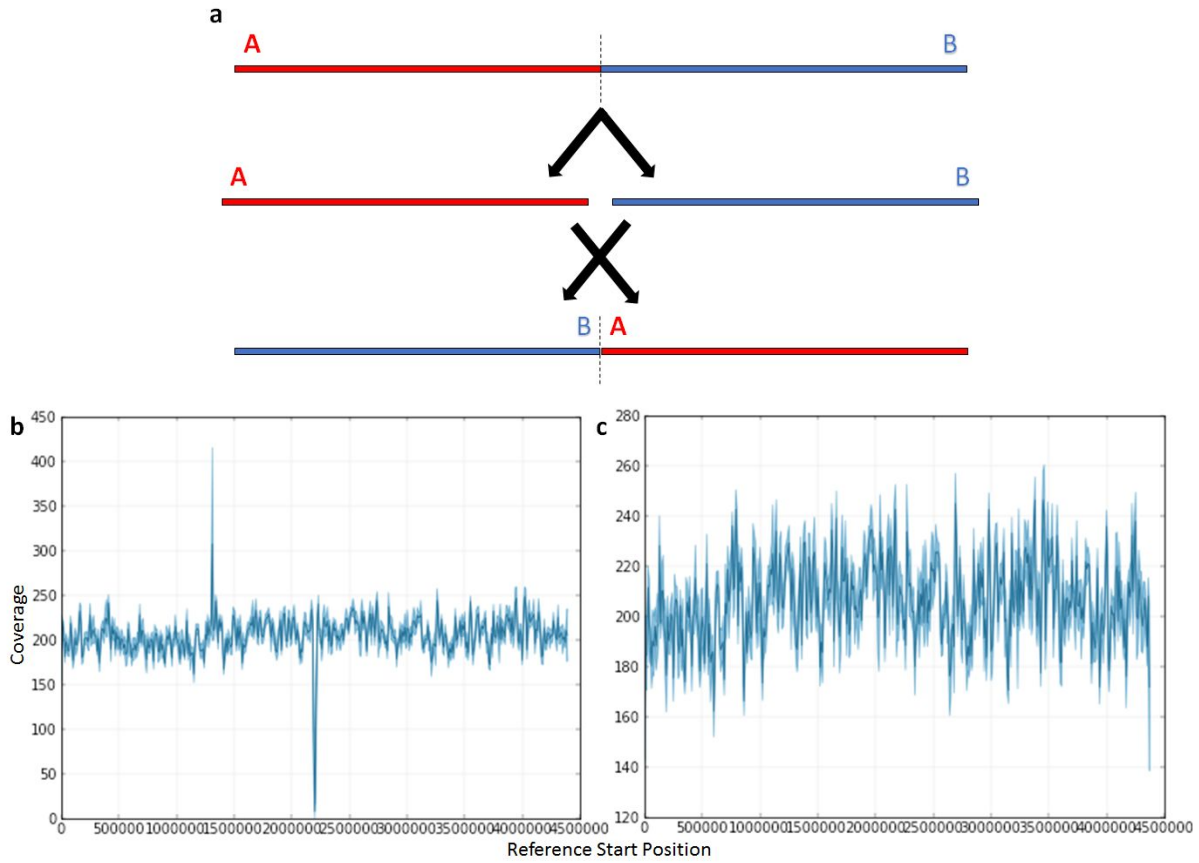

Figure 1: Testing whether contigs produced in an assembly attempt can be circularised.
a: Graphical representation of sequence preparation for the circularisation test.
For each assembly, each contig was cut in half, and rejoined in the opposite order; this was then used as a reference for mapping the original reads using SMRT Portal's RS_Resequencing.1 protocol [39].
b: Read coverage across a reference generated by reversal of an HGAP.2-derived assembly of pb_359_5, as demonstrated in Figure 1a. As patches of very irregular read coverage were found, in particular the trough at the central cut point, this assembly was rejected as it was unlikely to be circular. It should be noted that this assembly attempt produced a single contig, whereas the final pb_359_5 assembly contained two contigs.
c: Read coverage across a reference generated by reversal of a Canu-derived assembly of pb_359_5, as demonstrated in Figure 1a. As average read coverage across the contig is relatively consistent, this contig was considered circularisable. This figure represents the longer of the two contigs present in this assembly.
(Figures 1b and 1c generated in SMRT Portal [30])

3.2.5. Assembly selection

After attempting each of the above assemblers on all eight samples, using a variety of different parameter values, final assemblies were chosen based on a number of factors. Firstly, each contig had to circularise when run through the Resequencing protocol described in section 3.2.4, i.e. the coverage across the cut point had to be consistent with the rest of the contig. Secondly, the coverage across the whole contig also had to be reasonably consistent, particularly avoiding areas of very low coverage. Thirdly, the contigs had to be of a reasonable size, i.e. no contigs which were too small even to be plasmids. Lastly, if multiple assemblies fulfilled these criteria, the largest (in terms of number of base pairs, not number of contigs) was generally preferred, to minimise the risk of genomic information being lost. The assembler settings used for the final assemblies of each sample are shown in Table 1.

Table 1: Final assembler settings used for each genome assembly. Parameters not listed were left at their default values. length_cutoff and length_cutoff_pr are roughly analogous to 'minimum seed read length' in HGAP. genomeSize is an approximation of genome size.		
Sample	**Assembler used**	**Changes to default settings**
pb_359_2	Falcon v1.7.5	length_cutoff 7000; length_cutoff_pr 7000
pb_359_3	Falcon v1.7.5	length_cutoff 17100; length_cutoff_pr 17100
pb_359_4	Canu v1.3	genomeSize 4.5m
pb_359_5	Canu v1.3	genomeSize 4.5m
pb_359_6	Falcon v1.7.5	length_cutoff 10600; length_cutoff_pr 10600
pb_359_7	Canu v1.3	genomeSize 5.4m
pb_359_8	Falcon v1.7.5	length_cutoff 17000; length_cutoff_pr 17000
Note: sample pb_359_1 was suspected of being from the same species as pb_359_6 during early assembly attempts, and has been excluded from this table as a final assembly was never produced.		

3.3. First annotation

Based on the final assemblies, a preliminary annotation was performed using the prokaryotic genome annotation software Prokka version 1.12-beta [40], with the default UniProt database updated to the September 2016 release [41], and with the additional inclusion of the Pfam protein families database (version 30.0) [42]. Prokka has a variety of dependencies, each used to predict the location and identity of genetic elements; these are listed in Table 2.

The goal of this first annotation was to predict protein-coding genes and generate an amino acid fasta file as input for the microbial phylogenetic placement program PhyloPhlAn (see section 3.4), as well as to predict 16S ribosomal RNA (rRNA) sequences for additional phylogenetic analysis (see section 3.5); as the gene prediction algorithm Prodigal and the rRNA prediction program RNAmmer were used for both the first and second annotation runs (as core dependencies of Prokka), the loci identified in both runs will be the same irrespective of their predicted products. Therefore, the annotation was run in two steps. Firstly, Prokka was run using default settings (with the addition of the UniProt and Pfam databases mentioned above):

```
$ prokka --outdir pb_359_X --prefix pb_359_X input.fasta
```

Secondly, once a taxonomic identity was established, a more comprehensive annotation was obtained by providing Prokka with additional, taxonomic group-specific databases, as described in section 3.7.

Table 2: List of Prokka dependencies and their respective functions.		
Dependency	**Function**	**Notes**
Aragorn (version 1.2) [43]	Predicts tRNAs and tmRNAs	-
RNAmmer (version 1.2) [44]	Predicts rRNAs	-
Infernal cmscan (version 1.1) [45]	Predicts ncRNAs	-
Prodigal (version 2.6) [46]	Predicts prokaryotic genes	-
BLASTp (version 2.4) [37]	Annotates proteins based on sequence similarity to known proteins	Uses Prokka's default databases and those provided by the user
HMMER (version 3.1) [47]	Annotates proteins based on similarity to Hidden Markov Models (HMMs) [48]	Uses Prokka's default databases and those provided by the user

3.4. Phylotaxonomic analysis

PhyloPhlAn version 0.99 works by comparing informative amino acid positions in 400 conserved protein sequences to generate a phylogenetic tree [49]. The initial tree in this study was produced using the program's -i option, which integrates user-specified input organisms into a tree comprising the program's default collection of 3,171 bacterial genomes, using an amino acid fasta file for each input organism:

```
$ ./phylophlan.py -i my_genomes_directory --nproc n
```

The branch labels initially show ID numbers rather than taxonomic names, so the following command must be run to amend this [50]:

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

The resulting tree was then visualised using FigTree version 1.4.3 [51]. Once the closest one or two families had been identified for each species, amino acid fasta (.faa) files for all available species in these families were obtained from NCBI's ftp site [52], using a Python script written to automate this process - NCBI_Downloader.py (see section 3.12.2). The named species downloaded this way (i.e. those not labelled as *Genus* sp.) were then put into a PhyloPhlAn-generated tree with the relevant pb_359 species, using the -u option, in order to determine their taxonomic identities more accurately:

```
$ ./phylophlan.py -u my_genomes_directory
```

Trees were initially constructed to place the species within their respective families, with a pair of species from a different family of the same order acting as an outgroup. However, in some cases the families were only represented by a small number of genera within NCBI (i.e. few .faa files were available for species in the genus); in such cases the trees were expanded to include all available named species in the order, with the outgroup coming from a different order in the same class.

3.5. 16S rRNA analysis

16S rRNA has long been used as a measure of relatedness between bacterial species, given that it is ubiquitous among living things and the sequence changes slowly over time [53]; although within-species 16S rRNA sequence variability does exist, one study found an average 99.30±1.38% 16S rRNA similarity within bacterial species, and 95.56±3.68% similarity between species in the same genus [54]. On account of this useful property, 16S rRNA comparisons were used to complement the results of PhyloPhlAn in determining the taxonomy of the species in this study. 16S rRNA predictions were made by the RNAmmer version 1.2 rRNA predictor used by Prokka [44]. The resulting predictions were used as queries for BLASTn [37]- searching the website's default nr/nt database, as well as the databases for reference genome sequences (refseq_genomic) and bacterial/archaeal 16S rRNA sequences - in order to find the highest-scoring hits in terms of query cover and identity. If necessary, the ends of these hits were trimmed so that they would begin and end in the same places as the original query (or vice versa if the original query was longer), in order to obtain a percentage similarity between the two sequences using the EMBOSS Needle global nucleotide aligner [55].

3.6. Additional classification analyses

In a majority of cases, there was still some uncertainty as to the identity of the species being investigated following the PhyloPhlAn and 16S rRNA analyses. In these cases, some additional checks were carried out in order to enable a more confident identification of the species.
In addition to 16S rRNA, other marker genes are sometimes used to assess relatedness between species. The sequences of three such genes - the housekeeping genes *gyrB*, *rpoB* and *rpoD* - were used to compare each of this study's species with suspected relatives [56, 57]. Where possible, the nucleotide sequences for these species were downloaded from NCBI [58], and then aligned with those from the species being studied using EMBOSS Needle to provide a similarity score [55].
Where genomes exist for species presumed to be related to those being studied, the features of these genomes were compared - specifically, the genome size, G+C content and number of protein-coding genes were considered. Other characteristics which were also compared include cell morphology, colour, and fatty acid composition (Johansson, O.N., unpublished work), as in some instances these differed between species within a genus and provided a useful way to differentiate between otherwise highly-similar species.

3.7. Second annotation

Informed by the results of the phylotaxonomic and 16S rRNA analyses described above, additional information was included in Prokka's databases, and more of Prokka's options were used, to facilitate more family/genus-specific annotation; each of the points below resulted in the generation of a different version of the annotation.
Firstly, Prokka's `--proteins` option was used to specify a family-specific, non-hypothetical protein list for each species; this option allows the user to specify a trusted list of proteins for Prokka to check against first, before checking its other databases to annotate loci. The list was obtained by downloading GenBank (.gbff) files from NCBI's ftp site for the relevant family using the NCBI_Downloader.py script (see section 3.12.2) [52], then concatenating and converting them with Prokka's prokka-genbank_to_fasta_db script:

```
$ prokka-genbank_to_fasta_db *.gbff > fam1.fasta
$ fasta2tab fam1.fasta | grep -v "hypothetical protein" | tab2fasta > family.fasta
$ prokka --outdir pb_359_X --prefix pb_359_X --proteins family.fasta input.fasta
```

Secondly, the `--proteins` option was used as above, but proteins without an associated gene name were removed from the list, in an attempt to increase the number of named genes identified:

```
$ prokka-genbank_to_fasta_db *.gbff > fam1.fasta
$ fasta2tab fam1.fasta | grep -v "~~~~~~" | tab2fasta > family_named.fasta
$ prokka --outdir pb_359_X --prefix pb_359_X --proteins family_named.fasta
input.fasta
```

Thirdly, the full non-hypothetical protein list (named and unnamed) was merged with the default bacterial sprot database, in an attempt to prevent the acceptance of a poor hit over a better one:

```
$ cat /db/prokka/kingdom/Bacteria/sprot family.fasta >> sprot
$ mv sprot /db/prokka/kingdom/Bacteria/sprot
$ prokka --setupdb
$ prokka --outdir pb_359_X --prefix pb_359_X input.fasta
```

Fourthly, the list of named proteins was merged with the default bacterial sprot database, in an attempt to prevent the acceptance of a poor hit over a better one, while identifying more named genes:

```
$ cat /db/prokka/kingdom/Bacteria/sprot family_named.fasta >> sprot
$ mv sprot /db/prokka/kingdom/Bacteria/sprot
$ prokka --setupdb
$ prokka --outdir pb_359_X --prefix pb_359_X input.fasta
```

Finally, the named protein lists from all relevant families (Rhodobacteraceae, Sphingomonadaceae, Erythrobacteraceae, Alteromonadaceae and Flavobacteriaceae) were concatenated and specified using the `--proteins` option, so that all annotation attempts had access to the same information:

```
$ cat family_named.fasta family_named_2.fasta … family_named_X >> family_all.fasta
$ prokka --outdir pb_359_X --prefix pb_359_X --proteins family_all.fasta
input.fasta
```

3.8. GenBank consensus

As various combinations of options and databases were used to annotate the genomes (as detailed in section 3.7), in some cases a gene predicted to code for a ‘hypothetical protein’ in one version of the annotation was predicted to code for a known protein in another version, with the reverse being true at a different locus; therefore, it was decided that it would be useful to be able to combine all of these GenBank files to maximise informative gene predictions and minimise ‘hypothetical protein’ annotations. To that end, the Python script GenBank_Consensus.py was written to overwrite hypothetical entries in a GenBank with equivalent, informative entries from another annotation of the same sequence (see section 3.12.3). An alternative version of the script, GenBank_Consensus_Names.py, was also written in order to overwrite unnamed genes with named ones, in case this proved to be more informative in terms of the pathways predicted in the pathway analysis step (see section 3.9).

3.9. Pathway analysis

After annotation, the final GenBank files were run through Pathway Tools version 20.5 to obtain predictions for the pathways present in each organism [22]. Depending on the annotation method used (see sections 3.7 and 3.8), the number of predicted pathways differed, so the annotation containing the highest number of predicted pathways was selected in order to gain as much information as possible. To try and maximise the number of predictions, in cases where the PathoLogic module could not allocate a specific function to a protein, the UniProt Knowledgebase (UniProtKB) was queried with the gene name (or the protein product if a name was not present) [41]; if a seemingly-reliable EC number was found among the results (either associated with a manually-curated result or a large number of uncurated ones), this was added to the gene's record in PathoLogic. As Pathway Tools' predictions are informed by EC numbers, this can enable the program to fill more pathway holes.
Once this manual assignment was completed, another of PathoLogic's functions - Pathway Hole Filler - was used [24]. In each case, all of the organism's pathways were selected as training data, and if the probability of a correct hit was found to be greater than 0.9, then the highest-ranked candidate was chosen to fill the gap. When pathways were rescored following manual assignment of probable enzymes and pathway hole filling, pathways for which less than half of the necessary enzymes were predicted were deleted. Considering that a number of hypothetical proteins still remain in the annotations, incomplete pathways with a majority of elements present were still noted, with the rationale that it is easier to later remove a false-positive result than to retrieve a true result which has been mistakenly discarded.

3.10. Additional analyses

In addition to searching for metabolic pathways using Pathway Tools, gene clusters associated with secondary metabolite biosynthesis were also sought using the web tool antiSMASH v3.0.5, which uses multiple other tools in order to identify and analyse these clusters based on known and putative examples [59].
As well as analysing the bacterial genomes, the presence of potential prophage sequences was also assessed using the web tool PHASTER [60], an upgrade of the older prophage search tool PHAST [61]. Based on the size of a potential prophage region and the number of known phage genes within it, the region is given a score and labelled as 'intact', 'incomplete' or 'questionable' [61].

3.11. NCBI submission

Three of the samples examined in this study - pb_359_2, pb_359_3 and pb_359_5 - were submitted to NCBI prior to the completion of this report [58]. As part of the submission process, NCBI staff run additional checks and corrections on the annotations before they are made available to the public, therefore a small number of changes were made from the Prokka annotations (for example, no pseudogenes were predicted by Prokka, whereas several were identified by NCBI). The pathway analysis described in section 3.9 was subsequently rerun using these updated annotations.

3.12. Code written for this project

Several Python scripts were written during this study in order to facilitate certain tasks. These are all available in this study's GitHub repository [62].

3.12.1. Sequence_Reverse.py [63]

In order to test whether the contigs generated by the genome assemblers were feasibly circular, and therefore more likely to be true bacterial chromosomes or plasmids, the Sequence_Reverse.py script was written to cut each sequence in half and rejoin the two halves in the opposite configuration, as shown in Figure 1a. These sequences were then used as references for SMRT Portal's Resequencing protocol [39], and average read coverage across the cut point was used to determine whether a contig was likely to be circular, as described in section 3.2.4, and shown in Figures 1b and 1c.

3.12.2. NCBI_Downloader.py [64]

The NCBI ftp site contains an abundance of data [52], including fasta and GenBank files, which proved useful in annotating the genomes assembled in this study and determining each species' taxonomic placement; given the volume of data on the site, a script that automates the process of downloading the required files is essential. The file structure used on the ftp site is somewhat complex, however, and previous scripts written to automate the download process (even a script written as recently as 2015 [65]) no longer work once the file structure is changed.
With the current file structure in mind, and heavily inspired by this previous script, the NCBI_Downloader.py script was written; it searches for all species on the site which match a user-specified list of genera, downloads the compressed files of the specified format for these species, decompresses them and renames each file to match the relevant species name, rather than using NCBI's identifier. The script has been used to download .faa (amino acid fasta) and .gbff (GenBank) files so far, but with minor refinements could potentially be used to download other file types.

3.12.3. GenBank_Consensus.py (and GenBank_Consensus_Names.py) [66, 67]

A variety of different options were used in Prokka when annotating each genome; however, giving the program access to more databases did not always result in a strictly better annotation. For example, specifying one list of trusted proteins with the `--proteins` option would yield a functional annotation for gene X, whereas using a different list would give better results (i.e. less 'hypothetical protein' predictions) overall, but would result in gene X being labelled as producing a 'hypothetical protein'.
To try and resolve this problem, the GenBank_Consensus.py script was written; it copies each record from a primary GenBank file into an output file, but when it finds a gene with 'hypothetical protein' as its product, it checks against the equivalent record in a secondary GenBank file. If the entry in the secondary file is not hypothetical, this entry is copied to the output file in place of its counterpart from the primary file. This script can be run using several input files sequentially, to incorporate results from multiple GenBank files and minimise the number of hypothetical proteins predicted.
In addition, an alternative version of the script - GenBank_Consensus_Names.py - was written to maximise the number of named gene predictions; as well as overwriting hypothetical protein predictions with non-hypothetical ones, gene predictions without an associated gene name are overwritten by a corresponding named entry.

4. Results

4.1. Assembly results

When examining early assembly attempts, it was discovered that pb_359_1 and pb_359_6 were most likely from the same species; with one exception, all contigs differed in size from their counterpart by less than 0.1% (the remaining contig differed by less than 0.11%), and querying each assembly against the other using BLASTn showed that the sequence identity between each contig pair was ≥ 99.8% [37]. In addition, when portions of each contig were used as BLASTx queries [37], each contig's results matched its counterpart's. It was decided that only one of the two samples should be investigated further, and therefore pb_359_1 was disregarded in favour of pb_359_6, as circularised contigs could only be obtained for the latter. The results of all other assemblies are shown in Table 3.

pb_359_6 was found to have a very high number of suspected plasmids (seven, compared to three or fewer in all other samples); despite the use of all three assemblers with a variety of parameters, this number could not be lowered. In addition, they were all found to be circularisable, and attempts to combine contigs failed to yield circularisable results, therefore the high number of plasmids appears to be correct. There is a precedent for a *Sulfitobacter* species to have a high number of large plasmids - *Sulfitobacter* sp. AM1-D1 (assembly accession no. GCA_001886735.1) has five plasmids, three of which classify as megaplasmids (plasmids larger than 100kb [68]).

Table 3: Assembly sizes (in bp) and G+C content figures for the final assemblies of the *Skeletonema marinoi*-associated bacteria examined in this study.
Per-contig column **1** refers to the chromosome, and columns **2-8** refer to plasmids.

Sample	Assembly size and average G+C content	Per-contig size and G+C content (from longest to shortest contig)							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
pb_359_2	**4,381,426 60.9%**	4,170,996 60.9%	180,135 60.1%	30,295 58.3%	-	-	-	-	-
pb_359_3	**3,987,360 60.7%**	3,836,950 60.8%	111,030 57.4%	39,380 58.2%	-	-	-	-	-
pb_359_4	**3,479,724 58.0%**	3,479,724 58.0%	-	-	-	-	-	-	-
pb_359_5	**4,630,160 57.0%**	4,386,892 57.2%	243,268 53.7%	-	-	-	-	-	-
pb_359_6	**5,121,602 59.5%**	3,572,445 59.9%	428,095 58.4%	292,917 58.8%	284,777 58.4%	209,222 56.8%	142,107 60.3%	99,245 60.7%	92,794 58.9%
pb_359_7	**5,331,190 61.5%**	4,723,013 61.6%	372,263 60.3%	154,467 62.8%	81,447 60.4%	-	-	-	-
pb_359_8	**5,857,781 39.8%**	5,793,053 39.8%	64,728 43.8%	-	-	-	-	-	-

Note: sample pb_359_1 was suspected of being from the same species as pb_359_6 during early assembly attempts, and has been excluded from this table as a final assembly was never produced.

4.2. Phylotaxonomic analysis and classification

The full initial tree, containing the seven pb_359 samples along with the 3,171 species included by PhyloPhlAn, can be found in this study's GitHub repository [69]. The trees included in this section were subsampled by including only the group of clades nearest to the pb_359 sample(s), and clades which exclusively contain all or most of a single genus were compressed; complete trees for all samples can also be found in this study's GitHub repository [69]. As four of the seven samples were found to belong to the same family - Rhodobacteraceae - these are all presented in the same tree (Figure 2); each other sample has its own tree. The full predicted classification for each sample is shown in Table 4.

4.2.1. pb_359_2

pb_359_2 appears in a clade composed exclusively of species in the genus *Roseovarius*, its closest relative being *Roseovarius mucosus* (Figure 2) (the *R. mucosus* genome used in this analysis [type strain DSM 17069^T, accession no. NZ_AONH00000000.1] is currently listed as 'suppressed' on NCBI, but was still included in this analysis due to its proximity to pb_359_2 in earlier versions of this analysis). This result is backed up by the 16S rRNA analysis, as the 16S rRNA sequence of the type strain *R. mucosus* DFL-24^T (accession no. NR_042159.1) was found to have 99.8% identity to the two 16S rRNAs of pb_359_2 (the three 16S rRNAs of the unnamed species *Roseovarius* sp. TM1035 [accession no. NZ_ABCL00000000.1] are identical to those of pb_359_2). Therefore, I consider it most likely that pb_359_2 is a member of the genus *Roseovarius*; based on the high 16S rRNA sequence similarity, it is likely a strain of *Roseovarius mucosus*. This conclusion is supported by nucleotide sequence similarity in *gyrB*, *rpoB* and *rpoD* (>92%), similar G+C content (within 2% of the type strain), and a similar genome size (difference of around 0.14 Mbp); cell shape, colour and lipid composition are also similar (data not shown), but these appear to be general traits of multiple species in the genus [70].

4.2.2. pb_359_3

pb_359_3 appears in a clade composed exclusively of species in the genus *Loktanella*, its closest relative being *Loktanella vestfoldensis* (Figure 2). This result is backed up by the 16S rRNA analysis, as the three 16S rRNA sequences of the type strain *L. vestfoldensis* DSM 16212^T (accession no. NZ_ARNL00000000.1) were found to have 99.8% identity to the two 16S rRNA sequences of pb_359_3. Therefore, I consider it most likely that pb_359_3 is a member of the genus *Loktanella*; based on the high 16S rRNA sequence similarity, it is likely a strain of *Loktanella vestfoldensis*. This conclusion is supported by nucleotide sequence similarity in *gyrB*, *rpoB* and *rpoD* (>89%, comparable to the difference between *L. vestfoldensis* strains DSM 16212^T and SKA53 [accession no. NZ_AAMS00000000.1]), similar G+C content (within 3% of the type strain), a similar genome size (difference of around 0.27 Mbp), and a similar colouration; cell shape and lipid composition are also similar (data not shown), but these appear to be general traits of multiple species in the genus [71].

4.2.3. pb_359_4

pb_359_4 appears as sister species to *Sphingorhabdus marina* (classified as *Sphingopyxis marina* prior to 2013 [72]) (Figure 3). However, the most closely-related 16S rRNA sequence to be found was from *Sphingopyxis flavimaris* strain R-36742 (accession no. FR691421.1), with 99.9% identity to the

two 16S rRNA sequences of pb_359_4; in contrast, the similarity between the 16S rRNA sequences of pb_359_4 and the *Sphingorhabdus marina* type strain DSM 22363^T (accession no. NZ_FSQW00000000.1) is only 97.6%. *Sphingopyxis flavimaris* was reclassified as *Sphingorhabdus flavimaris* in 2013, along with reclassification of *Sphingorhabdus litoris* (formerly *Sphingopyxis litoris*) and the aforementioned *Sphingorhabdus marina* [72] (neither *S. flavimaris* nor *S. litoris* appear in Figure 3). Given these results, I consider it most likely that pb_359_4 is a member of the genus *Sphingorhabdus*; based on the high 16S rRNA sequence similarity, it is likely a strain of *Sphingorhabdus flavimaris*. This conclusion is supported by an identical G+C content [73], although as the genome of *S. flavimaris* has not previously been sequenced, other marker genes and genome traits cannot be compared. Cell shape, colour and lipid composition were also similar (data not shown), but these appear to be general traits of multiple species in the genus [72, 73].

Figure 2: Outgroup-rooted, subsampled phylogenetic tree showing the positions of pb_359_2, pb_359_3, pb_359_6 and pb_359_7 in the family Rhodobacteraceae (order Rhodobacterales); two members of the genus *Hyphomonas* (family Hyphomonadaceae; order Rhodobacterales) act as the outgroup. Node labels represent bootstrapping support. Tree generated by PhyloPhlAn version 0.99 [49] and visualised in FigTree version 1.4.3 [51].

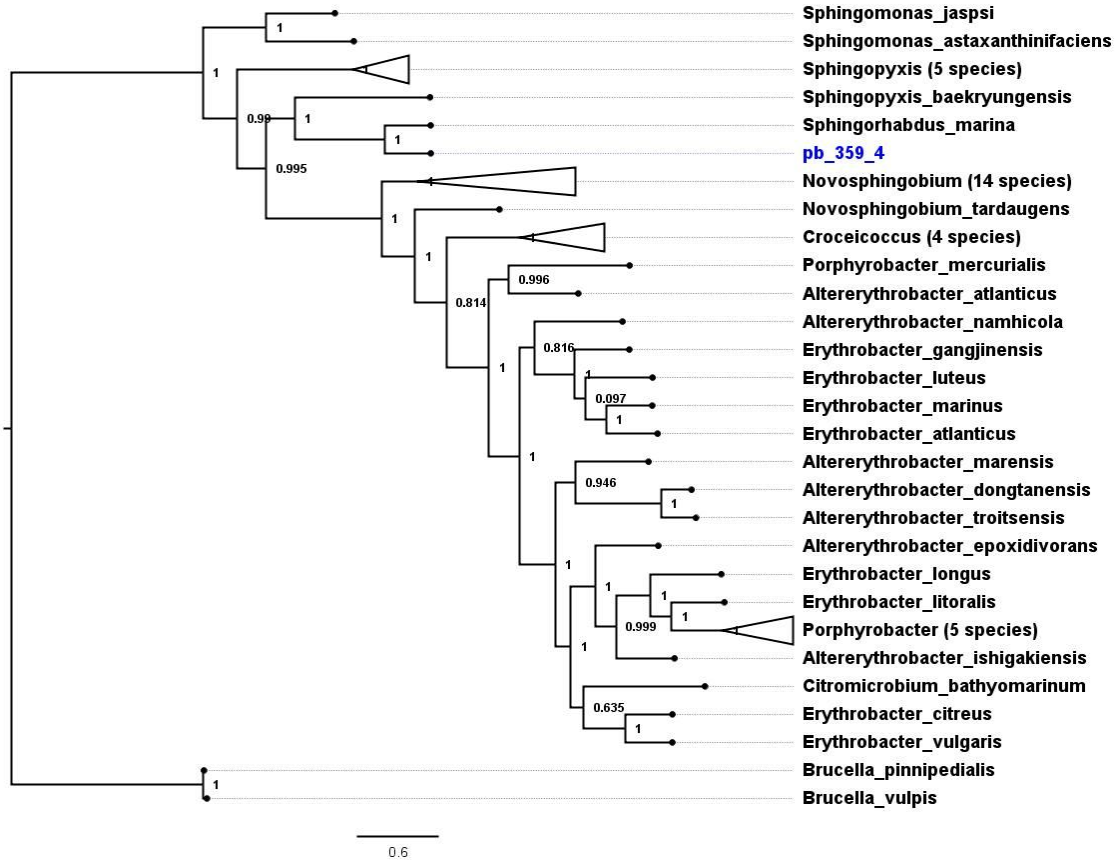

Figure 3: Outgroup-rooted, subsampled phylogenetic tree showing the position of pb_359_4 in the order Sphingomonadales (class Alphaproteobacteria); two members of the genus *Brucella* (order Rhizobiales; class Alphaproteobacteria) act as the outgroup. Node labels represent bootstrapping support. Tree generated by PhyloPhlAn version 0.99 [49] and visualised in FigTree version 1.4.3 [51].

4.2.4. pb_359_5

pb_359_5 was found in a clade together with *Marinobacter algicola* and *Marinobacter salarius* (Figure 4). However, depending on which species were included in the phylogenetic analysis (that is, which of the more distantly related clades of Alteromonadales were included), pb_359_5 would either appear as sister species to *M. algicola* (not shown), *M. salarius* (as in Figure 4), or as sister to the clade of *M. algicola* and *M. salarius* (as in the full tree). While the 16S rRNA sequence of the *M. salarius* type strain R9SW1^T (accession no. CP007152.1) does show high similarity to the three 16S rRNA sequences of pb_359_5 (99.4%/99.5% identity), the 16S rRNA sequence of the *M. algicola* type strain DG893^T (accession no. NZ_ABCP01000031.1) is even more similar, sharing 99.8%/99.9% identity with the 16S rRNA sequences of pb_359_5 (two of the three 16S rRNA sequences in pb_359_5 differ from the other by a single base substitution); the 16S rRNA similarity between *M. salarius* and *M. algicola* is mentioned in the paper first describing *M. salarius* [74]. In addition, several unnamed species of *Marinobacter* - *M.* sp. R-28770, *M.* sp. R-28768 and *M.* sp. C18 (accession nos. AM944523.1, AM944524.1 and NZ_LQXJ01000037.1, respectively) - share 99.9%/100% 16S rRNA sequence similarity with pb_359_5.

Considering the similarity to both *M. algicola* and *M. salarius* in terms of both the PhyloPhlAn and 16S rRNA analyses, the classification decision was determined by other factors. The G+C contents of all three bacteria are within 0.1% of each other, and they all have a similar colour and cell shape (data not shown). The genome size difference between pb_359_5 and the *M. salarius* type strain is 0.01 Mbp, whereas between pb_359_5 and the *M. algicola* type strain the difference is 0.22 Mbp. The number of protein-coding genes in pb_359_5 is much closer to *M. algicola* (difference of 293 genes) than to *M. salarius* (difference of 1290 genes); however, *M. salarius* has a large number of pseudogenes, and if these are also taken into account, then the difference between pb_359_5 and *M. salarius* falls to 55 genes, whereas the difference between pb_359_5 and *M. algicola* only falls to 230 genes.

The most compelling evidence in deciding the species identity of pb_359_5 came from the marker genes *gyrB*, *rpoB* and *rpoD*; the nucleotide sequence similarity of these genes between pb_359_5 and *M. algicola* is between 94.1% and 95.6%, whereas the similarity between pb_359_5 and *M. salarius* is between 99.5% and 99.8%.

Based on all of the evidence, I consider it most likely that pb_359_5 is a member of the genus *Marinobacter*, and a strain of *Marinobacter salarius*.

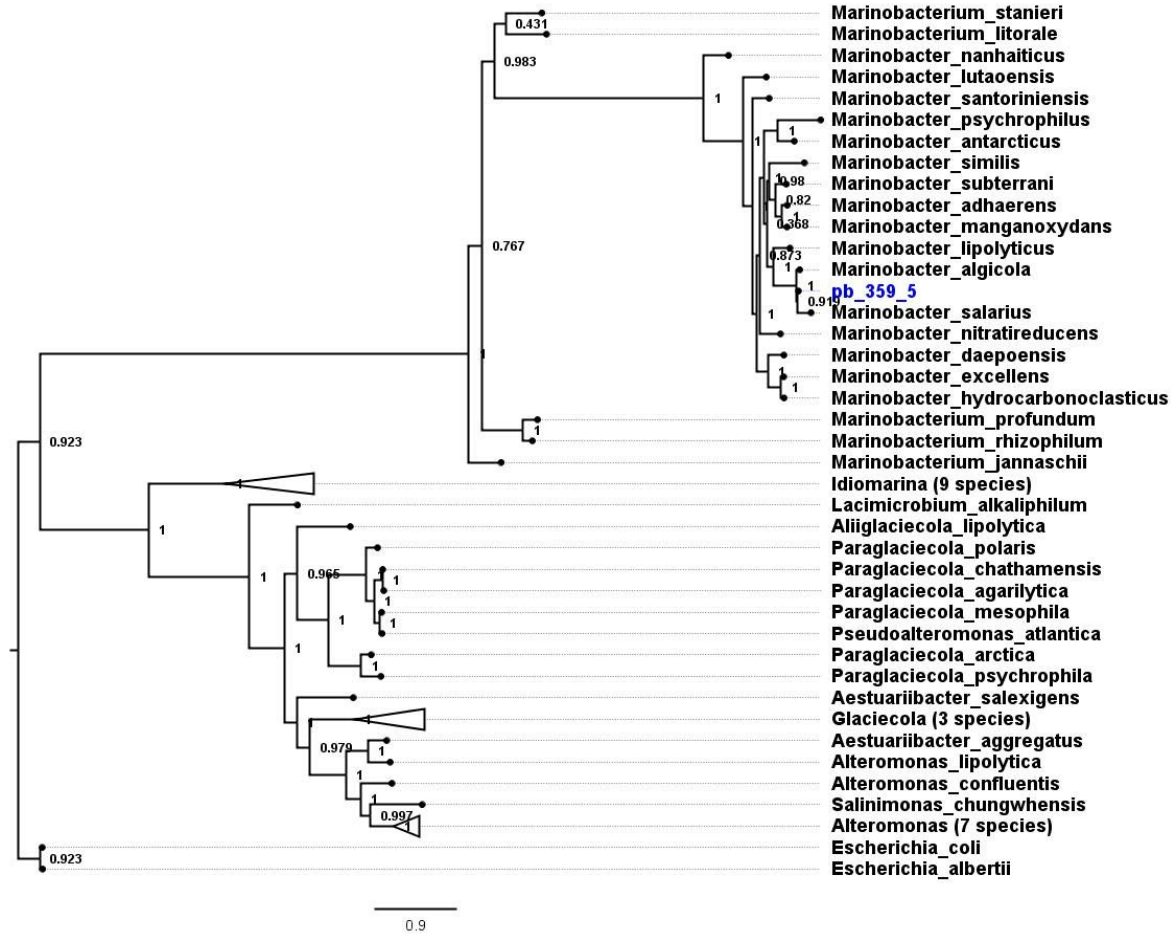

Figure 4: Outgroup-rooted, subsampled phylogenetic tree showing the position of pb_359_5 in the order Alteromonadales (class Gammaproteobacteria); two members of the genus *Escherichia* (order Enterobacterales; class Gammaproteobacteria) act as the outgroup. Node labels represent bootstrapping support. Tree generated by PhyloPhlAn version 0.99 [49] and visualised in FigTree version 1.4.3 [51].

4.2.5. pb_359_6

pb_359_6 appears in a clade composed primarily of species in the genus *Sulfitobacter*; its closest relative is *Sulfitobacter pseudonitzschiae* (Figure 2). This result is backed up by the 16S rRNA analysis, as the 16S rRNA sequence of the type strain *S. pseudonitzschiae* H3^T (also known as DSM 26824^T from inclusion in the DSMZ culture collection) (accession nos. NZ_JAMD00000000.1 and NZ_FQVP00000000.1, respectively) was found to have 99.2% identity to the two 16S rRNA sequences of pb_359_6 (one of which is seemingly located on a plasmid). The two 16S rRNA sequences of *Sulfitobacter* sp. 20_GPM-1509m (accession no. NZ_JIBC00000000.1) also share 99.2% 16S rRNA sequence identity with pb_359_6 (and are identical to the 16S rRNA sequence of the aforementioned *S. pseudonitzschiae* type strain), along with the 16S rRNA sequence of *Staleya guttiformis* strain R16 (accession no. AB607871.1) (reclassified as *Sulfitobacter guttiformis* in 2007 [75]); however, *S. guttiformis* appears to be more distantly related based on the phylogenetic tree (Figure 2). In addition, one better, unnamed result was found in the 16S rRNA comparison - *Sulfitobacter* sp. SAG13 (accession no. KX268604.1) shows 99.9% 16S rRNA sequence identity to pb_359_6.

Based on the results of the PhyloPhlAn and 16S rRNA analyses, I consider it most likely that pb_359_6 is a member of the genus *Sulfitobacter*; based on the high 16S rRNA sequence similarity, it is likely a strain of *Sulfitobacter pseudonitzschiae*. This conclusion is supported by nucleotide sequence similarity in *gyrB*, *rpoB* and *rpoD* (>89%), similar G+C content (within 2.2% of the type strain), a similar genome size (difference of around 0.17 Mbp), and a similar number of protein-coding genes (difference of 265 genes); cell shape and colour were also similar (data not shown), but these appear to be general traits of multiple species in the genus [76, 77].

4.2.6. pb_359_7

pb_359_7 was found in a clade of assorted species, the closest of which being *Mameliella alba* (Figure 2). This contradicts the results of the 16S rRNA analysis, which found the closest 16S rRNA match to be *Antarctobacter heliothermus*; the 16S rRNA sequences of the type strain *A. heliothermus* EL-219^T (also known as DSM 11445^T from inclusion in the DSMZ culture collection) (accession nos. NR_026406.1 and NR_115889.1, respectively) were found to have 99.5% and 99.9% identity, respectively, to the two identical 16S rRNA sequences of pb_359_7 (the 16S rRNA sequence of *Antarctobacter* sp. R14 [accession no. AB607870.1] was also found to be 99.5% similar to that of pb_359_7). This contradiction likely arises from the fact that no amino acid data was available from *Antarctobacter* for use in the PhyloPhlAn analysis, as the genome of the only species from this genus to be sequenced - *A. heliothermus* strain DSM 11445^T [78] - is in a permanent draft state (BioSample SAMN04488078). Furthermore, despite its proximity to pb_359_7 in the phylogenetic tree (Figure 2), *Mameliella alba*'s 16S rRNA sequences (accession nos. NZ_FMZI00000000.1 and NZ_JSUQ00000000.1) show only 96.3% identity to those of pb_359_7, a statistic shared by the 16S rRNA sequence of *Ruegeria* sp. PBVC088 (accession no. NZ_LZNT00000000.1). On this basis, I consider it most likely that pb_359_7 is a member of the genus *Antarctobacter*; based on the high 16S rRNA sequence similarity, it is likely a strain of *Antarctobacter heliothermus*. This conclusion is supported by nucleotide sequence similarity in *gyrB* and *rpoB* (>88% similarity to partial sequences from *A. heliothermus* strain DSM 11445^T when trimmed to the same length), similar G+C content (within 0.7% of the type strain), similar genome size (difference of around 0.16 Mbp) and a similar number of protein-coding genes (difference of 99 genes).

4.2.7. pb_359_8

pb_359_8 appears in a clade composed exclusively of species in the genus *Arenibacter*, its closest relative being *Arenibacter algicola* (Figure 5). The 16S rRNA analysis partially supports this result - while the three 16S rRNA sequences of the *A. algicola* type strain TG409^{T} (accession no. NZ_JPOO00000000.1) share 99.9% identity with the three 16S rRNA sequences in pb_359_8, so do the 16S rRNA sequences of *Flexibacter aggregans* strains BSs20185 and IFO 15975 (accession nos. DQ514301.1 and AB078039.1, respectively) (this species was reclassified as *Flexithrix dorotheae* in 2007 [79]). However, while this species is in the same phylum as *Arenibacter* (Bacteroidetes), it is not in the same class (*Flexithrix* is in Cytophagia, whereas *Arenibacter* is in Flavobacteriia). Taking the evidence from both the PhyloPhlAn and 16S rRNA analyses into account, I consider it most likely that pb_359_8 is a member of the genus *Arenibacter*; based on the high 16S rRNA sequence similarity, it is likely a strain of *Arenibacter algicola*. This conclusion is supported by nucleotide sequence similarity in *gyrB* and *rpoB* (>99%), similar G+C content (within 2.1% of the type strain), a fairly similar genome size (difference of around 0.3 Mbp) and a somewhat similar number of protein-coding genes (difference of 489 genes); cell shape and colour were also similar (data not shown), but these appear to be general traits of multiple species in the genus [80].

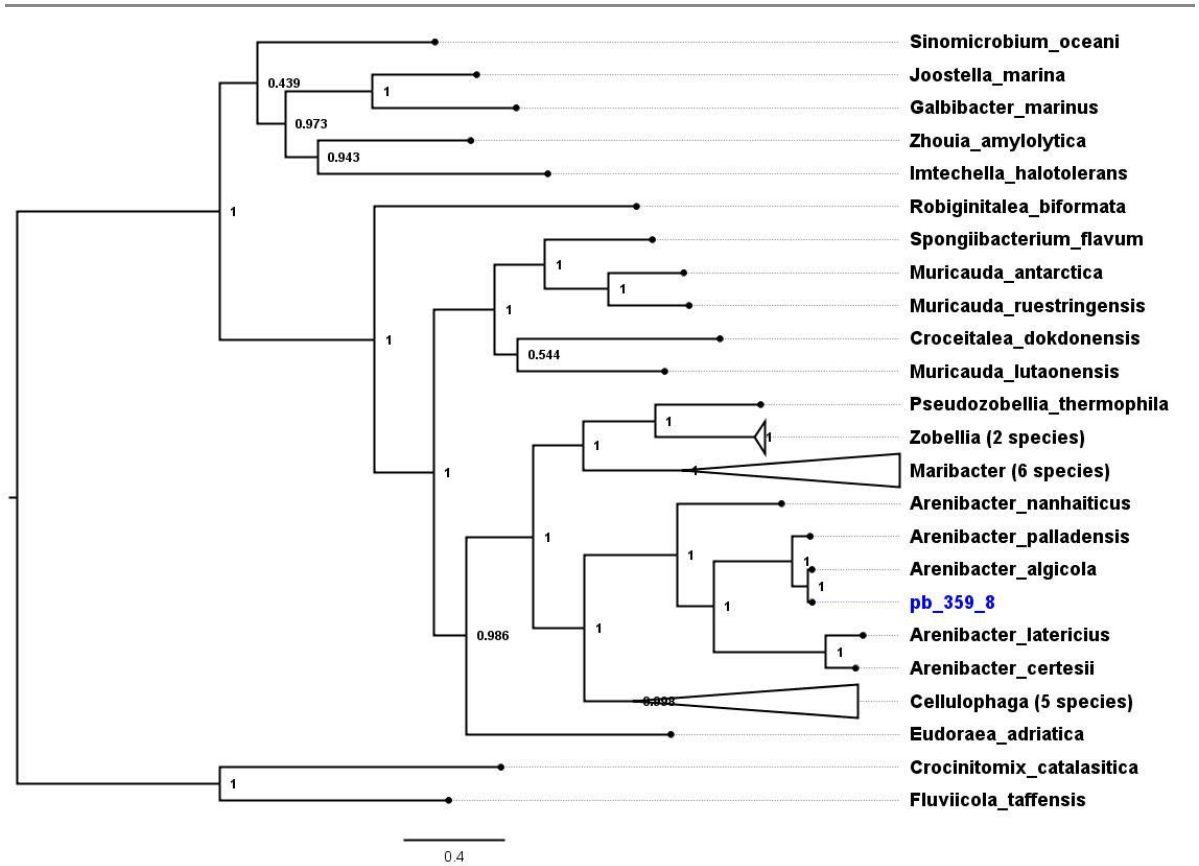

Figure 5: Outgroup-rooted, subsampled phylogenetic tree showing the position of pb_359_8 in the family Flavobacteriaceae (order Flavobacteriales); two members of the family Crocinitomicaceae (genera *Crocinitomix* and *Fluviicola*; order Flavobacteriales) act as the outgroup. Node labels represent bootstrapping support. Tree generated by PhyloPhlAn version 0.99 [49] and visualised in FigTree version 1.4.3 [51].

Table 4: Taxonomic predictions for each sample based on the PhyloPhlAn, 16S rRNA, and additional analyses.						
Sample	**Phylum**	**Class**	**Order**	**Family**	**Genus**	**Species**
pb_359_2	Proteobacteria	Alpha-proteobacteria	Rhodobacterales	Rhodobacteraceae	*Roseovarius*	*Roseovarius mucosus*
pb_359_3	Proteobacteria	Alpha-proteobacteria	Rhodobacterales	Rhodobacteraceae	*Loktanella*	*Loktanella vestfoldensis*
pb_359_4	Proteobacteria	Alpha-proteobacteria	Sphingomonadales	Sphingomonadaceae	*Sphingorhabdus*	*Sphingorhabdus flavimaris*
pb_359_5	Proteobacteria	Gamma-proteobacteria	Alteromonadales	Alteromonadaceae	*Marinobacter*	*Marinobacter salarius*
pb_359_6	Proteobacteria	Alpha-proteobacteria	Rhodobacterales	Rhodobacteraceae	*Sulfitobacter*	*Sulfitobacter pseudonitzschiae*
pb_359_7	Proteobacteria	Alpha-proteobacteria	Rhodobacterales	Rhodobacteraceae	*Antarctobacter*	*Antarctobacter heliothermus*
pb_359_8	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	*Arenibacter*	*Arenibacter algicola*

4.3. Annotation

Table 5 shows the number of features identified in each genome by both the Prokka annotation and the subsequent Pathway Tools analysis. In addition, pb_359_2, pb_359_3 and pb_359_5 underwent consistency checks performed by NCBI staff, which identified additional features such as pseudogenes. The number of coding sequences (CDS) in each replicon is shown in Table 6.

4.4. Pathway predictions and interesting genes

Based on the predictions of Pathway Tools, as well as searches for key terms in each sample's GenBank file, a number of potential features of interest were found in each species. These are summarised in Table 7.

The autoinducer AI-1 biosynthesis pathway, indicative of quorum sensing, was predicted in three of the four species identified in the family Rhodobacteraceae, except pb_359_7 [81].

A phosphatidylcholine biosynthesis pathway was predicted in six of the seven species. Interestingly, Pathway Tools predicts the type V pathway (producing phosphatidylcholine from phosphatidylethanolamine [82]) for pb_359_4 and pb_359_8, whereas all four Rhodobacteraceae species are predicted to have the type VI pathway (producing phosphatidylcholine from CDP-diacylglycerol and choline [83]).

As mentioned in section 2.3, bacteria have been found to provide vitamin B_{12} to many microalgae, including diatoms [11]. This phenomenon appears to be widespread among blooming algae, so it is perhaps unsurprising that six of the seven bacterial species in this study contain pathway components for vitamin B_{12} biosynthesis and salvage [84].

Bacteria-produced siderophores, particularly vibrioferrin, can be utilised by microalgae to facilitate iron uptake [12]. Two of the bacteria in this study showed evidence of siderophore biosynthesis pathways, one of which was identified as belonging to the genus *Marinobacter*, a genus known to include species which produce vibrioferrin [12]. However, Pathway Tools predicted biosynthesis pathways for the siderophores enterobactin and aerobactin, rather than vibrioferrin [85, 86].

Table 5: Features present in each genome according to the annotation predictions of Prokka. Predictions for the number of pathways were obtained from Pathway Tools, after running the Assign Probable Enzymes and Pathway Hole Filler tools. pb_359_2, pb_359_3 and pb_359_5 underwent additional checks by NCBI staff prior to genome submission.

Sample	Number of each feature found in the genome									Number of pathways
	CDS	Named genes	Hypothetical proteins	Pseudo-genes	tRNA	tmRNA	rRNA	ncRNA	Misc. binding	
pb_359_2	4178	2306	636	3	46	1	6	5	6	295
pb_359_3	3936	2180	669	8	45	1	6	19	0	317
pb_359_4	3348	1724	818	-	46	1	6	7	-	249
pb_359_5	4263	2451	755	7	51	1	9	11	0	270
pb_359_6	4967	2817	1048	-	44	1	6	19	-	321
pb_359_7	5098	2790	1150	-	48	0	6	12	-	320
pb_359_8	4984	2219	1469	-	47	1	9	22	-	268

Table 6: Number of CDS present in each replicon of the species examined in this study. An asterisk denotes a genome which was submitted to NCBI prior to the completion of this report. Due to the additional checks that occur prior to genome submission to NCBI, figures for genomes which have not yet been submitted may change slightly upon submission.

Sample	Total CDS	Number of CDS per replicon (from longest to shortest)							
		Chromosome	Plasmid 1	Plasmid 2	Plasmid 3	Plasmid 4	Plasmid 5	Plasmid 6	Plasmid 7
pb_359_2 *	4178	3953	189	36	-	-	-	-	-
pb_359_3 *	3936	3797	97	42	-	-	-	-	-
pb_359_4	3348	3348	-	-	-	-	-	-	-
pb_359_5 *	4263	3999	264	-	-	-	-	-	-
pb_359_6	4967	3497	412	255	277	193	152	93	88
pb_359_7	5098	4504	391	126	77	-	-	-	-
pb_359_8	4984	4901	83	-	-	-	-	-	-

As mentioned in section 2.3, DMSP and its breakdown products are used by some bacteria as a source of carbon and sulphur [15]. Four of the bacteria in this study were predicted to possess the superpathway of DMSP degradation, whereby DMSP can be broken down into either DMS and acrylate, or methanethiol and acetaldehyde [87]; one other species was predicted to only be able to break down DMSP into DMS and acrylate [88].

As a toxic element, cadmium can have a negative effect on diatom growth [89], but conversely, diatoms are also the only known organisms in which cadmium has been found to play a biological role; specifically, the diatom *Thalassiosira weissflogii* contains a carbonic anhydrase - CDCA1 - which can make use of cadmium instead of zinc [90]. No data on cadmium sensitivity appears to be available for *Skeletonema marinoi*, however, so it is unclear whether the relative abundance of cadmium-resistance genes in two of the associated bacterial species benefits the diatom, or solely the bacteria themselves.

Mercury is toxic to diatoms, as to other organisms [91]. Although all of the bacterial genomes in this study contain at least one gene from the *mer* operon, three of them contain a complete form of the operon on a plasmid [92], implying a potential role in detoxifying the environment around the diatom. Two of the bacterial species in this study also contain a component of an erythromycin biosynthesis pathway, albeit these pathways appear to be incomplete in both cases; this may be the result of either incomplete or incorrect annotation.
In addition, each of the genomes contained at least a few drug resistance genes, some of which were present in only one or two of the species studied; genes for multidrug resistance and bleomycin resistance were particularly prevalent across most of the species.

Table 7: Pathways and other features predicted to be present in the *Skeletonema marinoi*-associated bacterial species examined in this study. Pathways were predicted by Pathway Tools; other features were found by searching for key terms in the samples' respective GenBank files.

Sample	**pb_359_2**	**pb_359_3**	**pb_359_4**	**pb_359_5**	**pb_359_6**	**pb_359_7**	**pb_359_8**
Autoinducer AI-1 biosynthesis	✔	✔			✔		
Phosphatidylcholine biosynthesis	✔	✔	✔		✔	✔	✔
Vitamin B_{12} biosynthesis	✔	✔	✔	✔	✔	✔	
DMSP degradation	✔	✔		✔	✔	✔	
Siderophore biosynthesis				✔	✔		
Auxin (indole-3-acetate) biosynthesis	✔			✔	✔	✔	
Erythromycin biosynthesis			✔			✔	
Complete mercuric resistance (*mer*) operon	✔	✔		✔			
Relative abundance (6+) of Cd-resistance genes				✔			✔
Relative abundance (6+) of multidrug resistance genes	✔	✔	✔		✔	✔	✔
Relative abundance (6+) of bleomycin resistance genes			✔		✔	✔	✔

4.5. Other observations

Table 8 shows the results of the secondary metabolite gene cluster search performed with antiSMASH. Some of these results provide support for the predictions of Pathway Tools, for example the presence of homoserine lactone (the quorum sensing molecule produced by the autoinducer AI-1 biosynthesis pathway) in pb_359_2, pb_359_3 and pb_359_6, and the presence of siderophores in pb_359_5 and pb_359_6. However, the analysis also highlighted additional features, such as the presence of gene clusters for bacteriocins - a class of antimicrobial peptides [93]- in five of the seven species in the study.
The results of the prophage sequence search using PHASTER are shown in Table 9.
In addition to their interactions with *S. marinoi*, the bacteria of the microbiome also show evidence of interacting with each other. Specifically, pb_359_2 and pb_359_3 appear to share the same short plasmid, with the pb_359_3 plasmid being around 9kb larger than its pb_359_2 counterpart. This difference can be accounted for by a few additional genes, primarily those coding for transposons and DNA invertases, and aside from this additional ~9kb, the plasmids share 100% sequence identity.

Table 8: Results of secondary metabolite gene cluster search using antiSMASH v3.0.5 [59] (analyses performed February 2017).

Sample	**pb_359_2**	**pb_359_3**	**pb_359_4**	**pb_359_5**	**pb_359_6**	**pb_359_7**	**pb_359_8**
Homoserine lactone	2	1	-	-	4	2	-
Terpene	2	2	1	-	1	1	3
Ectoine	1	-	-	1	-	2	-
Type 1 PKS	1	-	-	-	-	1	-
Type 3 PKS	-	-	1	-	-	-	1
NRPS	-	-	-	1*	1	-	-
NRPS-Type 1 PKS	-	-	-	1	-	-	-
Type 1 PKS-NRPS	-	-	-	-	-	1	-
Bacteriocin	-	1	1	1	1	1	-
Lasso peptide	-	-	1	-	-	1**	-
Succinoglycan	-	1	-	-	-	-	-
Vicibactin	-	-	-	-	1	-	-
Aryl polyene-resorcinol	-	-	-	-	-	-	1***

(PKS = polyketide synthase, NRPS = nonribosomal peptide synthetase)
* antiSMASH identifies this NRPS as **serobactin**
** This lasso peptide gene cluster was found on a plasmid rather than on the chromosome
*** antiSMASH identifies this aryl polyene-resorcinol as **flexirubin**

Table 9: Results of PHASTER prophage search of the bacterial genomes [60] (analyses performed January 2017). For a full explanation of how PHASTER defines 'intact', 'questionable' and 'incomplete' phage regions, see Zhou *et al.* (2011) [61].

Sample	Predicted chromosomal prophage regions	Predicted plasmid prophage regions
pb_359_2	**3 intact regions**	-
pb_359_3	**2 intact**, 1 questionable, 3 incomplete	Plasmid 1 - 1 incomplete region
pb_359_4	**1 intact**, 1 incomplete	-
pb_359_5	1 incomplete region	-
pb_359_6	1 questionable region	**Plasmid 1 - 1 intact region** Plasmid 3 - 1 incomplete region Plasmid 5 - 1 incomplete region
pb_359_7	**1 intact**, 1 questionable, 2 incomplete	**Plasmid 1 - 1 intact**, 1 incomplete Plasmid 3 - 1 incomplete region
pb_359_8	2 questionable regions	-

5. Discussion

Amin *et al.* (2012) summarise the results of multiple studies on the composition of diatom-associated bacterial communities by noting that a majority of these bacterial species come from the phyla Proteobacteria and Bacteroidetes [7]. The seven *Skeletonema marinoi*-associated species investigated in this study belong to the Alphaproteobacteria (five species), Gammaproteobacteria (one species) and Bacteroidetes (one species), in agreement with this observation. In addition, the Amin *et al.* review notes a number of genera which appear to be particularly strongly associated with diatoms - *Sulfitobacter*, *Roseobacter*, *Alteromonas*, and *Flavobacterium* [7]; a *Sulfitobacter* species was among the bacteria identified in this study, as well as one species each in the same family as *Alteromonas* and *Flavobacterium*, and three members of the *Roseobacter* clade (a subgroup of the Alphaproteobacteria, noted for ≥89% 16S rRNA sequence identity to one another, and named after the first described members of the clade - *Roseobacter litoralis* and *Roseobacter denitrificans* [94]).

5.1. pb_359_2 - *Roseovarius mucosus* strain SMR3

Based on its similarity to the *Roseovarius mucosus* type strain DSM 17069^T, I conclude that pb_359_2 is most likely a strain of the same species, and due to its isolation from *Skeletonema marinoi* strain RO5AC, it has been designated ***Roseovarius mucosus* strain SMR3** (shown in Figure 6a). The aforementioned type strain has also been shown to associate with a marine phytoplankton, specifically the dinoflagellate *Alexandrium ostenfeldii*, from which it was initially isolated [70].

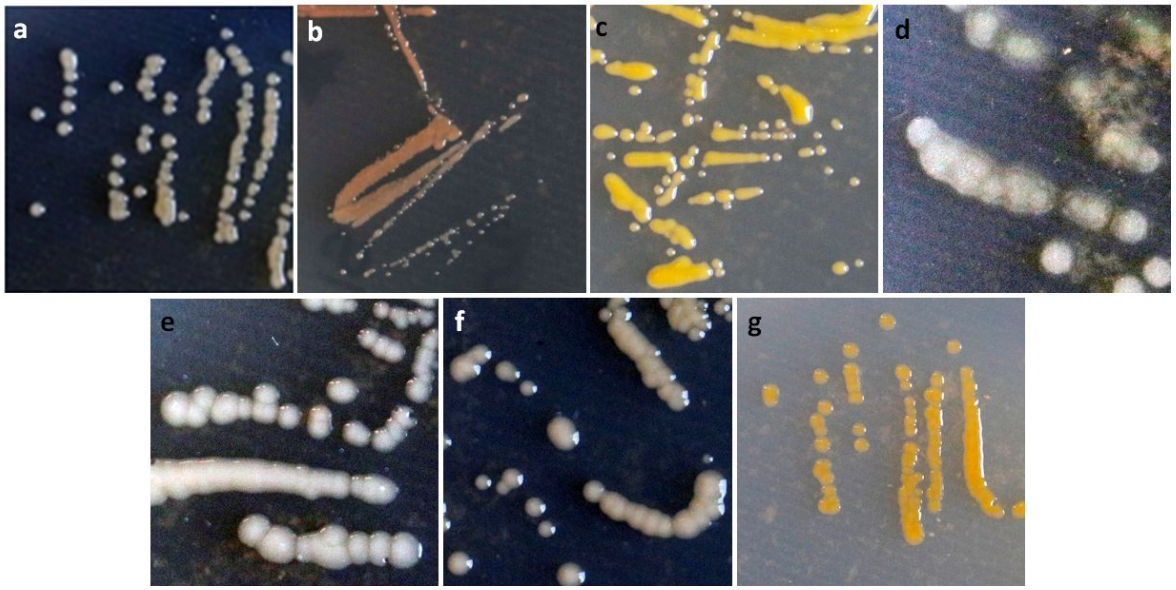

Figure 6: The *Skeletonema marinoi*-associated bacterial species investigated in this study.
a: *Roseovarius mucosus* strain SMR3 **b:** *Loktanella vestfoldensis* strain SMR4r
c: *Sphingorhabdus flavimaris* strain SMR4y **d:** *Marinobacter salarius* strain SMR5
e: *Sulfitobacter pseudonitzschiae* strain SMR1 **f:** *Antarctobacter heliothermus* strain SMS3
g: *Arenibacter algicola* strain SMS7
Images courtesy of Oskar Johansson, Department of Biological and Environmental Sciences, University of Gothenburg

---

Multiple elements of the *R. mucosus* strain SMR3 genome imply potential interactions with *S. marinoi*. Bacteria containing phosphatidylcholine often associate with eukaryotes [20], and two phosphatidylcholine synthase genes are present in this genome, along with a predicted three out of four components of the CDP-diacylglycerol biosynthesis II pathway upstream of the phosphatidylcholine biosynthesis VI pathway [83, 95]. Both Pathway Tools and antiSMASH also note the presence of biosynthesis genes for homoserine lactones - quorum sensing molecules [81]; quorum sensing is involved in the regulation of, among other things, virulence and biofilm production [18].
In terms of nutrient exchange, the *R. mucosus* strain SMR3 genome contains multiple examples of vitamin B_{12} synthesis and salvage pathways; all components of the 5,6-dimethylbenzimidazole biosynthesis I (aerobic) and adenosylcobalamin salvage from cobalamin pathways are predicted in the annotation [96, 97], with the cob(II)yrinate a,c-diamide biosynthesis II (late cobalt incorporation) and adenosylcobalamin salvage from cobinamide I pathways lacking only a single component each in the current annotation [98, 99]. It seems possible that *R. mucosus* strain SMR3 can make use of diatom-derived DMSP, as four of the five genes in the superpathway of DMSP degradation are predicted in the current annotation (*dmdD* is absent) [87]. In addition, both components of the indole-3-acetate biosynthesis IV (bacteria) pathway [100], resulting in auxin production, are predicted, although the elements upstream of this pathway (which result in the production of indole-3-acetonitrile) appear to be absent [101].
In terms of resistance genes, an abundance of genes pertaining to multidrug resistance were found, with eight annotated as producing a ‘multidrug resistance protein’, and a further six pertaining to multidrug export and transport.

Much of the material from the large plasmid - pSMR3-1 - appears to be present in the type strain (accession no. NZ_AONH00000000.1) on plasmid pRosmuc_A123 and scaffold14 [102], whereas only two short (> 5kb) hits were found when comparing the short plasmid - pSMR3-2 - and the type strain assembly using BLASTn [37].
Evidence of interaction with other elements of the microbiome was also observed. The short plasmid of both *R. mucosus* strain SMR3 and *Loktanella vestfoldensis* strain SMR4r (see section 5.2) were found to share 100% sequence identity along the majority of their length, with the *L. vestfoldensis* plasmid containing a handful of additional genes - three transposases, two DNA invertases, and an additional mercuric reductase. The main constituents of this short plasmid appear to be genes involved in mercuric resistance (*mer* operon components *merR1*, MerT, MerP, MerF and *merA* [gene names italicised; products of unnamed genes non-italicised]) and tripartite ATP-independent periplasmic (TRAP) transporters (involved in transport of a variety of carboxylate-group-containing substrates [103]), implying that these functions may be important in the environment these bacteria share.
A genome announcement has been published for this strain [104], and the data is available on NCBI (accession nos. CP020474.1-CP020476.1).

5.2. pb_359_3 - *Loktanella vestfoldensis* strain SMR4r

Based on its similarity to the *Loktanella vestfoldensis* type strain DSM 16212^T, I conclude that pb_359_3 is most likely a strain of the same species, and due to its isolation from *Skeletonema marinoi* strain RO5AC, it has been designated ***Loktanella vestfoldensis* strain SMR4r** (shown in Figure 6b). Although most species of *Loktanella* have been isolated from seawater and sediments, at least one - *Loktanella* sp. Gb03 - has been found in association with a marine phytoplankton, the dinoflagellate *Gambierdiscus belizeanus* [105]; this species has also been shown to possess algicidal capabilities.
Multiple elements of the *L. vestfoldensis* strain SMR4r genome imply potential interactions with *S. marinoi*. Bacteria containing phosphatidylcholine often associate with eukaryotes [20], and a phosphatidylcholine synthase gene is present in this genome, along with all four components of the CDP-diacylglycerol biosynthesis II pathway upstream of the phosphatidylcholine biosynthesis VI pathway [83, 95]. Both Pathway Tools and antiSMASH note the presence of biosynthesis genes for homoserine lactones - quorum sensing molecules [81]; quorum sensing is involved in the regulation of, among other things, virulence and biofilm production [18].
In terms of nutrient exchange, *L. vestfoldensis* strain SMR4r contains multiple examples of vitamin B_{12} synthesis and salvage pathways; all components of the 5,6-dimethylbenzimidazole biosynthesis I (aerobic) and adenosylcobalamin salvage from cobalamin pathways are predicted in the annotation [96, 97], with the cob(II)yrinate a,c-diamide biosynthesis II (late cobalt incorporation) and adenosylcobalamin salvage from cobinamide I pathways lacking only a single component each in the current annotation [98, 99]. It seems possible that *L. vestfoldensis* strain SMR4r can make use of diatom-derived DMSP, as after manual enzyme assignment and pathway hole filling, predictions for all elements of the superpathway of DMSP degradation were found [87]. In addition, secondary metabolite analysis highlighted the presence of a gene cluster for the biosynthesis of a bacteriocin (an antimicrobial peptide [93]), as well as for the biosynthesis of succinoglycan, a "symbiotically important exopolysaccharide" [106].
In terms of resistance genes, an abundance of genes pertaining to multidrug resistance were found, with ten annotated as producing a 'multidrug resistance protein', and a further three pertaining to multidrug efflux and transport. Uniquely among the species investigated in this study, *L. vestfoldensis*

strain SMR4r also contains one gene each pertaining to lincosamide resistance (*linA*) and fusaric acid resistance. Lincosamide targets the (bacterial) 50S ribosomal subunit [107], so this resistance gene is likely to be purely for the benefit of *Loktanella*; however, fusaric acid has been shown to negatively affect eukaryotic and bacterial cells, as well as inhibit bacterial quorum sensing, so fusaric acid resistance may prove beneficial to both *L. vestfoldensis* strain SMR4r and *S. marinoi* [108, 109].
Interestingly, siderophore biosynthesis (a process not predicted for this strain) may also improve fusaric acid resistance [108], therefore other species in the microbiome which do produce siderophores could exhibit fusaric acid resistance despite lacking the gene.
In addition, as mentioned in section 5.1, the short plasmid found in this strain - pSMR4r-2 - is near-identical to the short plasmid pSMR3-2 found in *Roseovarius mucosus* strain SMR3, showing evidence of interaction between the two species; as it carries this plasmid, *L. vestfoldensis* strain SMR4r also possesses a complete *mer* operon.
The genome data for this strain is available on NCBI (accession nos. CP021431.1-CP021433.1).

5.3. pb_359_4 - *Sphingorhabdus flavimaris* strain SMR4y

Based primarily on its 16S rRNA sequence similarity to *Sphingorhabdus flavimaris* strain R-36742, I conclude that pb_359_4 is most likely a strain of the same species, and due to its isolation from *Skeletonema marinoi* strain RO5AC, it has been designated ***Sphingorhabdus flavimaris* strain SMR4y** (shown in Figure 6c). Some evidence of interaction between *Sphingorhabdus* species and phytoplankton blooms has previously been documented [110].
Multiple elements of the *S. flavimaris* strain SMR4y genome imply potential interactions with *S. marinoi*. Bacteria containing phosphatidylcholine often associate with eukaryotes [20], and all elements of the phosphatidylcholine biosynthesis V pathway are predicted in this genome [82], along with all components of the upstream pathways of phosphatidylethanolamine biosynthesis I and CDP-diacylglycerol biosynthesis II [95, 111].
In terms of nutrient exchange, there is some evidence to suggest that *S. flavimaris* strain SMR4y may produce vitamin B_{12}; the pathway for adenosylcobalamin salvage from cobalamin is predicted to be present [97], along with seven of the eleven components of the cob(II)yrinate a,c-diamide biosynthesis II (late cobalt incorporation) pathway [98].
Sphingorhabdus flavimaris strain SMR4y is predicted to contain one of the two components of the erythromycin A biosynthesis pathway [112]. Secondary metabolite analysis also highlights a gene cluster associated with biosynthesis of a bacteriocin (an antimicrobial peptide [93]), along with another cluster associated with biosynthesis of a lasso peptide, a class of proteins which frequently play an antimicrobial role [113].
In terms of resistance genes, an abundance of genes pertaining to multidrug resistance were found, with twelve annotated as producing a ‘multidrug resistance protein’, and a further seven pertaining to multidrug export and efflux; fourteen genes pertaining to bleomycin resistance were also found. In addition, *S. flavimaris* strain SMR4y is one of two species examined in this study found to contain a methicillin resistance gene, in this case *mecR1*. As methicillin is a beta-lactam antibiotic of the penicillin class, and therefore affects bacterial cell wall synthesis, it doesn’t have a direct effect on diatoms due to differences in cell wall structure [114]; as a consequence, this gene is likely to benefit only the bacterium itself.

5.4. pb_359_5 - *Marinobacter salarius* strain SMR5

Based on its similarity to the *Marinobacter salarius* type strain R9SW1^T, I conclude that pb_359_5 is most likely a strain of the same species, and due to its isolation from *Skeletonema marinoi* strain RO5AC, it has been designated ***Marinobacter salarius* strain SMR5** (shown in Figure 6d). While the type strain of *M. salarius* was isolated from seawater [74], the type strain of the closely-related *Marinobacter algicola* - DG893^T - was discovered in association with marine phytoplankton, specifically the dinoflagellates *Gymnodinium catenatum* and *Alexandrium tamarense* [115]. *Marinobacter* species have also been found in association with other *Skeletonema* species; *Marinobacter* sp. strain MCTG268 was isolated from *Skeletonema costatum* [116].
Multiple elements of the *M. salarius* strain SMR5 genome imply potential interactions with *S. marinoi*. In terms of nutrient exchange, the genome contains multiple examples of vitamin B_{12} synthesis and salvage pathways, with all components of the 5,6-dimethylbenzimidazole biosynthesis I (aerobic), adenosylcobalamin salvage from cobalamin and adenosylcobalamin salvage from cobinamide I pathways present in the annotation [96, 97, 99]; six of the eleven components of the cob(II)yrinate a,c-diamide biosynthesis II (late cobalt incorporation) pathway were also found in the current annotation [98]. Two of the four genes involved in biosynthesis of the siderophore aerobactin were also found [86], as well as six of the eleven genes in the biosynthesis pathway of another siderophore - enterobactin [85]. Secondary metabolite analysis highlighted a gene cluster associated with production of a third siderophore - serobactin - implying a possible role in assisting with iron uptake in *S. marinoi*. Notably lacking from this list of siderophores is vibrioferrin, which has been demonstrated to be used by microalgae, and which is known to be produced by a close relative of *M. salarius* - the *M. algicola* type strain DG893^T [12]; it is possible that vibrioferrin biosynthesis genes have either been overlooked or misannotated in the current annotation.
In addition, after manual enzyme assignment and pathway hole filling, predictions for all elements of the superpathway of DMSP degradation were found, implying that the bacterium may be able to use diatom-derived DMSP [87]. The indole-3-acetate biosynthesis V (bacteria and fungi) pathway, resulting in auxin production, is also predicted to be present [117], although the elements upstream of this pathway (which result in the production of indole-3-acetonitrile) appear to be absent [101]; experimental work with this species, however, implies that it is indeed capable of producing auxin (Johansson, O.N. & Clarke, A.K., unpublished work). In addition, secondary metabolite analysis highlighted the presence of a gene cluster associated with biosynthesis of a bacteriocin - an antimicrobial peptide [93].
In terms of resistance genes, a complete *mer* operon (*merR1*, MerT, *merP*, *merA* and *merB* [gene names italicised; products of unnamed genes non-italicised]) was found on the plasmid, pSMR5, with a total of thirteen genes in the whole genome annotated as pertaining to mercury. Six genes were also found pertaining to cadmium resistance, with an additional gene encoding a ‘cadmium, cobalt and zinc/H(+)-K(+) antiporter’. Uniquely among the species investigated in this study, *M. salarius* strain SMR5 also contains two genes pertaining to methyl viologen (paraquat) resistance (*smvA* and *yddG*); while paraquat appears to have some impact on bacterial growth, its effect on photosynthesis (hence its use as a herbicide) may render it dangerous to *S. marinoi* as well, implying that both *M. salarius* strain SMR5 and its host could benefit from paraquat resistance [118, 119]. Compared to the other species in this study, however, *M. salarius* strain SMR5 contains a relatively low number of multidrug resistance genes; only four genes are annotated as encoding a ‘multidrug resistance protein’, with a further four pertaining to multidrug export, efflux and transport.
The genome data for this strain is available on NCBI (accession nos. CP020931.1 and CP020932.1).

5.5. pb_359_6 - *Sulfitobacter pseudonitzschiae* strain SMR1

Based on its similarity to the *Sulfitobacter pseudonitzschiae* type strain DSM 26824^T, I conclude that pb_359_6 is most likely a strain of the same species, and due to its isolation from *Skeletonema marinoi* strain RO5AC, it has been designated ***Sulfitobacter pseudonitzschiae*** **strain SMR1** (shown in Figure 6e). Multiple *Sulfitobacter* species have been found in association with diatoms; for example, this species was named after *Pseudo-nitzschia multiseries*, from which the type strain was isolated, and another *Sulfitobacter* species was discovered in association with the diatom *Amphiprora kufferathii* [76, 120].

Multiple elements of the *S. pseudonitzschiae* strain SMR1 genome imply potential interactions with *S. marinoi*. Bacteria containing phosphatidylcholine often associate with eukaryotes [20], and the phosphatidylcholine biosynthesis VI pathway is predicted to be present in this genome [83], along with three out of four components of the upstream CDP-diacylglycerol biosynthesis II pathway [95]. Both Pathway Tools and antiSMASH note the presence of biosynthesis genes for homoserine lactones - quorum sensing molecules [81]; quorum sensing is involved in the regulation of, among other things, virulence and biofilm production [18].

In terms of nutrient exchange, *S. pseudonitzschiae* strain SMR1 contains multiple examples of vitamin B_{12} synthesis and salvage pathways; all components of the adenosylcobalamin salvage from cobalamin and cob(II)yrinate a,c-diamide biosynthesis II (late cobalt incorporation) pathways are predicted in the annotation [97, 98], with the adenosylcobalamin salvage from cobinamide I and II pathways lacking only a single component each in the current annotation [99, 121]. Two of the four genes involved in biosynthesis of the siderophore aerobactin were also found [86], with the secondary metabolite analysis highlighting a gene cluster associated with production of another siderophore - vicibactin - implying a potential role in assisting with iron uptake in *S. marinoi*. It also seems possible that *S. pseudonitzschiae* strain SMR1 can make use of diatom-derived DMSP, as the DMSP degradation I (cleavage) pathway is predicted to be present [88].

In addition, two potential indole-3-acetate biosynthesis pathways (III and IV), resulting in auxin production, are predicted to be present [100, 122]; however, the elements upstream of variant IV of this pathway (which result in the production of indole-3-acetonitrile) appear to be absent [101]. Secondary metabolite analysis also highlights a gene cluster associated with biosynthesis of a bacteriocin - an antimicrobial peptide [93].

In terms of resistance genes, an abundance of genes pertaining to multidrug resistance were found, with twelve annotated as producing a ‘multidrug resistance protein’, and a further seven pertaining to multidrug export, efflux and transport; eight genes pertaining to bleomycin resistance were also found.

A noteworthy feature of the *S. pseudonitzschiae* strain SMR1 genome is the apparent abundance of plasmids it possesses; whereas the other bacteria in this study were found to contain a maximum of three plasmids, this strain appears to contain seven, five of which are larger than 100kb. Attempts to falsify this result were unsuccessful, however, and another *Sulfitobacter* species was recently sequenced which also contains a large number of megaplasmids (*Sulfitobacter* sp. AM1-D1, assembly accession no. GCA_001886735.1), lending support to the idea that this result is correct.

5.6. pb_359_7 - *Antarctobacter heliothermus* strain SMS3

Based primarily on its 16S rRNA sequence similarity to the *Antarctobacter heliothermus* type strain EL-219^T (= DSM 11445^T), I conclude that pb_359_7 is most likely a strain of the same species, and due to its isolation from *Skeletonema marinoi* strain ST54, it has been designated ***Antarctobacter heliothermus* strain SMS3** (shown in Figure 6f). However, as only one species has currently been described in this genus, this classification may require revision when more data becomes available. An *Antarctobacter* species has previously been found in association with a brown alga [123].
Multiple elements of the *A. heliothermus* strain SMS3 genome imply potential interactions with *S. marinoi*. Bacteria containing phosphatidylcholine often associate with eukaryotes [20], and the phosphatidylcholine biosynthesis VI pathway is predicted to be present in this genome [83], along with the upstream CDP-diacylglycerol biosynthesis pathway (variants I and II [95, 124]) being complete in the annotation. In addition, although not noted by the pathway analysis, secondary metabolite analysis highlighted the presence of two gene clusters related to production of homoserine lactones - quorum sensing molecules; quorum sensing is involved in the regulation of, among other things, virulence and biofilm production [18].
In terms of nutrient exchange, *A. heliothermus* strain SMS3 contains multiple examples of vitamin B_{12} synthesis and salvage pathways; all components of the 5,6-dimethylbenzimidazole biosynthesis I (aerobic), adenosylcobalamin salvage from cobalamin and cob(II)yrinate a,c-diamide biosynthesis II (late cobalt incorporation) pathways are present in the annotation [96, 97, 98], with the adenosylcobalamin salvage from cobinamide I pathway lacking only a single component in the current annotation [99].
It seems possible that *A. heliothermus* strain SMS3 can make use of diatom-derived DMSP, as all components of the superpathway of DMSP degradation are present in the genome [87]. In addition, two potential indole-3-acetate biosynthesis pathways (IV and V [100, 117]), resulting in auxin production, are predicted to be present, although the elements upstream of these pathway (which result in the production of indole-3-acetonitrile) appear to be absent [101].
Antarctobacter heliothermus strain SMS3 is predicted to contain two of the four components of the erythromycin D biosynthesis pathway [125]. Secondary metabolite analysis also highlights a gene cluster associated with biosynthesis of a bacteriocin - an antimicrobial peptide [93]. Another noteworthy feature highlighted by this analysis is that, uniquely among the species in this study, a secondary metabolite-associated gene cluster is present on a plasmid - in this case, a lasso peptide-associated gene cluster was identified on pSMS3-2; these peptides often play an antimicrobial role [113].
In terms of resistance genes, an abundance of genes pertaining to multidrug resistance were found, with seven annotated as producing a ‘multidrug resistance protein’, and a further five pertaining to multidrug efflux and transport; seven genes pertaining to bleomycin resistance were also found. Uniquely among the species investigated in this study, *Antarctobacter heliothermus* strain SMS3 contains a fosmidomycin resistance gene (*fsr*). Fosmidomycin is an inhibitor of the methylerythritol phosphate (MEP) pathway, responsible for the biosynthesis of some isoprenoids (such as carotenoid pigments) [126]; as this pathway is present in both diatoms and bacteria, the resistance gene may benefit both *A. heliothermus* strain SMS3 and *S. marinoi*.

5.7. pb_359_8 - *Arenibacter algicola* strain SMS7

Based on its similarity to the *Arenibacter algicola* type strain TG409^T, I conclude that pb_359_8 is most likely a strain of the same species, and due to its isolation from *Skeletonema marinoi* strain ST54, it has been designated ***Arenibacter algicola* strain SMS7** (shown in Figure 6g). The type strain was also discovered in association with a species of the *Skeletonema* genus - *Skeletonema costatum* [80]. Some elements of the *A. algicola* strain SMS7 genome imply potential interactions with *S. marinoi*. Bacteria containing phosphatidylcholine often associate with eukaryotes [20], and two of the three components of the phosphatidylcholine biosynthesis V pathway are predicted to be present in this genome [82], along with both elements of the upstream phosphatidylethanolamine biosynthesis I pathway [111], and three of the four elements of the CDP-diacylglycerol biosynthesis I pathway upstream of that [124]. Vitamin B_{12} salvage may also be possible, as the adenosylcobalamin salvage from cobalamin pathway is predicted to be present [97], although salvage alone may not be sufficient for supplying *S. marinoi* with the vitamin.
In terms of resistance genes, an abundance of genes pertaining to multidrug resistance were found, with six annotated as producing a 'multidrug resistance protein', and a further four pertaining to multidrug export, efflux and transport; six genes pertaining to bleomycin resistance were also found, as well as eight genes pertaining to cadmium resistance and an additional four genes whose annotations also refer to cadmium. In addition, *A. algicola* strain SMS7 is one of two species examined in this study found to contain a methicillin resistance gene, in this case *mecI*. As mentioned in section 5.3, methicillin doesn't have a direct effect on diatoms [114], so this resistance gene is likely to benefit only the bacterium itself.
Of note in the secondary metabolite analysis, whereas most results of this analysis gave only the class to which the gene cluster belonged (as shown in Table 8), *A. algicola* strain SMS7 was found to contain a gene cluster associated with a named compound - the pigment flexirubin, which likely accounts for the yellow colour of the colonies (as seen in Figure 6g) [127].

5.8. Conclusions

Based on the annotations of Prokka and the pathway predictions of Pathway Tools, a number of potential features have been found in each of the bacterial species in this study which may point to their respective roles in the *Skeletonema marinoi* microbiome.
With the possible exception of *Arenibacter algicola* strain SMS7, all of the species studied show the potential for supplying *S. marinoi* with nutrients - primarily vitamin B_{12}, but also iron and growth hormones; in return, five of the seven species appear to be able to make use of the carbon and sulphur source DMSP, known to be produced by *S. marinoi* [14]. Six of the seven species in this study show additional evidence of interaction with a eukaryote, in that they are predicted to produce the essential eukaryotic membrane phospholipid phosphatidylcholine [20]. All four Rhodobacteraceae species identified in this study also contain biosynthesis genes for quorum sensing molecules; among other roles, quorum sensing is involved in regulation of biofilm production and virulence, so the presence of these genes may relate to host adhesion or parasitism [18]. In addition, all seven species contain genes indicating potential antibiotic and/or heavy metal resistance mechanisms which, depending on their method of action, may benefit both bacterium and host. The abundance of antibiotic and mercury resistance genes in particular, along with the fact that six of the seven species harbour plasmids, appears to be consistent with a previous study which found a higher occurrence of these three features, along with pigmentation, in bacteria from surface water (sampling depth 2µm) than in those

taken from deeper waters (sampling depth 0.5m) [128]. As a photosynthetic organism, *S. marinoi* must come near to the surface for sunlight, and thus it stands to reason that its microbiome would have adapted to survive in this surface water environment.
It is important to bear in mind that the findings of Pathway Tools are predictions; in some cases certain parts of the pathways are missing, although this may be due to incomplete annotation as many 'hypothetical proteins' still remain in the current annotations. Further work is required to improve upon these annotations and fill in the pathway holes, as well as to determine the importance of the prophage sequences identified by PHASTER [60].
It is also important to note that these seven species do not constitute the full microbiome of *S. marinoi*; an earlier investigation into the microbiome of *S. marinoi* strain ST54 revealed three bacterial species of entirely different taxonomy to those found in this study, whose roles in the microbiome have yet to be determined [50].

---

6. Acknowledgements

Supervisor: Mats Töpel

Examiner: Anders Blomberg

Additional thanks to Tomas Larsson, Magnus Alm Rosenblad, Oskar N. Johansson, Adrian K. Clarke, Olga Kourtchenko, Anna Godhe, and the rest of the *Skeletonema marinoi* research group at the University at Gothenburg for their help and input during this study.

Analyses were performed using the Albiorix cluster at the University of Gothenburg (http://albiorix.bioenv.gu.se/).

The *Skeletonema marinoi* genome project, of which this is a derivative study, is part of IMAGO (Infrastructure for MArine Genetic model Organisms), and is funded by CeMEB (Centre for Marine Evolutionary Biology).

7. References

1. Townley, H.E., **Diatom frustules: Physical, optical, and biotechnological applications**. *The Diatom World*. Eds. Seckbach, J. & Kociolek, J.P.. Dordrecht: Springer, 2011. pp. 273-289
ISBN: 978-94-007-1327-7
(http://www.springer.com/gp/book/9789400713260)

2. Armbrust, E.V. (2009) **The life of diatoms in the world's oceans**. Nature, 459(7244): pp. 185-192
(https://dx.doi.org/10.1038/nature08057)

3. Glibert, P.M., Anderson, D.M., Gentien, P., Granéli, E. & Sellner, K.G. (2005) **The global, complex phenomena of harmful algal blooms**. Oceanography, 18(2): pp. 136–147
(https://dx.doi.org/10.5670/oceanog.2005.49)

4. Sarno, D., Kooistra, W.H.C.F., Medlin, L.K., Percopo, I. & Zingone, A. (2005) **Diversity in the genus *Skeletonema* (Bacillariophyceae). II. An assessment of the taxonomy of *S. costatum*-like species with the description of four new species**. J Phycol, 41(1): pp. 151–176
(https://dx.doi.org/10.1111/j.1529-8817.2005.04067.x)

5. Centre for Marine Evolutionary Biology website, *Skeletonema marinoi* page:
http://cemeb.science.gu.se/research/target-species-imago/skeletonema-marinoi
(Accessed 9 May 2017)

6. Kooistra, W.H.C.F., Sarno, D., Balzano, S., Gu, H., Andersen, R.A. & Zingone, A. (2008) **Global Diversity and Biogeography of *Skeletonema* Species (Bacillariophyta)**. Protist, 159(2): pp. 177-193
(https://dx.doi.org/10.1016/j.protis.2007.09.004)

7. Amin, S.A., Parker, M.S. & Armbrust, E.V. (2012) **Interactions between Diatoms and Bacteria**. Microbiol Mol Biol R, 76(3): pp. 667-684
(https://dx.doi.org/10.1128/MMBR.00007-12)

8. Schäfer, H., Abbas, B., Witte, H. & Muyzer, G. (2002) **Genetic diversity of 'satellite' bacteria present in cultures of marine diatoms**. FEMS Microbiol Ecol, 42(1): pp. 25-35
(https://dx.doi.org/10.1111/j.1574-6941.2002.tb00992.x)

9. Grossart, H-P., Levold, F., Allgaier, M., Simon, M. & Brinkhoff, T. (2005) **Marine diatom species harbour distinct bacterial communities**. Environ Microbiol, 7(6): pp. 860–873
(http://dx.doi.org/10.1111/j.1462-2920.2005.00759.x)

10. Guannel, M.L., Horner-Devine, M.C. & Rocap, G. (2011) **Bacterial community composition differs with species and toxigenicity of the diatom *Pseudo-nitzschia***. Aquat Microb Ecol, 64(2): pp. 117-133
(https://dx.doi.org/10.3354/ame01513)

11. Croft, M.T., Lawrence, A.D., Raux-Deery, E., Warren, M.J. & Smith, A.G. (2005) **Algae acquire vitamin B_{12} through a symbiotic relationship with bacteria**. Nature, 438(7064): pp. 90-93
(https://dx.doi.org/10.1038/nature04056)

12. Amin, S.A., Green, D.H., Hart, M.C., Küpper, F.C., Sunda, W.G. & Carrano, C.J. (2009) **Photolysis of iron-siderophore chelates promotes bacterial-algal mutualism**. PNAS, 106(40): pp. 17071-17076
(https://dx.doi.org/10.1073/pnas.0905512106)

13. Amin, S.A., Hmelo, L.R., van Tol, H.M., Durham, B.P., Carlson, L.T., Heal, K.R., Morales, R.L., Berthiaume, C.T., Parker, M.S., Djunaedi, B., Ingalls, A.E., Parsek, M.R., Moran, M.A. & Armbrust, E.V. (2015) **Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria**. Nature, 522(7554): pp. 98–101 (https://dx.doi.org/10.1038/nature14488)

14. Spielmeyer, A. & Pohnert, G. (2012) **Daytime, growth phase and nitrate availability dependent variations of dimethylsulfoniopropionate in batch cultures of the diatom *Skeletonema marinoi***. J Exp Mar Biol Ecol, 413: pp. 121-130 (https://dx.doi.org/10.1016/j.jembe.2011.12.004)

15. Reisch, C.R., Moran, M.A. & Whitman, W.B. (2011) **Bacterial Catabolism of Dimethylsulfoniopropionate (DMSP)**. Front Microbiol, 2:172 (https://dx.doi.org/10.3389/fmicb.2011.00172)

16. Yoshikawa, T., Nakahara, M., Tabata, A., Kokumai, S., Furusawa, G. & Sakata, T. (2008) **Characterization and expression of *Saprospira* cytoplasmic fibril protein (SCFP) gene from algicidal *Saprospira* spp. strains**. Fisheries Sci, 74(5): pp. 1109–1117 (https://dx.doi.org/10.1111/j.1444-2906.2008.01630.x)

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

18. Bassler, B.L. (1999) **How bacteria talk to each other: regulation of gene expression by quorum sensing**. Curr Opin Microbiol, 2(6): pp. 582-587 (https://dx.doi.org/10.1016/S1369-5274(99)00025-9)

19. Naviner, M., Bergé, J-P., Durand, P. & Le Bris, H. (1999) **Antibacterial activity of the marine diatom *Skeletonema costatum* against aquacultural pathogens**. Aquaculture, 174(1): pp. 15-24 (https://dx.doi.org/10.1016/S0044-8486(98)00513-4)

20. Aktas, M., Wessel, M., Hacker, S., Klüsener, S., Gleichenhagen, J. & Narberhaus, F. (2010) **Phosphatidylcholine biosynthesis and its significance in bacteria interacting with eukaryotic cells**. Eur J Cell Biol, 89(12): pp. 888-894 (https://dx.doi.org/10.1016/j.ejcb.2010.06.013)

21. Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., Verret, F., Berges, J.A., Brownlee, C., Cadoret, J-P., Chiovitti, A., Choi, C.J., Coesel, S., De Martino, A., Detter, J.C., Durkin, C., Falciatore, A., Fournet, J., Haruta, M., Huysman, M.J.J., Jenkins, B.D., Jiroutova, K., Jorgensen, R.E., Joubert, Y., Kaplan, A., Kröger, N., Kroth, P.G., La Roche, J., Lindquist, E., Lommer, M., Martin-Jézéquel, V., Lopez, P.J., Lucas, S., Mangogna, M., McGinnis, K., Medlin, L.K., Montsant, A., Secq, M-P.O-L., Napoli, C., Obornik, M., Parker, M.S., Petit, J-L., Porcel, B.M., Poulsen, N., Robison, M., Rychlewski, L., Rynearson, T.A., Schmutz, J., Shapiro, H., Siaut, M., Stanley, M., Sussman, M.R., Taylor, A.R., Vardi, A., von Dassow, P., Vyverman, W., Willis, A., Wyrwicz, L.S., Rokhsar, D.S., Weissenbach, J., Armbrust, E.V., Green, B.R., Van de Peer, Y. & Grigoriev, I.V. (2008) **The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes**. Nature, 456(7219): pp. 239-244 (https://dx.doi.org/10.1038/nature07410)

22. Karp, P., Paley, S. & Romero, P. (2002) **The Pathway Tools Software**. Bioinformatics, 18(suppl_1): pp. S225-S232
(https://dx.doi.org/10.1093/bioinformatics/18.suppl_1.S225)
Website: http://brg.ai.sri.com/ptools/
(Accessed 9 May 2017)

23. Caspi, R., Billington, R., Ferrer, L., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S. & Karp, P.D. (2016) **The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases**. Nucleic Acids Res, 44(1): pp. D471-D480
(https://dx.doi.org/10.1093/nar/gkv1164)
Website: https://metacyc.org/
(Accessed 9 May 2017)

24. Green, M.L. & Karp, P.D. (2004) **A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases**. BMC Bioinformatics, 5:76
(https://dx.doi.org/10.1186/1471-2105-5-76)

25. Zhu, L. & Wang, Q. (2015) **Numerical Mesocosm Experimental Study on Harmful Algal Blooms of Two Algal Species in the East China Sea**. Math Probl Eng, 2015: 8 pages
(https://dx.doi.org/10.1155/2015/169860)

26. SciLifeLab homepage: www.scilifelab.se
(Accessed 9 May 2017)

27. PacBio RS II website: http://www.pacb.com/products-and-services/pacbio-systems/rsii/
(Accessed 9 May 2017)

28. Chin, C-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., Turner, S.W. & Korlach, J. (2013) **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data**. Nat Methods, 10(6): pp. 563–569
(https://dx.doi.org/10.1038/nmeth.2474)
Website: https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP
(Accessed 9 May 2017)

29. Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H.J., Remington, K.A., Anson, E.L., Bolanos, R.A., Chou, H-H., Jordan, C.M., Halpern, A.L. Lonardi, S., Beasley, E.M., Brandon, R.C., Chen, L., Dunn, P.J., Lai, Z., Liang, Y., Nusskern, D.R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G.M., Adams, M.D. & Venter, J.C. (2000) **A Whole-Genome Assembly of *Drosophila***. Science, 287(5461): pp. 2196-2204
(https://dx.doi.org/10.1126/science.287.5461.2196)

30. SMRT Analysis Software website:
http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/
(Accessed 9 May 2017)

31. RS_HGAP_Assembly.2 protocol:
http://files.pacb.com/software/smrtanalysis/2.3.0/doc/smrtportal/help/Webhelp/CS_Prot_RS_HGAP_Assembly2.htm
(Accessed 9 May 2017)

32. RS_HGAP_Assembly.3 protocol:
http://files.pacb.com/software/smrtanalysis/2.3.0/doc/smrtportal/help/Webhelp/CS_Prot_RS_HGAP_Assembly3.htm
(Accessed 9 May 2017)

33. FALCON website: https://github.com/PacificBiosciences/FALCON
(Accessed 9 May 2017)

34. Comparison of FALCON and HGAP: https://github.com/PacificBiosciences/FALCON/issues/20
(Accessed 9 May 2017)

35. Gordon, D., Huddleston, J., Chaisson, M.J., Hill, C.M., Kronenberg, Z.N., Munson, K.M., Malig, M., Raja, A., Fiddes, I., Hillier, L.W., Dunn, C., Baker, C., Armstrong, J., Diekhans, M., Paten, B., Shendure, J., Wilson, R.K., Haussler, D., Chin, C-S. & Eichler, E.E. (2016) **Long-read sequence assembly of the gorilla genome**. Science, 352(6281): aae0344
(https://dx.doi.org/10.1126/science.aae0344)

36. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. & Phillippy, A.M. (2017) **Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation**. Genome Res, 27(5): pp. 722-736
(https://dx.doi.org/10.1101/gr.215087.116)
Website: https://github.com/marbl/canu
Documentation: http://canu.readthedocs.io/en/stable/
(Accessed 9 May 2017)

37. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) **Basic local alignment search tool**. J Mol Biol, 215(3): pp. 403-410
(https://dx.doi.org/10.1016/S0022-2836(05)80360-2)
Website: https://blast.ncbi.nlm.nih.gov/Blast.cgi
(Accessed 9 May 2017)

38. BlastViewer: Korilog SARL, Questembert, France (no longer officially available)
Obtained from http://download.cnet.com/BlastViewer/3000-2054_4-10703581.html
(Accessed 9 May 2017)

39. Nicholson, W.L., Leonard M.T., Fajardo-Cavazos P., Panayotova N., Farmerie W.G., Triplett E.W. & Schuerger A.C. (2013) **Complete Genome Sequence of *Serratia liquefaciens* Strain ATCC 27592**. Genome Announc, 1(4): e00548-13
(https://dx.doi.org/10.1128/genomeA.00548-13)

40. Seemann, T. (2014) **Prokka: rapid prokaryotic genome annotation**. Bioinformatics, 30(14): pp. 2068-2069
(https://dx.doi.org/10.1093/bioinformatics/btu153)
Website: http://www.vicbioinformatics.com/software.prokka.html
(Accessed 9 May 2017)

41. The UniProt Consortium (2015) **UniProt: a hub for protein information**. Nucleic Acids Res, 43(D1): pp. D204-D212
(https://dx.doi.org/10.1093/nar/gku989)
Website: http://www.uniprot.org/
(Accessed 9 May 2017)

42. Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J. & Bateman, A. (2016) **The Pfam protein families database: towards a more sustainable future**. Nucleic Acids Res, 44(D1): pp. D279-D285 (https://dx.doi.org/10.1093/nar/gkv1344)
Website: http://pfam.xfam.org/
(Accessed 9 May 2017)

43. Laslett, D. & Canback, B. (2004) **ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences**. Nucleic Acids Res, 32(1): pp. 11-16 (https://dx.doi.org/10.1093/nar/gkh152)
Website: http://mbio-serv2.mbioekol.lu.se/ARAGORN/
(Accessed 9 May 2017)

44. Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H-H., Rognes, T. & Ussery, D.W. (2007) **RNAmmer: consistent and rapid annotation of ribosomal RNA genes**. Nucleic Acids Res, 35(9): pp. 3100–3108 (https://dx.doi.org/10.1093/nar/gkm160)
Website: http://www.cbs.dtu.dk/services/RNAmmer/
(Accessed 9 May 2017)

45. Kolbe, D.L. & Eddy, S.R. (2011) **Fast Filtering for RNA Homology Search**. Bioinformatics, 27(22): pp. 3102-3109 (https://dx.doi.org/10.1093/bioinformatics/btr545)

46. Hyatt, D., Chen, G-L., LoCascio, P.F., Land, M.L., Larimer, F.W. & Hauser, L.J. (2010) **Prodigal: prokaryotic gene recognition and translation initiation site identification**. BMC Bioinformatics, 11:119 (https://dx.doi.org/10.1186/1471-2105-11-119)
Website: http://prodigal.ornl.gov/
(Accessed 9 May 2017)

47. Finn, R.D., Clements, J. & Eddy, S.R. (2011) **HMMER web server: interactive sequence similarity searching.** Nucleic Acids Res, 39(suppl_2): pp. W29-W37 (https://dx.doi.org/10.1093/nar/gkr367)

48. Bystroff, C. & Krogh, A. **Hidden Markov Models for Prediction of Protein Features**. *Protein Structure Prediction*. Eds. Zaki, M.J. & Bystroff, C.. Totowa: Humana Press, 2008. pp. 173–198 ISBN: 978-1-59745-574-9 (http://www.springer.com/la/book/9781588297525)

49. Segata, N., Börnigen, D., Morgan, X.C. & Huttenhower, C. (2013) **PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes**. Nat Commun, 4:2304 (https://dx.doi.org/10.1038/ncomms3304)
Website: https://huttenhower.sph.harvard.edu/phylophlan
(Accessed 9 May 2017)

50. Almstedt, A. (2016) **Phylotaxonomic analysis of bacteria associated with the diatom *Skeletonema marinoi***. Master's thesis (https://github.com/alvaralmstedt/mthesis/blob/master/PhylotaxonomicanalysisofbacteriaassociatedwithhediatomSkeletonemamarinoi_incl_titlepage_Almstedt_mar_2016.pdf)
(Accessed 9 May 2017)

51. FigTree (Rambaut, A.): http://tree.bio.ed.ac.uk/software/figtree/
(Accessed 9 May 2017)

52. NCBI ftp site: ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/
(Accessed 9 May 2017)

53. Woese, C.R. & Fox, G.E. (1977) **Phylogenetic structure of the prokaryotic domain: The primary kingdoms**. PNAS, 74(11): pp. 5088-5090
(https://dx.doi.org/10.1073/pnas.74.11.5088)

54. Větrovský, T. & Baldrian, P. (2013) **The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses**. PLOS ONE, 8(2): e57923
(https://dx.doi.org/10.1371/journal.pone.0057923)

55. Rice, P., Longden, I. & Bleasby, A. (2000) **EMBOSS: the European Molecular Biology Open Software Suite**. Trends Genet, 16(6): pp. 276-277
(https://dx.doi.org/10.1016/S0168-9525(00)02024-2)
Website: http://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html
(Accessed 9 May 2017)

56. Yamamoto, T. & Harayama, S. (1998) **Phylogenetic relationships of *Pseudomonas putida* strains deduced from the nucleotide sequences of *gyrB*, *rpoD* and 16S rRNA genes**.
Int J Syst Evol Microbiol, 48(3): pp. 813-819
(https://dx.doi.org/10.1099/00207713-48-3-813)

57. Tayeb, L.A., Lefevre, M., Passet, V., Diancourt, L., Brisse, S. & Grimont, P.A.D. (2008) **Comparative phylogenies of *Burkholderia, Ralstonia, Comamonas, Brevundimonas* and related organisms derived from *rpoB, gyrB* and *rrs* gene sequences**. Res Microbiol, 159(3): pp. 169-177
(https://dx.doi.org/10.1016/j.resmic.2007.12.005)

58. NCBI website: https://www.ncbi.nlm.nih.gov/
(Accessed 9 May 2017)

59. Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H.U., Bruccoleri, R., Lee, S.Y., Fischbach, M.A., Müller, R., Wohlleben, W., Breitling, R., Takano, E. & Medema, M.H. (2015) **antiSMASH 3.0 - a comprehensive resource for the genome mining of biosynthetic gene clusters**. Nucleic Acids Res, 43(W1): pp. W237-W243
(https://dx.doi.org/10.1093/nar/gkv437)
Website: http://antismash.secondarymetabolites.org/
(Accessed 9 May 2017)

60. Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y. & Wishart, D.S. (2016) **PHASTER: a better, faster version of the PHAST phage search tool**. Nucleic Acids Res, 44(W1): pp. W16-W21
(https://dx.doi.org/10.1093/nar/gkw387)
Website: http://phaster.ca
(Accessed 9 May 2017)

61. Zhou, Y., Liang, Y., Lynch, K.H., Dennis, J.J. & Wishart, D.S. (2011) **PHAST: A Fast Phage Search Tool**. Nucleic Acids Res, 39(suppl_2): pp. W347-W352 (https://dx.doi.org/10.1093/nar/gkr485)
Website: http://phast.wishartlab.com/
(Accessed 9 May 2017)

62. Repository of Python scripts written for this study:
https://github.com/MattPinder/MastersProject/Scripts
(Accessed 9 May 2017)

63. Sequence_Reverse.py (Pinder, M.):
https://github.com/MattPinder/MastersProject/blob/master/Scripts/Sequence_Reverse.py
(Accessed 9 May 2017)

64. NCBI_Downloader.py (Pinder, M.):
https://github.com/MattPinder/MastersProject/blob/master/Scripts/NCBI_Downloader.py
(Accessed 9 May 2017)

65. genome_downloader.py (Almstedt, A.):
https://github.com/alvaralmstedt/python_genome_ref_collab/blob/master/genome_downloader.py
(Accessed 9 May 2017)

66. GenBank_Consensus.py (Pinder, M.):
https://github.com/MattPinder/MastersProject/blob/master/Scripts/GenBank_Consensus.py
(Accessed 9 May 2017)

67. GenBank_Consensus_Names.py (Pinder, M.):
https://github.com/MattPinder/MastersProject/blob/master/Scripts/GenBank_Consensus_Names.py
(Accessed 9 May 2017)

68. Schwartz, E. & Steinbüchel, A., **Preface**. *Microbial Megaplasmids*. Ed. Schwartz, E.. Berlin Heidelberg: Springer-Verlag, 2009. pp. v-vi
(http://www.springer.com/gp/book/9783540854661)

69. Repository of full PhyloPhlAn-generated phylogenetic trees from this study:
https://github.com/MattPinder/MastersProject/tree/master/Phylotaxonomic_Analysis
(Accessed 9 May 2017)

70. Biebl, H., Allgaier, M., Lünsdorf, H., Pukall, R., Tindall, B.J. & Wagner-Döbler, I. (2005) ***Roseovarius mucosus* sp. nov., a member of the *Roseobacter* clade with trace amounts of bacteriochlorophyll *a***. Int J Syst Evol Microbiol, 55(6): pp. 2377–2383
(https://dx.doi.org/10.1099/ijs.0.63832-0)

71. Van Trappen, S., Mergaert, J. & Swings, J. (2004) ***Loktanella salsilacus* gen. nov., sp. nov., *Loktanella fryxellensis* sp. nov. and *Loktanella vestfoldensis* sp. nov., new members of the *Rhodobacter* group isolated from microbial mats in Antarctic lakes**. Int J Syst Evol Microbiol, 54(4): pp. 1263–1269
(https://dx.doi.org/10.1099/ijs.0.03006-0)

72. Jogler, M., Chen, H., Simon, J., Rohde, M., Busse, H., Klenk, H., Tindall, B. & Overmann, J. (2013) **Description of *Sphingorhabdus planktonica* gen. nov., sp. nov. and reclassification of three related members of the genus *Sphingopyxis* in the genus *Sphingorhabdus* gen. nov.** Int J Syst Evol Microbiol, 63(4): pp. 1342-1349 (https://dx.doi.org/10.1099/ijs.0.043133-0)

73. Yoon, J-H. & Oh, T-K. (2005) ***Sphingopyxis flavimaris* sp. nov., isolated from sea water of the Yellow Sea in Korea**. Syst Evol Microbiol, 55(1): pp. 369-373 (https://dx.doi.org/10.1099/ijs.0.63218-0)

74. Ng, H.J., López-Pérez, M., Webb, H.K., Gomez, D., Sawabe, T., Ryan, J., Vyssotski, M., Bizet, C., Malherbe, F., Mikhailov, V.V., Crawford, R.J. & Ivanova, E.P. (2014) ***Marinobacter salarius* sp. nov. and *Marinobacter similis* sp. nov., Isolated from Sea Water**. PLOS ONE, 9(9): e106514 (https://dx.doi.org/10.1371/journal.pone.0106514)

75. Yoon, J., Kang, S., Lee, M. & Oh, T. (2007) **Description of *Sulfitobacter donghicola* sp. nov., isolated from seawater of the East Sea in Korea, transfer of *Staleya guttiformis* Labrenz et al. 2000 to the genus *Sulfitobacter* as *Sulfitobacter guttiformis* comb. nov. and emended description of the genus *Sulfitobacter***. Int J Syst Evol Microbiol, 57(8): pp. 1788-1792 (https://dx.doi.org/10.1099/ijs.0.65071-0)

76. Hong, Z., Lai, Q., Luo, Q., Jiang, S., Zhu, R., Liang, J. & Gao, Y. (2015) ***Sulfitobacter pseudonitzschiae* sp. nov., isolated from the toxic marine diatom *Pseudo-nitzschia multiseries***. Int J Syst Evol Microbiol, 65(1): pp. 95–100 (https://dx.doi.org/10.1099/ijs.0.064972-0)

77. Sorokin, D.Y. (1995) ***Sulfitobacter pontiacus* gen. nov., sp. nov. – a new heterotrophic bacterium from the Black Sea, specialized on sulfite oxidation**. Microbiology, 64(3): pp. 295–305 (No DOI number)

78. Labrenz, M., Collins, M., Lawson, P., Tindall, B., Braker, G. & Hirsch, P. (1998) ***Antarctobacter heliothermus* gen. nov., sp. nov., a budding bacterium from hypersaline and heliothermal Ekho Lake**. Int J Syst Evol Microbiol, 48(4): pp. 1363-1372 (https://dx.doi.org/10.1099/00207713-48-4-1363)

79. Hosoya, S. & Yokota, A. (2007) **Reclassification of *Flexibacter aggregans* (Lewin 1969) Leadbetter 1974 as a later heterotypic synonym of *Flexithrix dorotheae* Lewin 1970**. Int J Syst Evol Microbiol, 57(5): pp. 1086-1088 (https://dx.doi.org/10.1099/ijs.0.64798-0)

80. Gutierrez, T., Rhodes, G., Mishamandani, S., Berry, D., Whitman, W.B., Nichols, P.D., Semple, K.T. & Aitken, M.D. (2014) **Polycyclic aromatic hydrocarbon degradation of phytoplankton-associated *Arenibacter* spp. and description of *Arenibacter algicola* sp. nov., an aromatic hydrocarbon-degrading bacterium**. Appl Environ Microbiol, 80(2): pp. 618-628 (https://dx.doi.org/10.1128/aem.03104-13)

81. MetaCyc website - autoinducer AI-1 biosynthesis: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6157 (Accessed 9 May 2017)

82. MetaCyc website - phosphatidylcholine biosynthesis V:
https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6825
(Accessed 9 May 2017)

83. MetaCyc website - phosphatidylcholine biosynthesis VI:
https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6826
(Accessed 9 May 2017)

84. Tang, Y.Z., Koch, F. & Gobler, C.J. (2010) **Most harmful algal bloom species are vitamin B_1 and B_{12} auxotrophs**. PNAS, 107(48): pp. 20756-20761
(https://dx.doi.org/10.1073/pnas.1009566107)

85. MetaCyc website - enterobactin biosynthesis:
https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=ENTBACSYN-PWY
(Accessed 9 May 2017)

86. MetaCyc website - aerobactin biosynthesis:
https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=AEROBACTINSYN-PWY
(Accessed 9 May 2017)

87. MetaCyc website - superpathway of dimethylsulfoniopropanoate degradation:
https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6049
(Accessed 9 May 2017)

88. MetaCyc website - dimethylsulfoniopropanoate degradation I (cleavage):
https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6046
(Accessed 9 May 2017)

89. Masmoudi, S., Nguyen-Deroche, N., Caruso, A., Ayadi, H., Morant-Manceau, A., Tremblin, G., Bertrand, M. & Schoefs, B. (2013) **Cadmium, Copper, Sodium and Zinc Effects on Diatoms: from Heaven to Hell — a Review**. Cryptogamie Algol, 34(2): pp. 185-225
(https://dx.doi.org/10.7872/crya.v34.iss2.2013.185)

90. Alterio, V., Langella, E., De Simone, G., & Monti, S. M. (2015) **Cadmium-Containing Carbonic Anhydrase CDCA1 in Marine Diatom *Thalassiosira weissflogii***. Mar Drugs, 13(4): pp. 1688–1697
(https://dx.doi.org/10.3390/md13041688)

91. Wu, Y., Zeng, Y., Qu, J.Y. & Wang, W-X. (2012) **Mercury effects on *Thalassiosira weissflogii*: Applications of two-photon excitation chlorophyll fluorescence lifetime imaging and flow cytometry**. Aquat Toxicol, 110-111: pp. 133-140
(https://dx.doi.org/10.1016/j.aquatox.2012.01.003)

92. Boyd, E.S. & Barkay, T. (2010) **The mercury resistance operon: from an origin in a geothermal environment to an efficient detoxification machine**. Front Microbiol, 3:349
(https://dx.doi.org/10.3389/fmicb.2012.00349)

93. Cotter, P.D., Ross, R.P. & Hill, C. (2013) **Bacteriocins — a viable alternative to antibiotics?** Nat Rev Micro, 11(2): pp. 95-105
(https://dx.doi.org/10.1038/nrmicro2937)

94. Buchan, A., González, J.M. & Moran, M.A. (2005) **Overview of the Marine Roseobacter Lineage**. Appl Environ Microbiol, 71(10): pp. 5665-5677 (https://dx.doi.org/10.1128/AEM.71.10.5665-5677.2005)

95. MetaCyc website - CDP-diacylglycerol biosynthesis II: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY0-1319 (Accessed 9 May 2017)

96. MetaCyc website - 5,6-dimethylbenzimidazole biosynthesis I (aerobic): https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5523 (Accessed 9 May 2017)

97. MetaCyc website - adenosylcobalamin salvage from cobalamin: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6268 (Accessed 9 May 2017)

98. MetaCyc website - cob(II)yrinate *a,c*-diamide biosynthesis II (late cobalt incorporation): https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-7376 (Accessed 9 May 2017)

99. MetaCyc website - adenosylcobalamin salvage from cobinamide I: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=COBALSYN-PWY (Accessed 9 May 2017)

100. MetaCyc website - indole-3-acetate biosynthesis IV (bacteria): https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5025 (Accessed 9 May 2017)

101. MetaCyc website - indole-3-acetate biosynthesis II: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-581 (Accessed 9 May 2017)

102. Riedel, T., Spring, S., Fiebig, A., Scheuner, C., Petersen, J., Göker, M. & Klenk, H-P. (2015) **Genome sequence of the *Roseovarius mucosus* type strain (DSM 17069^{T}), a bacteriochlorophyll *a*-containing representative of the marine *Roseobacter* group isolated from the dinoflagellate *Alexandrium ostenfeldii***. Stand Genomic Sci, 10(17) (https://dx.doi.org/10.1186/1944-3277-10-17)

103. Mulligan, C., Fischer, M. & Thomas, G.H. (2011) **Tripartite ATP-independent periplasmic (TRAP) transporters in bacteria and archaea**. FEMS Microbiol Rev, 35(1): pp. 68-86 (https://dx.doi.org/10.1111/j.1574-6976.2010.00236.x)

104. Töpel, M., Pinder, M.I.M., Johansson, O.N., Kourtchenko, O., Godhe, A. & Clarke, A.K. (2017) **Genome sequence of *Roseovarius mucosus* strain SMR3, isolated from a culture of the diatom *Skeletonema marinoi***. Genome Announc, 5(22): e00394-17 (https://dx.doi.org/10.1128/genomeA.00394-17)

105. Bloh, A.H., Usup, G. & Ahmad, A. (2016) ***Loktanella* spp. Gb03 as an algicidal bacterium, isolated from the culture of Dinoflagellate *Gambierdiscus belizeanus***. Vet World, 9(2): pp. 142-146 (https://dx.doi.org/10.14202/vetworld.2016.142-146)

106. Reuber, T.L. & Walker, G.C. (1993) **Biosynthesis of succinoglycan, a symbiotically important exopolysaccharide of *Rhizobium meliloti***. Cell, 74(2): pp. 269-280 (https://dx.doi.org/10.1016/0092-8674(93)90418-P)

107. Tenson, T., Lovmar, M. & Ehrenberg, M. (2003) **The Mechanism of Action of Macrolides, Lincosamides and Streptogramin B Reveals the Nascent Peptide Exit Path in the Ribosome**. J Mol Biol, 330(5): pp. 1005-1014 (https://dx.doi.org/10.1016/S0022-2836(03)00662-4)

108. Ruiz, J.A., Bernar, E.M. & Jung, K. (2015) **Production of Siderophores Increases Resistance to Fusaric Acid in *Pseudomonas protegens* Pf-5**. PLOS ONE, 10(1): e0117040 (https://dx.doi.org/10.1371/journal.pone.0117040)

109. Tung, T.T., Jakobsen, T.H., Dao, T.T., Fuglsang, A.T., Givskov, M., Christensen, S.B. & Nielsen, J. (2017) **Fusaric acid and analogues as Gram-negative bacterial quorum sensing inhibitors**. Eur J Med Chem, 126: pp. 1011-1020 (https://dx.doi.org/10.1016/j.ejmech.2016.11.044)

110. Parulekar, N.N., Kolekar, P., Jenkins, A., Kleiven, S., Utkilen, H., Johansen, A., Sawant, S., Kulkarni-Kale, U., Kale, M. & Sæbø, M. (2017) **Characterization of bacterial community associated with phytoplankton bloom in a eutrophic lake in South Norway using 16S rRNA gene amplicon sequence analysis**. PLOS ONE, 12(2): e0173408 (https://dx.doi.org/10.1371/journal.pone.0173408)

111. MetaCyc website - phosphatidylethanolamine biosynthesis I: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5669 (Accessed 9 May 2017)

112. MetaCyc website - erythromycin A biosynthesis: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-7108 (Accessed 9 May 2017)

113. Hegemann, J.D., Zimmermann, M., Xie, X. & Marahiel, M.A. (2015) **Lasso Peptides: An Intriguing Class of Bacterial Natural Products**. Acc Chem Res, 48(7): pp. 1909-1919 (https://dx.doi.org/10.1021/acs.accounts.5b00156)

114. D'Costa, P.M. & Anil, A.C. (2011) **The effect of bacteria on diatom community structure – the 'antibiotics' approach**. Res Microbiol, 162(3): pp. 292-301 (https://dx.doi.org/10.1016/j.resmic.2010.12.005)

115. Green, D.H., Bowman, J.P., Smith, E.A., Gutierrez, T. & Bolch, C.J.S. (2006) ***Marinobacter algicola* sp. nov., isolated from laboratory cultures of paralytic shellfish toxin-producing dinoflagellates**. Int J Syst Evol Microbiol, 56(3): pp. 523–527 (https://dx.doi.org/10.1099/ijs.0.63447-0)

116. Gutierrez, T., Whitman, W.B., Huntemann, M., Copeland, A., Chen, A., Kyrpides, N., Markowitz, V., Pillay, M., Ivanova, N., Mikhailova, N., Ovchinnikova, G., Andersen, E., Pati, A., Stamatis, D., Reddy, T.B.K., Ngan, C.Y., Chovatia, M., Daum, C., Shapiro, N., Cantor, M.N. & Woyke, T. (2016) **Genome Sequence of *Marinobacter* sp. Strain MCTG268 Isolated from the Cosmopolitan Marine Diatom *Skeletonema costatum***. Genome Announc, 4(5): e00937-16 (https://dx.doi.org/10.1128/genomeA.00937-16)

117. MetaCyc website - indole-3-acetate biosynthesis V (bacteria and fungi): https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5026 (Accessed 9 May 2017)

118. Dodge, A.D. & Harris, N. (1970) **The mode of action of paraquat and diquat**. Biochem J, 118(3): pp. 43P-44P (No DOI number)

119. Peterson, E., Fairshter, R., Morrison, J. & Cesario, T. (1981) **Effects of the herbicide paraquat dichloride on bacteria of human origin**. Appl Environ Microbiol, 41(1): pp. 327-328 (No DOI number)

120. Hünken, M., Harder, J., Kirst & G.O. (2008) **Epiphytic bacteria on the Antarctic ice diatom *Amphiprora kufferathii* Manguin cleave hydrogen peroxide produced during algal photosynthesis**. Plant Biol, 10(4): pp. 519-526 (https://dx.doi.org/10.1111/j.1438-8677.2008.00040.x)

121. MetaCyc website - adenosylcobalamin salvage from cobinamide II: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6269 (Accessed 9 May 2017)

122. MetaCyc website - indole-3-acetate biosynthesis III (bacteria): https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-3161 (Accessed 9 May 2017)

123. Tapia, J.E., González, B., Goulitquer, S., Potin, P., & Correa, J.A. (2016). **Microbiota Influences Morphology and Reproduction of the Brown Alga *Ectocarpus* sp.** Front Microbiol, 7:197 (https://dx.doi.org/10.3389/fmicb.2016.00197)

124. MetaCyc website - CDP-diacylglycerol biosynthesis I: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5667 (Accessed 9 May 2017)

125. MetaCyc website - erythromycin D biosynthesis: https://metacyc.org/META/NEW-IMAGE?type=PATHHWAY&object=PWY-7106 (Accessed 9 May 2017)

126. Possell, M., Ryan, A., Vickers, C.E., Mullineaux, P.M. & Hewitt, C.N. (2010) **Effects of fosmidomycin on plant photosynthesis as measured by gas exchange and chlorophyll fluorescence**. Photosynth Res, 104(1): pp. 49-59 (https://dx.doi.org/10.1007/s11120-009-9504-5)

127. Jehlička, J., Osterrothová, K., Oren, A. & Edwards, H.G.M. (2013) **Raman spectrometric discrimination of flexirubin pigments from two genera of *Bacteroidetes***. FEMS Microbiol Lett, 348(2): pp. 97–102 (https://dx.doi.org/10.1111/1574-6968.12243)

128. Hermansson, M., Jones, G.W. & Kjelleberg, S. (1987) **Frequency of antibiotic and heavy metal resistance, pigmentation, and plasmids in bacteria of the marine air-water interface**. Appl Environ Microbiol, 53(10): pp. 2338-2342 (No DOI number)