

Master Thesis

Genomic analysis of microbial gas vesicle gene clusters

Matthijs Smits^{1,*}, Ben Oyserman² and Marnix Medema^{3,*}

¹Department of Bioinformatics, Wageningen UR, Wageningen , 6708PB, Netherlands and ²Department of Microbiology, Leiden University, Leiden, 2311EZ, Netherlands.

* To whom correspondence should be addressed: Tel:+31653621814; E-mail:matthijs.smits@wur.nl

Received on 27 September 2019; revised on 15 October 2019; accepted on 15 November 2019

Abstract

Gas vesicle gene clusters (qvGCs) code for all proteins required to assemble a gas vesicle. Only a limited number of functional gvGCs have been established in literature, despite a wealth of gas vesicle protein (Gvps) homologues being found in the genomes of species previously not associated with gas vesicle production. By combining a variety of bioinformatic tools, clustering methods and a specialized prioritization algorithm, an attempt is made to find novel, functional gvGCs within these genomes. In this research, 315 gvGCs were found in a set of 512 genomes, using a newly developed search method and a modified version of antiSMASH. Subsequently, all qvGCs were trimmed and ranked based on several functional features, and in the ranking, many previously unidentified gvGCs came out on top. GvGCs were subjected to a BiG-SCAPE analysis, which resulted in a network and phylogeny showing the relationships among all gvGCs. The network revealed a clean separation of gvGC's occurring in Actinomycetes. In addition, a metagenomic sample from the TARA oceans project was investigated for presence of gvGCs, in which we found a significant quantity of DNA material of several specific Gvps, more so at higher depths than at lower depths. In conclusion, gvGCs are most likely present in a larger set of species than previously thought (200+), indicated by the considerable amount of Gvps occurring as clusters in a diverse set of genomes. Subsequent wet lab functional characterization of highly prioritized gvGCs can expand the knowledge surrounding gas vesicles and improve its potential in biotech applications.

1 Introduction

Microbial gas vesicles (see figure 1) are cylindrical organelles composed entirely out of protein, found in a wide range of prokaryotes, including but not limited to Cyanobacteria, Halobacteria and Bacillus bacteria (Zimmermann and Pfeifer, 2003). Micro-organisms containing gas vesicles are typically found in aqueous habitats, which is no surprise: gas vesicles reduce the cell's density and thereby provide buoyancy to the organism. Buoyancy allows the microbe to position itself vertically in the water column, so that it may optimize light and nutrient yields (Walsby, 1994). In addition to their ecological importance, buoyancy from gas vesicles may be exploited in biotechnology for harvesting biomass (DasSarma et al., 2016). Furthermore, gas vesicles have shown potential in biotech applications such as antigen presentation systems and drug delivery systems (Sremac and Stuart, 2008). Although promising for biotechnology, much of what is known about the structure and ecology of gas vesicles is from a small subset of model organisms, thereby perhaps limiting its potential in application. Hence, understanding the architecture of gvGCs and detecting previously non-established gas vesicle gene clusters is required. In turn, this will provide insight into the evolutionary, biosynthetic and phenotypic properties of these organelles.

Two proteins make up the bulk of the weight of the vesicle organelle: 90% GvpA (Gas Vesicle Protein A) and 10% GvpC (Gas Vesicle Protein C), however, there are many other genes/proteins that are involved in the production of gas vesicles (Knitsch et al., 2017). In model organisms, a gene cluster of about 8-14 genes is responsible for the formation of gas vesicles, with genes lying closely together on the chromosome or plasmid. The Gvp genes are usually located on two transcription units that are often oppositely oriented to each other, for instance in *Halobacterium salinarum*, the genes GvpACNO and GvpDEFGHIJKLM are oriented in such a way (Pfeifer, 2015). Of these genes, GvpC, GvpD, GvpE, GvpI, GvpH are not essential for gas vesicle formation, though they do affect their properties. GvpD and GvpE are shown to be regulatory proteins and affect the transcription level of Gvp genes (Scheuch et al., 2008). All other genes are hypothesized to be structurally part of the gas vesicle or aid in the structural formation of gas vesicles.

GvpA has received much interest from the scientific community due to its conserved sequence across phyla, its strongly hydrophobic nature (Walsby, 1994) and its prominent role in the gas vesicle structure. Of all Gvp genes, GvpA is the best candidate for being an indicator of gas vesicle production (Strunk et al., 2011). GvpA is a small protein of about 70-80 amino acids, and in forming the ribs of the gas vesicles, it makes up about 90% of the weight of the gas vesicle.

Gas vesicles are widely distributed and found in both bacteria and archaea. In archaea, gas vesicles are found in 2 out of 3 archaeal phyla, whereas in bacteria they are a found in 5 out of 11 bacterial phyla. In image 2, an overview of the distribution of gas vesicles in prokaryotes can be found. One striking fact is that not all of the gas vesicle forming microorganisms are exclusively aquatic organisms, take for instance *Bacillus megaterium*, which is often found in soils(van Keulen et al., 2005). Gas vesicle function in soil environments is not completely clear, though it might be related to spore dispersion (spores of bacilli species often contain gas vesicles) and maintenance of hydrostatic pressure (Pfeifer, 2012).

The current state of the art of gas vesicles in microbes is that a large number (150+) of the gas vesicle producing microbes have been observed and confirmed at being capable of gas vesicle production (Walsby, 1994), however, many of these microbes have not been sequenced. In addition, a larger number of microbes have annotated Gvp genes (400+), however, it is not clear whether these Gvp genes actually lead to gas vesicle formation. Furthermore, many of the annotated Gvp genes have not been organized nor have they been validated. For instance, a phylogenetic analysis of the full collection of annotated Gvp genes is absent from scientific literature.



Figure 1. Phylogenetic distribution of gas vesicles. Dotted lines are phyla where no gas vesicles are found whereas solid lines are phyla where gas vesicles are found.(Walsby, 1994)

Given the gaps in the research of microbial gas vesicles discussed above, we arrive at the goal of this research. In this research, we aim to create a broader understanding of the distribution and architecture of putative gvGCs in publicly available genomes. In order to achieve this, we intend to build a local data set encompassing all the genetic information associated with gas vesicles, and subsequently organize this data in such a way that it may be suitable for biological interpretation. The molecular mechanisms of a few model organisms have been well characterized, but predicting the presence or absence of gas vesicles by genomic sequences alone is not currently possible. Ideally, it would be possible to not only predict the presence of a fully functional gvGC, but also predict structural properties such as buoyancy and expression. Indeed, this is one of the greater challenges in modern biology.

2 Material and Methods

A search method was developed in bash, python and several software tools, in order to discover novel gas vesicle gene clusters. All associated code can be found on the gas vesicle github page (https://git.wur.nl/gas-vesicle-msc-thesis/gas_vesicle_git). The input consisted of a list of microbial species (appendix, table 1) that were known from literature to produce gas vesicles and a set of known, functionally characterized gas vesicle gene clusters in protein fasta format (git, combined_phyla_representers.fa). In addition, microbial species were selected using several gas vesicle Pfam architectures (appendix, table 2) from the Pfam database (Finn et al., 2013). Both species lists were combined and their associated GenBank files were downloaded. Only completely assembled records were considered, and plasmids were also included in the dataset.

2.1 Multigeneblast

A artificial cluster was created by combining a selection of diverse and functionally characterized gas vesicle proteins into one large cluster in fasta format (git, compiled-cluster.fasta). The selection revolved around picking phylogenetically distant Gvps, in an attempt to have each gas vesicle containing phylum represented in the search. The artificial cluster served as a search query for the program 'Multigeneblast' (Medema et al., 2013), which identifies multiple gene blast (Altschul et al., 1990) hits within close vicinity of each other, thereby attributing a certain likelihood to a region for it being a bio-synthetic cluster. A dataset was generated from over input GenBank genomes/plasmids using the tool MakeDB, which is required in order to run Multigeneblast. An architectural Multigeneblast run was performed by using a minimum sequence coverage of 30%, a minimum percent identity of 40% and a maximum distance between two blast hits (to be counted as belonging to the same locus) of 10kb. Multigeneblast creates an output region with all significant blast hits in it and also includes flanking regions of 20 kb on each side of the outer two blast hits. The Multigeneblast output was processed using (git, multiparser.py) by parsing the output file and referring back to the original GenBank files to extract all protein sequences found within the region.

2.2 Clustering

Many of the proteins present in the Multigeneblast output regions are probably unrelated to gas vesicles, as flanking regions were also included in the output. Flanking regions of the putative core clusters might or might not have Gyps in them. In this research, it was important to find orthologues that occur multiple times across clusters in different phylogenetic groups. In theory, all proteins that are related to structural formation of gas vesicles are likely to share some form of homology (Walsby, 1994), as their domain architectures are specialized for their function. Hence, a clustering step was performed using MMSEQS2 with a 50% sequence similarity threshold. The bins are then filtered using a custom script (git, filtering.py) based on the number of proteins in them (singletons will be unrelated to gas vesicle production), the diversity of species found within the bin, and presence of gas vesicle related keywords in annotations. Bins were assigned specific annotations based on a simple majority rule. For instance, if more than 50% of proteins carry annotations with 'GvpN' in them, the remainder will also get labelled 'GvpN'. In this manner, both location in a multigeneblast cluster and protein homology are take into account simultaneously. A fasta file was created in which all proteins from remaining bins are merged (git, all-gvps.fasta).

2.3 EFI-EST and HMM development

Identifying the right sequences that represent both previously characterized gas vesicles proteins and novel ones is key to the success of this research. For this reason, a visualization of a sequence similarity network from

the proteins from step 2.2 was made, using the EFI-EST pipeline with an E-value filter type and filter value of 10. The output network files were visualized using Cytoscape (git, combined_fastaFullNetwork_44441.xgmml). In doing so, it allowed for a manual curation of suitable protein sequences/protein nodes to be used in the creation of HMMs, paying specific attention to species diversity (at least 3 different genera should be found in the group), Gvp label consensus (more than 90% should carry similar labels) and number of sequences (at least 30). The selected 10 nodes were the best Gvp representors that could be found in this dataset based on said criteria. Selected sequences were aligned using ClustalW and pH-MMs were created using HMMER3, at least 30 sequences were used for each model, 10 models were created in total. HMMs were created in order to capture the essence of gvGC sequences, so that sensitive detection of novel gvGCs can be achieved in successive steps. The following annotations were represented by the pHMMs: GvpA, GvpG, GvpG1, GvpG2, GvpK, GvpL-F, GvpN, GvpO, GvpR and GvpU.

2.4 antiSMASH modification

AntiSMASH, a tool specialized in identifying bio-synthetic gene clusters for secondary metabolites, was installed, modified and run in a virtual environment. The modification involved removing all original rules for cluster detection, and replacing it by one rule that allowed for gas vesicle gene clusters to be detected. The rule combines all gas vesicle related pHMMs created in step 2.3 and 7 additional gas vesicle related pHMMs acquired from pFam, leading to a total of 17 pHMMs used. In order to qualify for a cluster, the antiSMASH rule determines whether two pHMM matches occurred within a vicinity of at least 20 genes from each other. The rule was formatted like so:

...(GVPG-2-cytoscape and GVPO-1-cytoscape) or (GVPO-1-cytoscape and GVPU-1-cytoscape)...etc

for each possible pair of pHMMs (17^2) . The modified version of antiSMASH was used to search throughout the same 512 genbank files that served as an input for the initial Multigeneblast cluster identification step in order to observe a potential increase/decrease in cluster detection sensitivity. A recall check was performed by checking whether all literature verified gas vesicle producers present in the genome dataset were also found in the antiSMASH output.

2.5 Trimming

AntiSMASH outputs a GenBank formatted region containing the cluster. By default, antiSMASH outputs an arbitrary flanking region of 20 genes on each side of the outer two HMM hits of the cluster. Therefore, the true gas vesicle cluster size is generally smaller than the region created in the output, which can have adverse consequences for further processing and provides an unrealistic representation of the true cluster (git, trim.py). Automated trimming was based on multiple criteria based on investigation of verified functional gvGC's, and included features such as length of intergenic regions (should not exceed 200 bp), sense (operons tend to be on the same strand), and presence of HMM hits. For each HMM hit a search window starts working outwards in a 'greedy' fashion to see whether the genes meet the proper conditions to be part of the cluster.

2.6 Hydropathicity Analysis

The bulk of the gas vesicle structure and function is made up by GvpA. GvpA is highly hydrophobic (Walsby, 1994), which is related to its function as a gas barrier: the hydrophobic forces are hypothesized to keep fluid out of the gas vesicle. As functional GvpAs are hydrophobic, this feature could provide a useful screen to identify functional gvGCs. Therefore, the hydropathicity index for proteins in the gas vesicle dataset were investigated further. Index calculation was based on the method developed by Kyte and Doolittle (1982). First, a genome wide calculation of hydropathicity was performed on proteins in all genomes in the dataset, as to gain insight into the typical distribution of protein hydropathicity in microbes present in the dataset. Second, specific proteins such as GvpA were tested for their hydrophaticity in comparison to typical proteins, in an attempt to prove or disprove the claim that GvpA's have a strongly hydrophobic character.



Figure 2. Scoring system applied in order to prioritize gvGC's

2.7 Prioritization

Most gas vesicle gene clusters in the dataset have never been functionally characterized. Therefore, to assess which clusters are likely to be functional, prioritization of gvGC is required, thereby paving the way for future wet lab gas vesicle verification. For prioritization, a score was assigned to each cluster based on criteria that were met in verified gas vesicle clusters. These criteria are: presence of GvpA or GvpA2; presence of GvpN; total number of HMM hits; presence of at least one strongly hydrophobic protein (≥ 0.6 , based on section 3.3); gas vesicle related annotations. For each criterion, points are awarded which are indicative for the presence of gas vesicle related features. The scoring system is presented as a flowchart in figure 2.6.

2.8 BiG-SCAPE

To visualize and phylogenetically characterize the trimmed antiSMASH output, a network was built using BiG-SCAPE (Navarro-Muñoz et al., 2018), a tool that is capable of defining a distance metric between gene clusters. The distance metric of BiG-SCAPE is based on Jaccard distance, Adjacency Index and Domain Sequence Similarity. Clan cutoff (cut-off parameters for which clustering families into clans will be performed in raw distance) was set to 0.1, 0.25, 0.3, 0.5 and 1.0. The output provides a network between all clusters and a visualization of phylogenetic relationships among clusters. Further exploration and manipulation of the network is possible using Cytoscape.

2.9 Mutual exclusivity

Mutual exclusivity between one ore more Gvps could indicate a similar function. Gas vesicles require specific building blocks in order to be assembled, these building blocks are encoded in gene clusters. Only if all building blocks are present does production of a gas vesicle become possible. The HMM hits were generated for sensitive detection of these



Figure 3. Quantity of Gvps found in different phyla. All Gvp occurrences are corrected by the amount of refseq genomes the given phylum has in the NCBI database. Top of y-axis is 100%

building blocks and can therefore provide insight into the composition of these clusters. To this end, an overview was generated to see which HMM hits are found in which clusters.

2.10 TARA oceans

Gas vesicles are an an abundant organelle in nature. Especially in aquatic environments, gas vesicles are hypothesized to be a common form of motility for many micro-organisms. In an attempt to demonstrate presence in nature of putative gvGCs detected in this research, a metagenomic sample from the TARA oceans project is used. In the TARA oceans project, a schooner went on a voyage and took over 50 metagenomic samples from the oceans, 12 samples from this project were used in this research. The 12 samples contain the genomic DNA content of the microbiome at 3 specific depths, 4 replicates were sequenced from each depth. All samples were taken at approximately -8.9971 latitude, -139.1963 longitude. In order to test whether the genomes of these microbes contain any gvGC's, the raw reads of the samples were mapped to antiSMASH output regions using a pipeline developed by Koen van den Berg (Thesis student). RPKM and coverage values are calculated and represented in the form of a heatmap.

3 Results

3.1 Statistical summary of preliminary analysis

In total, 2126 microbial species names were selected based on presence of pFam/HAMAP domains. In addition, 86 species from verified gas vesicle producers established in literature were added, of which only 14 had sequenced genomes. Running a script which downloads genomes based on species name, resulted in only 512 genomes downloaded due to the lack

of assemblies or incomplete assemblies. For future analysis incomplete assemblies could also be considered. Running multigeneblast resulted in a collection of approximately 3000 genes in genbank format. Clustering with MMSEQS2 yielded 504 clusters, though about 20 percent of them were singletons. After filtering (removing clusters<3 genes, removing monophyletic clusters, removing non-gvp clusters), about 340 cluster remained for further processing. Annotations were propagated among clusters, i.e. prominent keywords/annotations were appointed to hypothetical and unknown proteins, and all information gained in the clustering step was added to the annotations of the respective proteins. All clustering bins were then merged together into one fasta file (git, all-gvps.fasta), which now contains all the proteins. All proteins were subjected to the EFI-EST pipeline, which provided a sequence similarity network consisting of about 100 distinct groups (git, combined_fastaFullNetwork_44441.xgmml). From the network, groups were selected which provided the basis for 10 pHMMs, which are theorized to be sensitive to detection of Gvp's.

3.2 gasSMASH

By modifying the rule setup of antiSMASH, gasSMASH was created. GasSMASH is capable of detecting gas vesicle gene clusters in any given genbank file, and outputs the core clusters and flanking regions as genbank files. In order to test the efficiency of gasSMASH in comparison to Multigeneblast, gasSMASH was run on the same 512 input genomes used in step 2.1. Instead of 121 clusters (Multigeneblast), this resulted in 315 clusters (appendix, table 3). Over 90% of clusters in the gasSMASH output contained 3 or more HMM hits. In figure 3, an overview of the various Gvps in gvGCs detected by gasSMASH and their relative abundance per phylum can be found.



Figure 4. BiGSCAPE network containing all found gvGCs. The cutoff value used in this network is 0.3. Distinct groups are labeled by numbers. Group 1, 2, 3 and 9 are all at least 95% Streptomyces. Group 4 are Bacilli. Group 5 consists of a mixture of Actinomycetes: Rhodococcus and Mycobacterium. 7 contains Cyanobacteria. Group 8 are Euryarchaeota. The gvGC phylogenetic tree of group 1 is shown on the right hand side.

3.3 Trimming

Trimming was performed in order to reduce the gasSMASH output regions to the core clusters. On average, the gasSMASH regions were reduced from 46 genes to 11 genes (median=11, $\sigma = 5.3$). In literature, typical gvGC's are described to be around 8-14 genes long, though some can reach up to 20 genes. Therefore, the average of 11 genes seems in line with literature findings.

3.4 Hydropathicity

Hydropathicity was calculated for all proteins in all 512 genomes, which resulted in the histograms shown in figure 5, as can be seen, the overall average hydropathicity (based on the Doolittle et al. index) for these proteins is -0.3672. Next, all Non-GvpA Gvp proteins that were found inside clusters were scored using a similar approach, this resulted in a comparable average of -0.3577 (not shown in figure), indicating that they do not deviate significantly from ordinary proteins in terms of hydrophaticity. Finally, only the GvpA's, picked up by gaSMASH, were scored. This resulted in a statistically significant (Wilcoxon rank-sum, p<0.05) higher average of -0.00155. In addition, the histogram from the GvpA plot shows a binomial distribution. The left bell curve has a distribution of which the peak approximately sits on -0.35, the same value as the overall average (i.e. these proteins are not clearly hydrophobic). The right 'bell curve' on the other hand, has a peak of around 0.6, which is clearly strongly hydrophobic.

3.5 Prioritization

The clusters were all scored based on specific criteria, defined in section 2.8. The average score for all clusters, was found to be 46 (median=46, σ = 14.9). The highest scoring cluster was from Streptomyces violaceoruber, which had a score of 88, and was, remarkably enough, not a species

currently established as a gas vesicle producer in literature. The lowest scoring cluster had a score of 6.

3.6 Recall

In order to validate the approach, a recall check was performed. In order to do so, it was first established which species that had been verified as gas vesicle producers in literature were also in the input dataset. Consequently, it was found that only 14 species (from a list of 86 verified species) were also present in the input dataset. The small amount of verified gas vesicle input genomes was simply due to a lack or incomplete assembly of these remaining 72 species. Of the 14 species, all 14 had at least 1 cluster represented in the final output, thereby indicating a 100% recall. Furthermore, these clusters had an average prioritization score of 63 (vs. an overall average of all clusters at 46), and were thereby all ranked among the top 30% of clusters.

3.7 Mutual exclusivity analysis

An absence/presence overview was generated of all 315 clusters. A few conclusions can be drawn from this overview. First and foremost, it is true for any cluster found in the dataset, that if GvpA or GvpA2 is present, they are never found together. These two facts indicate that most likely, GvpA and GvpA2 are essential building blocks for the gas vesicle and perform roughly the same function, so that if one is absent, the other one is always present and vice versa. Another striking fact from the overview is that GvpN is present in all verified gas vesicle producers, this could indicate an essential role of GvpN in the production of gas vesicles.

3.8 Creation of phylogenetic network

Generation of a BiG-SCAPE network provided insight into the phylogenetic relationships among gvGC's that were found. The network shows 4



Figure 5. Histograms indicating the distribution of hydropathicity indices of proteins across two datasets (Top: all proteins in all input genomes. Bottom: proteins determined to be GvpA). The red line indicates the average of -0.3672 found in the overall average of all genomes.

groups that all belong mainly to the phylum Actinomycetes, indicating that gas vesicle gene clusters from this phylum were both conserved among themselves and dominantly present in this data set. Furthermore, at a raw distance cut-off of 0.3, 78 singletons were generated, and 1223 links were formed. *Bacilli, Cyanobacteria*, and *Euryarchaeota* all presented themselves as distinct clades within the network. 10 of the 14 verified gvGCs were found inside linked groups, the others showed up as singletons.

3.9 TARA oceans

Gas vesicle gene clusters were significantly present in the genomic content of TARA oceans datasets. In total, 30 gvGCs were shown to be found in significant quantities, though only with limited coverage (<10%) of untrimmed gasSMASH regions. Furthermore, a statistically significant difference of gvGCs was found at different depths, with more genetic gvGC content being present at higher depth (5m) than at lower depths (115m, 500m).

4 Discussion

Microbial gas vesicles are ecologically relevant and ubiquitous organelles that show potential for biotechnology applications. In this research, we show that there is a larger and more diverse set of clustering Gvps in publicly available genomes than previously thought. Actinomycetes, a phylum previously not associated with gas vesicles (until recently (Huang et al., 2019)), made a significant contribution to the overall quantity of found gvGCs. However, actinomycetes are over represented in the refseq database (Pruitt et al., 2007) (in relation to cyanobacteria and archaea), creating a bias towards this phylum. Nonetheless, even after correction for this bias, actinomycetes still make a significant contribution to all gvGCs found in this research.

Finding a novel cluster based on something known is a struggle, considering the feature that made them novel and undiscovered is exactly the feature that can't be defined in the known search query. By using a combination of genomic viscinity (MGB/antiSMASH), homology (MMSEQS2), annotation information (retained from GBK files), protein signatures (HMMER) and hydropathicity, we created a more sensitive search and prioritization method than can be achieved by for instance blasting a protein in an online database. In doing so, we made selection of gvGC contenders for future research more straightforward. By combining the information learned from the networks (figure 4 and appendix, figure 1) , scoring (section 2.7 and appendix, table 3) and hmm hit overview (section 2.9), the three most interesting for future research, both in terms of being novel and likelihood of being functional, are: *Streptomyces Violaceoruber*, *Bacillus pseudofirmus* and *Yangia pacifica*.

The search method developed here remains a non-validated and exploratory rather than confirmatory method, considering functional characterization would have to take place to verify cluster functionality. In addition, wet lab characterization might prove difficult due to the possibly cryptic nature of gvGCs, making expression of the gvGC a struggle, though this might be resolved by doing heterologous expression of Gvps (Ramsay et al., 2011). The only true validation of the search method came in the form of a recall check, which, although resulted in a 100% recall, was limited in scope (14 gvGCs). Furthermore, a more complete validation should be performed by including verified negatives, i.e. non-functional gvGC's.

A BiG-SCAPE network showed a separation of gvGCs based on phylum, indicating a mostly vertical evolutionary development of gvGCs over time. However, there is a wide distribution and conservation of specific Gvps (Beard et al., 2002), suggesting that either horizontal gene transfer or may have played a large role in determining the early distribution of gvGCs, or that gvGCs may have originated in the last common ancestor of all current gas vesicle producers. Considering the clear separation of Actinomycetes, Archaea and Cyanobacteria in the network, gvGCs in each phylum are likely to be specialized/phylogenetically distant from gvGCs in other phyla. Hence, based on this alone, clusters from the actinomycetes are not guaranteed to be functional. However, loss of gene function is generally associated with a high mutation and deletion rate in microbes (Koskiniemi et al., 2012), making identification of non-functional Gvp orthologues virtually impossible after a number of generations (which did actually happen in this research). This given, in combination with a recent finding where gas vesicle production was demonstrated for the first time in a streptomyces strain (Huang et al., 2019), renders the set of gvGCs in actinomycetes discovered here to be functional a likely possibility.

From the stacked bar plot (Figure 3), it is clear that not all phyla contain the building blocks (Gvps) in the same ratios, which is remarkable considering a defined set of these building block are required for the creation of a gas vesicle. The lack of consistency could be the result of the gvGCs actually being different for each phylum, or could be caused by a failure of the pHMMs to pick up all the genes located in the gvGC. However, lacking a GvpA (figure 3, Gemmatimonadetes) will most likely indicate either an undetectable GvpA (at least by the methods presented here) or a non-functional gvGC considering its conserved nature and essential role in gas vesicle creation.

In light of the fact that hydrophobicity is likely to play a key role in GvpA function, proteins in the right peak of figure 5 are likely to be true, functional GvpA's, whereas proteins in the left peak have lost their function or a carry a faulty annotation.

Abundance analysis of TARA oceans samples indicates presence of gvGCs in the south pacific ocean, though only with surprisingly limited coverage of the cluster (<10%). Upon further investigation, it was found that in the regions where reads did map, a read depth of 50 or more was achieved, though in neighbouring parts of the cluster, no reads were found at all. Therefore, this might indicate that the entire cluster might be present in the initial sampling step, but that it might be lost in the filtering or preparation steps leading up to the sequencing. Misalignment of reads to the gvGCs is also a possibility, though this is unlikely due to the stringent read mapping parameters used in Bowtie 2. It was proven that more gvGC genetic material was present at higher depths (5m) than at lower depths (600m), which is in line with expectation: gas vesicles allow the microbes to reach higher layers of the ocean due to buoyancy.

Considering the input genomes were limited to genomes containing specific pFAM domains, future attempts at generating gvGC overviews could try to involve more broad/general sets of genomes or metagenomic assemblies. In addition, more true positives are required for further validation of this research.

5 Conclusion and Future Perspectives

In summary, a search method for finding novel gas vesicle gene clusters in microbes was developed by combining a variety of bioinformatic tools, clustering methods and custom scripts. The search method was carried out and yielded 315 clusters, of which more than 90% were not characterized. Next, clusters were trimmed to the core clusters, thereby providing clear definitions of where the hypothetical clusters begin and end. In order to make wet lab characterization more straightforward, clusters were scored based in regards to how likely they are to be functional, based on criteria that were observed in known functional clusters. The highest scoring cluster was found to be one in the genome of Streptomyces violaceoruber. A recall check was performed which revealed that all verified clusters were also detected by the search method. GvpA and GvpA2 are mutually exclusive genes present in all clusters found in this research, indicating a potentially similar function of these two proteins. Proteins labelled GvpA in this research are more hydrophobic than the overall average protein, and the extent to which any given GvpA is hydrophobic could say something about its functionality. An overview of the phylogenetic relationships that exist among all found gas vesicle gene clusters was generated using BiG-SCAPE, which showed a separation of gas vesicle gene clusters correlating to phylum, though some were intertwined. Presence of gas vesicle genes was confirmed in metagenomic samples of the TARA oceans project taken from the south pacific ocean. A statistically significant larger number of gas vesicle genetic material was found a higher depths (5m) than at lower depths (115m,600m), In future research, further characterization of found gvGC's through heterologous expression and imaging using cry-EM can be performed to validate and expand on this research. In addition, expression analysis in (meta)genomic samples can provide insight into the conditions under which gas vesicle genes are expressed.

References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403 – 410.
- Beard, S. J., Hayes, P. K., Pfeifer, F., and Walsby, A. E. (2002). The sequence of the major gas vesicle protein, gvpa, influences the width and strength of halobacterial gas vesicles. *FEMS Microbiology Letters*, 213(2):149–157.
- DasSarma, P., Karan, R., Kim, J.-M., Pecher, W., and DasSarma, S. (2016). Bioengineering novel floating nanoparticles for protein and drug

delivery. Materials Today: Proceedings, 3(2):206-210.

- Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. (2013). Pfam: the protein families database. *Nucleic Acids Research*, 42(D1):D222–D230.
- Huang, R., Lin, J., Gao, D., Zhang, F., Yi, L., Huang, Y., Yan, X., Duan, Y., and Zhu, X. (2019). Discovery of gas vesicles in streptomyces sp. cb03234-s and potential effects of gas vesicle gene overexpression on morphological and metabolic changes in streptomycetes. *Applied Microbiology and Biotechnology*, 103.
- Knitsch, R., Schneefeld, M., Weitzel, K., and Pfeifer, F. (2017). Mutations in the major gas vesicle protein gvpa and impacts on gas vesicle formation in haloferax volcanii. *Mol Microbiol*, 106(4):530–542.
- Koskiniemi, S., Sun, S., Berg, O. G., and Andersson, D. I. (2012). Selection-driven gene loss in bacteria. *PLOS Genetics*, 8(6):1–7.
- Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–32.
- Medema, M. H., Takano, E., and Breitling, R. (2013). Detecting sequence homology at the gene cluster level with multigeneblast. *Molecular biology and evolution*, 30(5):1218–1223.
- Navarro-Muñoz, J. C., Selem-Mojica, N., Mullowney, M. W., Kautsar, S., Tryon, J. H., Parkinson, E. I., De Los Santos, E. L., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernandez-Guerra, A., Dias Cappelini, L. T., Thomson, R. J., Metcalf, W. W., Kelleher, N. L., Barona-Gomez, F., and Medema, M. H. (2018). A computational framework for systematic exploration of biosynthetic diversity from large-scale genomic data. *bioRxiv*.
- Pfeifer, F. (2012). Distribution, formation and regulation of gas vesicles. *Nature Reviews Microbiology*, 10:705.
- Pfeifer, F. (2015). Haloarchaea and the formation of gas vesicles. *Life* (*Basel*), 5(1):385–402.
- Pruitt, K., Tatusova, T., and R Maglott, D. (2007). Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35:D61–5.
- Ramsay, J. P., Williamson, N. R., Spring, D. R., and Salmond, G. P. C. (2011). A quorum-sensing molecule acts as a morphogen controlling gas vesicle organelle biogenesis and adaptive flotation in an enterobacterium. *Proceedings of the National Academy of Sciences*, 108(36):14932–14937.
- Scheuch, S., Marschaus, L., Sartorius-Neef, S., and Pfeifer, F. (2008). Regulation of gvp genes encoding gas vesicle proteins in halophilic archaea. *Archives of Microbiology*, 190(3):333.
- Sremac, M. and Stuart, E. S. (2008). Recombinant gas vesicles from halobacterium sp. displaying siv peptides demonstrate biotechnology potential as a pathogen peptide delivery vehicle. *BMC Biotechnology*, 8(1):9.
- Strunk, T., Hamacher, K., Hoffgaard, F., Engelhardt, H., Zillig, M. D., Faist, K., Wenzel, W., and Pfeifer, F. (2011). Structural model of the gas vesicle protein gvpa and analysis of gvpa mutants in vivo. *Mol Microbiol*, 81(1):56–68.
- van Keulen, G., Hopwood, D. A., Dijkhuizen, L., and Sawers, R. G. (2005). Gas vesicles in actinomycetes: old buoys in novel habitats? *Trends in Microbiology*, 13(8):350–354.
- Walsby, A. E. (1994). Gas vesicles. Microbiological Reviews, 58(1):94.
- Zimmermann, P. and Pfeifer, F. (2003). Regulation of the expression of gas vesicle genes in haloferax mediterranei: interaction of the two regulatory proteins gvpd and gvpe. *Molecular Microbiology*, 49(3):783–794.

Appendices

Table 1. input data

Organism	Refseq assembly ID
Bacillus megaterium de Bary ATCC® 19213	NZ_AVBB00000000.1
Dactylococcopsis salina	NC_019780.1
Desulfotomaculum acetoxidans	NC_013216.1
Enhydrobacter aerosaccus	NZ_FUWJ0000000.1
Halobacterium salinarum	NC_002607.1
Haloferax mediterranei	NC_017941.2
Haloferax volcanii	NC_013967.1
Isosphaera pallida	NC_014962.1
Microcystis aeruginosa PCC 7806	NZ_CP020771.1
Planktothrix agardhii	NZ_ASAK0000000.1
Planktothrix rubescens GV3	NZ_AVFW00000000.1
Polaribacter filamentus	NZ_MQUA0000000.1
Serratia sp. ATCC39006	NZ_CP025084.1
Thiocapsa rosea	NZ_RBXL00000000.1
Trichormus variabilis SAG 1403-4b	NZ_RSCM00000000.1
Chloroherpeton thalassium	NC_011026.1
Haloquadratum walsbyi	NC_017459.1
Ancylobacter aquaticus	NZ_RXIO0000000.1
Aphanizomenon flos-aquae	NZ_AZYY00000000.1

Table 2. Pfam architectures used

HMM name	Evidence Codes Ontology	RULEBASE / EMBL	Nr. of residues
Gas vesicle structural protein	0000256	RU000632	70
Gas vesicle synthesis GvpLGvpF	0000313	AHG88481.1	262
Gas vesicle protein GvpC	0000313	EMBL:BAI91032.1	284



Figure 1: Cytoscape network of gvGCs. Size of nodes corresponds to the index of most hydrophobic protein in gvGC, color represents overall score of gvGC (purple=high ,red=medium, yellow=low), an edge represents a significant similarity between the two gvGCs.



Figure 2: Heatmap of the DNA abundance in TARA oceans samples. Condition is 5m and 600m, each depth has 4 replicates. Household genes are portrayed on the right.

Table 3: All found gvGCs

Organism	Accession	Score	Coordinates cluster in genome	Nr of genes	Verified?
Streptomyces violaceoruber	NZ_CP020570.1.region001.gbk	88	280134 280614	26	
Rhodococcus jostii	NC_008268.1.region002.gbk	79	3719849 3720452	29	
Yangia pacifica	NZ_CP022189.1.region001.gbk	76	1799480 1799915	29	
Anabaena sp.	NZ_CP011456.1.region001.gbk	75	1793726 1793942	16	yes
Anabaena sp.	NC_019427.1.region001.gbk	74	3267138 3267354	15	
Methanosaeta thermophila	NC_008553.1.region001.gbk	73	49688 50306	19	yes
Trichodesmium erythraeum	NC_008312.1.region001.gbk	72	3597439 3597628	18	yes
Bacillus pseudofirmus	NC_013791.2.region001.gbk	70	3797360 3797753	16	
Bacillus horikoshii	NZ_CP020880.1.region001.gbk	69	3382769 3384140	16	
Salimicrobium jeotgali	NZ_CP011361.2.region001.gbk	68	610676 611150	20	
Planktothrix agardhii	AP017991.1.region001.gbk	67	2815273 2815492	14	yes
Bradyrhizobium oligotrophicum	NC_020453.1.region001.gbk	66	1729103 1730159	15	
Desulfobacca acetoxidans	NC_015388.1.region001.gbk	66	1991298 1991607	23	
Methanosarcina sp.	NZ_CP009524.1.region001.gbk	65	3858641 3858872	15	
Rhodobacter sphaeroides	NC_009050.1.region001.gbk	65	322890 323103	15	
Bradyrhizobium sp.	NC_009445.1.region001.gbk	65	1356537 1357557	17	
Rhodobacter sphaeroides	NC_011958.1.region001.gbk	65	683345 683558	15	
Desulfomonile tiedjei	NC_018025.1.region001.gbk	65	5705216 5706644	20	
Streptomyces sp.	NZ_CP026121.1.region001.gbk	65	7342802 7344836	25	
Virgibacillus phasianinus	NZ_CP022315.1.region001.gbk	64	1577763 1578024	14	
Streptomyces sp.	NZ_CP015098.1.region001.gbk	64	7614064 7616080	24	
Rhodobacter sphaeroides	NZ_CP030272.1.region001.gbk	64	668257 668470	15	
Methanosarcina barkeri	NZ_CP009517.1.region001.gbk	64	3757979 3758210	15	yes
Streptomyces actuosus	NZ_CP029788.1.region001.gbk	64	7910896 7912585	24	
Streptomyces lincolnensis	NZ_CP022744.1.region001.gbk	64	8289994 8291719	24	
Streptomyces chartreusis	NZ_LT962942.1.region001.gbk	64	1967216 1968038	24	
Geobacter uraniireducens	NC_009483.1.region001.gbk	63	3286966 3288424	16	
Halobacterium salinarum	NC_010368.1.region001.gbk	63	69452 69674	17	yes
Bacillus megaterium	NC_014103.1.region001.gbk	63	3216256 3216673	14	yes
Haloferax mediterranei	NC_017941.1.region001.gbk	63	1614946 1615618	16	yes
Bacillus megaterium	NZ_CP009920.1.region002.gbk	63	169062 169479	14	
Streptomyces griseorubiginosus	NZ_CP032427.1.region001.gbk	63	1937631 1938018	23	
Bacillus megaterium	NC_014019.1.region002.gbk	63	3217016 3217433	14	yes
Candidatus Desulforudis	NC_010424.1.region001.gbk	63	1570824 1571163	15	
Halobacterium sp.	NC_002608.1.region002.gbk	63	16450 16705	17	
Psychromonas ingrahamii	NC_008709.1.region001.gbk	63	2130457 2130676	17	
Bacillus aryabhattai	NZ_CP024035.1.region001.gbk	62	3181381 3181798	14	
Streptomyces leeuwenhoekii	NZ_LN831790.1.region001.gbk	62	1580712 1581114	22	
Streptomyces cyaneogriseus	NZ_CP010849.1.region001.gbk	62	6563986 6565549	22	
Streptomyces pactum	NZ_CP019724.1.region002.gbk	62	604456 604915	22	
Streptomyces ambofaciens	NZ_CP012382.1.region002.gbk	62	7419007 7421287	22	
Bacillus megaterium	NZ_CP032527.1.region001.gbk	62	2088811 2089072	14	
Streptomyces sp.	NZ_CP013743.1.region003.gbk	62	9601604 9601970	22	
Microcystis aeruginosa	NZ_CP020771.1.region001.gbk	62	1900766 1900982	14	yes
Burkholderia sp.	NZ_CP013387.1.region001.gbk	61	2008037 2008307	16	
Thermacetogenium phaeum	NC_018870.1.region001.gbk	61	1078045 1078282	15	
Nostoc sp.	NZ_CP023278.1.region001.gbk	61	1463867 1465706	13	
Methanosarcina vacuolata	NZ_CP009520.1.region001.gbk	61	3958044 3958590	14	
Burkholderia oklahomensis	NZ_CP013356.1.region001.gbk	61	773541 774264	16	
Desulfotomaculum acetoxidans	NC_013216.1.region001.gbk	61	3216949 3217165	16	
Microcystis aeruginosa	NC_010296.1.region001.gbk	61	3400499 3400715	13	
Haloquadratum walsbyi	NC_017459.1.region001.gbk	61	971795 972014	14	
Burkholderia sp.	NZ_CP013458.1.region001.gbk	61	799369 800128	16	
Methanosarcina barkeri	NC_007355.1.region001.gbk	60	386176 386407	14	yes
Streptomyces glaucescens	NZ_CP009438.1.region001.gbk	60	6248018 6249872	20	-
Streptomyces formicae	NZ_CP022685.1.region001.gbk	60	7969801 7970743	20	
Methanosarcina barkeri	NZ_CP009526.1.region001.gbk	60	796472 797018	14	yes
Streptomyces sp.	NZ_CP009754.1.region001.gbk	60	1368922 1369237	22	

Table	3:	All	found	gvGCs
-------	----	-----	-------	-------

Organism	Accession	Score	Coordinates cluster in genome	Nr of gapas	Varified?
Streptomyces pactum	NZ CP010724 1 region001 gbk	50	604456 604015	21	vermeu:
Natronomonas moolanensis	NC 020388 1 region001 gbk	59	727153 727531	15	
Microcystis sp	NZ_CP020664_1 region001_gbk	59	1961901 1962117	13	
Microcystis sp.	NZ_AP019314_1 region001_gbk	59	1313222 1313/38	14	
Streptomyces lividans	NZ_CP009124.1 region002 gbk	59	1316695 1316992	21	
Dactylococcopsis salina	NC 010780 1 region001 gbk	58	1300003 1400581	15	
Halogeometricum boringuense	NC_014735_1_region001_gbk	58	324844 325123	15	Vec
Streptomyces parvulus	NZ CP015866 1 region001 gbk	58	524644 525125	22	yes
Microcystis aeruginosa	NZ_CP011304_1 region001_gbk	58	2499010 2500855	13	Vec
Natronolimnobius aegyptiacus	NZ_CP019893.1 region001.gbk	58	184538 184862	13	yes
Halobacterium salinarum	NC 010366 1 region001 gbk	58	16450 16705	16	Vec
Halobacillus mangrovi	NZ_CP020772_1_region001_gbk	58	1203301 1203808	13	yes
Haloquadratum walshvi	NC 008212 1 region001 gbk	57	952003 952222	14	Vec
Oceanobacillus ibevensis	NZ_CP020357_1_region001_gbk	57	32003 352222	14	yes
Halobacillus halophilus	NC 017668 1 region001 gbk	57	3090/37 3090803	17	
Virgibacillus	NC_017000.1.1cgl01001.g0k	57	5070+57 5070805	12	
halodenitrificans	NZ_CP033049.1.region001.gbk	57	1601321 1601729	13	
Halobacterium sp	NC 001860 1 region 001 gbk	57	16450 16705	15	
Virgibacillus sp	NZ_CP007161_1 region001_gbk	57	1462715 1463201	13	
Virgibacinus sp.	NZ_CP00/101.1.1egion001.gbk	57	4402645 4402410	13	
	NZ_CP020370.1.1egion001.gbk	57	250625 260000	10	
Serratia sp.	NC_002608.1 region001.gbk	57	16450 16705	21	
Anghagna yarishilia	NC_007412.1 region001.gbk	57	01210 02052	13	
Anabaena variabilis	NC_00/413.1.region001.gbk	50 56	91210 93052	12	yes
Streptomyces sp.	NZ_CP013142.1.region002.gbk	50 56	7357300 7359814	20	
Streptomyces amboraciens	NZ_CP012382.1.region005.gbk	50	757900 759(22	16	
Burkholderia sp.	NZ_CP013388.1.region001.gbk	56	/5/899 /58622	16	
Burkholderia sp.	NZ_CP009548.1.region001.gbk	56	57325 57595	16	
Jeotgalibacillus malaysiensis	NZ_CP009416.1.region001.gbk	50	3043065 3043443	13	
Nostoc sp.	NC_003272.1.region001.gbk	55 55	2696511 2698353	12	
Oceanobacillus sp.	NZ_CP024848.1.region001.gbk	55 55	360222 360798	12	
Natrialba magadii	NC_013922.1.region001.gbk	55 55	319911 322665	14	
Nodularia spumigena	NZ_CP020114.1.region001.gbk	55 55	981434 983270	12	
Streptomyces parvulus	NZ_CP015866.1.region002.gbk	55 54	5982246 5982756	17	
Haloterrigena turkmenica	NC_013/43.1.region001.gbk	54	2435960 2436389	12	
Rhodococcus hoagii	NZ_CP02//93.1.region001.gbk	54	unknown unknown	20	
Halopiger xanaduensis	NC_015666.1.region001.gbk	54	130816 131146	12	
Streptomyces avermitilis	NC_003155.5.region002.gbk	54	746751 747162	18	
Halobiforma lacisalsi	NZ_CP019285.1.region001.gbk	54	2273404 2273839	13	
Virgibacillus necropolis	NZ_CP022437.1.region001.gbk	54	659985 660174	12	
Isosphaera pallida	NC_014962.1.region001.gbk	53	10/2338 10/3/5/	13	yes
Streptomyces alboflavus	NZ_CP021/48.1.region001.gbk	53	7/11333 7/12698	18	
Nostoc sp.	AP018318.1.region001.gbk	53	4204906 4205308	12	
Natrinema pellirubrum	NC_019962.1.region001.gbk	53	1590210 1590564	11	
Streptomyces atratus	NZ_CP02/306.1.region001.gbk	53	unknown unknown	16	
Haloterrigena daqingensis	NZ_CP019327.1.region001.gbk	53	14/5595 14/5892	12	
Natronococcus occultus	NC_0199/4.1.region001.gbk	52	2/19/68 2/20320	11	
Halostagnicola larsenii	NZ_CP00/055.1.region001.gbk	52	1882/93 1883366	11	
Nostoc punctiforme	NC_010628.1.region001.gbk	52	2620120 2620564	12	
Streptomyces griseochromogenes	NZ_CP016279.1.region002.gbk	52	5045657 5045831	16	
Mycobacterium goodii	NZ_CP012150.1.region001.gbk	52	2526164 2526968	14	
Natronolimnobius sp.	NZ_CP024047.1.region001.gbk	52	1540558 1540849	14	
Mycolicibacterium smegmatis	NZ_CP02/541.1.region001.gbk	52	5946604 5946850	14	
Natronolimnobius sp.	NZ_CP02/033.1.region001.gbk	52	2157044 2159777	14	
Pelotomaculum	NC_009454.1.region001.gbk	52	410386 410653	12	
thermopropionicum			// /// ///////////////////////////////		
Bacillus filamentosus	NZ_CP011974.1.region001.gbk	52	615712 616891	12	
Streptomyces puniciscabiei	NZ_CP017248.1.region001.gbk	51	2209180 2209462	13	
Streptomyces sp.	NZ_CP011492.1.region001.gbk	51	7708905 7709178	17	

Table 3: All found gvGCs

Organism	Accession	Score	Coordinates cluster in genome	Nr of genes	Verified?
Streptomyces hygroscopicus	NC_017765.1.region001.gbk	51	8489078 8490623	13	
Candidatus Koribacter	NC_008009.1.region001.gbk	51	2840287 2843683	19	
Amycolatopsis methanolica	NZ_CP009110.1.region001.gbk	51	3827330 3827702	13	
Streptomyces hygroscopicus	NZ_CP013219.1.region001.gbk	51	7141124 7142696	13	
Streptomyces hygroscopicus	NC_020895.1.region001.gbk	51	8183164 8184709	13	
Natronobacterium gregoryi	NC_019792.1.region001.gbk	51	3011898 3012180	11	
Calothrix sp.	NZ_AP018172.1.region001.gbk	50	4093659 4094376	9	
Halothece sp.	NC_019779.1.region001.gbk	50	542683 542944	10	
Streptomyces sp.	NZ_CP029338.1.region001.gbk	50	1839682 1839964	12	
Leptolyngbya sp.	NZ_AP017308.1.region001.gbk	50	4278460 4278679	13	
Streptomyces koyangensis	NZ_CP031742.1.region002.gbk	49	3239338 3239851	11	
Streptomyces violaceusniger	NC_015957.1.region002.gbk	48	3823302 3824157	14	
Streptomyces nigra	NZ_CP029043.1.region001.gbk	48	6177931 6178339	14	
Mycobacterium sp.	NC_009077.1.region001.gbk	48	unknown unknown	11	
Streptomyces sp.	NZ_CP021121.1.region001.gbk	48	5551478 5551760	10	
Myxococcus xanthus	NC_008095.1.region001.gbk	48	3220620 3221007	10	
Mycobacterium sp.	NC_008146.1.region001.gbk	48	2481832 2482249	11	
Mycobacterium sp.	NC_008705.1.region001.gbk	48	2500816 2501233	11	
Halalkalicoccus jeotgali	NC_014297.1.region003.gbk	48	1910294 1910558	15	
Streptomyces sp.	NZ_CP021118.1.region001.gbk	48	5753395 5753692	10	
Mycobacterium sp.	NC_019966.1.region001.gbk	48	unknown unknown	10	
Streptomyces avermitilis	NC_003155.5.region003.gbk	48	746751 747162	10	
Pseudonocardia dioxanivorans	NC_015312.1.region001.gbk	48	4566525 4566825	10	
Streptomyces rubrolavendulae	NZ_CP017316.1.region001.gbk	47	124387 124666	9	
Streptomyces lydicus	NZ_CP029042.1.region001.gbk	47	1556808 1556982	9	
Streptomyces lividans	NZ_CP009124.1.region001.gbk	47	1316695 1316992	13	
Streptomyces sp.	NZ_CP015849.1.region001.gbk	47	6189754 6191485	13	
Streptomyces olivaceus	NZ_CP016795.1.region001.gbk	47	6748676 6750491	13	
Streptomyces sp.	NZ_CP0300/3.1.region001.gbk	47	5830302 5830/13	10	
Streptomyces sp.	NZ_CP026730.1.region001.gbk	47	6377577 6379506	13	
Streptomyces sp.	NZ_CP009754.1.region002.gbk	47	1368922 1369237	13	
Plantactinospora sp.	NZ_CP028159.1.region002.gbk	47	/499663 /500491	10	
Mathylagystia hyverhile	NZ_LN99/842.1.region001.gbk	47	8042123 8043953	13	
Streate weeks a slikes	NZ_CP019948.1.region001.gbk	47	3462/31 3463239	14	
Streptomyces albus	NZ_CP0330/1.1.region002.gbk	47	4429301 4429625	13	
Muschasterium shuhuansa	NZ_CP029301.1.1egi01001.gbk	47	722600 724005	15	
Straptomygas ap	NZ_CP012742_1 region002_chk	47	723090 724003 9601604 9601970	9	
Terribacillus gorionsis	NZ_CP013743.1.1egion002.gbk	47	2108064 2108207	9	
Straptomygas ap	NZ_CP028260.1 region001.gbk	47	7517264 7517666	11	
Leptolyngbya boryana	NZ_AP018203_1_region001_gbk	47	A85177A A853568	13	
Streptomyces scabiei	NC 013929 1 region001 gbk	47	4317210 4317618	13	
Streptomyces sp	NZ_CP022545_1_region002_gbk	46	6897197 6897512	8	
Streptomyces autolyticus	NZ_CP019458 1 region002 gbk	46	6757891 6758197	8	
Pimelobacter simplex	NZ_CP009896.1 region001 gbk	46	2684240 2685212	9	
Archangium gephyra	NZ CP011509.1.region001.gbk	46	3994508 3994901	9	
Streptomyces malaysiensis	NZ CP023992.1.region002.gbk	46	2284346 2284649	8	
Streptomyces sp.	NZ CP029541.1.region002.gbk	46	1287691 1288000	12	
Streptomyces sp.	NZ CP025018.1.region001.gbk	46	2984276 2984588	12	
Streptomyces sp.	NZ CP022545.1.region003.gbk	46	6897197 6897512	12	
Streptomyces collinus	NC_021985.1.region003.gbk	46	7704825 7705230	12	
Streptomyces autolyticus	NZ_CP019458.1.region004.gbk	46	6757891 6758197	12	
Streptomyces sp.	NZ_CP015098.1.region002.gbk	46	7614064 7616080	8	
Streptomyces malaysiensis	NZ_CP023992.1.region004.gbk	46	2284346 2284649	12	
Amycolatopsis sp.	NZ_CP024894.1.region001.gbk	46	5573942 5574209	12	
Streptomyces scabiei	NC_013929.1.region002.gbk	46	4317210 4317618	15	
Streptomyces sp.	NZ_CP029342.1.region001.gbk	46	4066885 4067167	10	
Streptomyces nigra	NZ_CP029043.1.region002.gbk	46	6177931 6178339	8	

Table	3:	All	found	gvGCs
-------	----	-----	-------	-------

Organism	Accession	Score	Coordinates cluster in genome	Nr of genes	Verified?
Streptomyces sp.	NZ_CP025018.1.region003.gbk	46	2984276 2984588	8	
Streptomyces koyangensis	NZ_CP031742.1.region001.gbk	45	3239338 3239851	11	
Streptomyces sp.	NZ_CP019779.1.region001.gbk	45	724514 725474	11	
Streptomyces sp.	NZ_CP018870.1.region001.gbk	45	426387 426795	12	
Staphylococcus simulans	NZ_CP015642.1.region001.gbk	45	2307730 2308117	11	
Nonomuraea sp.	NZ_CP017717.1.region001.gbk	45	8826994 8827288	11	
Rhodococcus opacus	NZ_CP003949.1.region001.gbk	45	4749401 4749746	13	
Streptomyces niveus	NZ_CP018047.1.region001.gbk	45	699610 699934	11	
Staphylococcus hyicus	NZ_LR134264.1.region001.gbk	45	2343714 2344101	11	
Kutzneria albida	NZ_CP007155.1.region001.gbk	45	7716808 7717978	10	
Streptomyces sp.	NZ_CP033073.1.region003.gbk	44	1346573 1346876	10	
Streptomyces alfalfae	NZ_CP015588.1.region001.gbk	44	1594864 1595209	10	
Streptomyces sp.	NC_015953.1.region001.gbk	44	7274382 7274703	10	
Rhodococcus sp.	NZ_CP016819.1.region001.gbk	44	1318579 1319275	10	
Streptomyces tirandamycinicus	NZ_CP029188.1.region001.gbk	44	6672978 6673287	10	
Sphaerobacter thermophilus	NC_013523.1.region001.gbk	44	964503 965019	10	
Streptomyces vietnamensis	NZ_CP010407.1.region001.gbk	44	860467 860926	11	
Actinoalloteichus sp.	NZ_CP025990.1.region001.gbk	44	1700864 1701185	10	
Rhodococcus sp.	NZ_CP018063.1.region001.gbk	44	4235621 4236329	10	
Rhodococcus rhodochrous	NZ_LT906450.1.region002.gbk	44	2580857 2581061	10	
Pseudonocardia sp.	NZ_CP012184.1.region001.gbk	44	523646 523961	10	
Rhodococcus jostii	NC_008268.1.region001.gbk	44	3719849 3720452	10	
Oscillatoria acuminata	NC_019693.1.region001.gbk	44	193418 194357	7	
Streptomyces fradiae	NZ_CP032266.1.region001.gbk	44	1256950 1257295	10	
Streptomyces sp.	NZ_CP025407.1.region001.gbk	44	unknown unknown	11	
Streptomyces pratensis	NC_016114.1.region001.gbk	44	300815 301094	10	
Nocardiopsis dassonvillei	NC_014210.1.region001.gbk	44	2798507 2798981	10	
Rhodococcus aetherivorans	NZ_CP011341.1.region002.gbk	44	5779817 5780531	10	
Pseudonocardia sp.	NZ_CP013854.1.region001.gbk	44	78192 78504	10	
Rhodococcus ruber	NZ_CP023714.1.region002.gbk	44	1315050 1316121	10	
Prauserella marina	NZ_CP016353.1.region001.gbk	44	4478177 4479128	11	
Rhodococcus pyridinivorans	NC_023150.1.region001.gbk	44	3458908 3459649	10	
Streptacidiphilus sp.	NZ_CP031264.1.region001.gbk	44	unknown unknown	10	
Streptomyces sp.	NZ_CP019779.1.region003.gbk	44	724514 725474	10	
Rhodococcus biphenylivorans	NZ_CP022208.1.region001.gbk	44	540129 540825	10	
Rhodococcus opacus	NZ_CP003949.1.region002.gbk	44	4749401 4749746	10	
Plantactinospora sp.	NZ_CP028159.1.region001.gbk	44	7499663 7500491	10	
Streptomyces lydicus	NZ_CP029042.1.region002.gbk	43	1556808 1556982	9	
Streptomyces malaysiensis	NZ_CP023992.1.region001.gbk	43	2284346 2284649	9	
Streptomyces albulus	NZ_CP007574.1.region001.gbk	43	287915 288089	9	
Streptomyces sp.	NZ_CP025018.1.region002.gbk	43	2984276 2984588	9	
Streptomyces violaceusniger	NC_015957.1.region003.gbk	43	3823302 3824157	9	
Halovivax ruber	NC_019964.1.region001.gbk	43	1952933 1953269	10	
Rhodococcus opacus	NC_012522.1.region001.gbk	43	344566 345274	10	
Streptomyces pristinaespiralis	NZ_CP011340.1.region002.gbk	43	8069730 8070786	9	
Thermobispora bispora	NC_014165.1.region001.gbk	43	3954705 3955233	10	
Gemmatirosa kalamazoonesis	NZ_CP007128.1.region001.gbk	43	1066265 1067996	11	
Sandaracinus amylolyticus	NZ_CP011125.1.region001.gbk	43	8520609 8521041	13	
Streptomyces sp.	NZ_CP029541.1.region003.gbk	43	1287691 1288000	9	
Streptomyces autolyticus	NZ_CP019458.1.region003.gbk	43	6757891 6758197	9	
Streptomyces albireticuli	NZ_CP021744.1.region001.gbk	43	5975836 5977519	9	
Streptomyces sp.	NZ_CP025018.1.region004.gbk	43	2984276 2984588	9	
Streptomyces autolyticus	NZ_CP019458.1.region001.gbk	43	6757891 6758197	9	
Streptomyces peucetius	NZ_CP022438.1.region002.gbk	43	490583 491630	9	
Stackebrandtia nassauensis	NC_013947.1.region001.gbk	43	5007795 5008347	9	
Pseudonocardia sp.	NZ_CP011862.1.region001.gbk	43	5222236 5222995	10	
Streptomyces malaysiensis	NZ_CP023992.1.region003.gbk	43	2284346 2284649	9	
Streptomyces xiamenensis	NZ CP009922.2.region001.gbk	43	5351267 5351558	10	

Table 3: All found gvGCs

Sreptomyces sp. N2_CT0295411_region00_gbk 43 724514_r25474 9 Streptomyces gilvosporus N2_CT0205691_region01_gbk 43 724514_r25474 9 Streptomyces gilvosporus N2_C10205691_region01_gbk 43 15519691552143 9 Streptomyces idencensinger NC_C1029571_region01_gbk 42 N43443198443157 8 Streptomyces idencensinger NC_Q00155_region01_gbk 42 N054419844319844360 8 Streptomyces infinaseprialitis NC_Q00155_region01_gbk 42 8069730907766 8 Streptomyces infinaseprialitis NC_CP00200_region01_gbk 42 169062169479 8 Streptomyces induces NZ_CP002201_region01_gbk 42 169062169479 8 Streptomyces induces NZ_CP002201_region01_gbk 42 169062169479 8 Streptomyces induces NZ_CP012971_region01_gbk 42 1690611699514 8 Streptomyces induces NZ_CP012971_region01_gbk 42 12381120138915 9 Streptomyces ingeartium NC_CP012971_region001_gbk	Organism	Accession	Score	Coordinates cluster in genome	Nr of genes	Verified?
Steptomyces spinvopeneu N2_CP019779.1_region01.gbk 43 724514725474 9 Streptomyces pinvopeneu N2_CP020569.1_region01.gbk 42 3833023824157 8 Streptomyces lavendula N2_CP020569.1_region01.gbk 42 1083442 108 Streptomyces avernilia NC_00912.1_region01.gbk 42 4067340948 8 Streptomyces sp. N2_CP0205241gion01.gbk 42 4067340948 8 Streptomyces sp. N2_CP02025451gion01.gbk 42 19462169479 8 Streptomyces sp. N2_CP020202.1.region01.gbk 42 1467301747000 8 Streptomyces sp. N2_CP020202.1.region01.gbk 42 146730174000 8 Streptomyces sp. N2_CP012082.1.region01.gbk 42 146730174000 8 Streptomyces inverse N2_CP012082.1.region01.gbk 42 146730174000 8 Streptomyces aublaw NC_015561.region00.gbk 42 2088112089072 8 Streptomyces aublaw NC_015561.region00.gbk 42 20885112089072 8 <td>Streptomyces sp.</td> <td>NZ_CP029541.1.region001.gbk</td> <td>43</td> <td>1287691 1288000</td> <td>9</td> <td></td>	Streptomyces sp.	NZ_CP029541.1.region001.gbk	43	1287691 1288000	9	
Streptomyces glivosporeu NZ_CPU205091_region01.gbk 43 1551069	Streptomyces sp.	NZ_CP019779.1.region002.gbk	43	724514 725474	9	
Shreptomyces violaceusriger NC_01997.1region001.gbk 42 8323302 3833302 3833302	Streptomyces gilvosporeus	NZ_CP020569.1.region001.gbk	43	1551969 1552143	9	
Streptomyces lavendude NZ_CP024985.1.region001.gbk 42 84344316843420. 83 Streptomyces avermilitis NC_0001421.region001.gbk 42 46057346048 8 Streptomyces avermilitis NC_00155.5.region001.gbk 42 80697308070786 8 Bacillus megaterium NZ_CP01390.1.region001.gbk 42 80697308070786 8 Streptomyces pristinaespiralis NZ_CP013208.1.region001.gbk 42 8097308070786 8 Streptomyces shundingensis NZ_CP013208.1.region010.gbk 42 809610699934 8 Streptomyces shundingensis NZ_CP013207.1.region010.gbk 42 426387426795 8 Brachbacterium sp. NZ_CP01397.1.region001.gbk 42 426387426795 8 Bacillus megaterium NC_010395.1.region001.gbk 42 426387426795 8 Brachbacterium sp. NC_010395.1.region001.gbk 42 10386171033991 8 Streptomyces aruley NC_013955.1.region001.gbk 42 10386171033991 8 Streptomyces aruley NC_0139301.1.region001.gbk 42<	Streptomyces violaceusniger	NC_015957.1.region001.gbk	42	3823302 3824157	8	
Saccharopolyspon arythmea NC_001421.region001.gbk 42 1083470 1083762 10 Streptomyces avermitilis NZ_CP010798.tregion001.gbk 42 460573 460648 8 Streptomyces spritinnespinilis NZ_CP010798.tregion001.gbk 42 169062 169479 8 Streptomyces spritinnespinilis NZ_CP012001.gbk 42 169062 169479 8 Streptomyces spritinnespinilis NZ_CP012001.gbk 42 169062 169479 8 Streptomyces spritinnespinilis NZ_CP012071.region001.gbk 42 16387426779 8 Streptomyces spritinnespinilis NZ_CP012071.region001.gbk 42 612685 613126 8 Bacillus megaterium NC_015564.tregion001.gbk 42 120387120799 8 5 5 5 5 5 5 13126 8 5 1201766 8 1030570 8 5 5 5 13057 130510 5 16 3 4 120387 1201768 1201768	Streptomyces lavendulae	NZ_CP024985.1.region001.gbk	42	8434431 8434806	8	
Streptomyces avermilitis NC_003155.5.region001.gbk 42 746751747162 8 Streptomyces pristinasepiralis NZ_CP011340.1.region001.gbk 42 6809730 8070786 8 Bacillus negatarium NZ_CP010340.1.region001.gbk 42 6897197 6897512 8 Streptomyces sp. NZ_CP032698.1.region001.gbk 42 6897197 6897512 8 Streptomyces sp. NZ_CP032698.1.region001.gbk 42 6499610 699934 8 Streptomyces sp. NZ_CP018070.1.region002.gbk 42 642637 4267955 8 Hoyosella subflava NC_015564.1.region001.gbk 42 3127016 3217433 8 Streptomyces andus NC_015564.1.region001.gbk 42 325791 7554776 9 Bacillus negaterium NC_0135951.region001.gbk 42 326370 426795 8 Streptomyces aluga NC_017586.1.region001.gbk 42 326370 3294633 9 Streptomyces aluga NC_019821.region001.gbk 40 3296330 3294633 9 Frankia alui NC_008278.1.region001.gbk 40 342657 5043	Saccharopolyspora erythraea	NC_009142.1.region001.gbk	42	1083447 1083762	10	
Streptomyces sp. NZ_CP01998.1.region01.gbk 42 460573460948 8 Bacillus megaterium NZ_CP01998.1.region01.gbk 42 169062169479 8 Streptomyces sp. NZ_CP023698.1.region01.gbk 42 1691076897512 8 Streptomyces hundungensis NZ_CP01220.1.region01.gbk 42 17467301747000 8 Streptomyces hundungensis NZ_CP01227.1.region00.gbk 42 17381201338915 9 Brachybacterium sp. NZ_CP01297.1.region00.gbk 42 612685613126 8 Bacillus megaterium NC_015564.1.region00.gbk 42 612685613126 8 Streptomyces andungens call tregion00.gbk 42 612685613126 8 Streptomyces callsus NZ_CP03257.1.region00.gbk 42 1038677.1039991 8 Streptomyces callsus NZ_CP03307.1.1.region00.gbk 41 4429262 8 Frankia anin NC.00921.region00.gbk 40 3286330 9 Frankia anin NC.00921.region00.gbk 40 478644478787 9 Streptomyces al	Streptomyces avermitilis	NC_003155.5.region001.gbk	42	746751 747162	8	
Streptomyces prestnaesprafils NZ_C (P01220). Lregion001.gbk 42 80097.30. 80077.80.	Streptomyces sp.	NZ_CP019798.1.region001.gbk	42	460573 460948	8	
Bacillus megaerium NZ_CP022361. Lregion01.gbx 42 109002109479 8 Streptomyces hundungensis NZ_CP022361. Lregion01.gbk 42 17467301747000 8 Streptomyces niveus NZ_CP012202.1.region01.gbk 42 699610699934 8 Brachybacterium sp. NZ_CP017207.1.region00.gbk 42 639610699934 8 Bacillus megaterium NC_2CP01807.1.region00.gbk 42 612685613126 8 Bacillus megaterium NC_015564.1.region00.gbk 42 75535917554776 9 Bacillus megaterium NC_013595.1.region00.gbk 42 10366371038901 8 Streptomyces atlus NC_017091.Lregion00.gbk 42 20888112089072 8 Streptomyces atlus NZ_CP01307.1.region00.gbk 40 47868444787837 9 Streptomyces atlus NZ_CP016279.1.region00.gbk 40 47868444787837 9 Streptomyces atlus NZ_CP016279.1.region00.gbk 40 474663345274 17 Frankia ap NC_0.01666.1.region00.gbk 40 474565345274 <t< td=""><td>Streptomyces pristinaespiralis</td><td>NZ_CP011340.1.region001.gbk</td><td>42</td><td>8069730 8070786</td><td>8</td><td></td></t<>	Streptomyces pristinaespiralis	NZ_CP011340.1.region001.gbk	42	8069730 8070786	8	
Streptomyces product NZ_CP02369.1.region00.gbk 42 069/197089/312 5 Streptomyces product NZ_CP02702.1.region00.gbk 42 unknown 8 Streptomyces nivcus NZ_CP01870.1.region00.gbk 42 1338120 138915 9 Streptomyces sp. NZ_CP01870.1.region00.gbk 42 426387426795 8 Hoyosella subflava NC_0.15564.1.region00.gbk 42 312165 8 Streptomyces sp. NZ_CP01870.1.region00.gbk 42 3127716 9 Bacillus megaterium NC_0.015594.1.region00.gbk 42 10386371038991 8 Streptomyces athus NC_0.017586.1.region00.gbk 42 10366371038991 8 Streptomyces afterya NC_0.017596.1.region00.gbk 40 32965303296633 9 Frankia alni NC_0.008278.1.region00.gbk 40 374588874091 10 Rhodococcus opacus NC_0.01722.1.region00.gbk 40 74768874091 10 Rhodocaccus opacus NC_0.01722.1.region00.gbk 40 74768874091 10	Structure and an	NZ_CP009920.1.region001.gbk	42	169062 169479	8	
siterptomyces nutaunigensa VZ_CP02020:1:region001.gbt 42 17407.901170.000 8 Siterptomyces views NZ_CP02020:1:region001.gbt 42 699610699934 8 Brach/beaterium sp. NZ_CP018047.1:region002.gbt 42 426387426795 8 Hoyosella subliava NC_015564.1:region001.gbt 42 13381201338120133815 9 Sitreptomyces sp. NZ_CP01272.1:region001.gbt 42 426387426795 8 Bacillus megaterium NC_013955.1.region001.gbt 42 75535917554776 9 Bacillus megaterium NC_013952.1.region001.gbt 42 10386371038991 8 Streptomyces atus NZ_CP013272.1.region00.gbt 40 329633023266433 9 Frankia api NC_000271.1.region01.gbt 40 47864444787837 9 Streptomyces atus NZ_CP012327.1.region00.gbt 40 8744567954581 8 Halohasta linchfieldia NZ_CP0123481.1.region01.gbt 40 874428 874091 10 Rhodococcus opacus NC_014666.1.region001.gbt 40	Streptomyces sp.	NZ_CP022545.1.region001.gbk	42	089/19/ 089/012	8	
Streptomyces niveus NZ_CP018071.region00.gbk 42 unknown unkn	Streptomyces nundungensis	NZ_CP032098.1.1egion001.gbk	42	1/40/30 1/4/000	0	
Brachybackrimus NZ_CP012971.region001.gbt 42 13381201338915 9 Brachybackrimus NZ_CP0188701.region002.gbt 42 426887<426795	Streptomyces piveus	NZ_CP018047.1 region002 gbk	42	600610 60003/	0 8	
Streptomyces sp. NZ_CP018870.1.region002.gbk 42 426387426795 8 Hoyosella subflava NC_015564.1.region002.gbk 42 612685613126 8 Bacillus megaterium NC_01019.1.region001.gbk 42 75535917554776 9 Bacillus megaterium NZ_CP035271.region00.gbk 42 20888112089072 8 Streptomyces atleya NZ_CP03561.region01.gbk 42 20884112089072 8 Streptomyces atleya NZ_CP033071.1.region01.gbk 41 44293014429625 8 Frankia api NC_O08278.1.region00.gbk 40 32063303296633 9 Frankia sp. NC_O02221.1.region00.gbk 40 87484287837 9 Streptomyces griseochromogenes NZ_CP024451.region00.gbk 40 87456575045831 8 Halohasta litchfieldiae NZ_CP024451.region01.gbk 40 874563345274 17 Frankia inefficax, NC_O14661.region001.gbk 39 94755059476033 8 Streptomyces peucetius NZ_CP0243451.region002.gbk 35 196071603	Brachybacterium sp	NZ_CP017297.1 region001 gbk	42	1338120 1338915	9	
Hoyseila subflav NC_015564.1.region00.gbk 42 612685613126 8 Bacillus megaterium NC_01409.1.region00.gbk 42 32170163217433 8 Streptomycen angium roseum NC_01555.1.region00.gbk 42 20888112089072 8 Streptomyces atluy NC_01756.1.region00.gbk 42 20888112089072 8 Streptomyces atluy NC_002781.region00.gbk 44 4293014429625 8 Frankia alni NC_002921.1.region00.gbk 40 32963303296633 9 Frankia sp. NC_0016279.1.region00.gbk 40 874242874091 10 Rhodococcus opacus NC_016269.1.region00.gbk 40 874242874091 10 Rhodococcus opacus NC_016261.region00.gbk 40 34566345274 17 Frankia inefficar NC_016661.region00.gbk 39 9475030 8 Amycolatopsis mediterranei NC_01703.1.region00.gbk 37 1136345 1136345 Geitlerinema sp. NC_019703.1.region00.gbk 37 19463061 8 Jancolatops	Streptomyces sp.	NZ CP018870.1.region002.gbk	42	426387 426795	8	
Bacillus megaterium NC_014019.1.region001.gbk 42 3217016 3217433 8 Streptosporangium roseum NC_013595.1.region002.gbk 42 7553591 .7554776 9 Bacillus megaterium NC_017586.1.region001.gbk 42 2088811	Hovosella subflava	NC 015564.1.region001.gbk	42	612685 613126	8	
Streptosporangium roseum NC_013595.1.region001.gbk 42 75535917554776 9 Bacillus megaterium NZ_CP03527.1.region002.gbk 42 20888112089072 8 Streptomyces catleya NZ_CP033071.1.region001.gbk 41 4429025 8 Frankia alni NC_007871.1.region001.gbk 40 32960303296633 9 Frankia alni NC_008271.1.region001.gbk 40 4429421 47868444787837 9 Streptomyces griseochromogenes NZ_CP016279.1.region001.gbk 40 8743428874091 10 Rhodococcus opacus NC_01252.1.region001.gbk 40 7975809476033 8 Streptomyces peucetius NZ_CP024483.1.region001.gbk 39 94755059476033 8 Amycolatopsis mediiterranei NC_019703.1.region001.gbk 37 11363451136564 5 Planktothrix agardhii AP01791.1.region002.gbk 37 1136345 1136544 5 Streptomyces sp. NZ_CP013743.1.region002.gbk 33 1028154 5 5 Streptomyces optartrusis NZ_LT056042.1.regio	Bacillus megaterium	NC 014019.1.region001.gbk	42	3217016 3217433	8	
Bacillus megaterium NZ_CP032527.1.region00.gbk 42 2088811 2089072 8 Streptomyces altusy NZ_CP032527.1.region00.gbk 42 1038637 1038901 8 Streptomyces altusy NZ_CP032070.l.region00.gbk 44 4429301 4429625 8 Frankia ap. NC_00921.1.region00.gbk 40 3296330 3296633 9 Frankia sp. NZ_CP024845.1.region00.gbk 40 8748644 4787837 9 Streptomyces griseochromogene NZ_CP024845.1.region00.gbk 40 87428 874091 10 Rhodococcus opacus NC_012522.1.region00.gbk 40 87428 874091 10 Amycolatopsis mediterranei NC_018266.1.region01.gbk 39 9475505 9476033 8 Geitlerinema sp. NC_019703.1.region00.gbk 37 1136345 1136564 5 Planktothrix agarthii AP017991.1.region00.gbk 35 1967216 1968038 7 Synechococcus sp. NZ_019731.1.region00.gbk 35 1967216 1968038 7 Synechococcus sp. NZ_019703.1.region00.gbk 35 9016104 9601970 <td>Streptosporangium roseum</td> <td>NC_013595.1.region001.gbk</td> <td>42</td> <td>7553591 7554776</td> <td>9</td> <td></td>	Streptosporangium roseum	NC_013595.1.region001.gbk	42	7553591 7554776	9	
Streptomyces cattleya NC_017386.1.region001.gbk 42 10386371038991 8 Streptomyces albus NC_002371.1.region001.gbk 44 4429301422052 8 Frankia api NC_002921.1.region001.gbk 40 32963303296633 9 Streptomyces grisecdromogenes NC_0124245.1.region001.gbk 40 47868444787837 9 Streptomyces grisecdromogenes NC_0124245.1.region001.gbk 40 873428874091 10 Rhodococcus opacus NC_012522.1.region001.gbk 40 79705807970907 9 Amycolatopsis mediterranei NC_014266.1.region001.gbk 39 490553491603 8 Streptomyces peucetius NC_04703.1.region001.gbk 38 94655339466061 8 Geitlerinema sp. NC_019703.1.region002.gbk 37 1163644 5 Streptomyces spacetiterranei NC_019703.1.region002.gbk 35 19672161968038 7 Synechococcus sp. NC_007751.1.region002.gbk 35 19601249601970 22 Synechococcus sp. NC_019701.1.region002.gbk 34 569343	Bacillus megaterium	NZ_CP032527.1.region002.gbk	42	2088811 2089072	8	
Streptomyces albus NZ_CP033071.1.region001.gbk 41 4429301	Streptomyces cattleya	NC_017586.1.region001.gbk	42	1038637 1038991	8	
Frankia alni NC_008278.1.region001.gbk 40 32963303296633 9 Frankia sp. NC_009921.1.region001.gbk 40 47868444787837 9 Streptomyces griseochromogene NZ_CP016279.1.region001.gbk 40 50456575045831 8 Halohasta litchfieldiae NZ_CP024845.1.region001.gbk 40 7475807970907 9 Armycolatopsis mediterranei NC_012522.1.region001.gbk 39 94755059476033 8 Streptomyces peucetius NZ_CP024381.region001.gbk 39 94055339466061 8 Geitterinema sp. NC_019703.1.region002.gbk 37 28152732815492 11 Streptomyces chartreusis NZ_LT962942.1.region002.gbk 37 28152732815492 11 Streptomyces sp. NZ_LO19703.1.region002.gbk 35 19672161960782 6 Streptomyces sp. NZ_LO19703.1.region002.gbk 35 19672161960778 6 Synechococcus sp. NZ_LO19701.1.region002.gbk 32 2850579285079 5 Leptolyngbya sp. NZ_AP017360.1.region002.gbk 32	Streptomyces albus	NZ_CP033071.1.region001.gbk	41	4429301 4429625	8	
Frankia sp. NC_009921.1.region001.gbk 40 4786844 4787837 9 Streptomyces griseochnomogenes NZ_CP016279.1.region001.gbk 40 65456575045831 8 Halohasta linchfelidie NZ_CP024845.1.region001.gbk 40 37428874091 10 Rhodococcus opacus NC_012522.1.region001.gbk 40 79705807970907 9 Amycolatopsis mediterranei NC_018266.1.region001.gbk 39 9475505 9476033 8 Streptomyces peucetius NZ_CP02438.1.region001.gbk 38 9465533 9466661 8 Geitlerinema sp. NC_019703.1.region002.gbk 37 2115273 2815492 11 Streptomyces chartreusis NZ_LT962942.1.region002.gbk 35 1967216 1968038 7 Synechococcus sp. NZ_CP013743.1.region001.gbk 34 569343 569778 6 Synechococcus sp. NZ_AP017367.1.region002.gbk 32 2880579 2850912 4 Pseudanabaena sp. NZ_AP017367.1.region002.gbk 32 2084758 3084977 4 Pseudanabaena sp. NZ_AP017367.1.region002.gbk 32<	Frankia alni	NC_008278.1.region001.gbk	40	3296330 3296633	9	
Streptomyces griseochromogene NZ_CP016279.1.region001.gbk 40 50456575045831 8 Halohasta litchfieldiae NZ_CP024845.1.region001.gbk 40 873428874091 10 Rhodococcus opacus NC_015221.1region001.gbk 40 79705807970907 9 Amycolatopsis mediterranei NC_018266.1.region001.gbk 39 94755059476033 8 Amycolatopsis mediterranei NC_012438.1.region001.gbk 39 94055339466661 8 Geitlerinema sp. NC_019703.1.region002.gbk 37 1136345136564 5 Planktothrix agardhii AP017991.1.region002.gbk 35 19672161968038 7 Synechococcus sp. NZ_CP013743.1.region002.gbk 35 14895441490024 6 Streptomyces sp. NZ_CP013743.1.region002.gbk 35 1061649001970 22 Synechococcus sp. NC_019701.1.region002.gbk 32 2850579285012 4 Pseudanabaena sp. NZ_AP017560.1.region002.gbk 32 2850579285012 4 Pseudonabaena sp. NZ_CP020370.1.region002.gbk 32	Frankia sp.	NC_009921.1.region001.gbk	40	4786844 4787837	9	
Halohasta litchfieldiae NZ_CP024845.1.region001.gbk 40 873428.8.74091 10 Rhodococcus opacus NC_012522.1.region001.gbk 40 344566345274 17 Frankia inefficax, NC_014666.1.region001.gbk 39 94755059476033 8 Streptomyces peucetius NZ_CP022438.1.region001.gbk 39 9405839466061 8 Geitlerinema sp. NC_019703.1.region001.gbk 37 2115545 113554 5 Planktothrix agardhii AP017991.1.region002.gbk 37 2115545 1136544 5 Streptomyces chartreusis NZ_CP013743.1.region002.gbk 35 196716 1968038 7 Synechococcus sp. NC_007775.1.region002.gbk 35 1489544 1490024 6 Streptomyces sp. NZ_CP013743.1.region002.gbk 32 2850579 22 24 Synechococcus sp. NC_019702.1.region002.gbk 32 2850579 2350912 4 Pseudanabaena sp. NZ_AP017367.1.region002.gbk 32 23084758 3084977 4 Psychromonas	Streptomyces griseochromogenes	NZ_CP016279.1.region001.gbk	40	5045657 5045831	8	
Rhodococcus opacus NC_012522.1.region002.gbk 40 344566 345274 17 Frankia inefficax, NC_0114666.1.region001.gbk 39 9475505 977003 8 Amycolatopsis mediterranei NC_018266.1.region001.gbk 39 490583 947603 8 Amycolatopsis mediterranei NC_019703.1.region001.gbk 37 1136345 136564 5 Planktohrix agardhi AP017991.1.region002.gbk 37 215273 2815492 11 Streptomyces chartreusis NZ_LT962942.1.region002.gbk 35 1967216 1968038 7 Synechcococcus sp. NC_019701.1.region001.gbk 35 9601604 960178 6 Streptomyces sp. NC_019702.1.region002.gbk 33 1028154 1028499 5 Leptolyngbya sp. NZ_AP017360.1.region002.gbk 30 2130457 250971 4 Pseudanabaena sp. NC_01703.1.region002.gbk 30 2130457 2130676 5 Bacillus marisflavi NZ_CP033051.1.region002.gbk 30 120457 <td>Halohasta litchfieldiae</td> <td>NZ_CP024845.1.region001.gbk</td> <td>40</td> <td>873428 874091</td> <td>10</td> <td></td>	Halohasta litchfieldiae	NZ_CP024845.1.region001.gbk	40	873428 874091	10	
Frankii inefficax, NC_014666.1.region001.gbk 40 7970580	Rhodococcus opacus	NC_012522.1.region002.gbk	40	344566 345274	17	
Amycolatopsis mediterranei NC_018260.1 region001.gbk 39 94/75051.s94/0033 8 Streptomyces peucetius NZ_CP022438.1.region001.gbk 39 490583491630 8 Amycolatopsis mediterranei NC_014318.1.region001.gbk 37 1136345136564 5 Planktothrix agardhii AP017991.1.region002.gbk 37 28152732815492 11 Streptomyces chartreusis NZ_CP01775.1.region002.gbk 35 19672161968038 7 Synechococcus sp. NZ_C007775.1.region002.gbk 35 96016049601970 22 Synechococcus sp. NZ_C019702.1.region001.gbk 34 569343569778 6 Pseudanabaena sp. NZ_AP017367.1.region002.gbk 32 2850579 2850979 Leptolyngbya sp. NZ_AP017367.1.region002.gbk 30 21304572130676 5 Bacillus marisflavi NZ_CP023051.1.region002.gbk 30 21304572130676 5 Bacillus marisflavi NZ_CP020370.1.region001.gbk 28 31628564493410 5 Halatkalicoccus jostii NC_0029268.1.region003.gbk 28	Frankia inefficax,	NC_014666.1.region001.gbk	40	7970580 7970907	9	
Nz_CP02438.1.region001.gbt 39 491630 8 Amycolatopsis mediterranei NC_014318.1.region001.gbt 38 9465533 9406061 8 Geitlerinema sp. NC_019703.1.region002.gbt 37 2815273 2815492 11 Streptomyces chartreusis NZ_LT962942.1.region002.gbt 35 1967216 1968038 7 Synechococcus sp. NZ_CP013743.1.region001.gbt 35 1489544 1490024 6 Streptomyces sp. NZ_CP013743.1.region001.gbt 34 569343 569778 6 Pseudanabaena sp. NZ_O19701.1.region002.gbt 33 1028154 1028499 5 Leptolyngbya sp. NZ_AP017367.1.region002.gbt 30 2130457 2130676 5 Bacillus marisflavi NZ_CP033051.1.region002.gbt 30 2130457 2130676 5 Bacillus marisflavi NZ_CP033051.1.region002.gbt 28 3162856 3164239 15 Candidatus Thiodictyon NZ_CP020370.1.region001.gbt 28 1910294 1910558 14 Natronocccus jostii NC_0142971.region001.gbt 28 3719849 3720452 6 <td>Amycolatopsis mediterranei</td> <td>NC_018266.1.region001.gbk</td> <td>39</td> <td>9475505 9476033</td> <td>8</td> <td></td>	Amycolatopsis mediterranei	NC_018266.1.region001.gbk	39	9475505 9476033	8	
Aniyeotatopissi mediterianet NC_019703.1.region001.gbk 35 9400011 5 Geitlerinema sp. NC_019703.1.region002.gbk 37 2815273 2815492 11 Streptomyces chartreusis NZ_LT962942.1.region002.gbk 35 1967216 1968038 7 Synechococcus sp. NC_007775.1.region002.gbk 35 1489544 1490024 6 Streptomyces sp. NZ_CP013743.1.region001.gbk 35 9601604 9601970 22 Synechococcus sp. NC_019702.1.region002.gbk 33 1028154 1028499 5 Leptolyngbya sp. NZ_AP017367.1.region002.gbk 32 2850579 2850912 4 Pseudanabaena sp. NZ_AP017360.1.region002.gbk 32 2084758 3084977 4 Psychromonas ingrahamii NC_008709.1.region003.gbk 28 3162856 3164239 15 Candidatus Thiodictyon NZ_CP01370.1.region003.gbk 28 1910294 1910558 14 Natroncococcus occultus NC_019976.1.region003.gbk 28 1910294 1910558 14 Natroncococcus occultus NC_007777.1.region003.gbk 28 1910294 1910558 14 Natroncococcus occultus	Amugalatancia maditamanci	NZ_CP022438.1.region001.gbk	39 20	490583 491630	8	
Planktothrix agardhii AP017991.1.region002.gbk 37 2815273.2815492 11 Streptomyces chartreusis NZ_LT962942.1.region002.gbk 35 19672161968038 7 Synechococcus sp. NC_007775.1.region002.gbk 35 14895441490024 6 Streptomyces sp. NZ_CP013743.1.region001.gbk 35 96016049601970 22 Synechococcus sp. NC_019701.1.region001.gbk 34 569343569778 6 Pseudanabaena sp. NC_019701.1.region002.gbk 32 2850579<2850912	Geitlerinema sn	NC_019703_1_region001_gbk	30 37	9405555 9400001 1136345 1136564	0 5	
Streptomyces chartreusis NZ_LT962942.1.region002.gbk 35 1967216 1968038 7 Synechococcus sp. NZ_CP013743.1.region002.gbk 35 1489544 1490024 6 Streptomyces sp. NZ_CP013743.1.region001.gbk 35 9601604 .9601970 22 Synechococcus sp. NC_019701.1.region002.gbk 33 1028154 1028499 5 Leptolyngbya sp. NZ_AP017560.1.region002.gbk 32 3084758 3084977 4 Pseudanabaena sp. NZ_O0870.1.region002.gbk 30 2130457 2130676 5 Bacillus marisflavi NZ_CP020370.1.region001.gbk 28 3162856 3164239 15 Candidatus Thiodictyon NZ_CP020370.1.region001.gbk 28 1910294 1910558 14 Natronococcus occultus NC_019976.1.region003.gbk 28 3719849 3720452 6 Streptomyces anbofaciens NZ_CP01382.1.region001.gbk 26 7419007 7421287 4 Natronococcus ocultus NC_019777.1.region003.gbk 28 3719849 3720452 6 Streptomyces anbofaciens	Planktothrix agardhii	AP017991 1 region002 gbk	37	2815273 2815492	11	
Synechococcus sp. NC_007775.1.region002.gbk 35 1489544 1490024 6 Streptomyces sp. NZ_CP013743.1.region001.gbk 35 9601604 9601970 22 Synechococcus sp. NC_019702.1.region001.gbk 34 569343 569778 6 Pseudanabaena sp. NC_019701.1.region002.gbk 33 1028154 1028499 5 Leptolyngbya sp. NZ_AP017367.1.region002.gbk 32 2850579 2850912 4 Pseudanabaena sp. NZ_AP017560.1.region002.gbk 32 2850579 2850912 4 Pseudanabaena sp. NZ_AP01750.1.region002.gbk 30 2130457 2130676 5 Bacillus marisflavi NZ_CP020370.1.region003.gbk 28 3162856 3164239 15 Candidatus Thiodictyon NZ_CP020370.1.region001.gbk 28 1910294 1910558 14 Natronococcus occultus NC_008268.1.region001.gbk 28 3719849 3720452 6 Streptomyces parvulus NZ_CP01382.1.region001.gbk 28 3719849 3720452 6 Streptomyces parvulus NZ_CP001382.1.region003.gbk 22 287915 288089 3 Streptomyces parvulus	Streptomyces chartreusis	NZ_LT962942_1 region002.gbk	35	1967216 1968038	7	
Streptomyces sp. NZ_CP013743.1.region001.gbk 35 9601604 9601970 22 Synechococcus sp. NC_019702.1.region001.gbk 34 569343 569778 6 Pseudanabaena sp. NC_019701.1.region002.gbk 33 1028154 1028499 5 Leptolyngbya sp. NZ_AP017367.1.region002.gbk 32 2850579 2850912 4 Pseudanabaena sp. NZ_AP017560.1.region002.gbk 32 2850579 2850912 4 Psychromonas ingrahamii NC_008709.1.region002.gbk 32 2130457 2130676 5 Bacillus marisflavi NZ_CP020370.1.region001.gbk 28 3162856 3164239 15 Candidatus Thiodictyon NZ_CP020370.1.region001.gbk 28 1910294 1910558 14 Natronococcus occultus NC_019976.1.region003.gbk 28 3719849 3720452 6 Streptomyces ambfaciens NZ_CP012382.1.region003.gbk 26 7419007 7421287 4 Streptomyces ambfaciens NZ_CP001574.1.region003.gbk 26 5982246 5982756 4 Frankia casuarinae, NC_007771.1.region003.gbk	Synechococcus sp.	NC 007775.1.region002.gbk	35	1489544 1490024	6	
Synechococcus sp. NC_019702.1.region001.gbk 34 569343 569778 6 Pseudanabaena sp. NC_019701.1.region002.gbk 33 1028154 1028499 5 Leptolyngbya sp. NZ_AP017367.1.region002.gbk 32 2850579 2850912 4 Pseudanabaena sp. NZ_AP017560.1.region002.gbk 32 3084758 3084977 4 Psychromonas ingrahamii NC_008709.1.region002.gbk 30 2130457 2130676 5 Bacillus marisflavi NZ_CP033051.1.region001.gbk 28 3162856 3164239 15 Candidatus Thiodictyon NZ_CP020370.1.region001.gbk 28 1910294 1910558 14 Natronococcus occultus NC_019976.1.region001.gbk 28 1910294 1910558 14 Natronococcus jostii NC_002868.1.region001.gbk 28 3719849 3720452 6 Streptomyces parvulus NZ_CP012382.1.region001.gbk 26 7419007 7421287 4 Streptomyces puccetius NZ_CP015764.1.region003.gbk 22 198730 1199582 4 Streptomyces albulus NZ_CP007574.1.region003.gbk	Streptomyces sp.	NZ CP013743.1.region001.gbk	35	9601604 9601970	22	
Pseudanabaena sp. NC_019701.1.region002.gbk 33 1028154 1028499 5 Leptolyngbya sp. NZ_AP017367.1.region002.gbk 32 2850579 2850912 4 Pseudanabaena sp. NZ_AP017560.1.region002.gbk 32 3084758 3084977 4 Psychromonas ingrahamii NC_008709.1.region002.gbk 30 2130457 2130676 5 Bacillus marisflavi NZ_CP033051.1.region001.gbk 28 3162856 3164239 15 Candidatus Thiodictyon NZ_CP020370.1.region003.gbk 28 4492645 4493410 5 Halalkalicoccus jeotgali NC_019976.1.region001.gbk 28 unknown unknown 6 Rhodococcus jostii NC_009868.1.region001.gbk 28 unknown unknown 6 Streptomyces parvulus NZ_CP01382.1.region001.gbk 26 7419007 7421287 4 Streptomyces parvulus NZ_CP015866.1.region003.gbk 26 5982246 5982756 4 Frankia casuarinae, NC_007777.1.region001.gbk 22 1198730 1199582 4 Streptomyces abulus NZ_CP01342.1.region003.gbk <	Synechococcus sp.	NC_019702.1.region001.gbk	34	569343 569778	6	
Leptolyngbya sp. NZ_AP017367.1.region002.gbk 32 2850579 2850912 4 Pseudanabaena sp. NZ_AP017560.1.region002.gbk 32 3084758 3084977 4 Psychromonas ingrahamii NC_008709.1.region002.gbk 30 2130457 2130676 5 Bacillus marisflavi NZ_CP033051.1.region001.gbk 28 3162856 3164239 15 Candidatus Thiodictyon NZ_CP020370.1.region003.gbk 28 4492645 4493410 5 Halalkalicoccus jeotgali NC_014297.1.region001.gbk 28 1910294 1910558 14 Natronococcus occultus NC_019976.1.region001.gbk 28 unknown unknown 6 Rhodococcus jostii NC_008268.1.region003.gbk 28 3719849 3720452 6 Streptomyces ambofaciens NZ_CP015866.1.region003.gbk 26 7419007 7421287 4 Streptomyces albulus NZ_CP007574.1.region001.gbk 22 1198730 1199582 4 Streptomyces speucetius NZ_CP01342.1.region03.gbk 22 287915 28089 3 Streptomyces speucetius NZ_CP01342.1.region03.gbk 21 7357366 7359814 12	Pseudanabaena sp.	NC_019701.1.region002.gbk	33	1028154 1028499	5	
Pseudanabaena sp. NZ_AP017560.1.region002.gbk 32 3084758 3084977 4 Psychromonas ingrahamii NC_008709.1.region002.gbk 30 2130457 2130676 5 Bacillus marisflavi NZ_CP033051.1.region001.gbk 28 3162856 3164239 15 Candidatus Thiodictyon NZ_CP020370.1.region003.gbk 28 4492645 4493410 5 Halalkalicoccus jeotgali NC_014297.1.region001.gbk 28 1910294 1910558 14 Natronococcus occultus NC_019976.1.region003.gbk 28 unknown unknown 6 Rhodococcus jostii NC_008268.1.region003.gbk 28 3719849 3720452 6 Streptomyces ambofaciens NZ_CP012382.1.region001.gbk 26 7419007 7421287 4 Streptomyces parvulus NZ_CP015866.1.region003.gbk 26 5982246 5982756 4 Frankia casuarinae, NC_007777.1.region001.gbk 22 1198730 1199582 4 Streptomyces abulus NZ_CP023714.1.region003.gbk 22 490583 491630 5 Streptomyces sp. NZ_CP023714.1.region001.gbk 21 1315050 1316121 4 Rhod	Leptolyngbya sp.	NZ_AP017367.1.region002.gbk	32	2850579 2850912	4	
Psychromonas ingrahamii NC_008709.1.region002.gbk 30 2130457 2130676 5 Bacillus marisflavi NZ_CP033051.1.region001.gbk 28 3162856 3164239 15 Candidatus Thiodictyon NZ_CP020370.1.region003.gbk 28 4492645 4493410 5 Halalkalicoccus jeotgali NC_014297.1.region001.gbk 28 1910294 1910558 14 Natronococcus occultus NC_019976.1.region001.gbk 28 unknown 6 Rhodococcus jostii NC_008268.1.region003.gbk 28 3719849 3720452 6 Streptomyces ambofaciens NZ_CP012382.1.region001.gbk 26 7419007 7421287 4 Streptomyces parvulus NZ_CP015866.1.region003.gbk 22 1198730 1199582 4 Streptomyces albulus NZ_CP007574.1.region002.gbk 22 287915 288089 3 Streptomyces sp. NZ_CP01342.1.region003.gbk 21 7357366 7359814 12 Rhodococcus ruber NZ_CP023714.1.region001.gbk 21 5	Pseudanabaena sp.	NZ_AP017560.1.region002.gbk	32	3084758 3084977	4	
Bacillus marisflavi NZ_CP033051.1.region001.gbk 28 3162856 3164239 15 Candidatus Thiodictyon NZ_CP020370.1.region003.gbk 28 4492645 4493410 5 Halalkalicoccus jeotgali NC_014297.1.region001.gbk 28 1910294 1910558 14 Natronococcus occultus NC_019976.1.region001.gbk 28 unknown unknown 6 Rhodococcus jostii NC_008268.1.region003.gbk 28 3719849 3720452 6 Streptomyces ambofaciens NZ_CP012382.1.region001.gbk 26 7419007 7421287 4 Streptomyces parvulus NZ_CP015866.1.region003.gbk 26 5982246 5982756 4 Frankia casuarinae, NC_007777.1.region001.gbk 22 1198730 1199582 4 Streptomyces albulus NZ_CP022438.1.region003.gbk 22 287915 28089 3 Streptomyces sp. NZ_CP013142.1.region003.gbk 21 7357366 7359814 12 Rhodococcus ruber NZ_CP023714.1.region001.gbk 21 </td <td>Psychromonas ingrahamii</td> <td>NC_008709.1.region002.gbk</td> <td>30</td> <td>2130457 2130676</td> <td>5</td> <td></td>	Psychromonas ingrahamii	NC_008709.1.region002.gbk	30	2130457 2130676	5	
Candidatus Thiodictyon NZ_CP020370.1.region003.gbk 28 4492645 4493410 5 Halalkalicoccus jeotgali NC_014297.1.region001.gbk 28 1910294 1910558 14 Natronococcus occultus NC_019976.1.region001.gbk 28 unknown unknown 6 Rhodococcus jostii NC_008268.1.region003.gbk 28 3719849 3720452 6 Streptomyces ambofaciens NZ_CP012382.1.region001.gbk 26 7419007 7421287 4 Streptomyces parvulus NZ_CP015866.1.region003.gbk 26 5982246 5982756 4 Frankia casuarinae, NC_007777.1.region001.gbk 22 1198730 1199582 4 Streptomyces albulus NZ_CP007574.1.region002.gbk 22 287915 288089 3 Streptomyces spucetius NZ_CP023714.1.region003.gbk 21 7357366 7359814 12 Rhodococcus ruber NZ_CP013142.1.region001.gbk 21 1315050 1316121 4 Rhodococcus aetherivorans NZ_CP020372.1.region001.gbk 21 5779817 5780531 4 Candidatus Thiodictyon NZ_CP020372.1.region002.gbk 19 54256 55342 2	Bacillus marisflavi	NZ_CP033051.1.region001.gbk	28	3162856 3164239	15	
Halakalicoccus jeotgali NC_014297.1.region001.gbk 28 1910294 1910558 14 Natronococcus occultus NC_019976.1.region001.gbk 28 unknown unknown 6 Rhodococcus jostii NC_008268.1.region003.gbk 28 3719849 3720452 6 Streptomyces ambofaciens NZ_CP012382.1.region001.gbk 26 7419007 7421287 4 Streptomyces parvulus NZ_CP015866.1.region003.gbk 26 5982246 5982756 4 Frankia casuarinae, NC_007777.1.region001.gbk 22 1198730 1199582 4 Streptomyces albulus NZ_CP007574.1.region002.gbk 22 287915 288089 3 Streptomyces speucetius NZ_CP013142.1.region003.gbk 21 7357366 7359814 12 Rhodococcus ruber NZ_CP013142.1.region001.gbk 21 1315050 1316121 4 Rhodococcus aetherivorans NZ_CP01372.1.region001.gbk 21 5779817 5780531 4 Candidatus Thiodictyon NZ_CP020372.1.region001.gbk 19 54256 55342 2 2 Synechococcus sp. NC_007775.1.region003.gbk 19 1489544 1490024 9 5 <td>Candidatus Thiodictyon</td> <td>NZ_CP020370.1.region003.gbk</td> <td>28</td> <td>4492645 4493410</td> <td>5</td> <td></td>	Candidatus Thiodictyon	NZ_CP020370.1.region003.gbk	28	4492645 4493410	5	
Natronococcus occultus NC_019976.1.region001.gbk 28 unknown 6 Rhodococcus jostii NC_008268.1.region003.gbk 28 3719849 3720452 6 Streptomyces ambofaciens NZ_CP012382.1.region001.gbk 26 7419007 7421287 4 Streptomyces parvulus NZ_CP015866.1.region003.gbk 26 5982246 5982756 4 Streptomyces parvulus NZ_CP015866.1.region001.gbk 22 1198730 1199582 4 Streptomyces albulus NZ_CP007574.1.region002.gbk 22 287915 288089 3 Streptomyces peucetius NZ_CP013142.1.region003.gbk 21 7357366 7359814 12 Rhodococcus ruber NZ_CP023714.1.region001.gbk 21 1315050 1316121 4 Rhodococcus aetherivorans NZ_CP01342.1.region001.gbk 21 5779817 5780531 4 Candidatus Thiodictyon NZ_CP020372.1.region001.gbk 19 54256 55342 2 Synechococcus sp. NC_007775.1.region003.gbk 19 14	Halalkalicoccus jeotgali	NC_014297.1.region001.gbk	28	1910294 1910558	14	
Rhodococcus jostii NC_008268.1.region003.gbk 28 3719849 3720452 6 Streptomyces ambofaciens NZ_CP012382.1.region001.gbk 26 7419007 7421287 4 Streptomyces parvulus NZ_CP015866.1.region003.gbk 26 5982246 5982756 4 Frankia casuarinae, NC_007777.1.region001.gbk 22 1198730 1199582 4 Streptomyces albulus NZ_CP007574.1.region002.gbk 22 287915 288089 3 Streptomyces peucetius NZ_CP022438.1.region003.gbk 21 7357366 7359814 12 Rhodococcus ruber NZ_CP023714.1.region001.gbk 21 7357366 7359814 12 Rhodococcus ruber NZ_CP011341.1.region001.gbk 21 1315050 1316121 4 Rhodococcus aetherivorans NZ_CP020372.1.region001.gbk 21 5779817 5780531 4 Candidatus Thiodictyon NZ_007775.1.region002.gbk 19 569343 569778 3 Synechococcus sp. NC_007775.1.region003.gbk 19 1489544 1490024 9 Synechococcus sp. NC_007775.1.region004.gbk	Natronococcus occultus	NC_019976.1.region001.gbk	28	unknown unknown	6	
Streptomyces ambofaciens NZ_CP012382.1.region001.gbk 26 7419007 7421287 4 Streptomyces parvulus NZ_CP015866.1.region003.gbk 26 5982246 5982756 4 Frankia casuarinae, NC_007777.1.region001.gbk 22 1198730 1199582 4 Streptomyces albulus NZ_CP007574.1.region002.gbk 22 287915 288089 3 Streptomyces peucetius NZ_CP022438.1.region003.gbk 21 7357366 7359814 12 Rhodococcus ruber NZ_CP013142.1.region001.gbk 21 1315050 1316121 4 Rhodococcus aetherivorans NZ_CP013141.1.region001.gbk 21 5779817 5780531 4 Candidatus Thiodictyon NZ_CP020372.1.region002.gbk 19 54256 55342 2 Synechococcus sp. NC_019702.1.region003.gbk 19 569343 569778 3 Synechococcus sp. NC_007775.1.region003.gbk 19 1489544 1490024 9 Synechococcus sp. NC_007775.1.region004.gbk 19 1489544 1490024 3	Rhodococcus jostii	NC_008268.1.region003.gbk	28	3719849 3720452	6	
Streptomyces parvulus NZ_CP015866.1.region003.gbk 26 5982246 5982756 4 Frankia casuarinae, NC_007777.1.region001.gbk 22 1198730 1199582 4 Streptomyces albulus NZ_CP007574.1.region002.gbk 22 287915 288089 3 Streptomyces peucetius NZ_CP022438.1.region003.gbk 22 490583 491630 5 Streptomyces sp. NZ_CP013142.1.region003.gbk 21 7357366 7359814 12 Rhodococcus ruber NZ_CP023714.1.region001.gbk 21 1315050 1316121 4 Rhodococcus aetherivorans NZ_CP01341.1.region001.gbk 21 5779817 5780531 4 Candidatus Thiodictyon NZ_CP020372.1.region001.gbk 19 54256 55342 2 Synechococcus sp. NC_019702.1.region002.gbk 19 569343 569778 3 Synechococcus sp. NC_007775.1.region003.gbk 19 1489544 1490024 9 Synechococcus sp. NC_007775.1.region004.gbk 19 1489544 1490024 3	Streptomyces ambofaciens	NZ_CP012382.1.region001.gbk	26	7419007 7421287	4	
Frankia casuarinae, NC_007771.1.region001.gbk 22 1198730 1199582 4 Streptomyces albulus NZ_CP007574.1.region002.gbk 22 287915 288089 3 Streptomyces peucetius NZ_CP022438.1.region003.gbk 22 490583 491630 5 Streptomyces sp. NZ_CP013142.1.region003.gbk 21 7357366 7359814 12 Rhodococcus ruber NZ_CP023714.1.region001.gbk 21 1315050 1316121 4 Rhodococcus aetherivorans NZ_CP01341.1.region001.gbk 21 5779817 5780531 4 Candidatus Thiodictyon NZ_CP020372.1.region001.gbk 19 54256 55342 2 Synechococcus sp. NC_019702.1.region002.gbk 19 569343 569778 3 Synechococcus sp. NC_007775.1.region003.gbk 19 1489544 1490024 9 Synechococcus sp. NC_007775.1.region004.gbk 19 1489544 1490024 3	Streptomyces parvulus	NZ_CP015866.1.region003.gbk	26	5982246 5982756	4	
Streptomyces abulus NZ_CP0075/4.1.region002.gbk 22 287913 288089 5 Streptomyces peucetius NZ_CP022438.1.region003.gbk 22 490583 491630 5 Streptomyces sp. NZ_CP013142.1.region003.gbk 21 7357366 7359814 12 Rhodococcus ruber NZ_CP023714.1.region001.gbk 21 1315050 1316121 4 Rhodococcus aetherivorans NZ_CP011341.1.region001.gbk 21 5779817 5780531 4 Candidatus Thiodictyon NZ_CP020372.1.region001.gbk 19 54256 55342 2 Synechococcus sp. NC_019702.1.region003.gbk 19 1489544 1490024 9 Synechococcus sp. NC_007775.1.region004.gbk 19 1489544 1490024 3	Frankia casuarinae,	NC_00////.1.region001.gbk	22	1198/30 1199582	4	
Streptomyces sp. NZ_CP013142.1.region003.gbk 21 7357366 7359814 12 Rhodococcus ruber NZ_CP023714.1.region001.gbk 21 1315050 1316121 4 Rhodococcus aetherivorans NZ_CP011341.1.region001.gbk 21 5779817 5780531 4 Candidatus Thiodictyon NZ_CP020372.1.region001.gbk 19 54256 55342 2 Synechococcus sp. NC_019702.1.region002.gbk 19 569343 569778 3 Synechococcus sp. NC_007775.1.region003.gbk 19 1489544 1490024 9 Synechococcus sp. NC_007775.1.region004.gbk 19 1489544 1490024 3	Streptomyces abunds	NZ_CP007374.1.region002.gbk	22	287913 288089 400583 401630	5	
Rhodococcus ruber NZ_CP023714.1.region001.gbk 21 7357500 7359814 12 Rhodococcus aetherivorans NZ_CP011341.1.region001.gbk 21 1315050 1316121 4 Rhodococcus aetherivorans NZ_CP011341.1.region001.gbk 21 5779817 5780531 4 Candidatus Thiodictyon NZ_CP020372.1.region001.gbk 19 54256 55342 2 Synechococcus sp. NC_019702.1.region002.gbk 19 569343 569778 3 Synechococcus sp. NC_007775.1.region003.gbk 19 1489544 1490024 9 Synechococcus sp. NC_007775.1.region004.gbk 19 1489544 1490024 3	Streptomyces sp	NZ_CP013142.1 region003.gbk	22	7357366 7350814	12	
Rhodococcus nucli NZ_CP011341.1.region001.gbk 21 1515050 1510121 4 Rhodococcus aetherivorans NZ_CP011341.1.region001.gbk 21 5779817 5780531 4 Candidatus Thiodictyon NZ_CP020372.1.region001.gbk 19 54256 55342 2 Synechococcus sp. NC_019702.1.region002.gbk 19 569343 569778 3 Synechococcus sp. NC_007775.1.region003.gbk 19 1489544 1490024 9 Synechococcus sp. NC_007775.1.region004.gbk 19 1489544 1490024 3	Bhodococcus ruber	NZ_CP013142.1.1egion003.gbk	21	1315050 1316121	12	
Candidatus Thiodictyon NZ_CP020372.1.region001.gbk 19 54256 55342 2 Synechococcus sp. NC_019702.1.region002.gbk 19 569343 569778 3 Synechococcus sp. NC_007775.1.region003.gbk 19 1489544 1490024 9 Synechococcus sp. NC_007775.1.region004.gbk 19 1489544 1490024 3	Rhodococcus aetherivorans	NZ CP011341.1 region001 gbk	21	5779817 5780531	4	
Synechococcus sp. NC_019702.1.region002.gbk 19 569343 569778 3 Synechococcus sp. NC_007775.1.region003.gbk 19 1489544 1490024 9 Synechococcus sp. NC_007775.1.region004.gbk 19 1489544 1490024 3	Candidatus Thiodictyon	NZ CP020372.1.region001 ohk	19	54256 55342	2	
Synechococcus sp. NC_007775.1.region003.gbk 19 1489544 1490024 9 Synechococcus sp. NC_007775.1.region004.gbk 19 1489544 1490024 3	Synechococcus sp.	NC 019702.1.region002.gbk	19	569343 569778	3	
Synechococcus sp. NC_007775.1.region004.gbk 19 1489544 1490024 3	Synechococcus sp.	NC_007775.1.region003.gbk	19	1489544 1490024	9	
	Synechococcus sp.	NC_007775.1.region004.gbk	19	1489544 1490024	3	

Organism	Accession	Score	Coordinates cluster in genome	Nr of genes	Verified?
Candidatus Thiodictyon	NZ_CP020370.1.region002.gbk	19	4492645 4493410	2	
Rhodococcus opacus	NC_012520.1.region001.gbk	19	456007 456373	3	
Rhodococcus jostii	NC_008268.1.region004.gbk	16	3719849 3720452	3	
Streptomyces sp.	NZ_CP013743.1.region004.gbk	16	9601604 9601970	3	
Rhodococcus rhodochrous	NZ_LT906450.1.region001.gbk	16	2580857 2581061	3	
Rhodococcus opacus	NZ_CP003949.1.region004.gbk	16	4749401 4749746	3	
Pseudonocardia dioxanivorans	NC_015312.1.region002.gbk	15	4566525 4566825	2	
Streptomyces actuosus	NZ_CP029788.1.region002.gbk	15	7910896 7912585	2	
Streptomyces sp.	NZ_CP033073.1.region001.gbk	15	1346573 1346876	2	
Streptomyces sp.	NZ_CP033073.1.region002.gbk	15	1346573 1346876	2	
Pseudanabaena sp.	NZ_AP017560.1.region003.gbk	15	3084758 3084977	2	
Streptomyces collinus	NC_021985.1.region001.gbk	14	7704825 7705230	1	
Streptomyces collinus	NC_021985.1.region004.gbk	14	7704825 7705230	1	
Halalkalicoccus jeotgali	NC_014297.1.region002.gbk	14	1910294 1910558	1	
Streptomyces collinus	NC_021985.1.region002.gbk	14	7704825 7705230	1	
Haloterrigena turkmenica	NC_013743.1.region002.gbk	14	2435960 2436389	1	
Streptomyces sp.	NZ_CP013142.1.region001.gbk	14	7357366 7359814	1	
Streptomyces sp.	NZ_CP029541.1.region004.gbk	14	1287691 1288000	1	
Leptolyngbya sp.	NZ_AP017367.1.region001.gbk	12	2850579 2850912	3	
Geitlerinema sp.	NC_019703.1.region002.gbk	9	1136345 1136564	3	
Pseudanabaena sp.	NZ_AP017560.1.region001.gbk	7	3084758 3084977	4	
Rhodococcus opacus	NZ_CP003950.1.region001.gbk	6	107605 107965	2	
Synechococcus sp.	NC_007775.1.region001.gbk	5	1489544 1490024	2	
Synechococcus sp.	NC_019702.1.region003.gbk	5	569343 569778	2	
Pseudanabaena sp.	NC_019701.1.region003.gbk	5	1028154 1028499	2	
Rhodococcus opacus	NZ_CP003949.1.region003.gbk	4	4749401 4749746	1	
Pseudanabaena sp.	NC_019701.1.region001.gbk	4	1028154 1028499	1	

Table 3: All found gvGCs