POLYPHASIC CHARACTERIZATION OF AN EPILITHIC BIOFILM FROM A LAVA CAVE IN KĪLAUEA CALDERA, HAWAIʻI

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

MICROBIOLOGY

DECEMBER 2012

By
Jimmy Hser Wah Saw

Dissertation Committee:

Stuart P. Donachie, Chairperson
Gernot G. Presting
Tung T. Hoang
Paul Q. Patek
Hongwei Li

We certify that we have read this dissertation and that, in our opinion, it is satisfactory in scope and quality as a dissertation for the degree of Doctor of Philosophy in Microbiology.

DISSERTATION COMMITTEE

_____
Chairperson

_____

_____

_____

_____

Dedicated to

my parents and late grandparents.

Acknowledgments

I would like to thank my advisor Dr. Stuart Donachie for his mentorship and for giving me an opportunity to work on this project in his lab. This work would not have been possible without his help and support, both intellectually and in other aspects of my life. I also thank him for being a friend and for supporting me personally. It would have been a really difficult journey without Stuart's support, both as a mentor and as a friend.

I would like to acknowledge Dr. Tung Hoang and his lab for supporting me with their time and supplies, Dr. Shaobin Hou for help with DNA sequencing, and Dr. Sean Callahan and his lab for letting me use their light incubator in which to cultivate cyanobacteria. I am grateful to Dr. Maqsudul Alam for his mentorship and guidance when I was an undergraduate and graduate student in his lab. Thank you to Dr. Paul Patek for years of advice in his role as graduate chair, and for letting me work on the Department of Microbiology website; to Dr. Gernot Presting for advice with bioinformatics work, and Dr. Dennis Kunkel for assistance with electron microscopy. I would also like to acknowledge Dr. Michael Schatz for advice on genome assembly and verification, and Dr. Thijs Ettema for advice on evolutionary genomics and phylogenomics related questions. Further thanks to Dr. Durrell Kapan for his serving earlier as one of my committee advisors, and Dr. Hongwei Li for serving as a new committee member. I also thank Siobhan Burns for proofreading a dissertation draft.

This dissertation was generated in LaTeX thus enabling code syntax highlighting and other useful features such as links to references, figures and tables. Thanks to Robert Brewer and the UH Manoa Information and Computer Sciences department for making the LaTeX dissertation template, thus making my life so much easier. Thanks to Pavel Senin for help with LaTeX related questions.

I am grateful to my friend Jason Aung and his family for helping me out with my living situation and for financial help. I cannot imagine how difficult my life would have been without their support and friendship. I would also like to thank Mark Chaplin for helping me with my living situation, especially as I would have been homeless without his help towards the end of my dissertation writing. Many thanks to my brother Ronald Saw for supporting my parents for many

years when I could not, and also for helping me out when times were hard. And last but definitely not least I thank my girlfriend Karli for giving me the motivation to finish my PhD.

Abstract

The microbial community in an epilithic biofilm on an lava cave wall in Kīlauea Caldera, Hawai'i, was characterized by a polyphasic approach. Ribosomal-pyrotag and metagenomic sequencing revealed phylogenetic diversity rivaling that in a Guerrero Negro hypersaline microbial mat. Targeted cultivations led to the isolation, characterization, and genome sequencing of a deeply divergent novel cyanobacterium. Diverse *Bacteria* and *Archaea* lineages were detected. The most abundant sequences, representing $\sim$24% of the metagenomic reads analyzed, affiliated with *Burkholderia*. Comparative metagenomic analyses revealed community composition and function most similar to those in soils. Two novel cyanobacteria detected in metagenomic data were cultivated; JS1 is related to *Gloeobacter violaceus* PCC 7421^T, the only cultivated *Gloeobacter* species. JS2 may represent a new genus in the Oscillatoriales since it shares $<$95% 16S rRNA gene sequence identity with its nearest neighbor, a *Leptolyngbya* sp. A third cultivated cyanobacterium (JS3) not detected in clone libraries, ribosomal-pyrotag or metagenomic data sets, belongs in the true-branching filamentous Stigonematales; JS3 shares 98.1% 16S rRNA gene sequence identity with *Fischerella muscicola* PCC 7414, and may be a new *Fischerella* sp.

Comparing the complete genome sequence of JS1 with that of *G. violaceus* PCC 7421^T revealed JS1 represents a new species, despite sharing 98.7% 16S rRNA gene sequence identity with PCC7421^T. The name *Candidatus* Gloeobacter kilaueaensis is proposed, with JS1^T the Type strain. Maximum likelihood phylogenetic trees based on 16S rRNA gene sequences and 43 concatenated ribosomal proteins showed *Candidatus* Gloeobacter kilaueaensis JS1^T places in the deep-branching *Gloeobacter* clade, but is less basal than *G. violaceus*. Divergence times based on Bayesian analyses suggested these *Gloeobacter* species diverged 150-300 MYA. The isolation, characterization, and genome sequencing of a deeply divergent novel *Gloeobacter* is significant given that for forty years we have known only one species in the entire order. Of broader significance is confirmation that by integrating molecular and cultivation methods we can target for cultivation specific *Bacteria* and or *Archaea* only detected in molecular analyses; a range of scripts was also developed to analyze and visualize sequence data.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Microbes appear to be ubiquitous on Earth, inhabiting features characterized by persistent and extreme cold, to those which are hot and perhaps additionally defined by extreme pH. Bacteria have even been found in places once thought most unlikely to host viable bacterial communities, such as deep subsurface mines, and the Atacama desert [1, 2, 3, 4, 5, 6]. Life on Earth is considered to have begun as structurally and functionally simple unicellular organisms. Biological activities, however, have shaped Earth's geosphere for billions of years, including involvement in the dissolution and precipitation of minerals. Bacterial impacts on the atmosphere are most prominent in the rise of oxygen gas (O_2), which is directly attributed to oxygenic photosynthetic microbes, and later also to chloroplasts in plants [7]. Even today, microbial metabolic processes and their interactions with organic and inorganic compounds is crucial in the functioning of ecosystems [8, 9, 10, 11]

Traditionally, the study of microbes was based only upon those those that grew on a nutrient medium, and which could be maintained for further study as live cultures. With the advent of molecular, or more specifically DNA-based methods, it became possible to detect and study microbes that have (or had) no known cultivated representatives [12, 13, 14, 15]. Such observations led to the opinion that not many microbes are 'culturable'; others might suggest that any microbe which grows in an environment must be culturable if the right conditions are provided. More recently, and even with considerable advances in molecular biology and DNA sequencing technology, it appears that culture-independent methods also fail to detect some microorganisms only detected by cultivation approaches [16]. Microbial ecologists might now advocate combining both 'traditional' cultivation approaches and molecular methods in an integrated manner to maximize detection of taxa in the environment [17]. It must be said that if we do not know how much diversity is in a sample to begin with, we are unlikely to know what fraction we are detecting [18].

Hawai‘i is famous for its biological diversity, yet there have been few studies of microbial diversity in the archipelago. Some work has focused on the few lakes in Hawai‘i [16], while other studies have focused on volcanic habitats [19, 20, 21, 22, 23, 24, 25, 26, 27]. Efforts to characterize microbial diversity in Hawaiian lava caves have been limited in terms of the geographic range of the caves involved [28, 29]. Even fewer studies have considered relatively young caves in volcanically active features in Hawai‘i, although phototrophic microbial mats in caves and low-light habitats have been reported [30, 31].

Caves present sites whose study provides insights into analogous systems on other planets, such as Mars [32]. Caves could protect life from harmful solar and cosmic radiations from space, as might be the case on planets such as Mars that lack a magnetosphere to deflect charged particles from the Sun, or an ozone layer that absorbs UV radiation [33]. Had life evolved on Mars, the best chances for its survival, according to our experience of terrestrial life, would be in or around caves; caves provide not only shelter, but moisture and warmth due to heat trapped inside the cave. Whether or not life exists or did exist on Mars remains to be determined, but understanding cave systems on Earth, and the nature of life in these caves, will be crucial as we move towards manned exploration of Mars and other planets. Caves tend also to be unexplored because they are difficult to reach, may be remote, and are often just dangerous to explore; microbial community structure and function in different cave systems remains poorly described.

1.1 Ecological surveys and community genomics/metagenomics of similar habitats

This dissertation focuses on an epilithic biofilm on an indirectly illuminated entrance wall of a lava cave in a volcanically active crater. To put the biofilm in question into context, I chose to compare it with those in habitats or features that are comparable either in a geomorphological or microbiological context, e.g., caves in other areas, volcanic soils, microbial mats from hypersaline ponds, and microbial mats from hot springs.

Literature searches for cave microbial communities show contributions from just a handful of researchers. For example, Northup et al. (2003) explored microbial communities from ferromanganese deposits in Lechugilla and Spider Caves in New Mexico [28], and reported a microbial community comprising iron- and manganese-oxidizing and reducing bacteria from diverse lineages. Along with geochemical analyses, the authors supported hypotheses that microbes contribute to mineral deposition in these habitats. Spanish researchers also described cyanobacteria in the Gelada Cave [34].

Microbial communities have been investigated in lava caves in New Mexico, Hawai'i, and the Azores through 16S rRNA gene clone libraries and electron microscopy [29]. Clones prepared from genomic DNA extracted from epilithic mats of different colors and secondary mineral deposits showed a total of 14 *Bacteria* phyla occurred in samples from 12 caves from three sites. The *Actinobacteria* phylum was ubiquitous, occuring in all these caves. The microbial diversity profile from the caves was said to be closely related to those of soil microbial communities. Copper-containing blue-green mineral deposits from Kipuka Kanohina Cave Preserve in Mauna Loa, and gold-colored mineral deposits from Thurston Lava Tube in Kilauea that appeared nonbiological contained filamentous microbes according to a visual interpretation of scanning electron micrographs [29]. However, all these samples were collected from the unlit sections of the lava tubes in question.

Macalady et al. (2012) explored community structure, physiology and biogeochemistry of extremely acidophilic sulfur-oxidizing biofilm (snottites) in the Frasassi cave system in Italy [35]. Bacteria from the genus *Acidithiobacillus* were found to be responsible for the formation and morphology of extremely acidic snottites, and were also found to be the primary producers in this highly specialized microbial community. This work also highlighted the role of a single phylum in shaping the snottite community in unlit sections of the cave.

Extensive work related to microbial diversity has been conducted around the Kīlauea volcanic zone [19, 20, 21, 22]. Studies of most relevance to this work include those of both microbial diversity and of carbon monoxide (CO) oxidizers in Kīlauea caldera's volcanic soils [19, 20, 21, 22, 23] but none of the cited work involved caves within the Kīlauea Caldera itself.

Other microbial habitats or features analagous to that in this study are mats from hypersaline habitats and hot springs. The Guerrero Negro hypersaline mat is a phototrophic-based microbial mat in hypersaline lagoons in Baja California Sur, Mexico [36, 37]. The mat was reported to contain unprecedented *Bacteria* diversity [36]. Although cyanobacteria formed most of the mat's biomass, *Chloroflexi* dominated the clone libraries, followed by *Proteobacteria* and *Bacteroidetes*. Previously undetected candidate phyla also occured in the clone libraries. Biological complexity in this mat was said to exceed that in other complex habitats, such as human and mouse distal guts. A further detailed metagenomic characterization of the mat system was conducted in 2008 using millimeter-scale mapping, revealing differences in microbial diversity at different depths [37].

Other similar habitats in which microbial mats have been characterized are hot springs in Yellowstone National Park (YNP), which have been investigated by both metagenomic and metatranscriptomic approaches [38, 39, 40]. These studies, however, showed the the mat communities

were dominated by unicellular cyanobacteria of the genus *Synechococcus* that are more closely related to the deeply divergent *Gloeobacter*.

This study focuses on an epilithic microbial biofilm that differs from the slime and ooze-type mats in caves described by Northup et al. (2011) [29]. Here, the biofilm's purple color was tentatively attributed to the presence of a photopigment. This habitat likely does not fall strictly into one defining category; initial assumptions might be that the biofilm could host microbes dispersed by air from the nearby volcanic soil, and also from the plant rhizosphere above the cave through percolation of meteoric water, or from grasses in soil near the cave entrance (Figure 2.3). Combining cultivation and molecular approaches in microbial diversity studies enables a better characterization of the community structure and function than if just one such method were applied [16]. Thus, this approach here would provide a better understanding of the diversity and roles played by microbes in this unique habitat.

1.2 Role of cyanobacteria in rock alteration and mineral formation

The significance of cyanobacteria in global biogeochemical cycles and maintenance of life on Earth is well-known. They are primary producers responsible for much of the gaseous oxygen in Earth's atmosphere. However, their role in a more geological context is often poorly defined and even largely overlooked. Due to the ability of some to form mucilaginous sheaths and biofilms, and to promote the precipitation of calcium carbonate, some cyanobacteria form structures known as stromatolites. These are well known in the fossil record [41] and are useful in geology and biology for estimating emergence times of different microbial lineages across geological timeframes [42].

A few recent papers have highlighted the role of cyanobacteria in intracellular carbonate [43] and iron [44] precipitations. Couradeau et al. (2012) discovered an early diverging cyanobacterium, *Candidatus* Gloeomargarita lithophora that forms, intracellularly, a type of carbonate known as benstonite. Carbonates such as calcites have been known to be formed extracellularly by microbes, but intracellular carbonate formation was unknown until recently [44]. Mineralized or calcified cyanobacteria are occasionally found in the microfossil record, but they are relatively rare [45, 46]. Only in microfossils younger than 1200 million years are well-calcified cyanobacterial sheaths described; they are rarer in older rock records [46]. Benstonite precipitates in *Candidatus* Gloeomargarita lithophora contain as much barium, magnesium, and strontium as they contain calcium, and it has been said it would be more sensible to look for these minerals in similar proportions in microfossils rather than microscopically identifying the tiny inclusion bodies, which are simply

more difficult to detect visually. Elsewhere, *Marsacia ferruginose* JSC-1 (*Leptolyngbya*, order Oscillatorales) is involved in iron redox cycling and deposition, and was originally isolated from an iron-depositing hot spring in Yellowstone National Park [44].

1.3 Cyanobacteria genomics

Since the advent of shotgun sequencing, the number of fully sequenced cyanobacteria genomes has increased steadily. To date, 45 complete genomes from a total of 2,329 complete microbial genomes are from cyanobacteria (Table 1.1) and along with 30 draft genomes of 3,967 draft microbial genomes (Table 1.2) (as of 07/17/2012) (`ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/`). Cyanobacteria genomes thus represent less than 2% of all completely sequenced microbial genomes to date. A need clearly exists to fully sequence genomes of more cyanobacteria because they are among the most morphologically diverse microbes, and they perform important fractions of global primary production and nitrogen fixation. More reference cyanobacteria genomes are also needed if we are to fully understand the origin and evolution of photosynthesis.

The best represented genus among the cyanobacteria in terms of number of genomes sequenced, with twelve complete genomes to date, is *Prochlorococcus*, followed by *Synechococcus* with eleven complete genomes (Table 1.1). Among the genus *Gloeobacter*, however, *Gloeobacter violaceus* PCC 7421 is the only representative of the genus and thus the only one whose genome has been sequenced. However, the *Gloeobacter violaceus* genome is significant because it has been used to root most cyanobacteria phylogenetic trees [47, 48, 49] or trees used to investigate the evolution of specific genes in cyanobacterial or photosynthetic lineages [50, 51, 52].

Gloeobacter is in a strategic position in the *Cyanobacteria* lineage, and any information related to its origin and evolution would be of enormous interest to the scientific community; not only for cyanobacteria research, but also for evolution of photosynthesis and developmental biology, such as the origin of thylakoid membranes in *Cyanobacteria*. The aim here was to cultivate the *Gloeobacter* previously identified in a clone library constructed from genomic DNA extracted from the epilithic biofilm, sequence its genome, and interpret this information in the context of metagenomic data also generated from the biofilm.

Table 1.1. Complete *Cyanobacteria* genome sequences available from NCBI

Species name	genome size (Mbp)	NCBI Bioproject number
Prochlorococcus marinus subsp. pastoris str. CCMP1986	1.66	PRJNA57761
Prochlorococcus marinus str. MIT 9313	2.41	PRJNA57773
Prochlorococcus marinus subsp. marinus str. CCMP1375	1.75	PRJNA57995
Prochlorococcus marinus str. MIT 9303	2.68	PRJNA58305
Prochlorococcus marinus str. AS9601	1.70	PRJNA58307
Prochlorococcus marinus str. MIT 9211	1.69	PRJNA58309
Prochlorococcus marinus str. MIT 9515	1.70	PRJNA58313
Prochlorococcus marinus str. MIT 9312	1.71	PRJNA58357
Prochlorococcus marinus str. NATL2A	1.84	PRJNA58359
Prochlorococcus marinus str. NATL1A	1.86	PRJNA58423
Prochlorococcus marinus str. MIT 9301	1.64	PRJNA58437
Prochlorococcus marinus str. MIT 9215	1.74	PRJNA58819
Synechococcus elongatus PCC 7942	2.74	PRJNA58045
Synechococcus elongatus PCC 6301	2.70	PRJNA58235
Microcystis aeruginosa NIES-843	5.84	PRJNA59101
Nostoc punctiforme PCC 73102	9.06	PRJNA57767
Thermosynechococcus elongatus BP-1	2.59	PRJNA57907
Trichodesmium erythraeum IMS101	7.75	PRJNA57925
Gloeobacter violaceus PCC 7421	4.66	PRJNA58011
Acaryochloris marina MBIC11017	8.36	PRJNA58167
Arthrospira platensis NIES-39	6.79	PRJDA42161
Anabaena variabilis ATCC 29413	7.11	PRJNA58043
'Nostoc azollae' 0708	5.49	PRJNA49725
cyanobacterium UCYN-A	1.44	PRJNA43697
Synechococcus sp. CC9311	2.61	PRJNA58123
Synechococcus sp. CC9605	2.51	PRJNA58319
Synechococcus sp. CC9902	2.23	PRJNA58323
Synechococcus sp. JA-3-3Ab	2.93	PRJNA58535
Synechococcus sp. JA-2-3B'a(2-13)	3.05	PRJNA58537
Synechococcus sp. PCC 7002	3.41	PRJNA59137
Synechococcus sp. WH 8102	2.43	PRJNA61581
Synechococcus sp. WH 7803	2.37	PRJNA61607
Synechococcus sp. RCC307	2.22	PRJNA61609
Nostoc sp. PCC 7120	7.21	PRJNA57803
Synechocystis sp. PCC 6803	3.95	PRJNA57659
Synechocystis sp. PCC 6803 substr. PCC-P	3.57	PRJDA72557
Synechocystis sp. PCC 6803 substr. GT-I	3.57	PRJNA158059
Synechocystis sp. PCC 6803 substr. PCC-N	3.57	PRJNA159835
Synechocystis sp. PCC 6803	3.70	PRJNA159873
Cyanothece sp. PCC 7822	7.84	PRJNA52547
Cyanothece sp. ATCC 51142	5.46	PRJNA59013
Cyanothece sp. PCC 7424	6.55	PRJNA59025
Cyanothece sp. PCC 8801	4.79	PRJNA59027
Cyanothece sp. PCC 8802	4.80	PRJNA59143
Cyanothece sp. PCC 7425	5.79	PRJNA59435

Table 1.2. Draft *Cyanobacteria* genome sequences available from NCBI

Species name	Estimated genome size (Mbp)	NCBI Bioproject number
Prochlorococcus marinus str. MIT 9202	1.69039	PRJNA54709
Crocosphaera watsonii WH 8501	6.23816	PRJNA54123
Crocosphaera watsonii WH 0003	5.8905	PRJNA61839
Nodularia spumigena CCY9414	5.31626	PRJNA54171
Arthrospira platensis str. Paraca	4.99756	PRJNA55907
Microcoleus chthonoplastes PCC 7420	8.65162	PRJNA54695
Leptolyngbya valderiana BDU 20041	0.089264	PRJNA54785
Arthrospira maxima CS-328	6.00331	PRJNA55093
Cylindrospermopsis raciborskii CS-505	3.87903	PRJNA42983
Raphidiopsis brookii D9	3.18651	PRJNA42981
Microcoleus vaginatus FGP-2	6.69893	PRJNA67389
Moorea producta 3L	8.38942	PRJNA66849
Synechococcus sp. WH 7805	2.62037	PRJNA54217
Synechococcus sp. WH 5701	3.04383	PRJNA54219
Synechococcus sp. RS9917	2.57954	PRJNA54221
Synechococcus sp. RS9916	2.66446	PRJNA54223
Synechococcus sp. BL107	2.28338	PRJNA54225
Synechococcus sp. PCC 7335	5.96411	PRJNA54731
Synechococcus sp. WH 8109	2.11849	PRJNA55973
Synechococcus sp. CB0205	2.42731	PRJNA61893
Synechococcus sp. CB0101	2.6864	PRJNA61895
Synechococcus sp. WH 8016	2.69484	PRJNA74433
Oscillatoria sp. PCC 6506	6.67671	PRJNA50611
Fischerella sp. JSC-11	5.38	PRJNA75099
Lyngbya sp. PCC 8106	7.03751	PRJNA54161
Acaryochloris sp. CCMEE 5410	7.87548	PRJNA78283
Cyanothece sp. CCY0110	5.88053	PRJNA54615
Cyanothece sp. ATCC 51472	5.42819	PRJNA75093
Cyanobium sp. PCC 7001	2.8327	PRJNA54675
Arthrospira sp. PCC 8005	-	PRJNA49969

1.4 Review of recent approaches in genomics

Approaches to sequencing microbial genomes have their inherent advantages and disadvantages, but the nature of the starting material can significantly affect the outcome. For example, we might sequence the genome of cells from a pure culture (cf. isolate genome sequencing), or in single cells only, or by reconstruction of complete genomes from metagenomic sequence data.

The 'isolate genome sequencing' approach in this context refers to traditional shotgun sequencing of microbes based on genomic DNA isolated from a pure culture. This straightforward method has been used in most assemblies of complete microbial genomes. The method relies on having a high concentration of genomic DNA for sequencing libraries, concentrations generally only obtainable from a large volume of culture. In this respect, many environmental microbes may be difficult to cultivate, or their doubling times may be very long [53], so the genomes of such recalcitrant organisms have been difficult to sequence.

Metagenomic and single-cell genome sequencing methods can bypass cultivation steps needed in the isolate genome sequencing approach. However, each has its limitations, so no one method is perfect for all applications. The metagenomic sequencing approach can quickly reveal the genomic potential and metabolic diversity of a community, but assembling complete genomes from metagenomic data is rarely possible [54, 55]. Assembly of near-complete genomes from metagenomic data is possible, but is not a trivial task, and has been only feasible in low-complexity microbial communities [54]. In a recent paper, however, new algorithms were developed to isolate and assemble the complete genome of an *Archaea* poorly represented in metagenomic data [55]. This was achieved using state-of-art computational algorithms and deep sequencing of the microbial community. Although quite promising, this approach is not yet suitable for all sequencing applications.

Single-cell genome sequencing approaches have become popular in recent years, as they can circumvent the need for cultivation while succeeding with small samples. Advances in single-cell transcriptomics by Kang et al. (2011) have shown it is possible to study individual bacterial cells and their role in disease pathogenesis [56]. Several key technologies are crucial in isolating single bacterial cells from an environmental sample, such as flow cytometry, microfluidic devices, and mechanical/optical manipulation devices. Isolating eukaryotic cells is straightforward given their relatively large size, but isolating single cells of specific *Bacteria* and *Archaea* from mixed cultures can be difficult as they may have similar cell sizes and shapes.

Environmental samples may contain many different organisms in the sample matrix, and it can be difficult to pinpoint the organism of interest. Fluorescent labeling methods can distinguish target organisms, and such an approach has been used to isolate and sequence genomes of marine microbes [57, 58, 59, 60]. However, few among such reports have managed to assemble complete genomes from single cells [60, 59, 57]; indeed, a population of cells has generally been collected as opposed to a single bacterial cell. The single-cell sequencing approach is not without its problems, such as the introduction of chimeric sequences during multiple displacement amplification (MDA), and contamination due to the high sensitivity of the ϕ29 DNA polymerase [61]. An absolutely clean environment is needed for this work, while assemblies based on libraries comprising MDA-amplified DNA also tend to miss some regions in a genome, so assemblies of complete genomes are still quite rare [58].

The Genomic Encyclopedia of Bacteria and Archaea (GEBA) project was initiated by the Department of Energy (DOE) Joint Genome Institute (JGI) in 2009 [62] (`http://www.jgi.doe.gov/programs/GEBA/pilot.html`). This project is driven primarily by phy-

logenomics, and intends to fill the void left by early genome sequencing projects that tended to choose organisms based on physiology (*e.g.*, disease-causing pathogens) rather than their presumed place in the Tree of Life. As such, many microbes that are evolutionarily interesting were left out of early sequencing projects, and their importance is only now being addressed in metagenomic sequencing projects. By revealing a large array of organisms never before detected in the environment, however, we find that very few 'reference' genomes exist for comparison when a metagenome sample is analyzed. For example, 'recruitment' of sequences, a process that gathers related sequences of reference organisms from metagenomes, often shows very little recruitment in such circumstances because there are few such appropriate 'reference' genomes [63, 64, 58].

GEBA aims to fill this gap by giving more attention to less well-known but still interesting microbes. Most of the genome sequencing for the GEBA project was in fact carried out on cultivated strains, largely due to the fact that this approach yields high-quality draft assemblies; building complete genomes is also possible because high-quality genomic DNA can be extracted from only pure cultures. Complete genomes are crucial in phylogenomics, where one compares exact numbers of gene gains or losses due to evolution, and to confirm why some genes are present in an organism and not in another. This highlights why we should not abandon traditional cultivation-based approaches, and why we should increase our efforts to further knowledge of how to cultivate environmental microbes.

1.5 Scope of current work and specific aims of the dissertation

This is the first metagenomic study of a phototrophic epilithic biofilm in a lava cave. Studies of phototrophic mats elsewhere have considered those from hot springs or hypersaline systems [38, 39, 37]. The microbiology of cave systems in Hawai'i has been reported, but no such work has applied metagenomic sequencing for diversity or considered lava caves in Kīlauea Caldera [29]. The epilithic biofilm discussed here, however, is quite unique; it is in part analagous to hotspring microbial mats because of the availablity of light (albeit of lower intensity here), copious water, and heat, yet a mat of this type (pigmentation, presence of photoautrophs, location) appears never to have been described. The discovery of this epilithic biofilm led first to a brief survey of its bacterial diversity through a 16S rRNA gene clone library, and subsequently to the cultivation and complete genome sequencing of a novel *Gloeobacter* detected in the clone library. This work is significant because the *Gloeobacter* species cultivated is just the second species reported in the genus in 38 years.

Due to the rarity of such epilithic biofilms in cave systems in Kīlauea Caldera, with this purple type being seen in just one of the ∼200 caves in the caldera, initial work here aimed to apply single-cell genome sequencing to novel microbes therein; this approach needs very little sample for isolation of bacterial cells, and thus helps conserve the material. Single-cell genome sequencing was attempted here, but the attempts failed; traditional cultivation-based approaches to provide sufficient material for genome sequencing were used instead. The results of the single-cell genome sequencing approaches are not included in this dissertation.

Specific aims of the research described in this dissertaion:

1. **Aim 1:** To describe phylogenetic diversity and metabolic potential of microorganisms in the epilithic biofilm through 16S rDNA variable sequence (pyrotag) and metagenomic data.
2. **Aim 2:** To target for cultivation potentially novel microbes identified in molecular data from the epilithic biofilm.
3. **Aim 3:** To isolate and sequence a novel *Gloeobacter* sp. identified in preliminary studies of the epilithic biofilm.

Rationale for each specific aim:

Aim 1: To describe phylogenetic diversity and metabolic potential of microorganisms in the epilithic biofilm through 16S rDNA variable sequence (pyrotag) and metagenomic data. Metagenomic analyses allow the nature and extent of the community's metabolic diversity to be determined, and to shed light on whether or not the community is unique. Metabolic diversity and adaptations to the environment in this community are expected to be similar to those in other cyanobacterial mats, such as that described in the hypersaline Guerrero Negro [37, 36]. Comparing results from these two environments should reveal genes and pathways both common and unique to each. It is likely that the lava cave epilithic biofilm community has not been adequately sampled, but given the number and size of the sequences thus far available, a well-informed estimate about the extent of sampling might be derived, as might whether or not the community's estimated metabolic diversity can be meaningfully compared to metabolic diversity in other environments. Through metabolic pathway annotationss, comparative metagenomics indicates if a community's physiologies and functions are more closely related to one environment than others.

Aim 2: To target for cultivation potentially novel microbes identified in molecular data from the epilithic biofilm. Several sequences in the Kīlauea 16S rDNA clone library affiliate with

potentially new species or genera. As either or both of culture and genomic information from these new species would be valuable contributions to the community, some of the potentially novel lineages of *Bacteria* and possibly *Archaea* were targeted for cultivation. Cultivation is an essential part of such research, since a microbe's complete genome can be sequenced without cultivation (e.g., single-cell genome sequencing), but by doing so we derive little information about the microbe's physiology and role *in situ*.

Aim 3: To isolate and sequence a novel *Gloeobacter* sp. identified in preliminary studies of the epilithic biofilm. It is known that *Gloeobacter* is one of the earliest diverging cyanobacteria. Having genomic information for the strain or species cultivated here would alone be an outstanding result, especially as only one other *Gloeobacter* strain and species is known. Thus, having in culture only the second known *Gloeobacter* species, or even a strain of the type species, and then information from its complete genome, would provide valuable information for cyanobacteria researchers globally. Not only could this newly cultivated species provide a resource for cyanobacteria genomics, it might also provide material for genetic experiments, such as mutational analyses, or directed evolution to investigate how the photosynthesis apparatus adapted to higher light intensities. Ultimately, both molecular and genomics methods may be combined with traditional cultivation-based approaches to target novel microbes in the environment for detailed characterization.

Combining these three approaches may provide the basis for significantly advancing microbial ecology for some time to come. Such a polyphasic approach not only facilitates rapid surveys of microbial diversity in a community, but also simplifies the cultivation of specific organisms. For example, a microbe of biotechnological interest can be quickly identified using metagenomics, cultivated (perhaps with some imagination, creativity and patience), and then its complete genome sequence might reveal genes of interest that can be engineered to improve function and yield of desired substrates. This dissertation provides an example of how the polyphasic approach can characterize a functionally interesting microbe from the environment.

Chapter 2

Metagenomics sequencing and analysis of an epilithic phototrophic microbial mat from Kīlauea, Hawai‘i

Jimmy Saw, Mark Brown, Jamie Foster, Yuri Wolf, Michael Galperin, Eugene Koonin, Stephan Kempe, Harry Schick, Keali‘imanaulu‘okeahe Taylor, Stuart Donachie. *In preparation. To be submitted to PLoS ONE or Geobiology*

2.1 Abstract

We present a metagenomic study of a previously unreported epilithic biofilm from a lava cave wall in Kīlauea Caldera, Hawai‘i. Pyrotag-sequencing of a variable region of the 16S rRNA gene revealed unprecedented diversity among *Bacteria* and *Archaea*. Metagenomic analysis based on 454 Pyrosequencing revealed a highly complex community dominated by *Proteobacteria*, predominantly beta*Proteobacteria*. In addition to the *Proteobacteria*, other *Bacteria* phyla are prominent, *e.g.*, *Acidobacteria*, *Actinobacteria*, *Chloroflexi*, and *Cyanobacteria*. *Archaea* sequences belonging to the *Euryarchaeota* and *Crenarchaeota* were detected. The *Euryarchaeota* sequences were the most common of these *Archaea* lineages. Comparison with metagenomes from other habitats showed the epilithic biofilm is functionally related to soil microbial communities. This work describes the first metagenomic analysis of a phototrophic microbial mat in a volcanic lava cave.

2.2 Introduction

Caves are important sites in which to study how organisms adapt to their surroundings, and also how they might evolve in what are generally rather stable environments. The lack of light, temperatures lower on average than those outside the cave, and higher relative humidity, all create micro-climates that tend to support unique fauna that have both adapted to and in some cases evolved in caves. Caves are also important from an astrobiological perspective, since Mars especially is known to have caves [32].

Microbial biofilms occur in habitats as diverse as hypersaline ponds, hot springs, the human mouth, and hydrothermal vents [65, 66]. Communities in such habitats are often complex, with certain functions taken by different members of the community. The roles of such microbes in different habitats are unclear and require further investigation. Studies of microbial diversity in caves in Hawai‘i are rare. For example, attempts have been made to relate the biogeochemistry of mineral deposits to the presence of certain cell types and morphology, and some molecular work [28, 29]. Northup et al. (2011) investigated similarities and differences between microbial diversity in caves of Hawai‘i, New Mexico and the Azores [29] in preliminary studies of microbial and cyanobacterial diversity.

In 2005 a purple and green pigmented epilithic biofilm was observed on the downward facing wall of a cave entrance in Kīlauea Caldera. Samples of the biofilm were collected for cultivation and DNA-based analyses. Additional samples were collected during susequent visits until 2009, when increasingly hazardous gas emissions from the nearby Halema‘uma‘u crater closed access to the entire caldera. However, the material collected was the first of a lava cave epilithic biofilm in Hawai‘i to be analyzed by metagenomic and pyrotag-sequencing. The analyses revealed unprecedented taxonomic and metabolic diversity in the community, and that rare microbes could be targeted for single-cell genome analyses or cultivation.

2.3 Materials and Methods

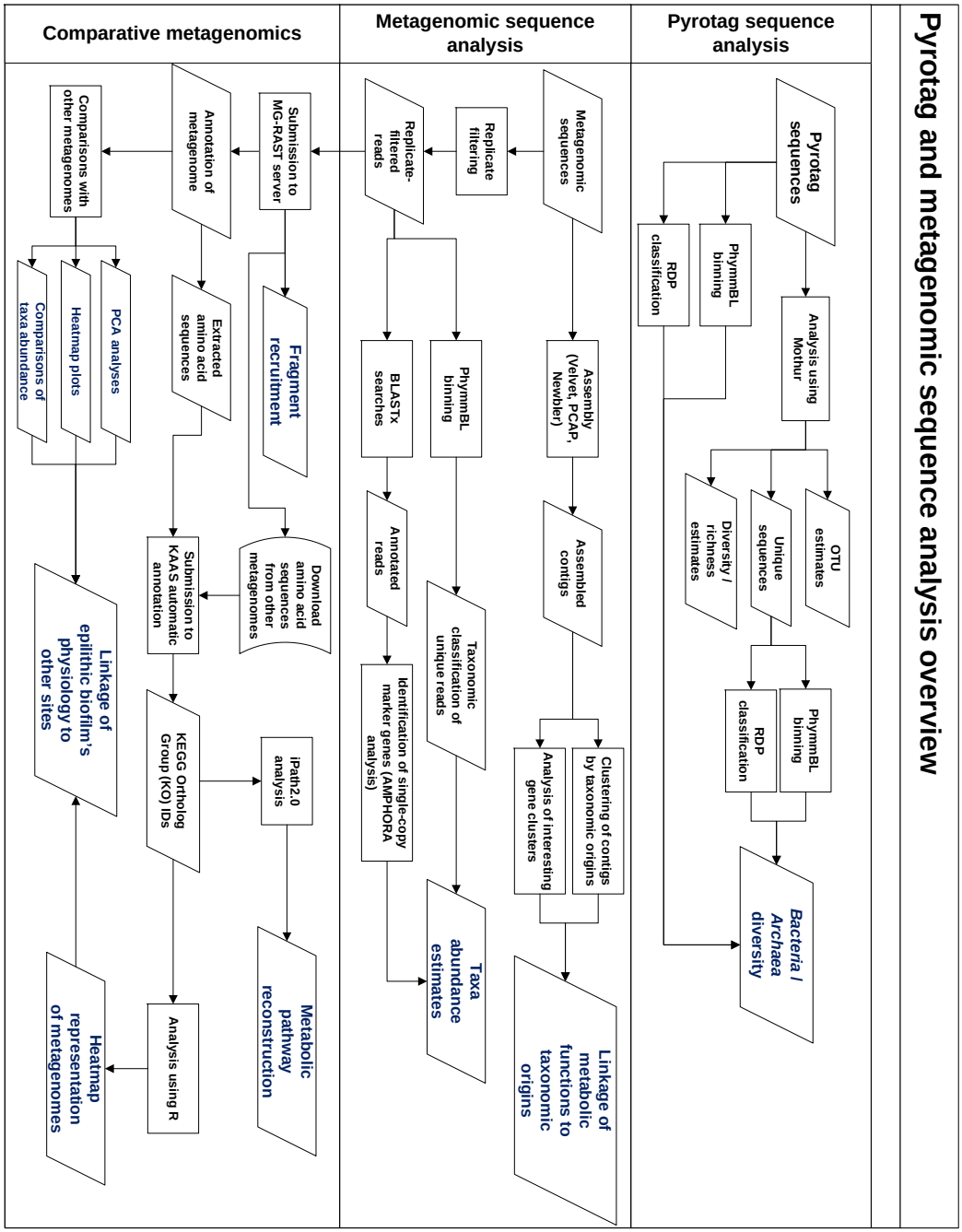


Figure 2.1. Flowchart of steps involved in the analysis of pyrotag and metagenomic sequences. Flowchart comprises three sections: pyrotag sequence analysis, metagenomic sequence analysis, and comparative metagenomics.

The analyses of pyrotag and metagenomic sequences from the cave biofilm community is summarized in the flowchart in Figure 2.1.

2.3.1 Field observations, sample collection and sequencing

Kīlauea Crater is a ∼5 x 3 km depression some 1300 m above sea level, in the Hawaii Volcanoes National Park. Multiple lava fields representing different eruption events since 1885 cover various parts of the crater floor. Lava caves form in many ways. For example, the surface of the lava may solidify while lava below drains away (as in the case of lava tubes), or the lava surface solidifies before gas bubbles escape; trapped bubbles can thus form lava domes. The cave described here is located in the 1919 lava flow.

The cave is located below ground level, as opposed to being of the type one might simply walk into, as in a cliff or hillside. The cave's only entrance is a ∼1 x 0.5 m horizontal fissure at ground level, along the ground level margin of slight dome that forms the roof of a bubble or fold in the lava (Figures 2.2 and 2.3). In the cave, part of the ceiling extends immediately below the entrance; although this surface is not in darkness nor oriented towards the inside of the cave, it is also not in a position to be directly illuminated by the sun. Conditions in the cave are very different from those immediately outside the entrance, and may be considered 'extreme'. A persistent winds sweeps from within the cave as warm air rises from within and escapes through the entrance. Parts of the cave floor at depths of ∼3 cm have reached ∼90°C (measured by temperature probe). Air temperature ∼20 cm above the floor was consistently 35 to 45°C, and relative humidity has exceeded 100% (Table 2.1). During several visits between 2006 and 2009, a heavy mist of condensation was present, and fell like a light rain from surfaces, including that of the epilithic biofilm.

Table 2.1. Site data

	Big Ell (1*)	Big Ell (2*)
T (min)	17.9°C	29.1°C
T (max)	45.9°C	38.3°C
T (mean)	27.1°C	33.1°C
Rh (min)	52.3	F
Rh (max)	102.3	F
Rh (mean)	95.3	F

T - temperature, Rh - relative humidity, F - failed

*Data was recorded every two minutes at cave entrance in May - August, 2006 (1) and September - November, 2006 (2).

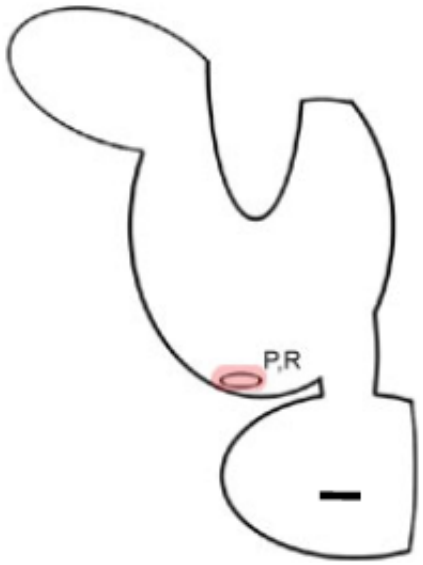


Figure 2.2. Schematic drawing of Big Ell cave outline. Cave entrance is highlighted in pink. P - temperature and relative humidity probes, R - rain gauge. Scale bar is $\sim$5 m

Although the biofilm was situated fully in the cave, its location near the entrance exposed it to low levels of light ($\sim$5 μEm^{-2}s^{-1}) at noon on a clear day; it was reasonable to assume that the biofilm thus comprised photoautotrophs as well as heterotrophs. The biofilm's appearence changed with time, especially evident in the diminishing fraction occupied by the purple material compared with the dark green material. Between 2006 and 2009, an approximately 50:50 split between green and purple parts of the $\sim$2 m^2 biofilm changed to just a few small patches of purple. Reasons for such a change are not known, but may well reflect changes in the level of volcanic activity, with concomitant changes in temperature, amount of groundwater and subsequent humidity in the cave, as well as increased concentrations of sulfur dioxide and carbon dioxide (See Figures 2.4(a) and 2.4(b)). Prior to 2009, the biofilm seemed to be dampened by condensation, but during the visit in 2009 the air was much less humid.

No lamination was evident in plugs of the mat collected into small tubes. This contrasts with the structure of phototrophic mats in hotsprings and salterns [36]. The mat was only a few millimeters thick, but there seemed a tendency for parts of it to flow or perhaps grow downwards, albeit slowly, through gravity (Figure 2.4(b)). Late in 2009, much of the top of the mat was green,

followed by a milky-white layer, and a dark bottom layer (Figure 2.4(b)). The purple sections of the mat occupied a few small patches among the predominantly green top layer.

Figure 2.3. Big Ell cave entrance to the right of the rock in the center of the frame. Note the vegetation on lava rocks surrounding the cave entrace.

For pyrotag and metagenomic sequencing, punch cores of purple areas of the epilithic biofilm were collected directly into sterile 50 mL polypropylene tubes and placed on dry-ice for return to the laboratory within eight hours of collection. In the laboratory, these samples were transferred to a -80°C freezer and stored until community genomic DNA was extracted in the MO BIO Ultraclean® Soil DNA isolation kit.

For pyrotag sequencing, variable regions in the *Bacteria* 16S rRNA gene were amplified by PCR with Taq DNA polymerase and primers designed to target V6-9 regions of 16S rRNA gene. PCR products were cleaned with the Qiagen DNA purification kit. Both purified PCR products of the amplicons and community genomic DNA were sent to a commercial sequencing company for sequencing of the pyrotags and the metagenome using a 454 GS FLX DNA sequencr (454 Life Sciences, Branford, CT).

(a)

(b)

Figure 2.4. Epilthic biofilm on cave wall. (a) Epilithic biofilm on cave entrance wall in 2007 (Hammer included for scale.), (b) Epilithic biofilm on cave entrance wall in 2009. Note purple patches.

2.3.2 Analysis of pyrotag sequences

The resultant pyrotag sequences were trimmed by custom scripts and analyzed in the MOTHUR program, version 1.25.1 [67], and processed according to the MOTHUR wiki page

(http://www.mothur.org/wiki/Schloss_SOP) to generate rarefaction curve of diversity and to measure richness in the biofilm.

Briefly, the following sets of commands were run in Mothur:

```
unique.seqs(fasta=MATTAGstrim.fasta)
align.seqs(template=prok.fasta, candidate=MATTAGstrim.unique.fasta, ksize=9, processors=2,
  flip=true)
filter.seqs(fasta=MATTAGstrim.unique.align, vertical=T)
dist.seqs(fasta=MATTAGstrim.unique.filter.fasta, cutoff=0.1, calc=eachgap, processors=2)
cluster(column=MATTAGstrim.unique.filter.dist, name=MATTAGstrim.names)
chimera.uchime(fasta=MATTAGstrim.unique.fasta, reference=prok.fasta, processors=2)
classify.seqs(fasta=MATTAGstrim.unique.filter.fasta, template=trainset7_112011.pds.fasta,
  taxonomy=trainset7_112011.pds.tax, cutoff=80, processors=2)
system(cp MATTAGstrim.unique.filter.pick.fasta mat.final.fasta)
system(cp MATTAGstrim.unique.filter.pds.taxonomy mat.final.taxonomy)
system(cp MATTAGstrim.names mat.final.names)
dist.seqs(fasta=mat.final.fasta, cutoff=0.15, processors=2)
cluster(column=mat.final.dist, name=mat.final.names)
classify.otu(list=mat.final.an.list, name=mat.final.names, taxonomy=mat.final.taxonomy,
  label=0.03)
```

The steps described above perform the following:

- Unique sequences are identified from trimmed sequences
- Unique sequences are then aligned against *Bacteria* and *Archaea* sequences using a built-in alignment tool
- Columns within alignments that are not useful were removed
- Distances between the sequences are then calculated
- OTUs are clustered based on default parameters, and
- Sequences are classified by searching against the traning set provided with Mothur tutorials

2.3.3 Analysis of metagenomic sequences

A total of 386,217 metagenomic sequence reads were generated by the Genome Sequencer FLX system. This metagenomic data is subsequently referred to as 'HAVO', *i.e.*, short for Hawai'i Volcano. Exact duplicates in the data were removed by the 454 Replicate Filter tool [68], which left 349,106 sequences for further analyses. G+C % and tetranucleotide frequencies were calculated in custom Python scripts (See Section 5.2.2). Dereplicated metagenomic reads

were classified using the PhymmBL binning tool (version 3.2) [69, 70]. Parts of metabolic pathways present in the metagenome were determined by submitting MG-RAST annotated amino acid sequences to the KAAS automatic annotation server to retrieve KEGG ortholog (KO) numbers. These KO numbers were submitted to the iPath2.0 web server [71] to draw KEGG pathway atlases highlighting pathway components catalyzed by orthlogs found in the metagenome.

2.3.4 Metagenomic sequence assembly

The 386,217 metagenomic reads were assembled with default parameters in Newbler, Velvet [72], and PCAP [73] assemblers.

2.3.5 Comparative metagenomic analyses

The MG-RAST metagenomic sequence analysis tool is an online tool that allows comparison of metagenomic data sets from diverse environments [74, 75]. As it is not practical to download all available metagenomic data sets to a personal laptop computer (many require gigabytes of hard disk space) to perform comparisons, the convenience of tools already implemented in MG-RAST, and the availability of a large number of data sets already annotated for comparison, the MG-RAST server was used to compare the cave biofilm metagenome with several other samples from different habitats. The cave biofilm metagenome is hosted at the MG-RAST website and will be made publicly available pending submission of a manuscript to a peer-reviewed journal. The MG-RAST metagenomic sequence analysis pipeline use the following procedures: Dereplication of identical reads by an MG-RAST filter to keep only a single representative of reads whose first 50 bases are identical. Reads passing this filter are then compared in the M5NR database (MIGS 5 Non Redundant database) of proteins to assign function and taxonomic affiliation to each read.

The cave epilithic biofilm metagenome was compared with several environmental samples listed in Table 2.2 to determine whether or not it relates in particular to one or another. Some of the sites chosen are similar to the cave epilithic biofilm (e.g., phototrophic mats). Comparisons were conducted through the MG-RAST metagenomic web server. Metagenomes from habitats such as hydrothermal vents and soils were included to determine if the HAVO results correlate with particular niche specializations.

Table 2.2. Metagenomic samples compared with the epilithic biofilm metagenome

Habitat	Number of samples
Guerrero Negro hypersaline microbial mat ([37])	10
Hot spring microbial mats from Yellowstone National Park ([39])	6
Hot spring microbial mat from Diamond Fork hot spring, Utah	1
Puerto Rico forest soil	1
Netherlands forest soil ([76])	4
Lost City hydrothermal vent microbial mat	1
Mariana trough vent fluid samples	1
Microbial mat from hydrothermal vent at an unknown site	1
Total	25

For the principle coordinate analysis (PCoA) plots, the data were compared to the MG-RAST M5NR (M5 Non Redundant) protein database using a maximum BLASTx E-value of 1^{-5}, a minimum identity of 60%, and a minimum alignment length of 15 amino acids. The plot was drawn using normalized values and Bray-Curtis distance. M5NR complies with the Genomic Standards Consortium (GSC) [77] guidelines. Heatmaps and dendrograms comparing habitats were created in theMG-RAST comparison tool, which performs analyses on the basis of two criteria: 1) organism abundance, or 2) abundance of metabolic categories, between the metagenomic samples compared.

To compare environments in which metabolic functions are closely related, amino acid sequences identified in these environments (25, including the cave biofilm) were downloaded from the MG-RAST server and submitted to the KEGG Automatic Annotation server (KAAS) to assign KEGG Ortholog IDs (KOs) to these sequences. The abundance of 14,054 KOs was then counted in each habitat, and a 14055×25 matrix was created. The matrix was loaded to the R statistical tool to calculate Pearson Correlation values from which clustered heatmaps were created.

2.3.6 Calculation of Effective Genome Size (EGS)

Calculations of Effective Genome Size (EGS) use the following formula, from Raes et al. [78]:

$$EGS = \frac{a + b \times L^{-c}}{x} \tag{2.3.1}$$

where a = 21.2, b = 4230, and c = 0.733. x is marker gene density, and L is average read length of the metagenomic sequences, and a, b, and c are previously determined parameter estimates.

2.4 Results and Discussions

2.4.1 Summary of sequence data

Different analyses were performed to corroborate community diversity in the HAVO epilithic biofilm. Sampling variable regions of the 16S rRNA gene through pyrotag sequencing provides a good estimate of species richness, diversity, and evenness in a given community. However, a PCR bias is usually associated with such studies, such that diversity might be over or underestimated [79]. It is thus important to use an alternative method to test if the observed diversity through one approach can be confirmed by another. In this section, results obtained in different data set are compared to determine if the observed community structure and diversity is close to the 'actual' state.

Metagenomic sequencing lacks biases associated with PCR and is useful for estimating community structure and diversity. Sequencing artifacts are known in 454 pyrosequencing, but bioinformatics tools have been written to deal with these, such as artificial sequence duplicates [68]. Thus, both pyrotag and metagenomic sequences were used to estimate species diversity and richness in the biofilm community.

2.4.1.1 Pyrotag data

A total of 64,206 pyrotag sequences from the amplified V6-9 region were obtained for the HAVO epilithic biofilm. In Mothur, 5,383 of these sequences determined to be unique were used for further analysis. For example, they were aligned with the built-in alignment tool in Mothur with curated SILVA *Bacteria* and *Archaea* alignments to calculate distances between the sequences. The statistics associated with trimmed and unique pyrotag data are shown in Tables 2.3 and 2.4.

Table 2.3. Trimmed pyrotag sequence statistics

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	52	52	0	2	1
2.5%-tile:	1	58	58	0	3	1606
25%-tile:	1	60	60	0	3	16052
Median:	1	61	61	0	3	32104
75%-tile:	1	62	62	0	4	48155
97.5%-tile:	1	65	65	0	5	62601
Maximum:	1	192	192	0	24	64206
Mean:	1	61.2161	61.2161	0	3.52937	
Number of Seqs:	64206					

Table 2.4. Unique trimmed pyrotag sequence statistics

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	52	52	0	2	1
2.5%-tile:	1	57	57	0	3	135
25%-tile:	1	60	60	0	3	1346
Median:	1	61	61	0	4	2692
75%-tile:	1	62	62	0	4	4038
97.5%-tile:	1	67	67	0	5	5249
Maximum:	1	192	192	0	24	5383
Mean:	1	61.4505	61.4505	0	3.69719	
Number of Seqs:	5383					

2.4.1.2 Metagenomic data

The 386,217 metagenomic sequence reads averaged 247 bp in length, representing a total of 95,386,202 bp (∼95Mbp). The summary and statistics of the metagenome, including of raw sequences, sequences remaining after quality checks, and final sequences for further analysis after processing in the MG-RAST metagenomics server were determined (Table 2.5). A tool [68] designed to filter sequencing artifacts in 454 pyrosequencing technology was used to remove sequence duplicates; this identified 371,106 unique sequences and are used for further analyses.

Table 2.5. Metagenome statistics. Raw metagenomic sequences obtained and reads that remained after removal of replicates.

Sequenced reads	386,217
Total sequence length	95,386,202 bp
Average read length	246 $\pm$ 24 bp
Mean GC percent	59 $\pm$ 7 %
Artificial Duplicate Reads	15,111
Reads after removal of replicates	371,106
Post QC: bp Count	87,122,035 bp
Post QC: Sequences Count	349,101
Post QC: Mean Sequence Length	249 $\pm$ 15 bp
Predicted Protein Features	303,839
Predicted rRNA Features	35,681
Identified Protein Features	125,191
Identified rRNA Features	120
Identified Functional Categories	116,510

2.4.2 Community diversity, richness, and evenness

2.4.2.1 Analysis using Mothur

A rarefaction curve based on tag sequences processed with Mothur showed a graph that had not reached its asymptote (Figure 2.5). This suggests that more sampling coverage is required to detect more unique taxa. Clearly, low abundance bacterial taxa may be undetected by this method, although they might be qualitatively important members of the biofilm community. Usually, species accumulation curves tend to rise slowly if the species assemblage in a given sample is highly uneven (with high abundance of one or few species and low abundance of others) [80]. A steeply rising curve indicates a more even species abundance in the community [80].

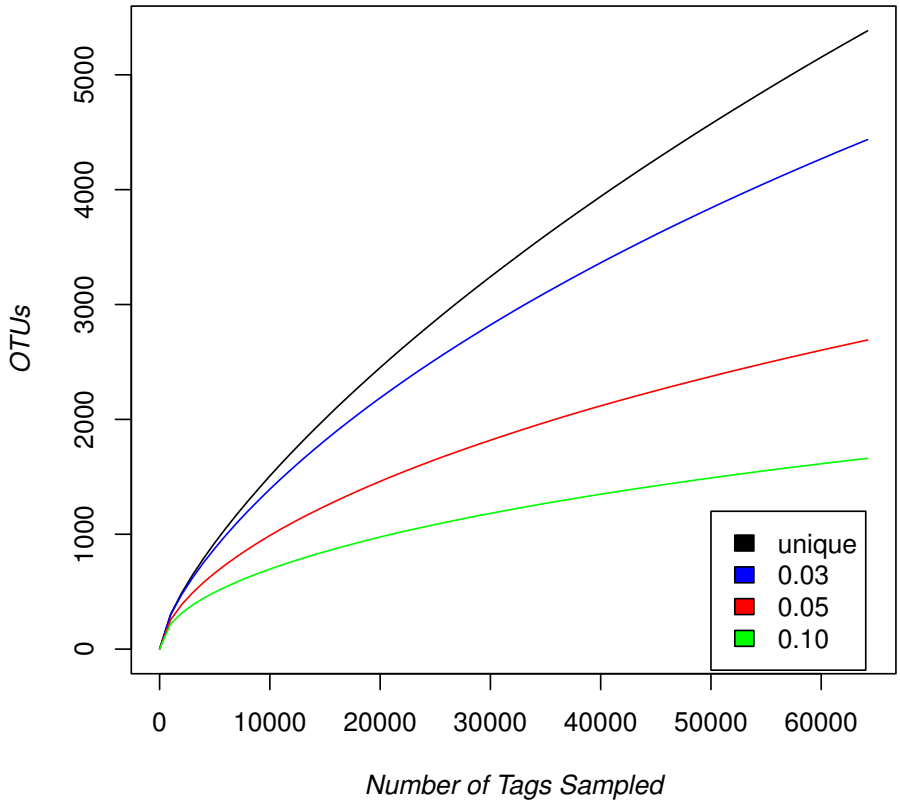


Figure 2.5. Rarefaction curve of tag sequences. Clustering of OTUs (Operational Taxonomic Units) was done with Mothur at cutoff values of 3%, 5%, and 10% and shown in different colored curves; black - unique, blue - 3%, red - 5%, and green - 10%.

Diversity estimates predicted by Mothur are shown in Table 2.6. The table shows Chao and ACE richness estimates and Shannon Diversity Index.

Table 2.6. Diversity and abundance estimates.

label	unique	0.02	0.03	0.04
nseqs	5383.00	5383.00	5383.00	5383.00
OTUs	5382.00	4382.00	3390.00	2215.00
shannon	8.59	8.33	7.88	6.97
shannon_lci	8.57	8.31	7.86	6.93
shannon_hci	8.61	8.35	7.91	7.01
chao	7242827.00	10783.31	10066.53	5410.85
chao_lci	2869238.14	10195.84	9298.52	4963.58
chao_hci	18295619.94	11430.14	10934.38	5930.91
ace	14485653.00	12047.14	17129.31	9324.93
ace_lci	51715.86	11416.71	16357.16	8846.59
ace_hci	4525384027.11	12734.08	17947.44	9837.77

2.4.2.2 Classification of tag sequences using RDP Classifier and PhymmBL binning tools

The RDP Classifier and PhymmBL binning tools were used to assign taxonomy to each Pyrotag sequence. Distribution and abundance of taxa identified by RDP Classifier and PhymmBL tool were visualized (Figures 2.6 and 2.7). The pie charts generated depict taxonomic ranks collapsed at a certain level to show the overall distribution of taxa and their abundances. Note that classification systems and naming convention by these two tools differ.

RDP classification shows the HAVO community in terms of unique sequences is dominated by *Chloroflexi*. The nine most abundant phyla are *Acidobacteria*, *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Cyanobacteria*, *Planctomycetes*, *Verrucomicrobia*, *Chlorobia*, and *Bacteroidetes*. On the other hand, a PhymmBL classification of tag sequences showed the community is dominated by *Proteobacteria*, with the next nine most abundant phyla being *Acidobacteria*, *Chloroflexi*, *Actinobacteria*, *Cyanobacteria*, *Firmicutes*, *Euryarchaeota*, *Deinococcus-Thermus*, *Bacteroidetes*, and *Thermoprotei*. The largest difference between the two classification systems is in that of *Proteobacteria* abundances; RDP computed an abundance estimate of $\sim$18%, whereas PhymmBL produced $\sim$50%. *Chloroflexi* abundances were also markedly different by these two methods, with RDP estimating 25% and PhymmBL estimating 9%. Conversely, abundance estimates by both methods for *Cyanobacteria* are similar, with 5% and 6% for RDP and PhymmBL, respectively.

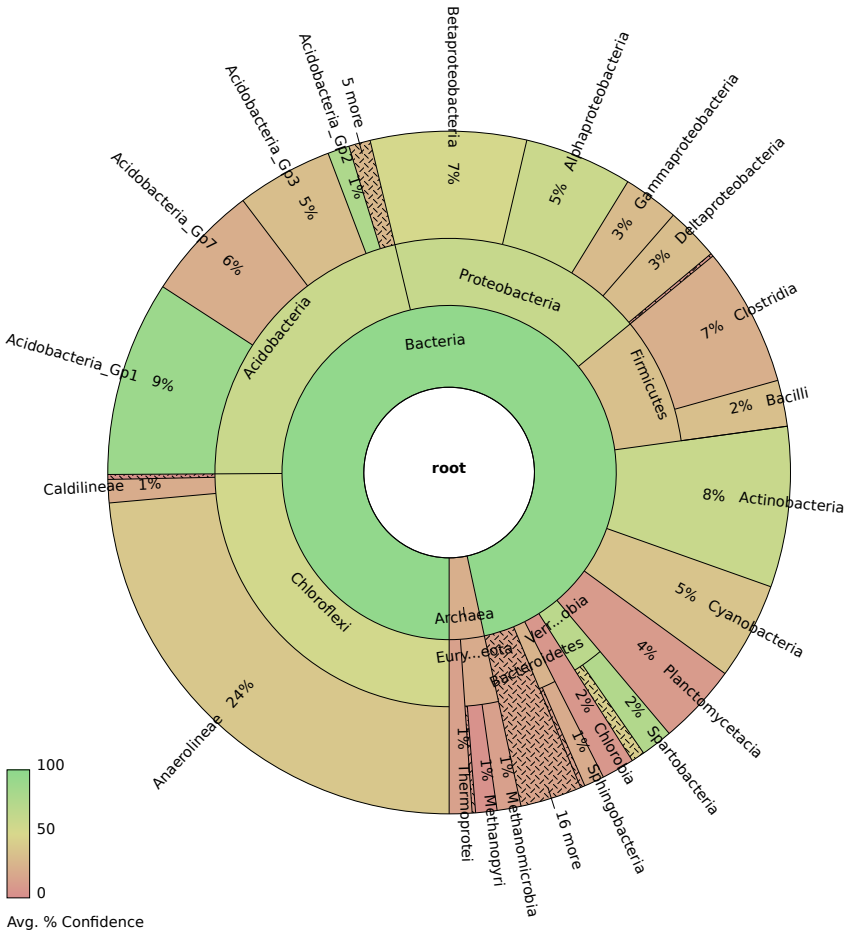


Figure 2.6. Microbial diversity and abundance based on pyrotag sequences classified by RDP Classifier. All 64,206 sequences were classified to estimate abundance. Note the confidence scores produced by the RDP Classifier. Confidence scores are on a scale of 0 to 100, and color-coded. The interactive web page can be accessed at http://www.hawaii.edu/microbiology/donachie/cave.tags.RDP.krona.html

Differences in the abundance estimates generated by these programs may be due to the nature of the pyrotag sequences. Specifically, they are extremely short sequences (50 - 100 bp), and these programs may well handle their classification in different ways. Accuracy may also suffer from variations in sequence lengths, so it is best to use all available methods when classifying pyrotag sequence data, and then compare the respective outputs. These results were also compared with those from a PhymmBL classification of metagenomic sequences (Figure 2.7).

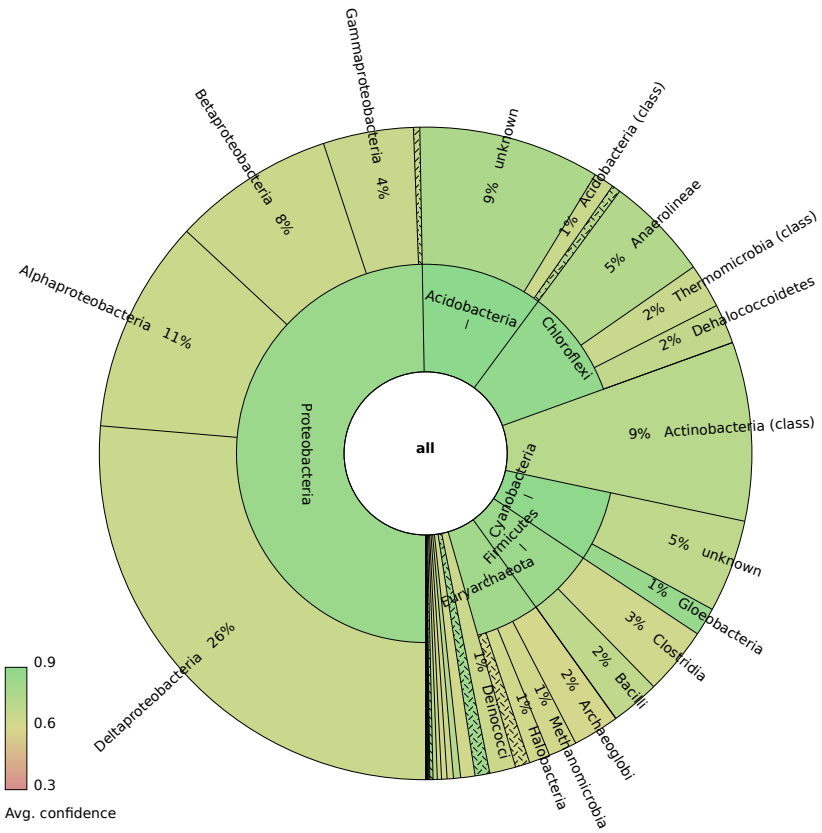


Figure 2.7. Microbial diversity and abundance based on pyrotag sequences classified by PhymmBL. All 64,206 sequences were classified to estimate abundance. Note the confidence scores produced by the PhymmBL program. Confidence scores are provided on a scale of 0.3 to 0.9 and color-coded. The interactive web page can be accessed at http://www.hawaii.edu/microbiology/donachie/cave.tags.phymmBL.krona.html

2.4.3 Phylogenetic diversity of the biofilm community on the basis of metagenomic data

Metagenomic sequences are useful in providing unbiased estimates of a community's diversity, abundance, and metabolic potential. To determine whether or not species diversity and abundances estimated from pyrotag sequences agrees with diversity estimates from the metagenomic data, HAVO metagenomic sequences were subjected to PhymmBL binning, MG-RAST analysis, and AMPHORA analyses.

2.4.3.1 Organism abundance in the community revealed by PhymmBL binning

Replicate-filtered 454 metagenomic sequences were checked for taxonomic affiliation with PhymmBL [69, 70] and visualized with Krona tool [81]. The PhymmBL binning tool classified metagenomic sequence reads and calculated abundances of organisms based on this classification system. PhymmBL uses Interpolated Markov Models (IMM) and BLAST [82] to compare metagenomic sequences against known reference organisms and Markov models. To visualize organism abundances, the Krona tool was used to create interactive pie charts that can collapse, expand, or display abundances of organisms based on different taxonomic ranks [81], such as phylum level (Figure 2.8).

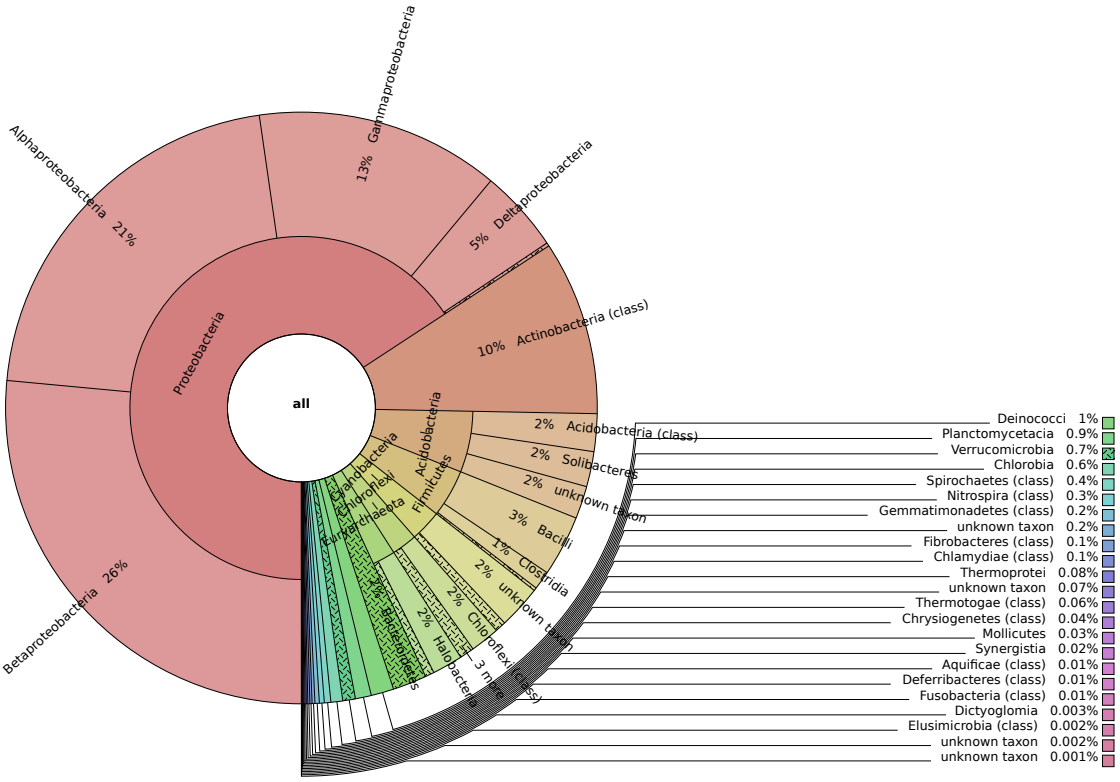


Figure 2.8. Distribution of phyla detected in the HAVO biofilm. Distribution of *Bacteria* and *Archaea* phyla after binning by the PhymmBL program. Phyla are sorted by most to least abundant. The interactive figure can be accessed at http://www.hawaii.edu/microbiology/donachie/phymmbl.krona.html.

Although initially thought to be dominated by a novel *Gloeobacter* species because of the prominent purple pigmentation, the epilithic biofilm's sequence pool comprised 66% *Proteobacte-*

ria-affiliated sequences. This indicates a few possible things; *Proteobacteria* may be numerically dominant in the community or that the community DNA extracted comprised mostly sequences of *Proteobacteria* origin. The other prominent phyla include *Actinobacteria* (10%), *Acidobacteria* (6%), *Firmicutes* (5%), *Cyanobacteria* (3%), *Chloroflexi* (2%), *Euryarchaeota* (2%), *Bacteroidetes* (2%), and *Deinococcus-Thermus* (1%). Detailed diversity and relative abundances of organisms at the rank of genus for the top 10 most abundant phyla are shown in Figures 2.10 through 2.18, and explained in detailed in Section 2.4.3.3.

2.4.3.2 Organism abundance in the community revealed by AMPHORA

To estimate phylogenetic diversity in the HAVO community, the cave biofilm metagenome was analyzed in the AMPHORA program [83]. This program aligns 31 single-copy marker genes from reference genomes to assign taxomic rank to single-copy marker genes identified in the metagenome. Taxonomic assignment using AMPHORA is advantageous compared to other methods. For example, it primarily identifies conserved marker genes that are usually present in single copies in microbes, so each copy represents a single organism. This method aims to obtain an unbiased estimate of organism abundance in the community based on metagenomic data. However, if the metagenome is under-sampled, the method would easily miss conserved marker genes that are usually present in low copy numbers.

AMPHORA confirmed the distribution of a few dominant bacterial groups in the HAVO community that was revealed by binning with PhymmBL. Although it was expected that the *Cyanobacteria* will be the most abudant members of the community, the *Proteobacteria*, *Acidobacteria*, and *Chloroflexi* are clearly the most abundant organisms (Figure 2.9). The results generally seem to agree with the PhymmBL metagenomic sequence binning results.

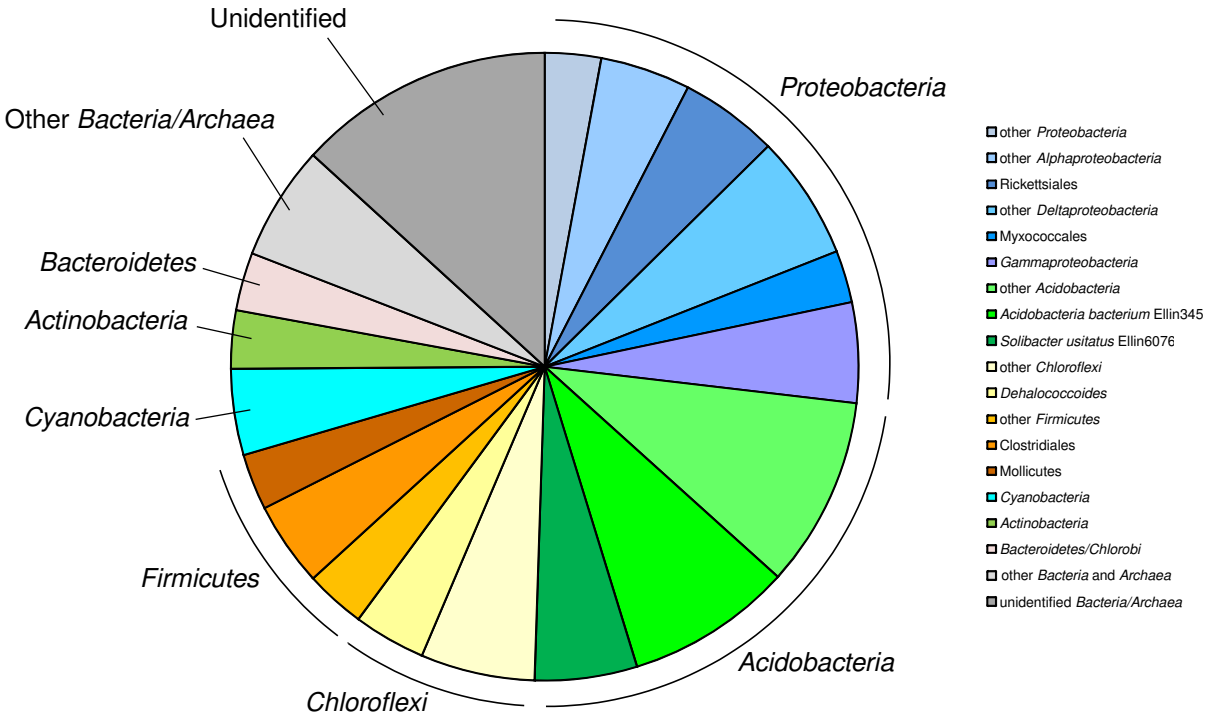


Figure 2.9. Amphora analysis of single-copy genes showing relative abundances of *Bacteria* and *Archaea* in the metagenomic data.

2.4.3.3 Detailed analysis of *Bacteria* diversity by PhymmBL binning

Organisms identified in the 10 most abundant *Bacteria* phyla are presented and discussed in detail, while pie charts enable visualization of relatives of dominant taxa identified in each phylum. Diversity abundance estimates are correlated with metabolic diversity in Section 2.4.4.

***Proteobacteria* diversity and abundance:**

Proteobacteria accounted for ∼66% of all metagenomic sequences. Among this phylum's component sub-classes, the *Betaproteobacteria* was the most abundant (∼40%) followed by the *Alphaproteobacteria* (∼32%). Sequences affiliating with the *Gammaproteobacteria* and *Deltaproteobacteria* subclasses comprised 20% and 7% respectively of the phylum. Sequences affiliating

with the *Burkholderia* comprised 68% of all *Betaproteobacteria* sequences total here (see the interactive pie chart).

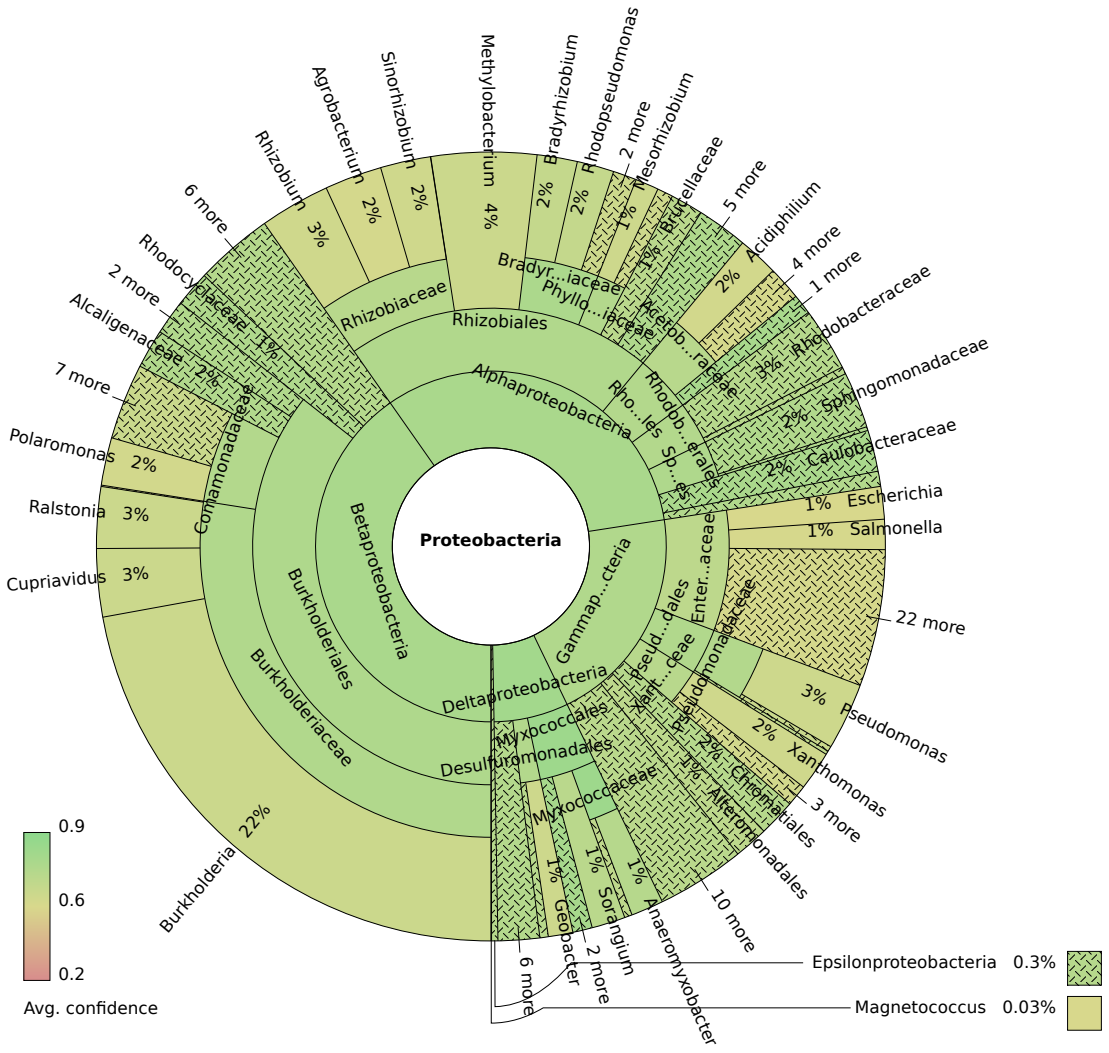


Figure 2.10. Genus level taxonomic diversity among *Proteobacteria* sequences after binning by PhymmBL. Dominance is heavily skewed towards beta- and alphaproteobacteria.

Actinobacteria diversity and abundance:

Actinobacteria accounted for ∼10% of the total metagenomic sequences. In the *Actinobacteria*, Mycobacteriaceae were the most abundant family, and accounted for 14% of all *Actinobacteria* sequences. The next 10 most abundant families were *Nocardiaceae*, *Micrococcaceae*,

Corynebacteriaceae, *Pseudonocardiaceae*, *Microbacteriaceae*, *Streptomycetaceae*, *Frankiaceae*, *Micromonosporaceae*, *Nocardioidaceae*, and *Streptosporangiaceae* (Figure 2.11).

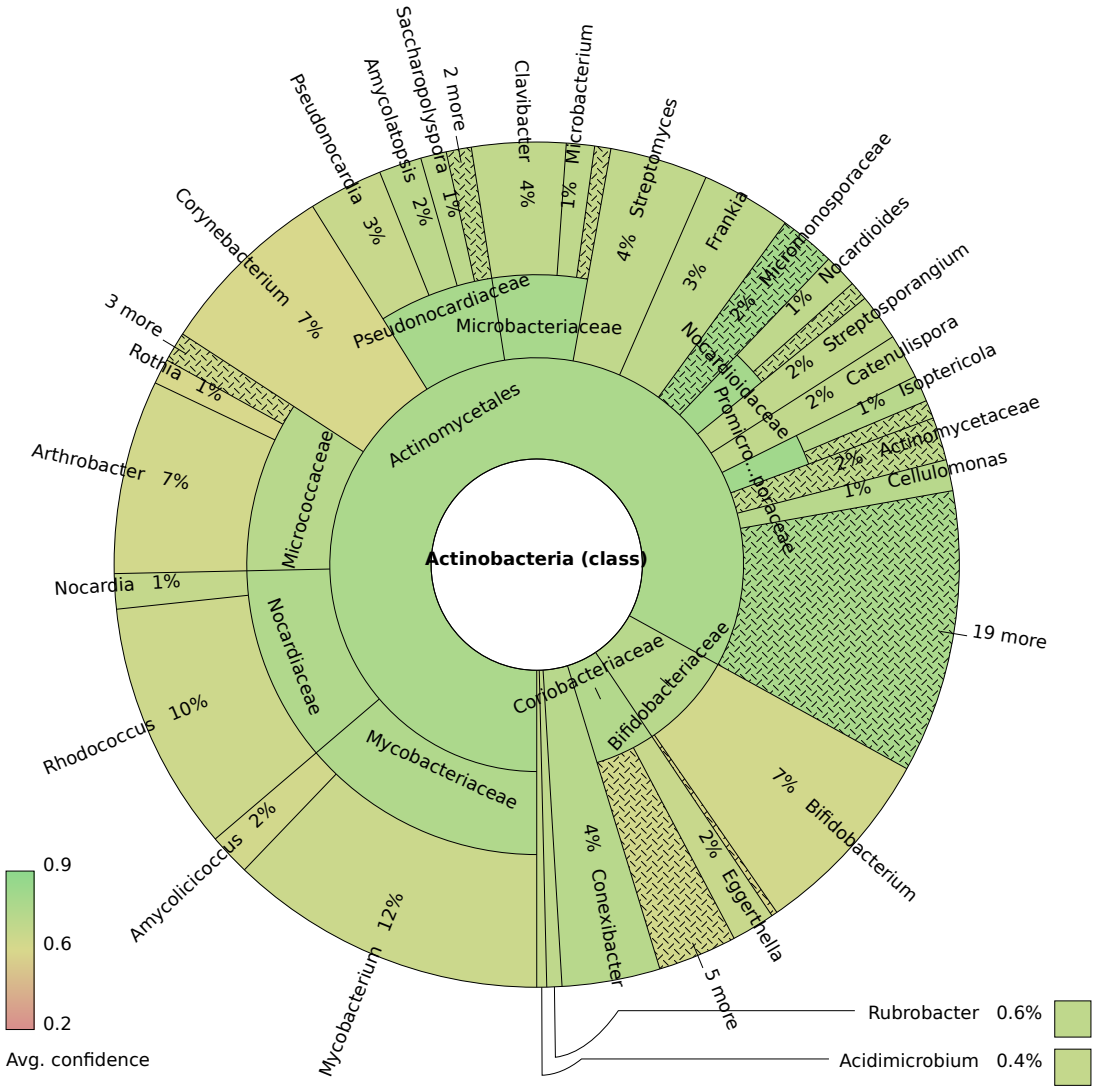


Figure 2.11. Diversity of *Actinobacteria* sequences at the genus level, after binning by PhymmBL.

Acidobacteria diversity and abundance:

Acidobacteria affiliates accounted for about 6% of all metagenomic sequences (Figure 2.12). Relatively few *Acidobacteria* cultivated representatives exists (*i.e.*, 4), and the most abun-

dant hits here are with the genus *Candidatus* Solibacter (33%), followed by *Candidatus* Koribacter (31%), and *Acidobacterium* (28%).

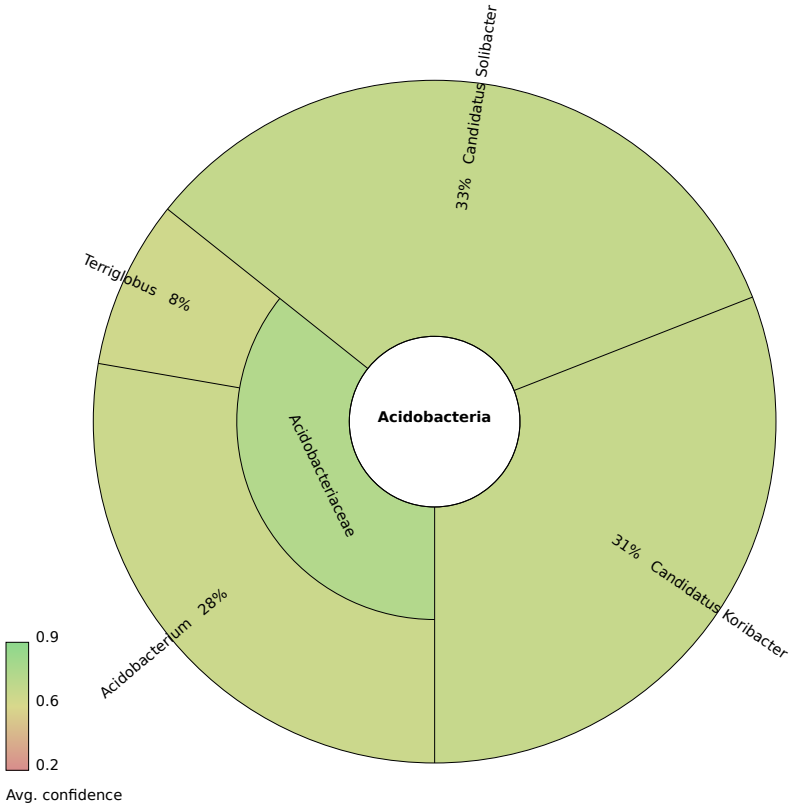


Figure 2.12. Diversity of *Acidobacteria* sequences at the genus level after binning by the PhymmBL.

Firmicutes **diversity and abundance:**

Firmicutes accounted for 5% of the total metagenomic sequences. The five most abundant *Firmicutes* genera are *Alicyclobacillus* (22%), *Geobacillus* (13%), *Bacillus* (9%), *Lactobacillus* (8%), and *Paenibacillus* (6%) (Figure 2.13).

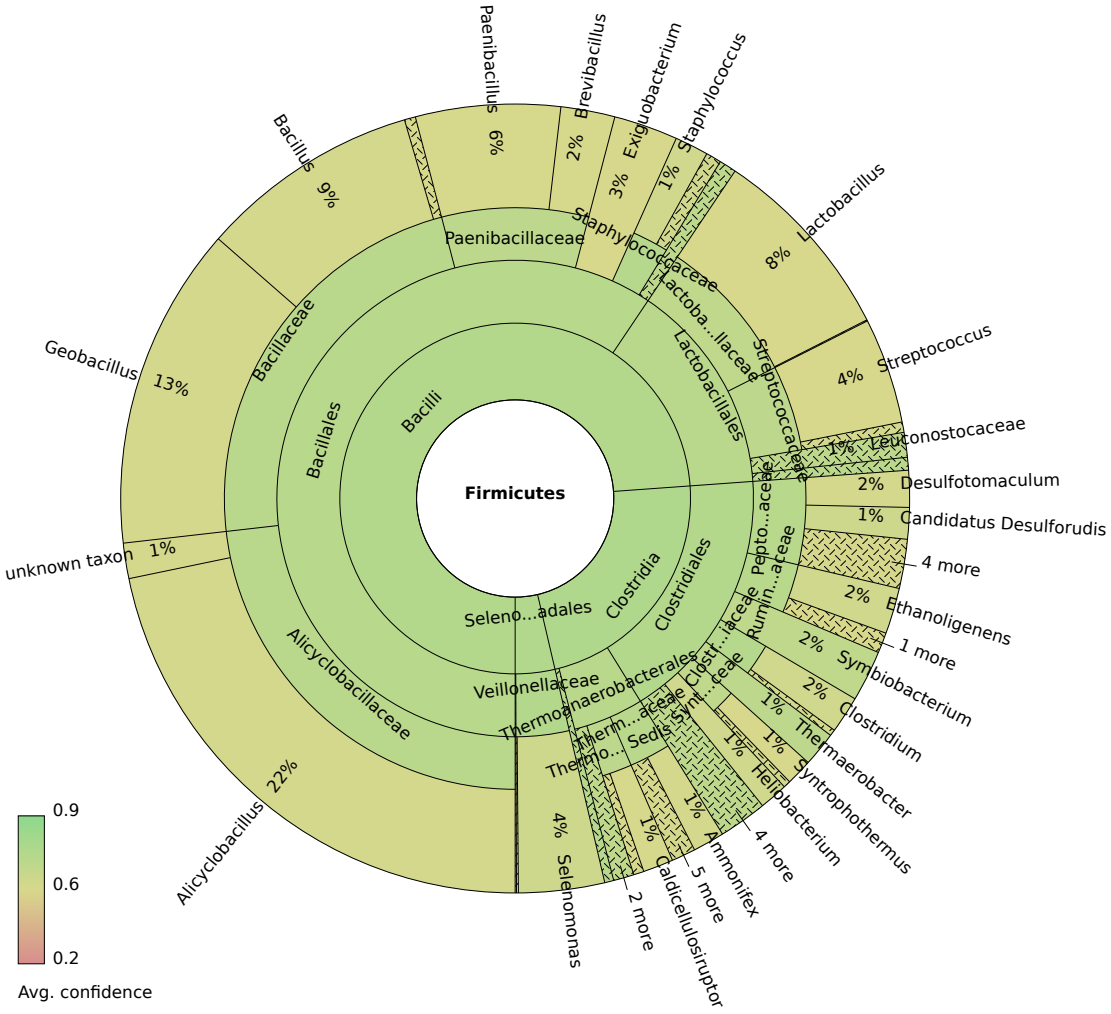


Figure 2.13. Diversity of *Firmicutes* sequences at the genus level after binning by PhymmBL.

Cyanobacteria diversity and abundance:

In contrast to how abundant were the most abundant *Proteobacteria*, all the *Cyanobacteria* sequences combined accounted for only 3% of the total metagenomic sequences; among the *Cyanobacteria* sequences, however, 40% affiliated with the *Chroococcales*, 25% with *Nostocales*, 20% with the *Gloeobacterales*, 3% *Prochlorales*, and the remainder are from unknown taxa (Figure 2.14). At the species level, 20% of reads were classified as *Gloeobacter violaceus* PCC 7421, indicating that the genus *Gloeobacter* is the most numerous of all cyanobacteria detected in the cave

biofilm metagenome. *Gloeobacter* is a special case because only one genus (and species) has been sequenced to date and PhymmBL classification may suffer from lack of sequences belonging to this genus in genome databases.

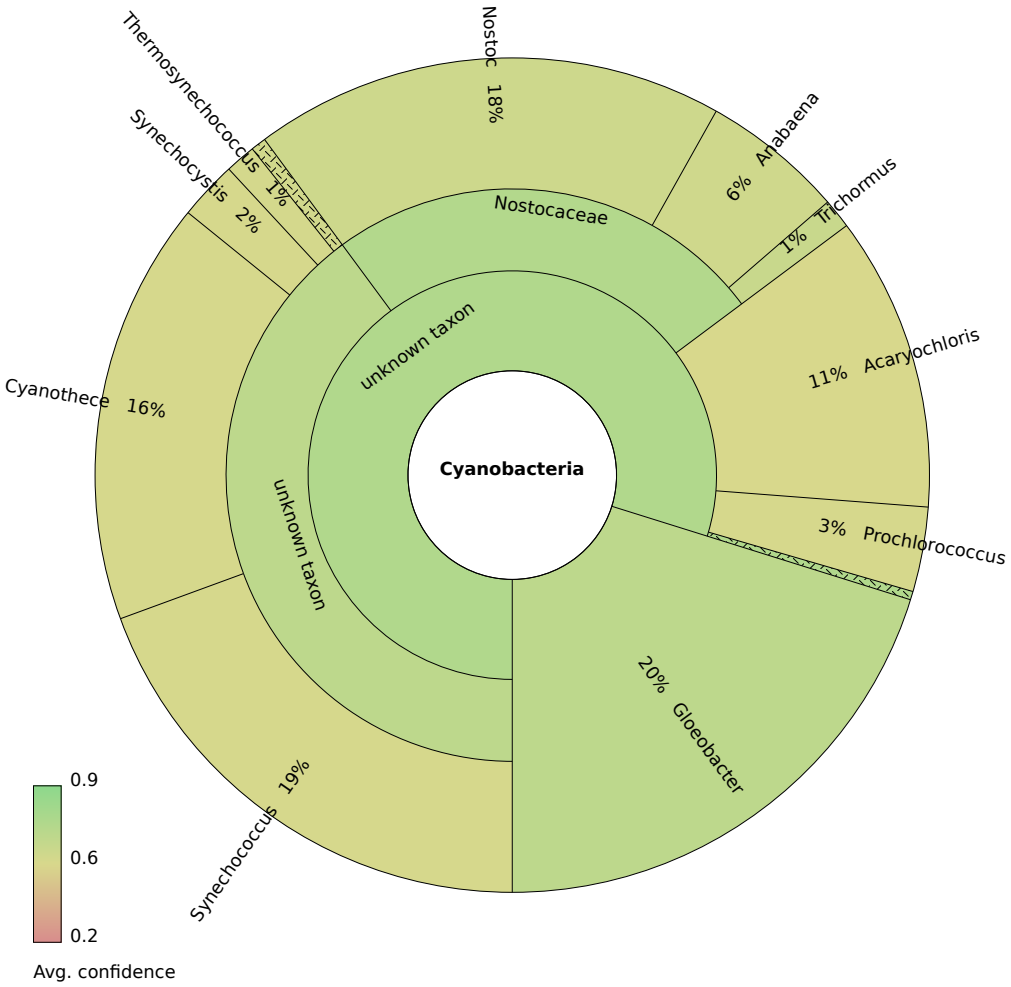


Figure 2.14. Genus level diversity of *Cyanobacteria* sequences after binning by PhymmBL.

Chloroflexi diversity and abundance:

Although numerically well represented in the tag sequence data, the *Chloroflexi* accounted for only 2% of the total metagenomic data based on the PhymmBL classification (Figure 2.15). This may be due to their low representation in complete genomes compared to the *Proteobacteria*, as well as low sampling depth. Among them, however, the five most abundant genera are: *Roseiflexus*

(36%), *Sphaerobacter* (20%), *Chloroflexus* (18%), *Herpetosiphon* (12%), and *Thermomicrobium* and *Anaerolinea* (6%).

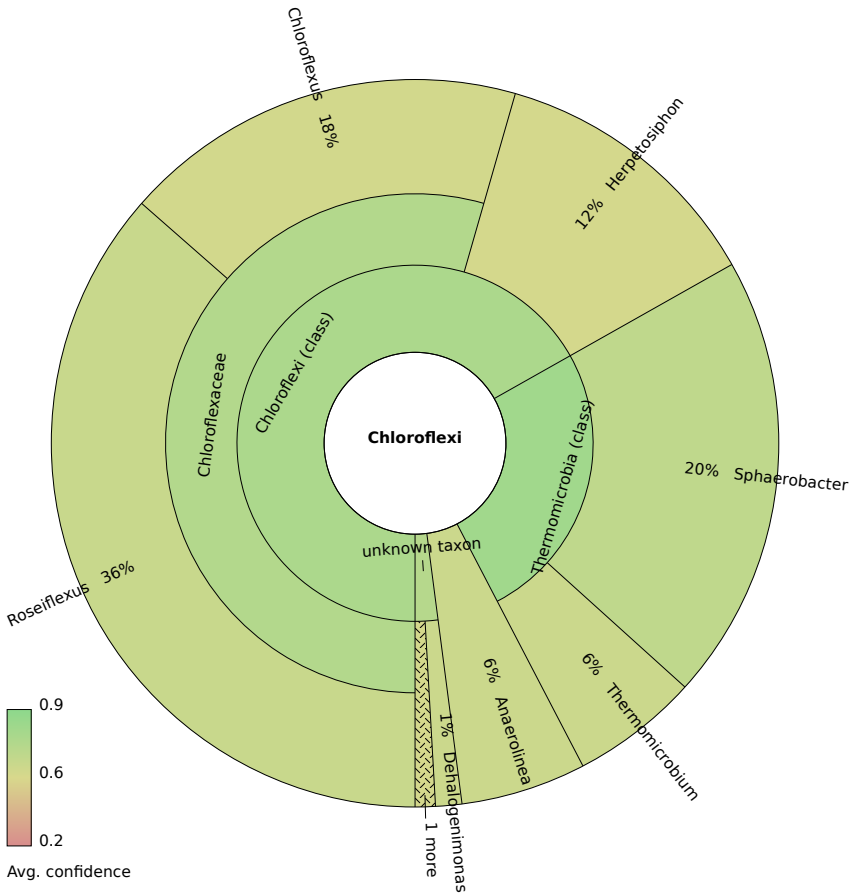


Figure 2.15. Genus level diversity among the *Chloroflexi* sequences after binning by PhymmBL.

Bacteroidetes diversity and abundance:

Among the Bacteroidetes, *Salinibacter* and *Spirosoma* are most numerous at 18% and 17% respectively, followed by *Rhodothermus* at 11% and *Haliscomenobacter* at 8%. *Bacteria* from the phylum *Bacteroidetes* occupy diverse habitats ranging from the human gut [84] to beach sand [85] and they are known to be prolific degraders of complex polymeric substances [86, 87]. The role these bacteria play in HAVO biofilm remains to be determined.

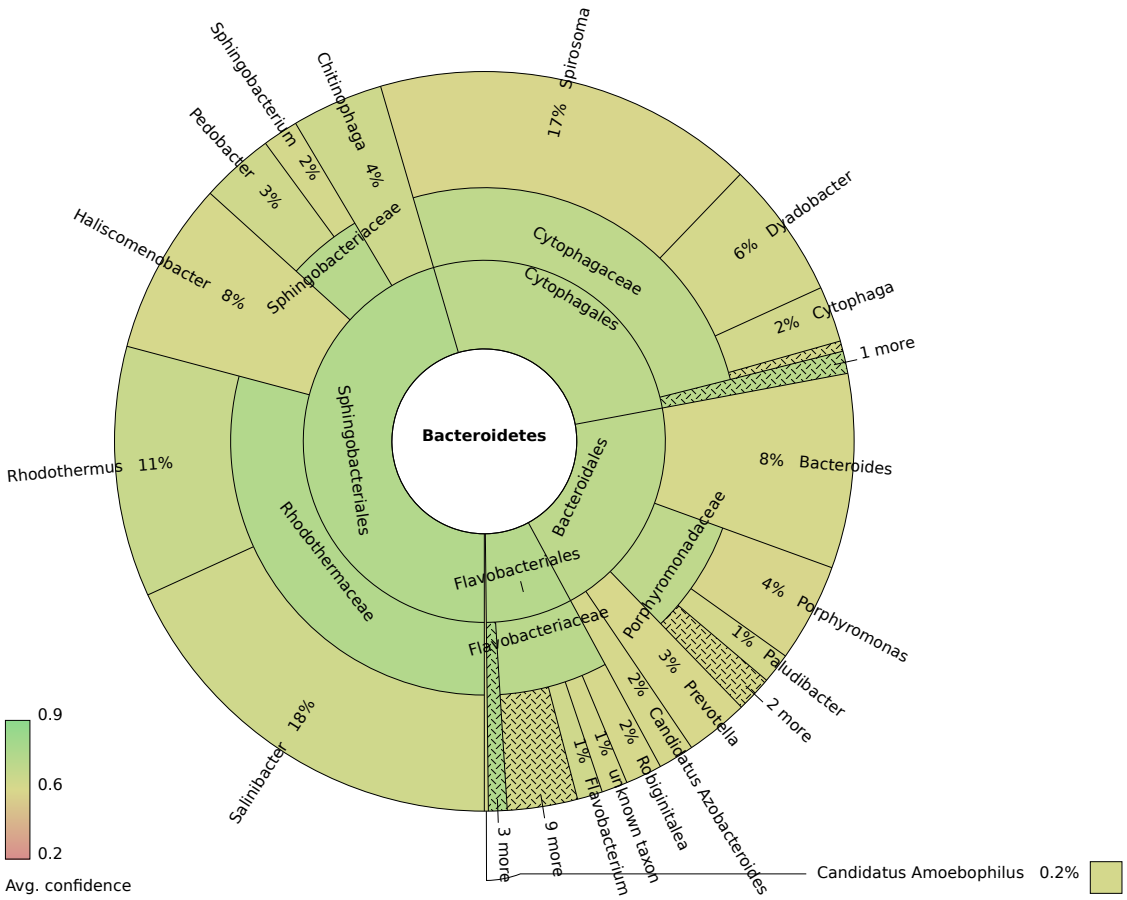


Figure 2.16. Genus level diversity among the *Bacteroidetes* after binning by PhymmBL.

Deinococcus-Thermus diversity and abundance:

The *Deinococus-Thermus* phylum comprises of mostly thermophilic bacteria known for their ability to withstand high temperatures [88] and/or gamma radiation [89, 90]. This phylum also hosts very few known taxa. The most abundant genera identified here were *Deinococcus* (64%) and *Meiothermus* (16%) (Figure 2.17). It is not surprising to have detected thermophilic bacteria from a volcanic cave but cultivation efforts could certainly be intensified in order to isolate thermophilic bacteria from the epilithic biofilm.

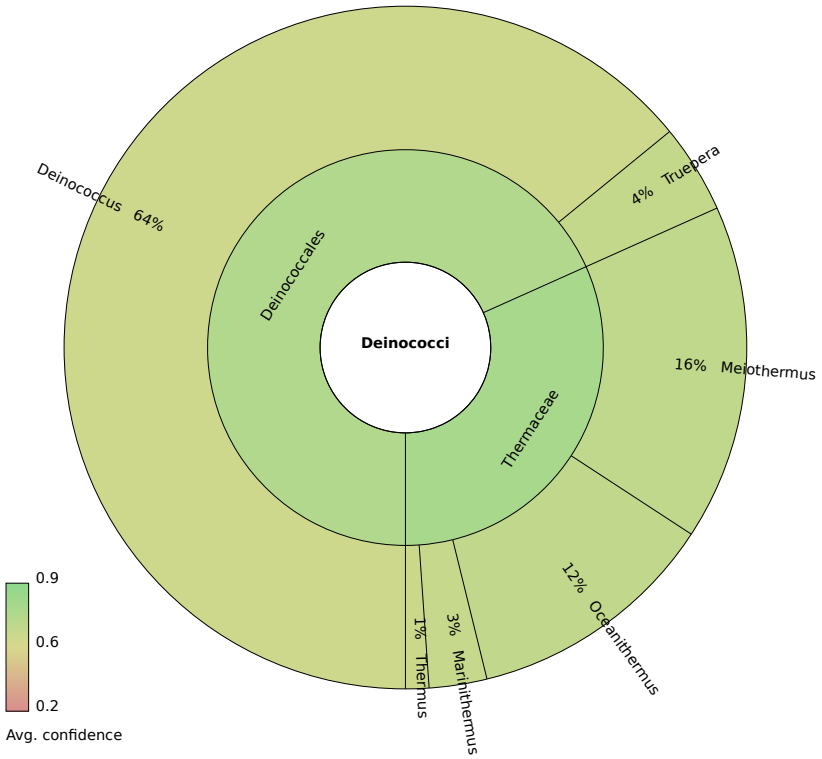


Figure 2.17. Genus level diversity among the *Deinococcus-Thermus* sequences after binning by PhymmBL.

***Planctomycetes* and *Verrucomicrobia* diversity and abundance:**

The phyla *Planctomycetes* and *Verrucomicrobia* are part of the PVC superphylum (*Planctomycetes/Verrucomicroba/Chlamydiae*) that hosts metabolically diverse bacteria from very different habitats [91]. Their roles invovle infection in humans [92] to methane oxidization in geothermal habitats [93, 94, 95]. Among the sequences classified as *Planctomycetes*, 40% belong in the genus *Planctomycetes*, 24% in the *Pirellula*, 21% in the *Rhodopirellula*, and 15% in the *Isosphaera* (Figure 2.18). The two abundant close relatives of the *Planctomycetes* identified are *Planctomyces brasiliensis* DSM 5305 (22%) from a marine alga [96] and *Planctomyces limnophilus* DSM 3776 (18%), a stalked and budding bacterium from a freshwater lake [97, 98].

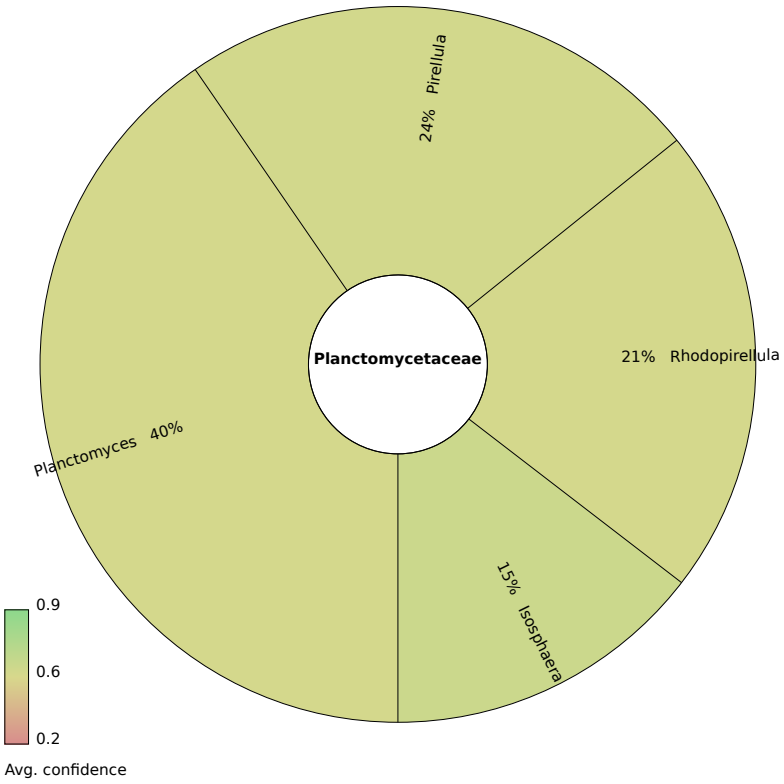


Figure 2.18. Genus level diversity among the *Planctomycetes* sequences after binning by PhymmBL.

Among the sequences identified as belonging to the phylum *Verrucomicrobia*, the majority belong in the *Opitutus* (70%) (Figure 2.19). The rest were *Coraliomargarita* (24%), *Akkermansia* (3%), and *Methylacidiphilum* (2%). *Opitutus terrae* is an anaerobic bacterium usually found in rice paddy soils, and is known for propionate production from plant polysaccharides [99]. The presence of its relatives in the HAVO biofilm indicates that they may occupy anaerobic zones in the biofilm, where they may perform similar functions.

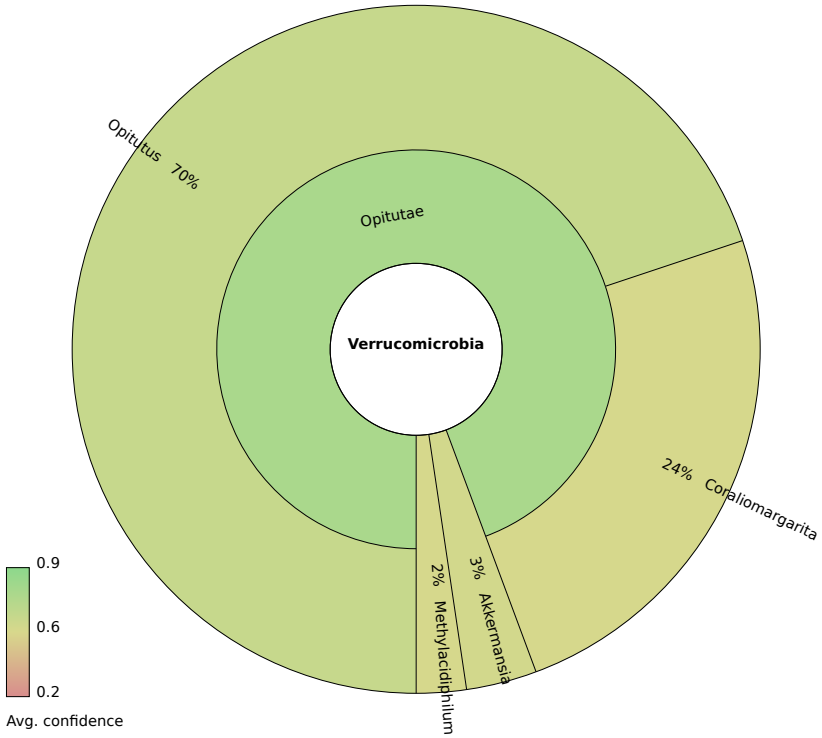


Figure 2.19. Genus level diversity among the *Verrucomicrobia* sequences after binning by PhymmBL.

Again, due to a lack of representative genomes from these two phyla, classification by PhymmBL may have been quite general and could have missed previously uncharacterized organisms.

2.4.3.4 *Archaea* diversity in the epilithic biofilm after binning by PhymmBL

Archaea taxonomic groups were identified in the metagenomic and pyrotag data sets by the PhymmBL binning tool (Tables 2.7 and 2.8; Figure 2.20). Pyrotags also classified with the RDP Classifier are not presented here in order to maintain consistency among the taxonomic assignments. Abundances do not always match in both data sets as they sampled sequences differently, and one is PCR-based, while the other one is not. Where insufficient data resulted in the phylum not being represented in one data set, it was omitted from the figures. *Crenarchaeota* genera were detected in both data sets, with taxonomic assignment at rank (Table 2.7).

Table 2.7. Comparison of diversity and abundance of *Crenarchaeota* in metagenomic and pyrotag data sets.

Genus names	fractions of the total in the metagenomic data set (%)	fraction in the unique pyrotag data set (%)
Acidianus	3 (1.00)	-
Acidilobus	8 (2.67)	6 (5.04202)
Aeropyrum	10 (3.33)	8 (6.72269)
Desulfurococcus	13 (4.33)	-
Hyperthermus	29 (9.67)	-
Ignicoccus	16 (5.33)	-
Metallosphaera	28 (9.33)	-
Pyrobaculum	52 (17.3)	-
Staphylothermus	3 (1.00)	-
Sulfolobus	49 (16.3)	-
Thermofilum	33 (11.0)	-
Thermoproteus	23 (7.67)	-
Thermosphaera	24 (8.00)	-
Vulcanisaeta	9 (3.00)	-

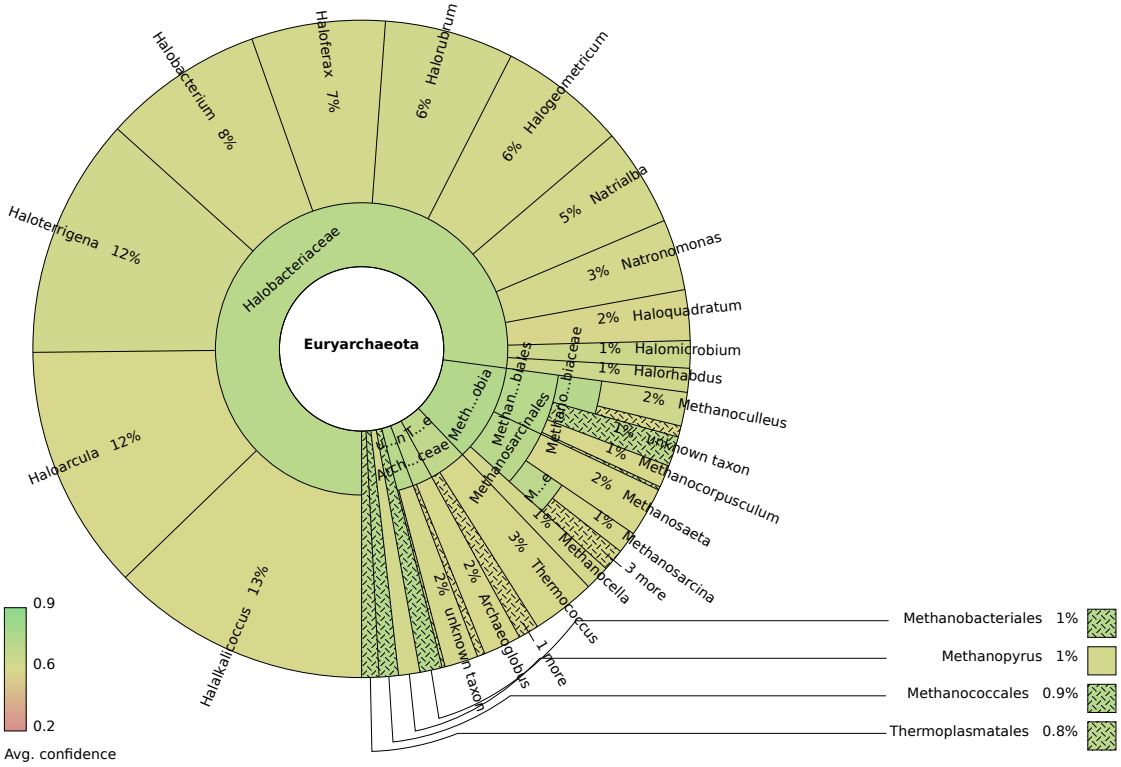


Figure 2.20. Genus level diversity among the *Euryarchaeota* sequences after binning by PhymmBL.

Table 2.8. Comparison of diversity and abundance of *Euryarchaeota* in metagenomic and pyrotag data sets

Genus names	fractions of the total in the metagenomic data set (%)	fraction in the unique pyrotag data set (%)
Aciduliprofundum	12 (0.15)	-
Archaeoglobus	149 (1.82)	26 (9.09)
Ferroglobus	30 (0.37)	-
Halalkalicoccus	1023 (12.5)	15 (5.24)
Haloarcula	967 (11.8)	17 (5.94)
Halobacterium	634 (7.76)	1 (0.35)
Haloferax	527 (6.45)	30 (10.5)
Halogeometricum	502 (6.14)	6 (2.09)
Halomicrobium	108 (1.32)	3 (1.05)
Haloquadratum	199 (2.44)	2 (0.70)
Halorhabdus	94 (1.15)	1 (0.35)
Halorubrum	509 (6.23)	2 (0.70)
Haloterrigena	949 (11.6)	8 (2.80)
Methanobacterium	42 (0.51)	-
Methanobrevibacter	14 (0.17)	-
Methanocaldococcus	22 (0.27)	2 (0.70)
Methanocella	98 (1.20)	-
Methanococcoides	24 (0.29)	-
Methanococcus	51 (0.62)	-
Methanocorpusculum	88 (1.08)	1 (0.35)
Methanoculleus	133 (1.63)	3 (1.05)
Methanohalobium	37 (0.45)	1 (0.35)
Methanohalophilus	26 (0.32)	22 (7.69)
Methanoplanus	44 (0.54)	1 (0.35)
Methanopyrus	84 (1.03)	7 (2.45)
Methanoregula	67 (0.82)	10 (3.50)
Methanosaeta	196 (2.40)	11 (3.85)
Methanosarcina	368 (4.50)	42 (14.7)
Methanosphaera	7 (0.09)	1 (0.35)
Methanosphaerula	51 (0.62)	-
Methanospirillum	13 (0.16)	4 (1.40)
Methanothermobacter	20 (0.24)	59 (20.6)
Methanothermus	7 (0.09)	-
Methanotorris	3 (0.04)	-
Natrialba	388 (4.75)	3 (1.05)
Natronomonas	280 (3.42)	3 (1.05)
Picrophilus	3 (0.04)	-
Pyrococcus	81 (0.99)	3 (1.05)
Thermococcus	252 (3.08)	2 (0.70)
Thermoplasma	68 (0.83)	-

2.4.4 Metabolic potential and metabolic pathway analysis of the biofilm community

According to the COG functional categories identified by the MG-RAST server, the majority of the metabolic functions of the HAVO community is devoted to energy metabolism (Figure 2.21). Amino acid transport and metabolism is the most abundant category, comprising 11% of the identified COG functional groups. The top ten metabolic functional categories after amino acid transport and metabolism are energy production and conversion (9%), replication, recombination and repair (7%), carbohydrate transport and metabolism (6%), translation, ribosomal structure

and biogenesis (6%), inorganic ion transport and metabolism (5%), cell wall/membrane/envelope biogenesis (5%), lipid transport and metabolism (4%), coenzyme transport and metabolism (4%), post-translational modification, protein turnover, chaperones (4%), and transcription (4%).

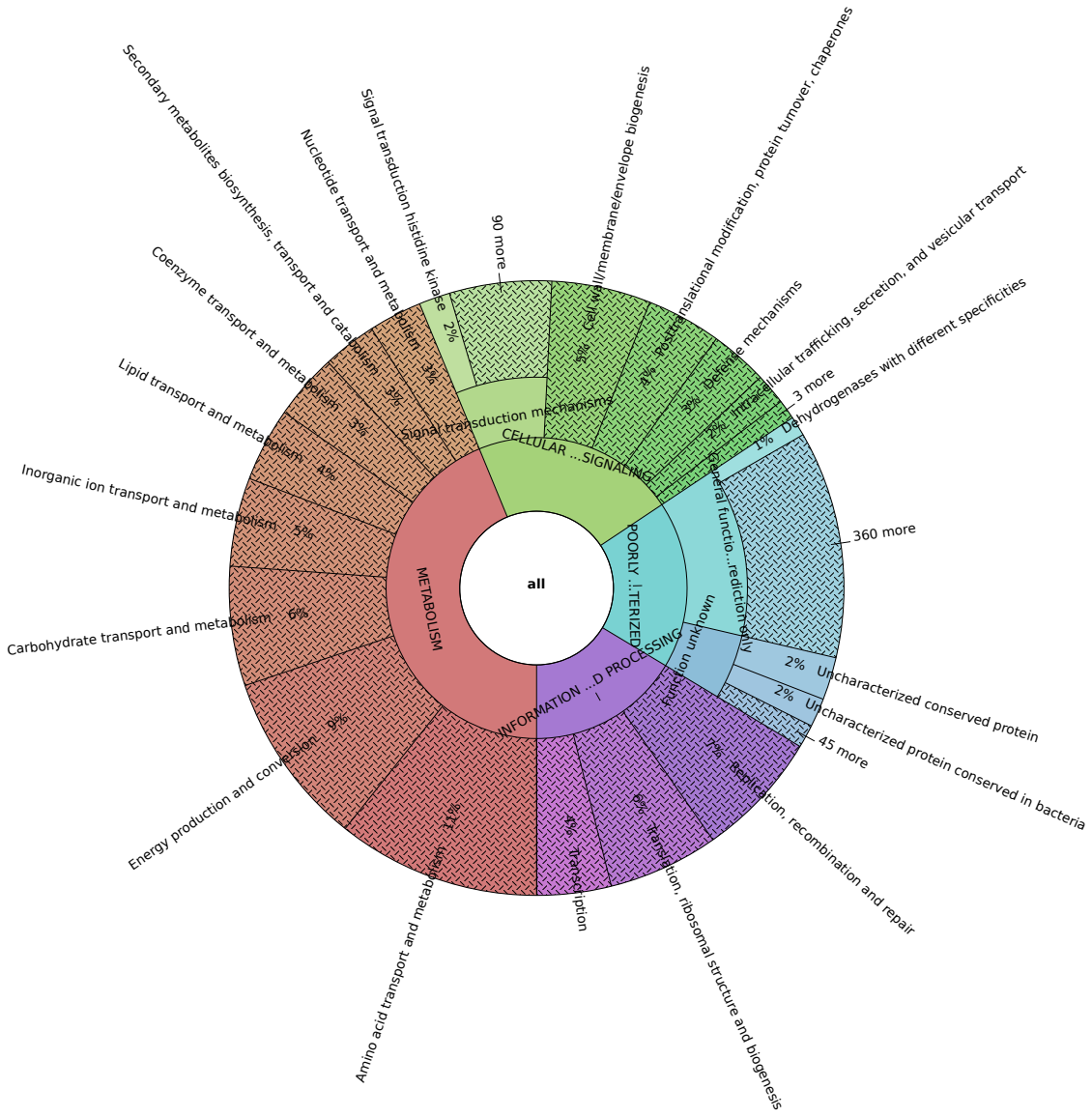


Figure 2.21. COG metabolic functional categories identified in the epilithic biofilm metagenome.

Several tools were used to perform a detailed analysis of the pathways in the HAVO community, including MG-RAST, KAAS, PhymmBL, and iPath2.0. This analysis aimed first to

determine which pathways were present, and then to determine which taxonomic groups were responsible for these pathways. To this end, amino acid sequences predicted in the de-replicated metagenomic data set processed by the MG-RAST server were downloaded; the amino acid sequences were then submitted to the KAAS annotation server to retrieve KEGG orthologous group numbers (KO numbers). A custom Python script was used to parse taxonomic assignments predicted by the PhymmBL binning tool. This approach allowed separation of KO numbers based on a given taxonomic rank. KO numbers are separated here at the rank of order.

Multiple metabolic pathways were detected in the HAVO sample (Figure 2.22), including secondary metabolite biosynthesis pathways identified from the metagenome (Figure 2.23). The HAVO community appears collectively able to perform a diverse range of metabolic functions, although some apparently complete pathways may be false given the fact that all genomic data were combined into one sequence pool. A better and more accurate portrayal of metabolic pathways would involve separating the enzymes identified in the metagenome and binning them based on their taxonomic affiliations. For example, combining all sequences belonging to the class *Betaproteobacteria* first, and then reconstructing metabolic pathways specific to each bin. This was completed here, and the individual pathways for each taxonomic group at the rank of phylum have been reconstructed. However, it is impractical to display all the pathway maps in this dissertation, so only overall pathway maps are presented, alongside the abundances of KEGG orthologous groups contributed by each taxonomic group (Table 2.9).

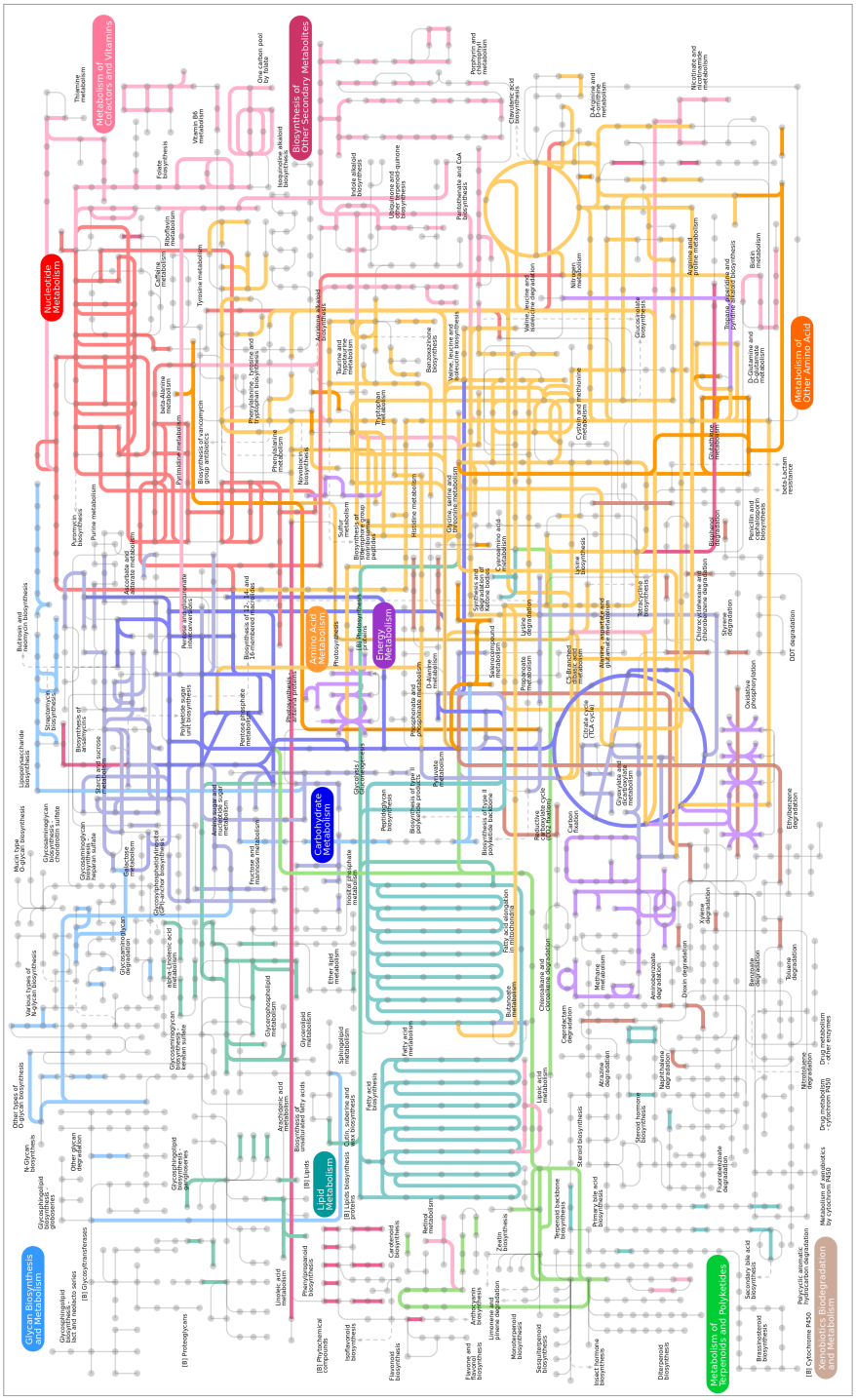


Figure 2.22. Metabolic pathways identified from the HAVO metagenome. Note, this is cumulative and does not differentiate between different taxonomic lineages that contributed to a pathway. The figure highlights in different colors pathway components catalyzed by enzymes identified in the HAVO sample.

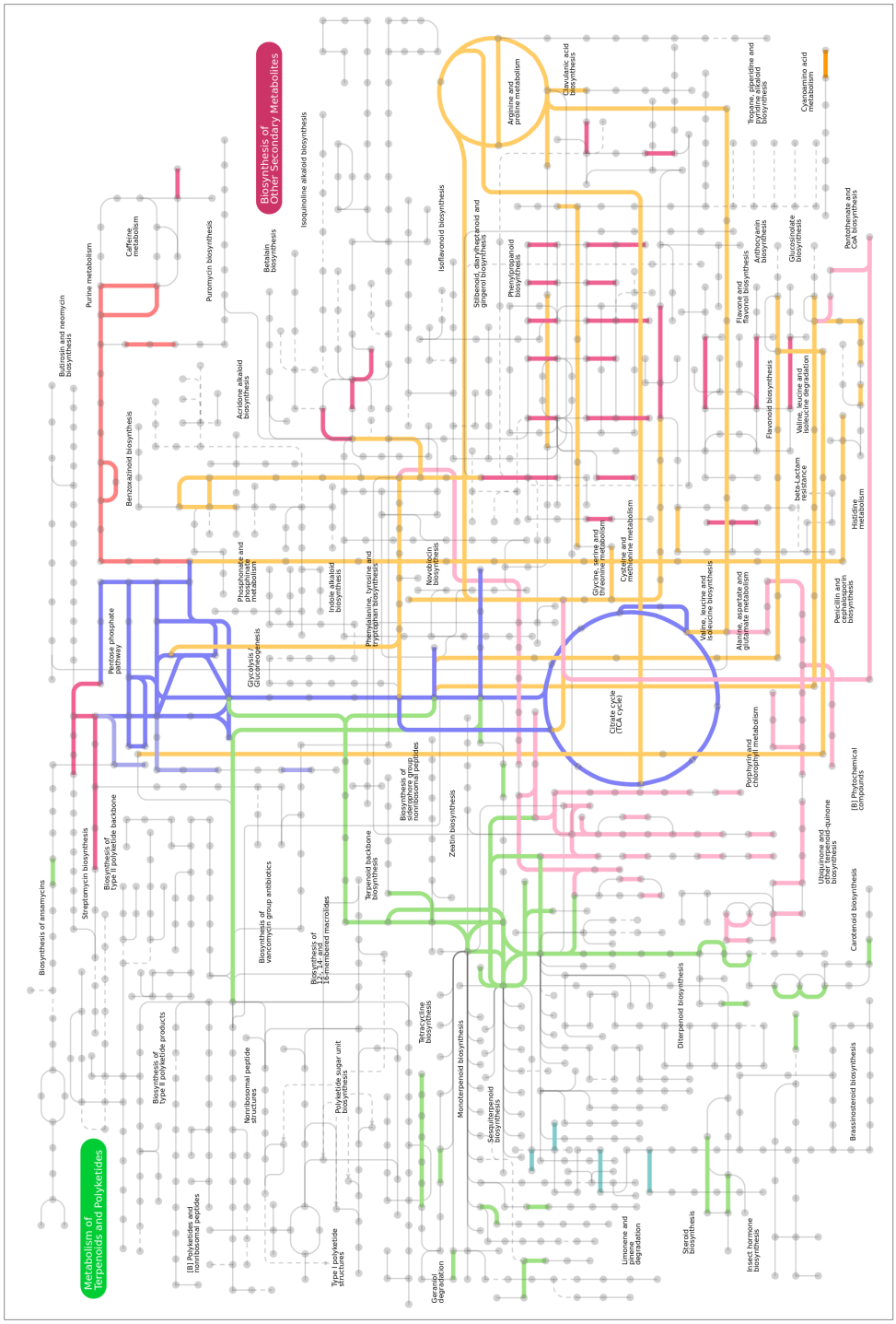


Figure 2.23. Secondary metabolite pathways identified in the HAVO epilithic biofilm metagenome. Note, this is cumulative and does not differentiate between different taxonomic lineages that contributed to the pathway.

Table 2.9. Top 30 taxonomic groups represented in the metabolic pathways at the rank of Order

Taxon (Order)	Number of KEGG ortholog groups identified	% of total (10,732)
Burkholderiales	2010	18.7290
Rhizobiales	1395	12.9985
Actinomycetales	967	9.0104
Myxococcales	504	4.6962
Solibacterales	388	3.6154
Acidobacteriales	327	3.0470
Enterobacteriales	287	2.6742
Nostocales	283	2.6370
Bacillales	270	2.5158
Chroococcales	250	2.3295
Rhodospirillales	231	2.1524
Gloeobacterales	231	2.1524
Caulobacterales	213	1.9847
Pseudomonadales	193	1.7984
Sphingobacteriales	165	1.5375
Desulfuromonadales	162	1.5095
Rhodobacterales	158	1.4722
Clostridiales	155	1.4443
Rhodocyclales	153	1.4256
Sphingomonadales	145	1.3511
Chloroflexales	140	1.3045
Chromatiales	129	1.2020
Xanthomonadales	124	1.1554
Thermales	109	1.0156
Deinococcales	107	0.9970
Halobacteriales	102	0.9504
Planctomycetales	98	0.9132
Sphaerobacterales	87	0.8107
Desulfovibrionales	87	0.8107
Nitrosopumilales	75	0.6988

Although individual pathways contributed by a specific taxon can be investigated, this approach has many problems. In an undersampled metagenome, for example, only the most abundant taxa will be represented in the pool. Thus, if certain pathway components are missing from a particular taxon, this does not necessarily mean the pathway is missing, but might simply indicate that it was not detected due to the low abundance of its members in the community. Such pathway analysis does not show all the pathways present in a given taxon, but is useful for detecting important pathways contributed by a taxon of interest (*e.g.*, methanogenesis). It is thus not wise to conclude that a certain pathway identified in the metagenomic data means that a spedcific taxon is the only one conducting this metabolic task; the lack of sequencing depth can also overlook organisms that may also perform the same function, but which are not represented in the sequence pool due to a lack of sequence coverage.

Table 2.10. Taxonomic groups represented in the metabolic pathways at the rank of Phylum

Taxon (Order)	Number of KEGG ortholog groups identified	% of total (11,328)
Proteobacteria	6392	56.4266
Actinobacteria	1151	10.1607
Acidobacteria	1082	9.5516
Cyanobacteria	811	7.1593
Firmicutes	497	4.3874
Chloroflexi	288	2.5424
Bacteroidetes	261	2.3040
Deinococcus-Thermus	216	1.9068
Euryarchaeota	158	1.3948
Planctomycetes	98	0.8651
Verrucomicrobia	88	0.7768
Thaumarchaeota	77	0.6797
Chlorobi	62	0.5473
Gemmatimonadetes	54	0.4767
Nitrospirae	49	0.4326
Spirochaetes	14	0.1236
Chlamydiae	10	0.0883
Thermotogae	5	0.0441
Crenarchaeota	5	0.0441
Synergistetes	3	0.0265
Fibrobacteres	2	0.0177
Aquificae	2	0.0177
Fusobacteria	1	0.0088
Deferribacteres	1	0.0088
Chrysiogenetes	1	0.0088

2.4.5 Effective Genome Size of the community

The Effective Genome Size (EGS) is termed as an "ecologically more meaningful measure of genome size", a measure of average genome size of a given community extrapolated by counting the number of single-copy marker genes present in a given sample [78]. This concept of EGS was introduced as a way to quantitatively estimate functional diversity of a community; to correlate environmental complexity and the diversity of genes that is required to adapt to environmental conditions [78].

The EGS of the HAVO epilithic biofilm community was determined by counting single-copy marker genes and normalizing them by gene length after Raes et al. [78]. By doing so, an EGS

estimate of 4.2 Mbp was obtained for the cave biofilm community. This is an effort to determine complexity of the community structure, since it has been reported that the effective genome size of a community varies with the type of environment; more oligotrophic environments are said to show smaller EGS and less diversity than environments with higher concentration of nutrients [78].

Thus, EGS is an indicator of community complexity, and more complex communities such as those in soils tend to have a larger EGS than less complex communities one might find in low nutrient features such as open ocean surface water [78]. By analyzing marker genes from the cave epilithic biofilm metagenome and using the estimated marker gene density value in the EGS formula 2.3.1, an estimated EGS of 4.2 Mbp was derived. Elsewhere, EGS values have ranged from 1.6 Mbp for the bacterial fraction of the Sargasso Sea metagenome, to 6.3 Mbp for the bacterial fraction of soil communities [78]. Since the HAVO sample clearly ranks towards more complex microbial communities on the basis of EGS, one might conclude that greater sampling effort would define the actual functional diversity of the community.

2.4.6 Metagenome assembly

Table 2.11. Metagenome assembly statistics

Total number of 454 reads	386,217
Total number of 454 bases	95,386,202
Number of Newbler contigs (≥500bp)	4,884
Largest Newbler contig size	13,678
Total size of Newbler contigs (≥500bp)	5,575,315 bp
Total size of Newbler contigs (all contigs)	6,101,596 bp
Number of Velvet contigs (≥500bp)	6,904
Largest Velvet contig size	8,729 bp
Total size of Velvet contigs (≥500bp)	5,922,262 bp
Total size of Velvet contigs (all contigs)	21,592,282 bp
Number of PCAP contigs (≥500bp)	7,388
Largest PCAP contig size	22,554 bp
Total size of PCAP contigs (≥500bp)	9,253,524 bp
Total size of PCAP contigs (all contigs)	20,303,079

The results of three different assemblies of the metagenome are presented (Table 2.11). Assemblies were attempted with three different assemblers: Newbler, Velvet, and PCAP. The metagenome was produced by a second generation 454 sequencer and lacked paired-end capability to resolve repeats. Without accompanying mate-pair information, assemblies can be difficult and result in misassemblies. Nonetheless, in a complex microbial community such as the HAVO biofilm, co-assembly of sequences would be expected to be rare due to the highly diverse nature of the sequence pool, and under-sampling of the community genomic DNA. Since the metagenomic library was not pair-ended, scaffolding of contigs could not be accomplished and contigs remain as 'singletons' without linkage information between them.

2.4.7 Metagenome recruitment analysis

Fragment recruitment (FR) can reveal abundant and sometime divergent organisms in metagenomic data sets, and organisms that are distant relatives of known organisms whose genomes have been completely sequenced [63, 58]. This process differs from binning or classification of sequence reads using known pre-conditions such as tetra-nucleotide frequencies or *k*-mers; it simply compares metagenomic data with known reference genomes, and where sequence reads match these references, one can infer that a close relative of the reference organism is represented in the metagenome. The abundance of matching species does not reflect the actual abundance of organisms in the metagenomic data set because, most often, there are no close relatives in public genome databases of organisms in the metagenomic sample, so only sequences with close relatives amomng reference genomes are detected through these searches.

The fragment recruitment tool in MG-RAST was utilized to identify organisms similar to known reference organisms. Although FR can be carried out with MUMMER and BLAST, the availability and frequent updates of MG-RAST means more reference genomes are available against which one can compare metagenomic data, and that less time is needed to manually sift through the results. The MG-RAST fragment recruitment uses the following criteria: BLASTn with maximum E-value cutoff of $1e^{-3}$.

Several different species of *Bacteria* and some *Archaea* were recruited from the metagenomic sequences (Table 2.12). Many species of top-recruiting organisms are from the *Acidobacteria* and *Chloroflexi* phyla. The top-recruiting *Cyanobacteria* is *Gloeobacter violaceus* PCC 7421, as was expected from the purple color of the cave epilithic biofilm. An unexpected finding was recruitment of sequence reads matching the genome of *Nitrosopumilus maritimus* SCM1. This is a member of the phylum *Thaumarchaeota*, that hosts of Cenarchaeales, Nitrosopumilales, Nitrososphaerales,

and organisms from other unclassified environmental samples. It is an ammonia-oxidizing archaeon first cultivated in 2005 from an aquarium tank [100] and whose complete genome was sequenced in 2010 [101].

Table 2.12. Top 62 reference species recruited from the epilithic biofilm metagenome (>300 reads)

Organism	Phylum	Counts
Candidatus Solibacter usitatus Ellin6076	*Acidobacteria*	5587
Candidatus Koribacter versatilis Ellin345	*Acidobacteria*	4372
Ktedonobacter racemifer DSM 44963	*Chloroflexi*	2780
Roseiflexus castenholzii DSM 13941	*Chloroflexi*	1772
Roseiflexus sp. RS-1	*Chloroflexi*	1728
Sphaerobacter thermophilus DSM 20745	*Chloroflexi*	1591
Chthoniobacter flavus Ellin428	*Verrucomicrobia*	1484
Acidobacterium capsulatum ATCC 51196	*Acidobacteria*	1462
Herpetosiphon aurantiacus ATCC 23779	*Chloroflexi*	1453
Acidobacterium sp. MP5ACTX8	*Acidobacteria*	1322
Acidobacterium sp. SP1PR4	*Acidobacteria*	1318
Acidobacterium sp. MP5ACTX9	*Acidobacteria*	1242
Chloroflexus aurantiacus J-10-fl	*Chloroflexi*	1242
Chloroflexus sp. Y-400-fl	*Chloroflexi*	1151
Gloeobacter violaceus PCC 7421	*Cyanobacteria*	1019
Chloroflexus aggregans DSM 9485	*Chloroflexi*	1013
Thermobaculum terrenum ATCC BAA-798	Unclassified	880
Thermomicrobium roseum DSM 5159	*Chloroflexi*	840
Sorangium cellulosum So ce 56	*Proteobacteria*	827
Myxococcus xanthus DK 1622	*Proteobacteria*	771
Gemmatimonas aurantiaca T-27	*Gemmatimonadetes*	725
Candidatus Nitrospira defluvii	*Nitrospirae*	618
Nostoc punctiforme PCC 73102	*Cyanobacteria*	614
Anabaena variabilis ATCC 29413	*Cyanobacteria*	612
Rubrobacter xylanophilus DSM 9941	*Actinobacteria*	600
Anaeromyxobacter dehalogenans 2CP-C	*Proteobacteria*	593
Anaeromyxobacter sp. Fw109-5	*Proteobacteria*	573
Meiothermus silvanus DSM 9946	*Deinococcus-Thermus*	533
Cyanothece sp. PCC 7425	*Cyanobacteria*	500
Bradyrhizobium japonicum USDA 110	*Proteobacteria*	489
Geobacter metallireducens GS-15	*Proteobacteria*	406
Microcoleus chthonoplastes PCC 7420	*Cyanobacteria*	400
Opitutus terrae PB90-1	*Verrucomicrobia*	394
Meiothermus ruber DSM 1279	*Deinococcus-Thermus*	385
Cyanothece sp. PCC 7424	*Cyanobacteria*	363
Verrucomicrobium spinosum DSM 4136	*Verrucomicrobia*	359
Nostoc sp. PCC 7120	*Cyanobacteria*	355
Acaryochloris marina MBIC11017	*Cyanobacteria*	351
Blastopirellula marina DSM 3645	*Planctomycetes*	340
Truepera radiovictrix DSM 17093	*Deinococcus-Thermus*	336
Nitrosopumilus maritimus SCM1	*Thaumarchaeota*	335
Moorella thermoacetica ATCC 39073	*Firmicutes*	327
Carboxydothermus hydrogenoformans Z-2901	*Firmicutes*	326
Pelobacter carbinolicus DSM 2380	*Proteobacteria*	325
Streptosporangium roseum DSM 43021	*Actinobacteria*	319
Pelobacter propionicus DSM 2379	*Proteobacteria*	319
Chitinophaga pinensis DSM 2588	*Bacteroidetes*	318
Thermus thermophilus HB27	*Deinococcus-Thermus*	318
Pelotomaculum thermopropionicum SI	*Firmicutes*	316
Planctomyces limnophilus DSM 3776	*Planctomycetes*	314
Mesorhizobium loti MAFF303099	*Proteobacteria*	304

2.4.8 Analysis of metabolic genes of interest in the epilithic biofilm metagenome

Metagenomes can reveal the presence of a wide range of genes, especially those that by definition go undetected in 16S rDNA clone libraries, or by PCR with degenerate primers that rely on known conserved genes. The HAVO metagenome, even though undersampled, contains a wealth of information and shows unprecedented diversity of genes in many pathways. To identify the most abundant genes in the HAVO community, a custom Python script retrieved gene names based on KEGG Orthologous (KO) group numbers identified by KAAS annotation (See 5.1.16). The abundances of these genes were tabulated, and those of interest extracted for further analysis. The fifty most abundant genes and their descriptions are recorded here (Table 2.13)

The most abundant genes belong to functional groups devoted to generation of energy, or for cellular functions such as ribosomal proteins or elongation factors. The most abundant gene, *atoB* (acetyl-CoA C-acetyltransferase) is involved in several metabolic pathway modules: ketone body biosynthesis, C5 isoprenoid biosynthesis (mevalonate pathway), ethylmalonyl pathway, dicarboxylate-hydroxybutyrate cycle, and hydroxypropionate-hydroxybutylate cycle. As it is a widely utilized enzyme, its abundance in the metagenome is not unexpected.

The analysis of individual genes identified in the HAVO metagenome did contain some surprises. First, the *mcrA* gene that encodes for methyl-coenzyme M reductase alpha subunit was not found, although a qPCR (data not shown) detected the *mcrA* gene, thus indicating the presence of methanogens. Similarly, *nifH*, a marker gene for nitrogen fixation was not found in the metagenome, although other components of nitrogen metabolism (*nifA* and *nifJ*) were detected. Nitrogen fixing bacteria such as *Anabaena* and *Rhizobium* were detected in the metagenome (Figures 2.14 and 2.10, respectively) but the absence of *nifH* in the metagenome may simply suggest that the community was undersampled.

Noteworthy, too, was the apparent absence of carbon monoxide dehydrogenase genes (*coxS*, *coxL*, and *coxM*) from the HAVO metagenome. Furthermore, only two copies of the carbon-monoxide dehydrogenase small subunit (*coxS*) were detected. Other subunits *coxM* and *coxL* were not found in the metagenome, although these genes would have been expected given the abundance of Beta*proteobacteria* and specifically *Burkholderia* who are known to harbor carbon monoxide oxidation genes [19, 21, 23, 102]. An archaeal ammonia monooxygenase subunit B (AmoC) gene having 97.4% amino acid sequence identity (77/79 amino acid sequences) to that of *Nitrosopumilus maritimus* SCM1 was detected in the metagenome, strongly suggesting the presence of a close relative of *N. maritimus* SCM1 in the cave epilithic biofilm community. However, methane oxi-

dation genes (*pmo*, *mmo*) were not found, although several taxa known to oxidize methane (such as *Methylacidiphilum* and *Methylobacterium*) were detected in the metagenome (Figures 2.19 and 2.10, respectively).

Table 2.13. Fifty most abundant genes detected in the epilithic biofilm metagenome

Gene name	Gene description	Gene counts (% of total)
atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	41 (0.454545)
ndh	NADH dehydrogenase [EC:1.6.99.3]	33 (0.365854)
fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	31 (0.343681)
copA	Cu^{2+}-exporting ATPase [EC:3.6.3.4]	30 (0.332594)
dnaK	molecular chaperone DnaK	28 (0.310421)
rpoE	DNA-directed RNA polymerase subunit delta	27 (0.299335)
guaB	IMP dehydrogenase [EC:1.1.1.205]	26 (0.288248)
acnA	aconitate hydratase 1 [EC:4.2.1.3]	26 (0.288248)
rpsA	small subunit ribosomal protein S1	24 (0.266075)
mfd	transcription-repair coupling factor (superfamily II helicase) [EC:3.6.4.-]	24 (0.266075)
lpd	dihydrolipoamide dehydrogenase [EC:1.8.1.4]	24 (0.266075)
ilvB	acetolactate synthase I/II/III large subunit [EC:2.2.1.6]	24 (0.266075)
fusA	elongation factor G	24 (0.266075)
valS	valyl-tRNA synthetase [EC:6.1.1.9]	23 (0.254989)
tuf	elongation factor Tu	22 (0.243902)
lon	ATP-dependent Lon protease [EC:3.4.21.53]	22 (0.243902)
gcvP	glycine dehydrogenase [EC:1.4.4.2]	22 (0.243902)
gcp	O-sialoglycoprotein endopeptidase [EC:3.4.24.57]	22 (0.243902)
dnaE	DNA polymerase III subunit alpha [EC:2.7.7.7]	22 (0.243902)
uvrB	excinuclease ABC subunit B	21 (0.232816)
metK	S-adenosylmethionine synthetase [EC:2.5.1.6]	21 (0.232816)
isp	major intracellular serine protease [EC:3.4.21.-]	21 (0.232816)
acd	acyl-CoA dehydrogenase [EC:1.3.8.7]	21 (0.232816)
uvrD	DNA helicase II / ATP-dependent DNA helicase PcrA [EC:3.6.4.12]	20 (0.221729)
groEL	chaperonin GroEL	20 (0.221729)
glmS	glucosamine–fructose-6-phosphate aminotransferase (isomerizing) [EC:2.6.1.16]	20 (0.221729)
gatA	aspartyl-tRNA(Asn)/glutamyl-tRNA (Gln) amidotransferase subunit A [EC:6.3.5.6 6.3.5.7]	20 (0.221729)
acs	acetyl-CoA synthetase [EC:6.2.1.1]	20 (0.221729)
thrS	threonyl-tRNA synthetase [EC:6.1.1.3]	19 (0.210643)
serA	D-3-phosphoglycerate dehydrogenase [EC:1.1.1.95]	19 (0.210643)
pnp	polyribonucleotide nucleotidyltransferase [EC:2.7.7.8]	19 (0.210643)
pgm	phosphoglucomutase [EC:5.4.2.2]	19 (0.210643)
metG	methionyl-tRNA synthetase [EC:6.1.1.10]	19 (0.210643)
map	methionyl aminopeptidase [EC:3.4.11.18]	19 (0.210643)
gyrB	gyrase subunit B [EC:5.99.1.3]	19 (0.210643)
fabH	3-oxoacyl-[acyl-carrier-protein] synthase III [EC:2.3.1.180]	19 (0.210643)
fabG	3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100]	19 (0.210643)
dnaJ	molecular chaperone DnaJ	19 (0.210643)
tktA	transketolase [EC:2.2.1.1]	18 (0.199557)
rplB	large subunit ribosomal protein L2	18 (0.199557)
pgk	phosphoglycerate kinase [EC:2.7.2.3]	18 (0.199557)
hemN	oxygen-independent coproporphyrinogen III oxidase [EC:1.3.99.22]	18 (0.199557)
gnd	6-phosphogluconate dehydrogenase [EC:1.1.1.44]	18 (0.199557)
gltX	glutamyl-tRNA synthetase [EC:6.1.1.17]	18 (0.199557)
carB	carbamoyl-phosphate synthase large subunit [EC:6.3.5.5]	18 (0.199557)
asnB	asparagine synthase (glutamine-hydrolysing) [EC:6.3.5.4]	18 (0.199557)
sdhA	succinate dehydrogenase flavoprotein subunit [EC:1.3.99.1]	17 (0.18847)
rpoC	DNA-directed RNA polymerase subunit beta' [EC:2.7.7.6]	17 (0.18847)
prsA	ribose-phosphate pyrophosphokinase [EC:2.7.6.1]	17 (0.18847)
pps	pyruvate, water dikinase [EC:2.7.9.2]	17 (0.18847)

Analysis of specific marker genes led to the conclusion that the metagenome was probably not sampled to great enough depth because organism abundance and expected marker gene abundance do not overlap. Carbon monoxide dehydrogenase genes were expected to be found in large numbers due to higher abundance of *Betaproteobacteria* and specifically *Burkholderia* known to encode a large number of *coxS*, *coxM*, and *coxL* genes.

2.4.9 Comparative metagenomic analyses

The availability of metagenomic data sets in publicly accessible resources such as MG-RAST [75], CAMERA [103], and IMG/M [104] permits comparison of the HAVO biofilm metagenome with data sets from other habitats. However, given that few habitats represented in publicly available metagenomic data sets are comparable with that in which the HAVO biofilm is located, the HAVO metagenome was compared with those from habitats likely to contain physiologically similar and dissimilar taxa. This approach allowed the closest neighbor of the HAVO community to be determined. MG-RAST and its associated databases was used for these comparative analyses. Environments selected for these comparisons are listed in Table 2.2.

2.4.9.1 Comparison of species richness and evenness

A measure known as α diversity describes species diversity and richness in a community. These diversity estimates were obtained here through the MG-RAST server and presented graphically (Figure 2.24). The MG-RAST server uses the Shannon diversity index to calculate α diversity (See 2.4.1).

$$\alpha = exp(H) = exp\left(-\sum_{i=1}^{m} p_i ln(p_i) \right) \tag{2.4.1}$$

The Guerrero Negro microbial mat presents the most diverse community based on α diversity estimates, although the HAVO epilithic biofilm is clearly very close (Figure 2.24). Rank abundance plots can show species richness and evenness in a community. Such plots produced by the MG-RAST server to compare community structures in different habitats show the HAVO mat sample has an abundance profile most similar to that of a Puero Rico forest soil (Figs. 2.25, 2.26, 2.31). This is also evident when abundance profiles are compared side-by-side to show which taxonomic groups have similar abundances (Figure 2.32. Yellowstone hot spring samples are dominated by *Cyanobacteria*, and there are fewer representatives of other phyla, so it is a very uneven community. By contrast, the HAVO epilithic biofilm, Netherlands soil, and Puerto Rican forest soil

show gently sloping rank abundance plots, indicating a more even distribution of phyla among their respective communities than in the Yellowstone hot spring community.

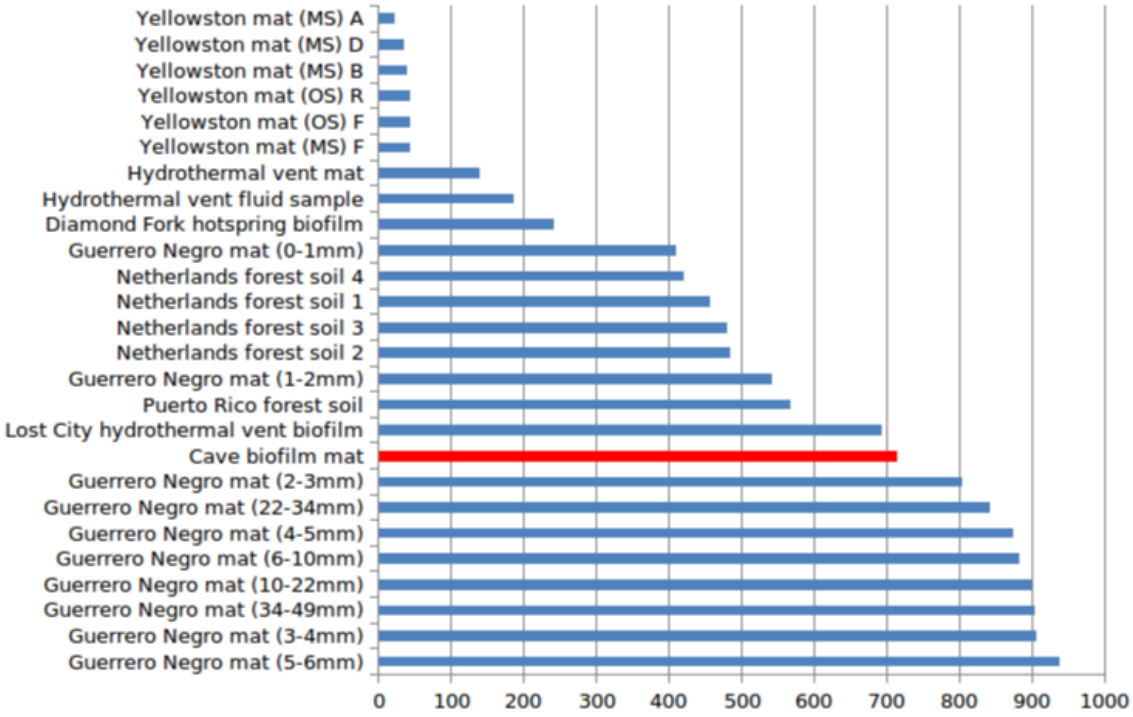


Figure 2.24. Comparison of α diversity between similar habitats based on metagenomic data.

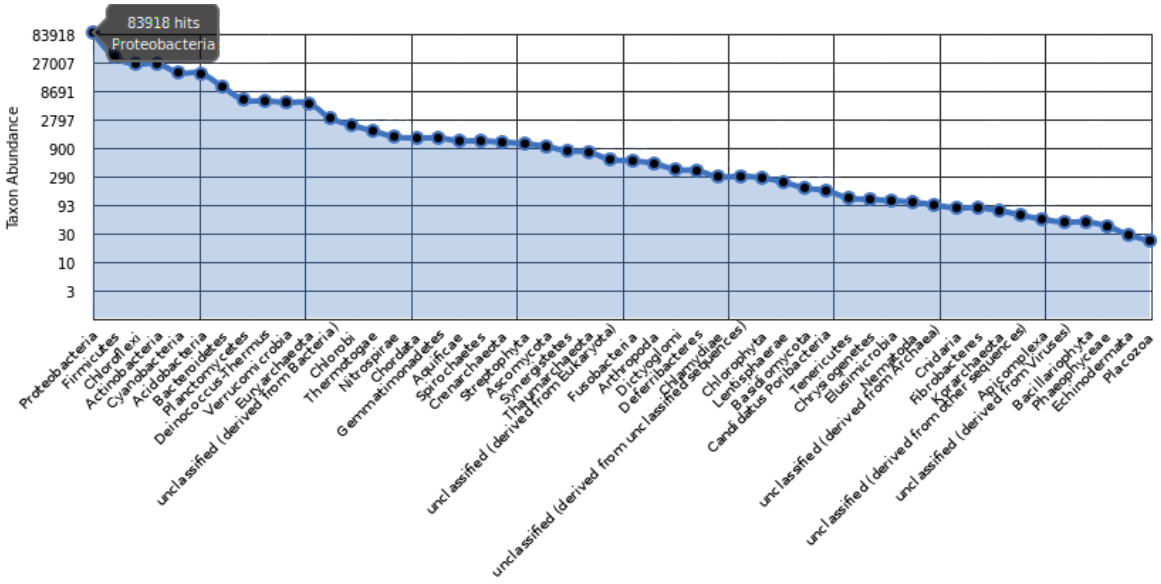


Figure 2.25. Rank abundance plot of taxa in the HAVO epilithic biofilm based on metagenomic reads. The most abundant phylum is *Proteobacteria*.

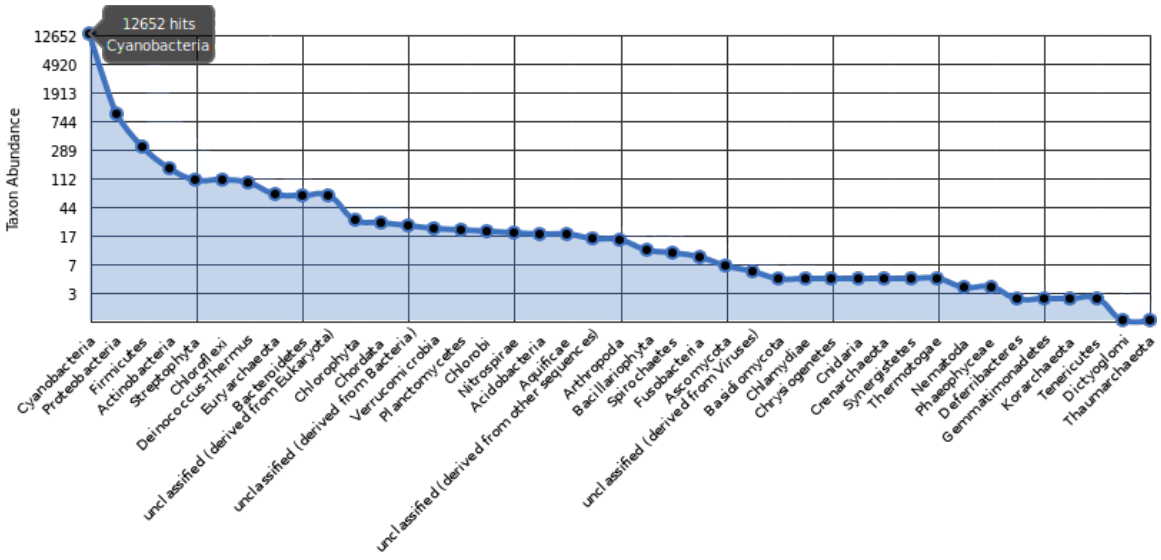


Figure 2.26. Rank abundance plot of taxa in Mushroom Springs mat core samples.

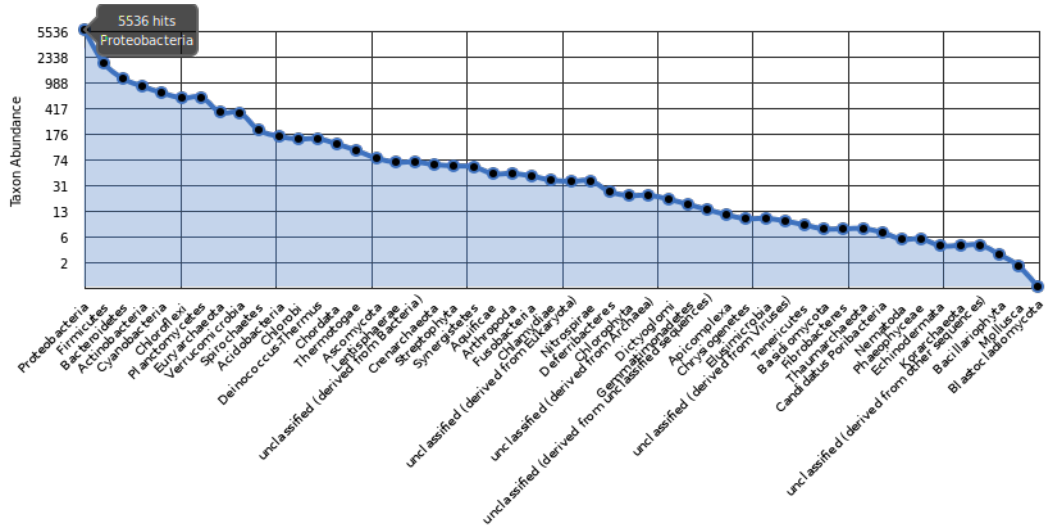


Figure 2.27. Rank abundance plot of taxa in the Guerrero Negro mat (5-6 mm).

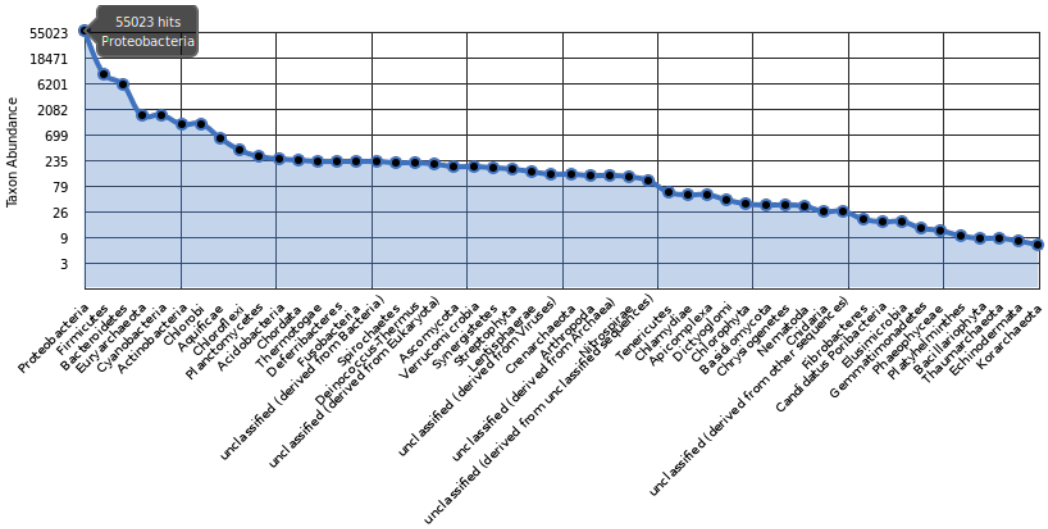


Figure 2.28. Rank abundance plot of taxa in a Lost City hydrothermal vent sample.

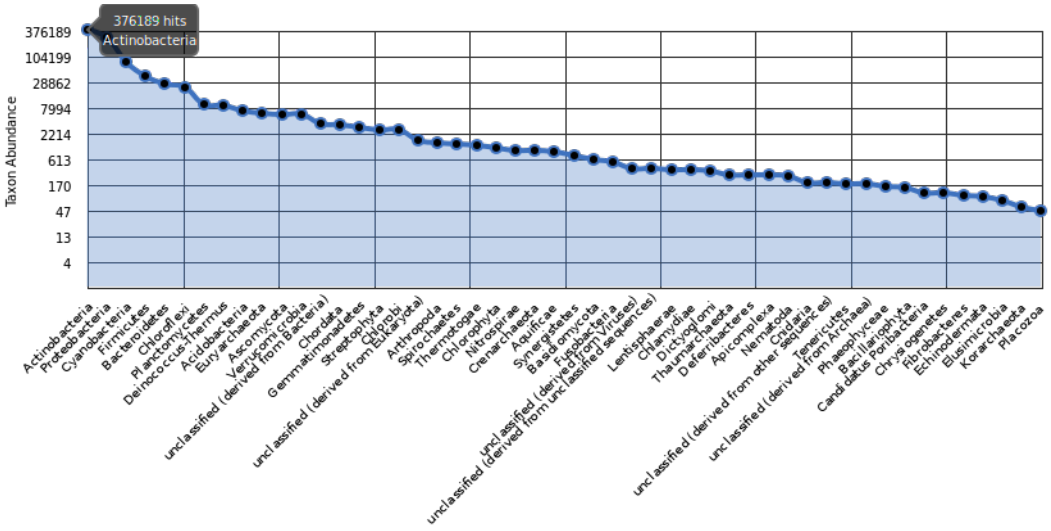


Figure 2.29. Rank abundance plot of taxa in a Netherland forest soil.

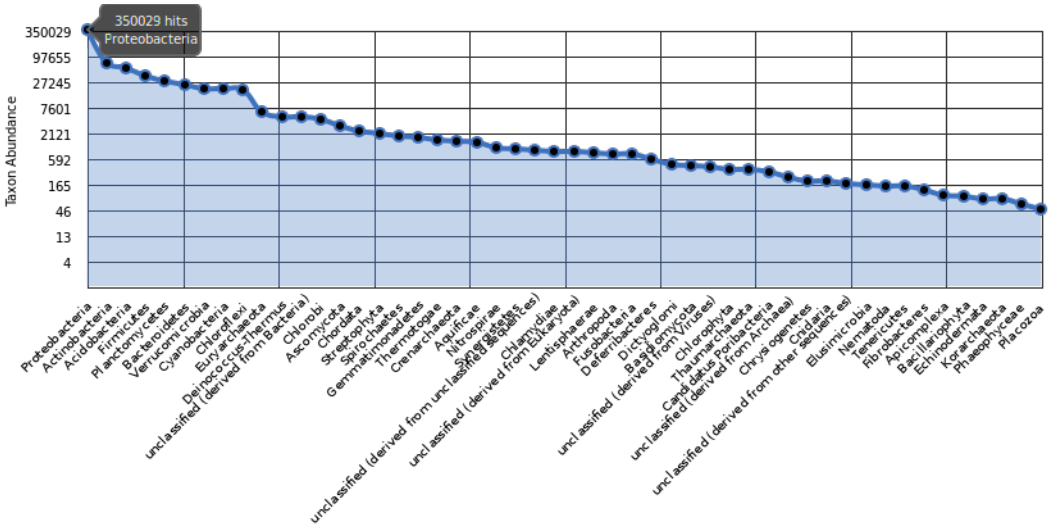


Figure 2.30. Rank abundance plot of Puerto Rican forest soil.

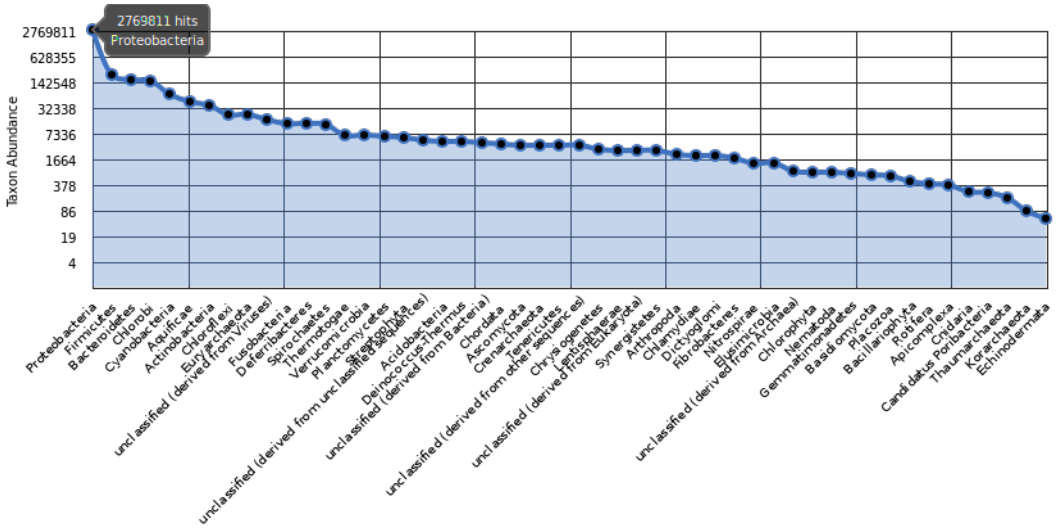


Figure 2.31. Rank abundance plot of taxa in a Diamond Fork hot spring biofilm.

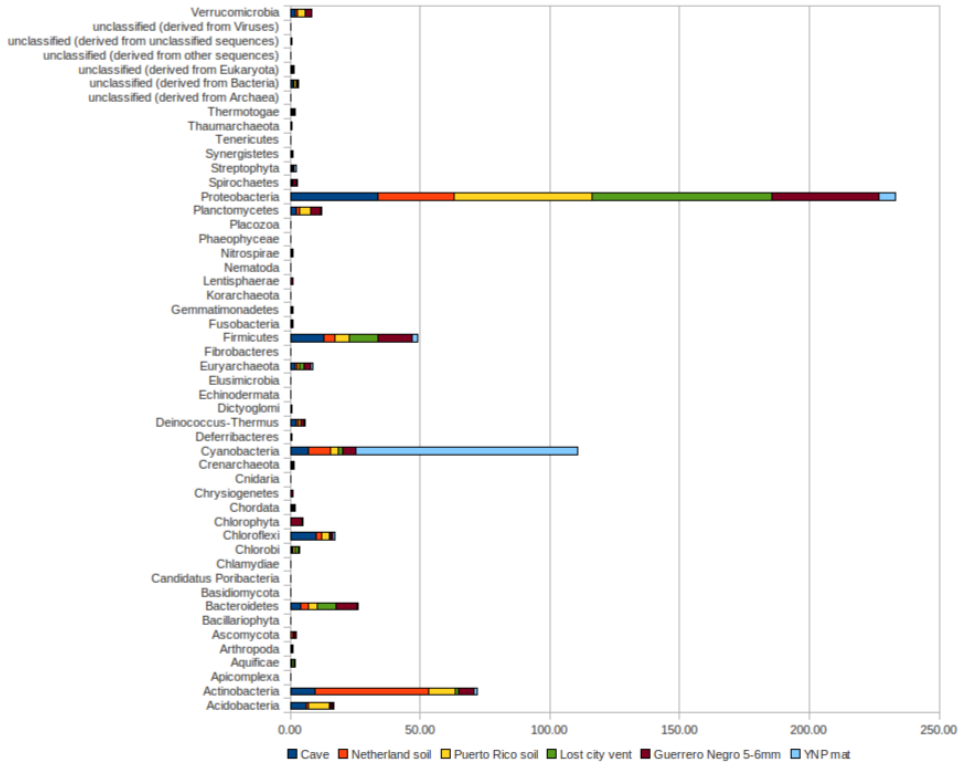


Figure 2.32. Comparison of rank abundances.

From the comparison of rank abundances, it was tentatively concluded that the HAVO epilithic biofilm community may share characteristics of those in soil rather than microbial mat communities.

2.4.9.2 Principal Component Analysis (PCA) of habitats based on species and metabolic abundance

Principal Component Analysis (PCA) is a widely used method to reduce dimensions in large data sets so they can be more easily visualized for interpretation. PCA was used to compare taxonomic or functional abundance profiles predicted by the MG-RAST server in different habitats, including the HAVO biofilm (Figures 2.33 to 2.37). The four plots illustrate all available options for PCA analysis with the MG-RAST server. MG-RAST uses several databases to annotate protein matches, so PCA was computed for each annotation (Figure 2.33 (KEGG), Figure 2.34 (eggNOG), Figure 2.35 (M5NR), and Figure 2.37 (Refseq)). PCA analysis based on a KEGG-based annotation

showed that the HAVO biofilm clustered closely with both microbial mat samples (green dots) and soil samples (purple dots) (Figure 2.33).

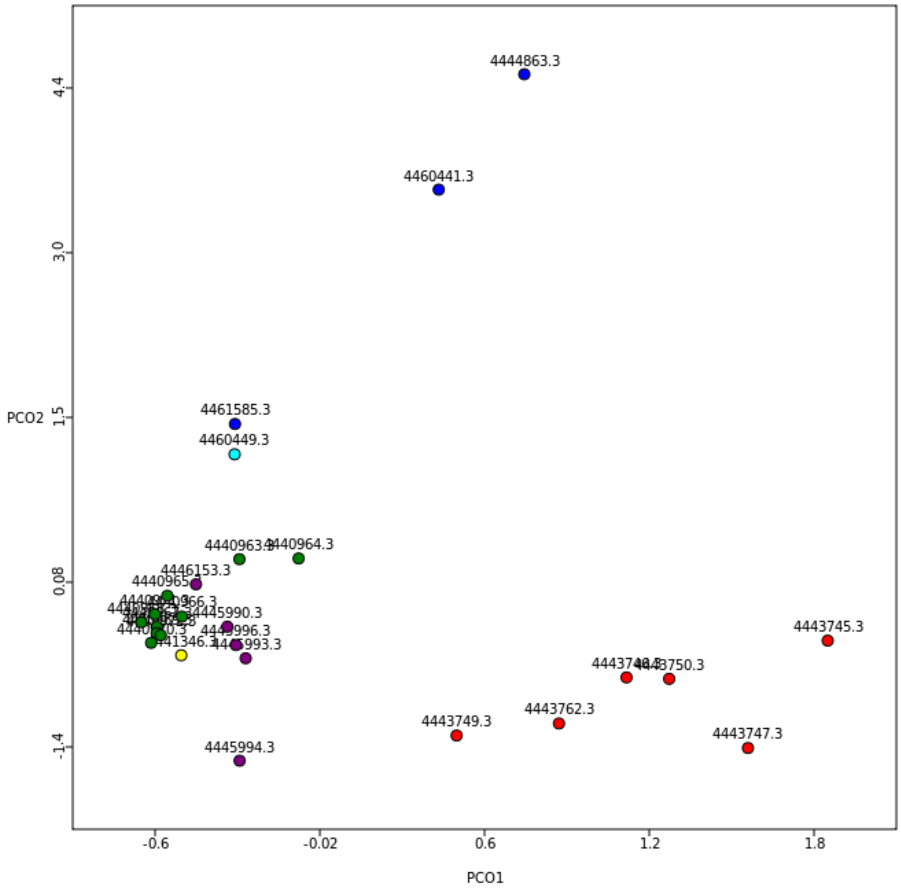


Figure 2.33. PCA plot based on taxonomic abundance profiles of 26 metagenomic samples using KEGG functional annotations. Different colors represent different biomes (Purple: soil, cyan: hot spring, green: microbial mat, blue: hydrothermal vents, red: hot spring microbial mat). HAVO sample in yellow.

A PCA analysis based on eggNOG-based annotation also showed the HAVO sample clustering with both microbial mat samples (green dots) and soil samples (purple dots) (Figure 2.34).

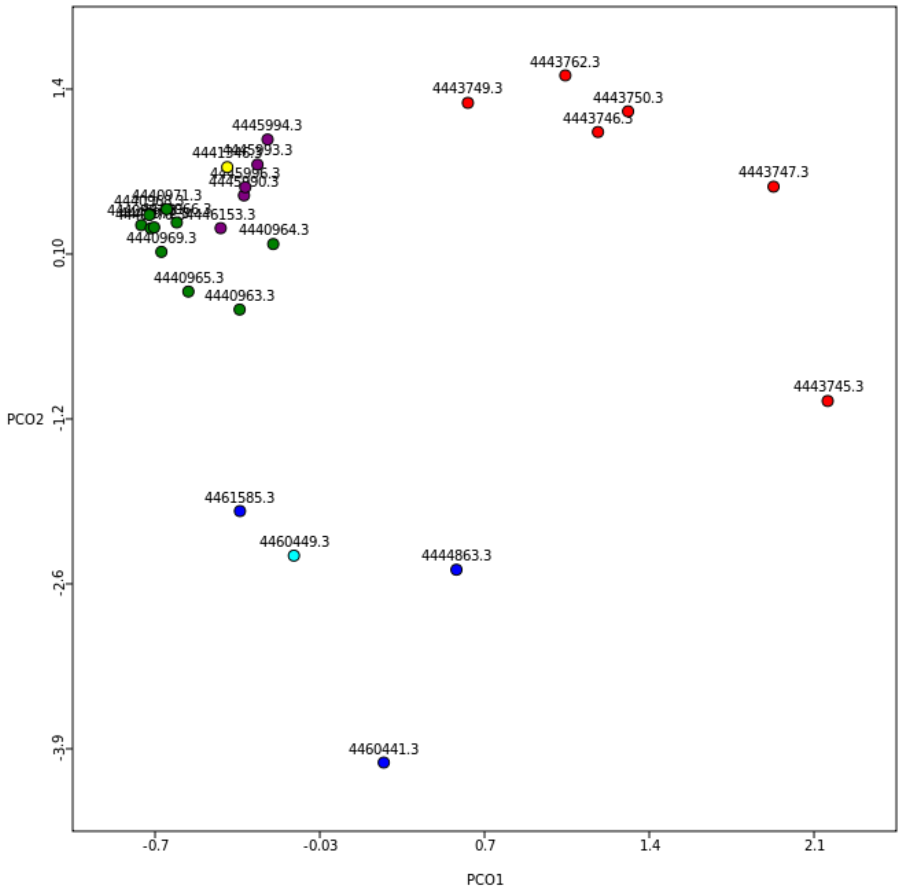


Figure 2.34. PCA plot based on taxonomic abundance profile of 26 metagenomic samples using eggNOG functional annotations. Different colors represent different biomes (Purple: soil, cyan: hot spring, green: microbial mat, blue: hydrothermal vents, red: hot spring microbial mat). HAVO sample is shown in yellow.

PCA analysis based on an M5NR-based annotation showed that HAVO sample clustered closely with soil samples (purple dots) and a hot spring sample (cyan dot) (Figure 2.35).

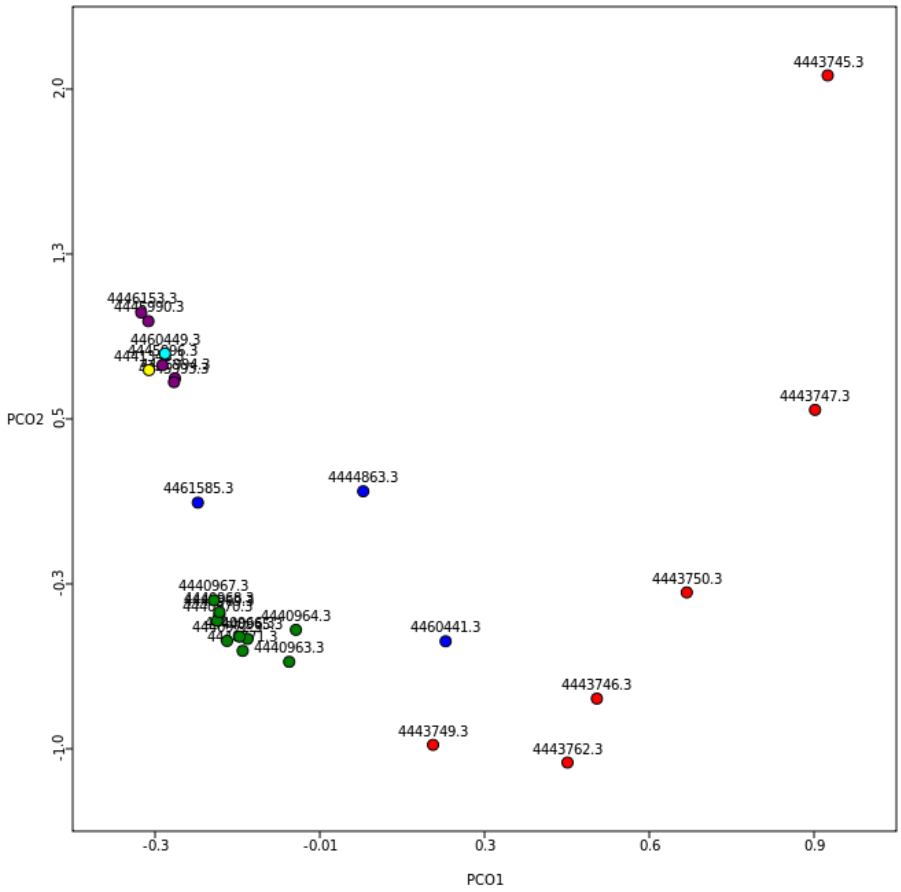


Figure 2.35. PCA plot based on taxonomic abundance profile of 26 metagenomic samples using M5NR functional annotations. Different colors represent different biomes (Purple: soil, cyan: hot spring, green: microbial mat, blue: hydrothermal vents, red: hot spring microbial mat). HAVO sample is shown in yellow.

Finally, PCA analysis based on Refseq-based annotation also showed the HAVO sample clustering closely with soil samples (purple dots) and a hot spring sample (cyan dot), as with the M5NR-based annotation (Figure 2.37).

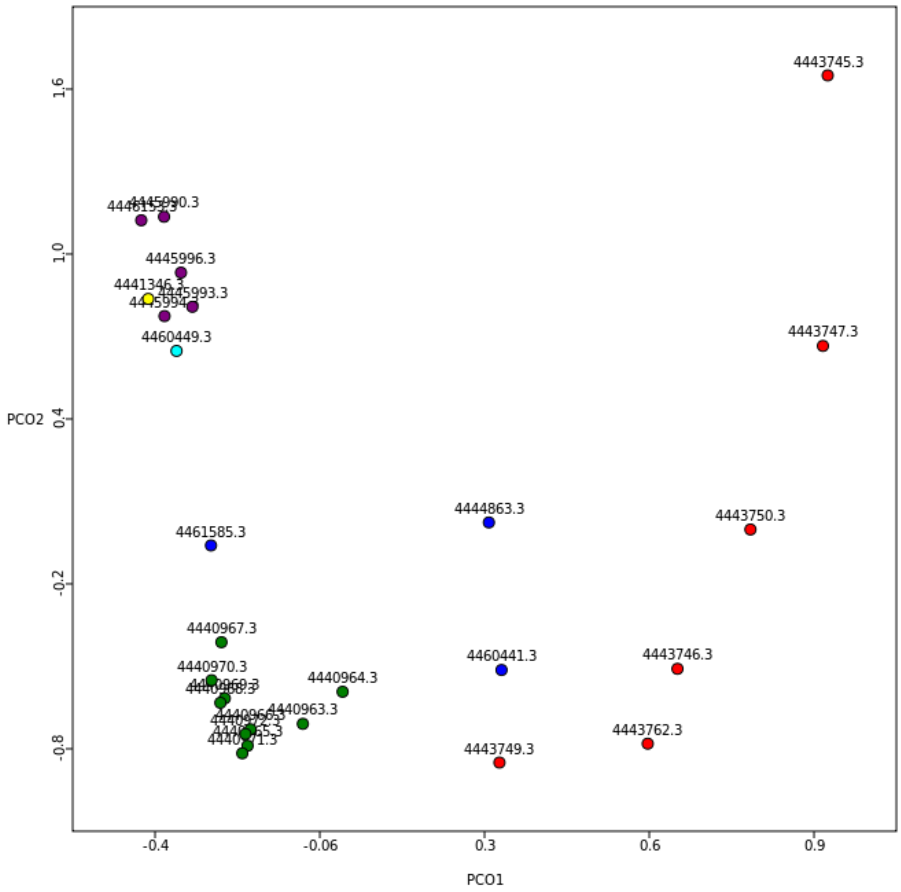


Figure 2.36. PCA plot based on taxonomic abundance profile of 26 metagenomic samples using Refseq functional annotations. Different colors represent different biomes (Purple: soil, cyan: hot spring, green: microbial mat, blue: hydrothermal vents, red: hot spring microbial mat). HAVO sample is shown in yellow.

Thus, PCA analysis using all four annotation methods indicated that the microbial community in the HAVO biofilm was closely related to that in soil on the basis of abundances of taxa identified in these samples. A PCA analysis was then used to compare the HAVO sample with the same habitats above but on the basis of metabolic functional gene abundances (instead of taxa abundances). PCA here used abundances of KEGG ortholog (KO) groups as the basis for comparison, and again revealed the HAVO sample (in red) grouped most closely with soil samples (yellow).

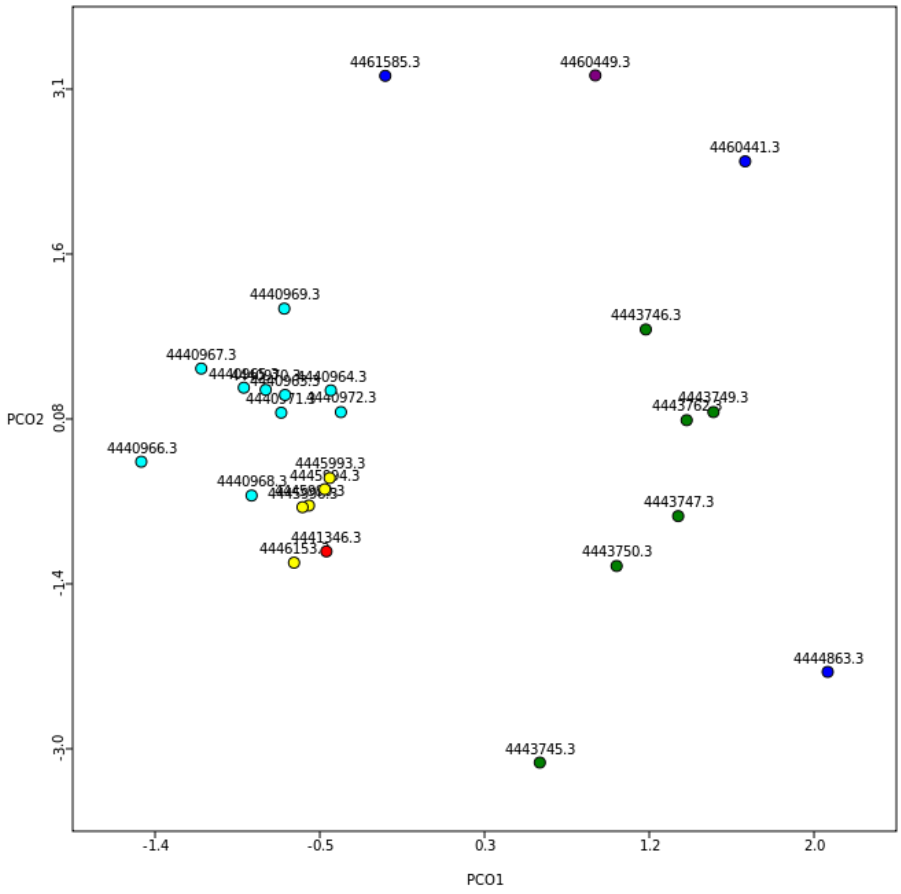


Figure 2.37. PCA plot based on metabolic abundance profiles of 26 metagenomic samples using KEGG orthologous groups. Different colors represent different biomes (Yellow: soil, cyan: microbial mat, purple: hot spring, blue: hydrothermal vents, green: hot spring microbial mat). HAVO sample in red.

2.4.9.3 Heatmap clustering of habitats based on metabolic diversity and abundance

MG-RAST may also cluster habitats on the basis of abundances of metabolic functional categories in these habitats. Therefore, this functionality was used to determine if it would group together similar habitats based on KO abundances (Figure 2.38). The analysis showed the HAVO sample grouped with Puerto Rico Forest soil, providing further support to the results of the PCA analyses.

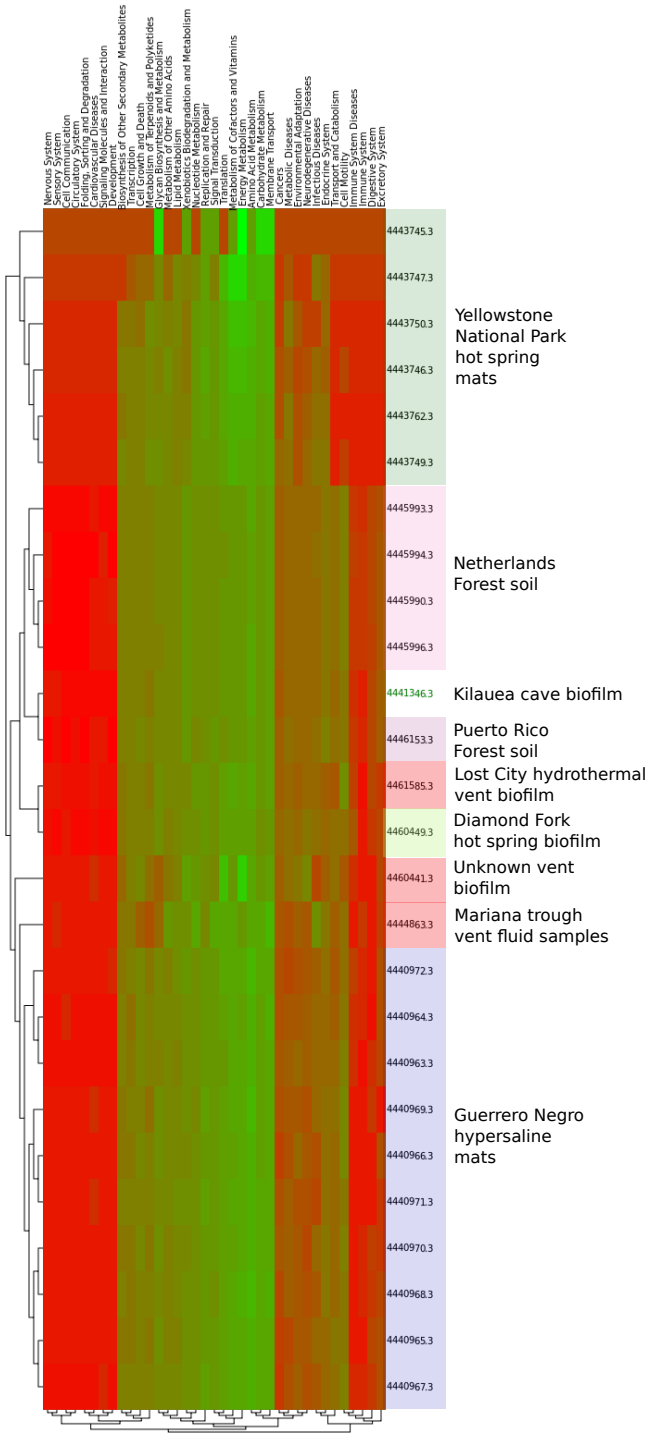


Figure 2.38. Heatmap plot based on abundance of KEGG functional categories among 26 metagenomic samples. Red represents 0 and green represent 1. Intermediate colors represent values between 0 and 1. HAVO sample is highlighted in green.

To further test if the HAVO sample is indeed closer in a microbial community structure content to soil samples than to other microbial mat samples, a Pearson correlation was calculated for KO abundances in these habitats. This analysis used abundances of KEGG functional categories in a similar way to PCA and MG-RAST heatmap analyses, but instead creates a Pearson correlation matrix based on these abundances. The similarities between the habitats based on this correlation matrix are then presented. The analysis again revealed the HAVO sample is closer to Puerto Rico forest soil and Netherland forest soil samples (sample names starting with NTS) (Figure 2.39). It is also interesting to see that Diamond Fork hot spring biofilm and Lost City hydrothermal vent microbial mat grouped (though distantly) with the HAVO sample.

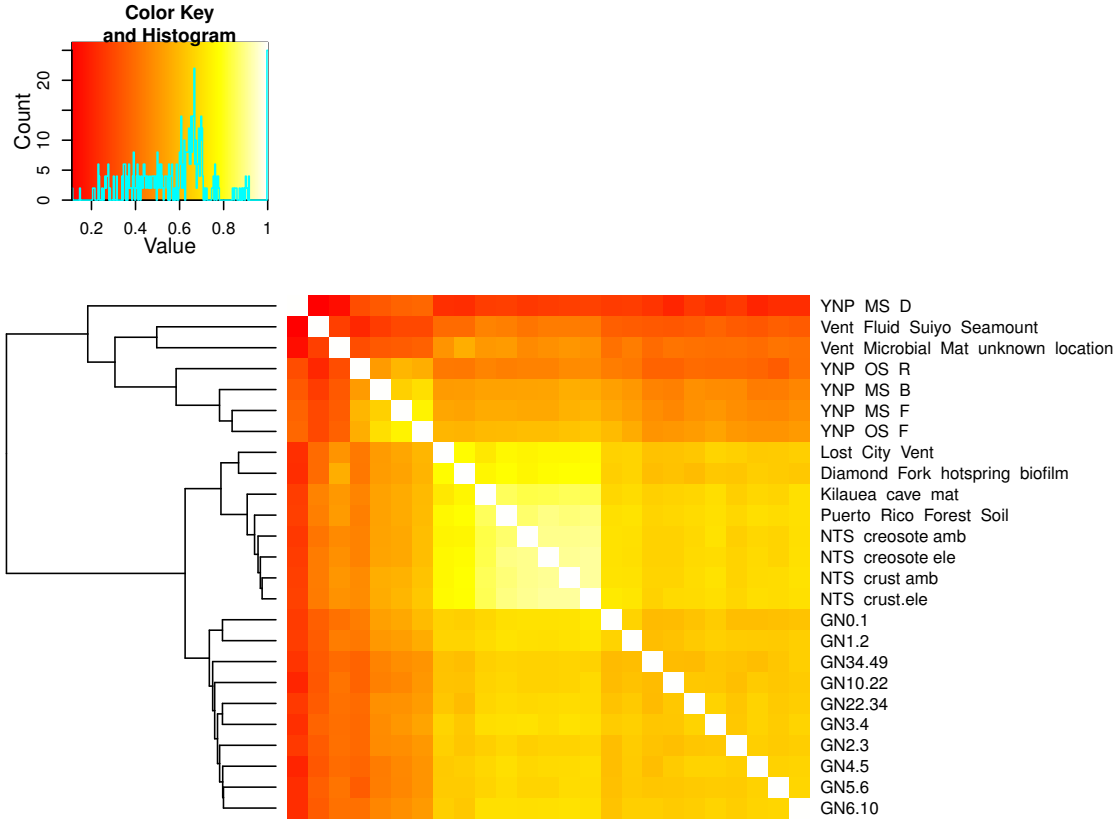


Figure 2.39. Heatmap of Pearson correlation matrix between 25 metagenomic samples based on abundance of KEGG functional categories. Note that Yellowstone Hot Spring Mat Core A sample was not included because its amino acids were not available for download from the MG-RAST server. White represents perfect correlation of 1 and red represents zero correlation. The matrix is a symmetric square matrix and dendrograms on the mirroring columns are omitted.

2.5 Conclusions

Ribosomal pyrotag sequencing revealed the HAVO microbial community was undersampled, and that more species would have been detected with greater sampling effort. The community rivals in complexity to those in soil and that in the Guerrero Negro hypersaline microbial mat. Based on unique tag-sequences, the community appears to be dominated by *Proteobacteria* (mostly alpha and beta), but prominent *Acidobacteria*, *Actinobacteria*, *Chloroflexi*, and *Cyanobacteria* fractions are also present.

Metagenomic sequencing showed that taxonomic abundance by this method was comparable to that determined by pyrotag sequencing, but with differences in the taxa detected. The most abundant taxa belonged in the *Proteobacteria*, mostly of the class *Betaproteobacteria*. On the basis of comparative metagenomic analyses, the HAVO community is also more closely related to communities in soil than to microbial mat communities. In this respect, the HAVO community may have originated, or at least be partially populated by microbes originating from nearby soils.

Recruitment analysis revealed the HAVO community contained close relatives of members of the *Chloroflexi* and *Acidobacteria*, whose genomes have been completely sequenced. An unexpected finding was the recruitment of a close relative of *Nitrosopumilus maritimus* SCM1 in the HAVO biofilm. This ammonia-oxidizing archaeon is a marine species, but a close relative appears to be in the HAVO biofilm as further evident by the detection of ammonia monooxygenase gene (AmoC) having 97.4% identity to the *Nitrosopumilus maritimus* SCM1 AmoC at the amino acid sequence level. Clearly, cultivating this organism would be of interest in order to characterize its physiology and to determine its role in the HAVO community. There were a few surprises in the metagenomic analysis, in that several metabolic marker genes such as *nifH* and *mcrA* were not detected although they were expected so. Lack of carbon monoxide dehydrogenase genes was also surprising, given the abundance of betaproteobacteria detected in the metagenome. This could be attributed to the fact that the sequencing effort was not deep enough to cover the true metabolic potential of the HAVO epilithic biofilm.

Chapter 3

Targeted cultivation of novel *Bacteria* from the HAVO cave epilithic biofilm

3.1 Abstract

Three new cyanobacteria were cultivated from the HAVO microbial mat, one each in the *Gloeobacter*, *Leptolyngbya*, and *Mastigocladus*. The *Gloeobacter* isolated shared 98.6% 16S rRNA gene sequence identity with *Gloeobacter violaceus* PCC 7421, the only cultivated species in the genus, and whose genome has been completely sequenced. The *Leptolyngbya* shares less than 95% 16S rRNA gene nucleotide identity with any known cyanobacteria, and highest sequence identity with a clone from a Greenland hot spring microbial mat (accession number: DQ431005.1). The *Mastigocladus* shares 98.12% 16S rRNA gene nucleotide sequence identity with a clone from Greenland (accession number: DQ431003.1), and 97.99% identity with *Fischerella* sp. JSC-11, a cultivated species currently in the draft sequencing stage. The genome of the novel *Gloeobacter* has been sequenced, and is described in Chapter 4 of this dissertation.

3.2 Introduction

The bacteria identified in a preliminary 16S ribosomal gene clone library and in metagenomic data indicate that multiple and diverse phyla are present. For example, a gene matching that of an ammonia-oxidation gene (*amoC* (ammonia monooxygenase subunit C; read name: EM7JFSU01BNU5Y) in *Nitrosopumilus maritimus* was detected. *Nitrosopumilus maritimus* is an ammonia-oxidizing archaeon (AOA) first isolated from an aquarium [100], and whose genome revealed a unique nitrogen metabolism and that AOA may be important in global nitrogen cycling

[101]. This is a marine archaeon and it is surprising to find a close relative in such a cave epilithic biofilm. Attempts were thus made to cultivate this archaeon from the biofilm using the medium described by Konneke et al. [100].

Genes matching cellulose degradation genes (endo-1,4-beta-glucanase) from *Acidothermus cellulolyticus* and other organisms were also identified in the cave biofilm metagenome. *Acidothermus cellulolyticus*, isolated from acidic hot springs in Yellowstone National Park is of particular interest because it can tolerate high temperatures, and its class of diverse cellulolytic enzymes could be of use in biotechnology [105]. Considering this potential application, an attempt was made to cultivate this close relative of an *Acidothermus* using LPBM acido-thermophile medium [105] (see Appendix B.3). Recently, Stott et al. [106] cultivated diverse bacterial phyla including *Proteobacteria*, *Firmicutes*, *Thermus/Deinococcus*, *Actinobacteria*, *Bacteroidetes*, *Chloroflexi*, *Acidobacteria*, and previously uncultured candidate division OP10 from geothermal soil in New Zealand [106]. The medium described by Stott et al. was used to here in an attaempt to cultivate bacteria from similar phyla detected in the HAVO biofilm metagenome.

Finally, diverse lineages of *Cyanobacteria* were detected in both the clone library and the metagenome from the HAVO sample, and attempts were made to cultivate them. Cyanobacteria may be important members of microbial communities, and here will make use of what little light is available in the cave entrance. Notably, they seem to be important contributors to biofilm structure due to the filamentous nature of some species, and perhaps by the production of mucoid extracellular material. Among the *Cyanobacteria* detected in the clone library, the most interesting finding was a relative of *Gloeobacter violaceus*. *Gloeobacter* is a deep-branching ancient cyanobacterium that lacks thylakoid structures, which requires it to instead carry out photosynthesis in the cell membrane; all other cyanobacteria have thylakoid membranes where photosynthesis is carried out [107, 108]. The only known cultivated representative is *Gloeobacter violaceus* PCC 7421, isolated in 1974 [107]. Cultivating only the second known *Gloeobacter*, more than thirty years since the only Type strain was first described, might be a significant contribution to many aspects of cyanobacteria research, and to the evolution of photosynthesis in general.

Targeted cultivation of novel microbes is an important process that is largely overlooked. As the majority of microbes in environments may be difficult to cultivate because they require highly specialized media compositions, most ecological studies tend to sequence clone libraries, pyrotags, or metagenomes and give little attention to cultivating bacteria in the laboratory. Although the availability of cheaper and higher throughput sequencing technologies is revolutionizing microbial ecology, the ultimate goal of these studies is to understand the physiologies of these *Bacteria* or

Archaea in diverse habitats. Studies that performed targeted cultivation include that of a *Leptospirillum* known to fix nitrogen (*Leptospirillum ferrodiazotrophum*), detected in acid mine drainage, and subsequently cultivated based on metagenomic data [109]. This approach represents quite a leap in microbial ecology, to identify metabolically interesting and important microbes from the environment and then cultivate them. In another example, Teske et al. were able to co-culture an *Arcobacter* and a *Desulfovibrio* from a cyanobacterial mat from Solar Lake (Sinai, Israel) that were always found together in a 16S rDNA sequence data based on DGGE (Denaturing Gradient Gel Electrophoresis) [110]. This study specifically designed media to satisfy nutritional requirements of both organisms based on predicted physiologies. Such examples highlight the importance of cultivation in microbial ecology and how molecular data can be used to characterize organisms of interest from the environment.

In this chapter I will present the results of targeted cultivation of novel taxa from the HAVO epilithic biofilm.

3.3 Materials and Methods

3.3.1 Cultivation media and their recipes

Media used in this work were prepared according to recipes in a standard text [111] and from papers describing isolations of specific organisms. Some modifications and other recipes are described here (Table 3.1; Appendix B).

Table 3.1. Cultivation media and their targets

Media name	Target organisms	Recipe
R2A	Heterotrophs	[111]
Nutrient Agar	Heterotrophs	[111]
AOAM	Ammonia-oxidizing *Archaea*	Appendix B.1 [100]
BG11M	*Cyanobacteria*	Appendix B.2 [111]
LPBM	Cellose-degrading bacteria	Appendix B.3 ATCC
FS1 and FS2	OP10 and *Acidobacteria*	Appendix B.4 [106]

3.3.2 Growth conditions

HAVO epilithic biofilm samples were aseptically dissected in the laboratory with a sterile scalpel. Small sections ($\sim$1 mm^3) were vortexed for several minutes in a particular liquid medium

in sterile 15 mL polypropylene tubes, and serially diluted before spread plating on different Petri plates. Plates were incubated at four incubation temperatures (29°C, 30°C, 45°C, and 50°C).

R2A and Nutrient Agar (NA) were used to cultivate heterotrophic bacteria. To cultivate ammonia oxidizing *Archaea*, inocula were spread on AOAM agar and incubated in darkness at 30°C until colonies formed. Similar conditions were used for the isolation of cellulose-degrading bacteria but with LPBM agar incubated at 45°C. FS1 and FS2 media cater to diverse bacteria listed by Stott et al. [106], and were incubated in darkness at 30°C. The goal of using FS1 and FS2 media was to target OP10 and *Acidobacteria* that are known to be difficult to cultivate but which are interesting nonetheless. To isolate thermophiles, a water bath was used to maintain a temperature of 50 $\pm$ 2 °C; and inoculated plates were placed above the water on on a plastic rack in the water bath.

Modified BG11 medium (BG11M) (Appendix B.2) was used to target cyanobacteria for cultivation. Inoculated culture tubes were wrapped in white paper towels to attenuate light, to mimic the light intensity in the cave entrance. Photosynthetically available radiation above the HAVO epiltihic biofilm was measured at $\sim$6.5 μEm^{-2}s^{-1} before noon on a clear day. Inoculated culture tubes were shaken under light in an incubator at 29°C for up to several months before sufficient material was obtained for microscopy and DNA extraction. Cyanobacteria were also cultivated on solid BG11 in Petri plates also wrapped in white paper towels.

3.3.3 Scanning electron microscopy

Log phase bacteria cultures (0.5 - 1.0 McFarland density, or 0.7 OD_{600}) in modified liquid BG11 were fixed by addition of 2.5% v/v EM grade 70% glutaraldehyde (Ted Pella Co.). Cells were then filtered onto 13 mm Isopore filters (0.8μm pore size) in Swinnex filter holders (Millipore Corp.). Subsequent steps (up to 70% ethanol) were completed with the filters in the Swinnex filter holders, but having first passed the solutions through a 0.22 μm pre-filter. Cells were rinsed in 0.1 M sodium cacodylate buffer (3 x 10 minutes; pH 7.4), then post fixed in 1% osmium tetroxide (OsO_4; Ted Pella Co.) in 0.1 M sodium cacodylate buffer for 40 minutes. This step was followed by three 10 minute rinses in the same buffer. Cells were dehydrated in an ascending ethanol concentration series from 10% to 70%. After the 70% ethanol rinse, the filters were removed from the Swinnex filter holders and placed in hand made lens tissue bags. These bags were placed in a stainless steel basket and dehydrated to 100% ethanol. Cells were critical point dried using liquid CO_2 (Tousimis Critical Point Dryer), and mounted and metal coated (gold:palladium; Hummer Sputter Coater II). Samples were examined using an Hitachi S-4800 field emission scanning electron microscope.

3.3.4 DNA extraction and 16S rRNA gene sequencing

Genomic DNA from all cultures deemed pure by microscopy and Gram stain was extracted using the MO BIO Ultraclean®Soil DNA isolation kit, according to the manufacturer's instructions. The quantity of genomic DNA from each extraction was estimated by gel electrophoresis. Bacterial primers (27F and 1492R) were used in polymerase chain reactions (PCR) to amplify a fragment of the 16S rRNA gene. Each PCR reaction contained 5 μl of 10$\times$ *Pfu* Buffer, 1 μl of 10 μM dNTP mixture, 5 μl of *Pfu* DNA polymerase, 1 μl of 10mM primer, 1 μl of DNA template, and nuclease-free water for a total of 50 μl. PCR conditions were 95°C (5 min), followed by 35 cycles of 95°C (30 sec), 52°C (30 sec), 72°C (30 sec), and a final extension of 72°C (7 min). PCRs were run in a Bio-Rad Thermal Cycler (Bio-Rad Laboratories, Hercules, CA). PCR products were cleaned with the MO BIO UltraClean® PCR Clean-Up Kit according to the manufacturer's instructions. Purified PCR products were sequenced using 27F, 533F, and 1492R primers, and assembled in the Seqman program (DNAstar Inc, Madison, WI) to produce near-full length 16S rDNA sequences.

3.3.5 Analysis of chlorophyll and carotenoid pigments by HPLC

An isolate tentatively identified through 16S rRNA gene sequencing as a *Gloeobacter* sp. was grown on modified BG11 agar plates for 3 weeks. Colonies were then extracted in HPLC-grade acetone (4°C, 24 hours). Extracts were warmed to room temperature, vortexed, and centrifuged for 5 minutes to remove cellular debris. Aliquots (1 ml) of the supernatant were combined with HPLC grade water (0.3 ml) in opaque autosampler vials, and injected (200 μL) onto a Varian 9012 HPLC system equipped with a Varian 9300 autosampler, a Timberline column heater (26°C), and a Waters Spherisorb® 5 μm ODS-2 analytical (4.6 x 250 mm) column and corresponding guard cartridge (7.5 x 4.6 mm). Pigments were detected with a ThermoSeparation Products UV2000 detector (λ1 = 436, λ2 = 450). A ternary solvent system was used for pigment analysis: Eluent A (methanol:0.5 M ammonium acetate, 80:20, v/v), Eluent B (acetonitrile:water, 87.5:12.5, v/v), and Eluent C (100% ethyl acetate). Solvents A and B contained an additional 0.01% 2,6-di-ter-butyl-p-cresol (0.01% BHT, w/v; Sigma-Aldrich) to prevent conversion of chlorophyll *a* into chlorophyll *a* allomers. The linear gradient used for pigment separation was a modified version of that described by Wright et al. (1991) [112] : 0.0' (90% A, 10% B), 1.00' (100% B), 11.00' (78% B, 22% C), 27.50' (10% B, 90% C), 29.00' (100% B), 30.00' (100% B), 31.00' (95% A, 5% B), 37.00' (95% A, 5% B), and 38.00' (90% A, 10% B) [113]. Eluent flow rate was maintained at 1.0 mL min^{-1}. Pigment peaks were identified by comparison of retention times with those of pure standards and extracts prepared

from algal cultures of known pigment composition. Whole extracts of JS1 were scanned between 350 and 800 nm in a Beckman DU800 spectrophotometer.

3.3.6 Phylogenetic analyses

Assembled and quality-checked 16S rRNA gene sequences were searched against the NCBI nt database using BLASTn to determine the closest neighbors. BLASTn results were manually checked to select taxa and Type strains for use in phylogenetic trees. Sequences chosen to build the 16S rRNA gene tree were aligned using the program Muscle [114] and trimmed using Gblocks [115]. Aligned and trimmed sequences were used for maximum likelihood analysis using the RAxML program [116] with 100 bootstrap replicates to build phylogenetic trees. The GTR + Γ model of nucleotide substitution was used, and the resulting trees were visualized with FigTree program. Exact parameters for Muscle, Gblocks, and RAxML are:

```
muscle -in toalign.fasta -out toalign.muscle
Gblocks toalign.muscle -t=d -e=-gb -b4=2
dissertation_ConvertAlignment.py toalign.muscle-gb fasta toalign.muscle-gb.phy phylip
raxmlHPC-PTHREADS-SSE3 -s toalign.muscle-gb.phy -n 16STree -T 2 -f a -x 12345 -# 100
-m GTRGAMMA
```

3.4 Results and Discussions

An early intent here was to cultivate and formally describe as many novel *Bacteria* and *Archaea* from the HAVO mat as time would permit. However, most of the organisms detected in the initial clone and metagenomic data sets that were specifically targeted did not grow on the media used. Furthermore, of those that did grow, most were not novel, sharing >98% 16S rRNA gene sequence identity with known species; they were thus not pursued further in terms of characterization. Only those cultivated organisms which are likely novel or of interest from an ecological perspective are considered further here, *e.g.*, cyanobacteria, which may be important in the HAVO mat's formation. A number of thermophilic bacteria were cultivated on R2A agar at $\sim$50°C, but they have yet to be identified and are not discussed further here.

3.4.1 Cultivation of *Cyanobacteria*

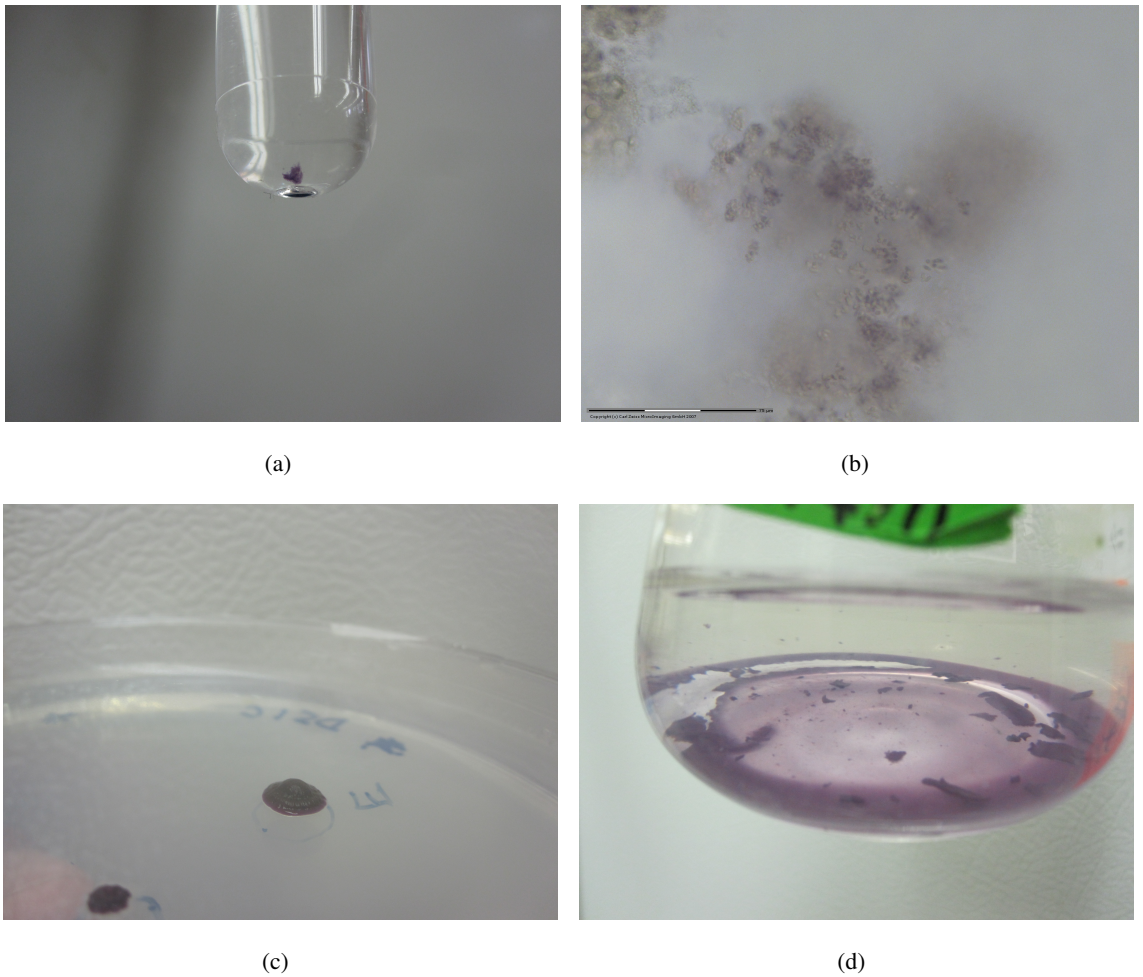
Brown et al. (unpublished) constructed a 16S rRNA gene clone library based on community genomic DNA extracted from the same HAVO epilithic biofilm in 2006. Partial-to-full

16S rDNA sequences were obtained by PCR amplification of community genomic DNA, and deposited in GenBank under 'Microbial diversity in a Hawaiian lava cave microbial mat' (Popset id number: 118084489). This data set contains 53 sequences of which 11 were considered cyanobacterial in origin. A sequence with 98.7% nucleotide identity to *Gloeobacter violaceus* PCC 7421 was identified in this popset (Accession No. EF032784.1. Finding this sequence encouraged an attempt to cultivate this rare, and here, potentially unique cyanobacterium from the biofilm. Ten other sequences of cyanobacterial origin were identified in the same clone library (accession numbers: EF032779.1, EF032780.1, EF032781.1, EF032782.1, EF032783.1, EF032785.1, EF032786.1, EF032787.1, EF032788.1, EF032789.1), so attempts were made to cultivate all cyanobacteria detected.

Using modified BG11 medium, HAVO biofilm samples collected in 2009 were shaken in culture tubes in a light incubator (in a continuous light cycle). Tubes were wrapped in a white paper towel to mimic the low-light levels in the cave entrance. Incubation times varied for each cyanobacterium. *Gloeobacter* was the slowest growing, generally taking from 2 weeks to a few months to produce visible purple clumps in the culture tubes or colonies on agar plates. Filamentous green cyanobacteria such as *Leptolyngbya* and *Fischerella* grew more quickly, with green filaments usually appearing within 1-2 weeks, leading eventually to biofilms on the walls of the culture tubes (Figure 3.5(c)). In order to increase chances of obtaining axenic samples, mixed cyanobacteria observed in liquid media were periodically inoculated on to solid media by spot inoculation to identify distinct colonies not mixed with other heterotrophic bacteria.

3.4.1.1 Cultivation of *Gloeobacter* sp. JS1

Purple *Gloeobacter* cells formed visible clumps at the bottom of culture tubes. These cells tended to form tightly associated cells surrounded by a sheath-like material resembling biofilm (Figure 3.1(b)). This material was prominent in SEMs of the cells (Figures 3.3 3.3(a) 3.3(b). At this point, the cultivated *Gloeobacter* strain was referred to as *Gloeobacter* sp. JS1



(a) (b)

(c) (d)

Figure 3.1. Light micrographs and photographs of cultivated *Gloeobacter* sp. cells. (a) First *Gloeobacter* sp. JS1 cells cultivated after collection of HAVO mat sample in 2009. Clumps of purple cells in a culture tube after being shaken for several weeks. (b) *Gloeobacter* sp. JS1 cells under a Zeiss PALM laser microdissection microscope. *Gloeobacter* cells tend to form clusters covered in capsule-like material. Scale bar is 75μm. (c) *Gloeobacter* sp. JS1 colony on BG11 showing raised colony morphology. (d) *Gloeobacter* sp. JS1 biofilm on the bottom of a conical flask.

Cells of *Gloeobacter* sp. JS1 are non-motile, unicellular rods of $\sim$0.8-1μm in width and 1-3μm in length (Figure 3.3). Cell division occurs by transverse binary fission in a single plane. *Gloeobacter* sp. JS1 cultures were routinely incubated at 29°C, but neither the optimal temperature for growth, nor the temperature range over which growth occurs, was investigated. On BG11M agar, *Gloeobacter* JS1 colonies are dark purple when the culture has been incubated only for up to several weeks. This color seems to be an indicator of the health status of the culture

(Figure 3.2). *Gloeobacter* JS1 cells autofluoresce when illuminated with a green laser (Figure 3.4). Autofluorescent cells were brighter at their poles.

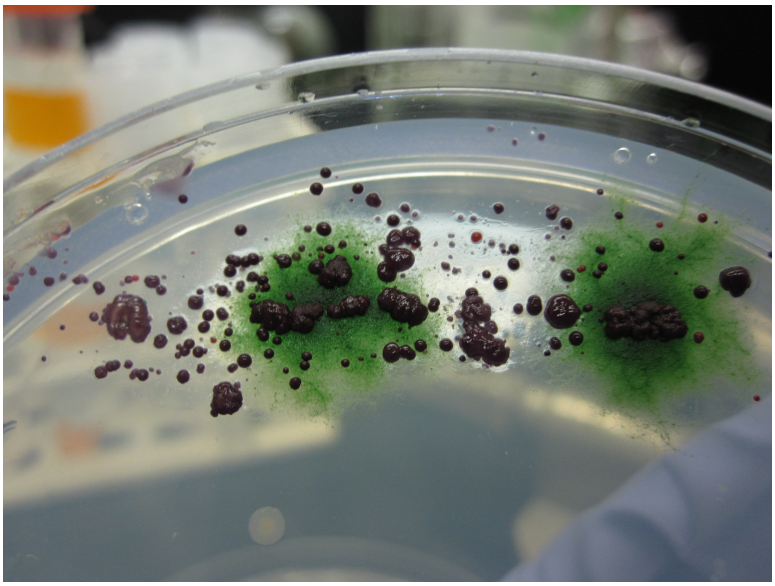
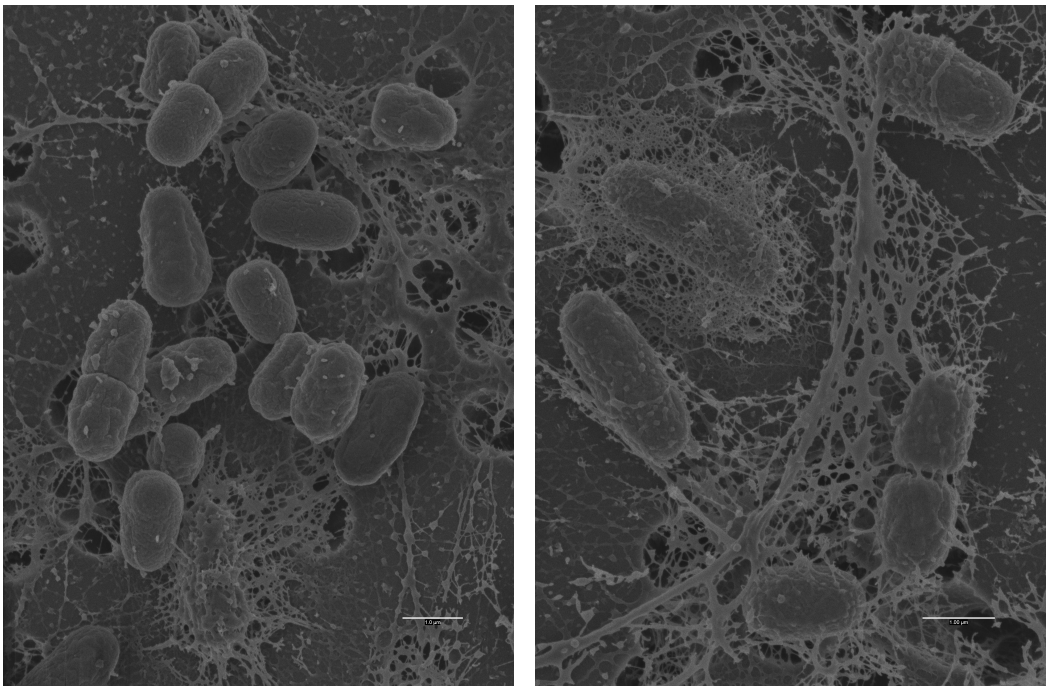


Figure 3.2. Non-axenic *Gloeobacter* JS1 on BG11M agar.



(a) (b)

Figure 3.3. SEMs of cultivated *Gloeobacter* sp. JS1 cells. (a) Cells enveloped in mucilaginous material are noticible near the bottom of the figure. (b) Mucilaginous material can be seen surrounding the cells. Scale bar is 1μm long.

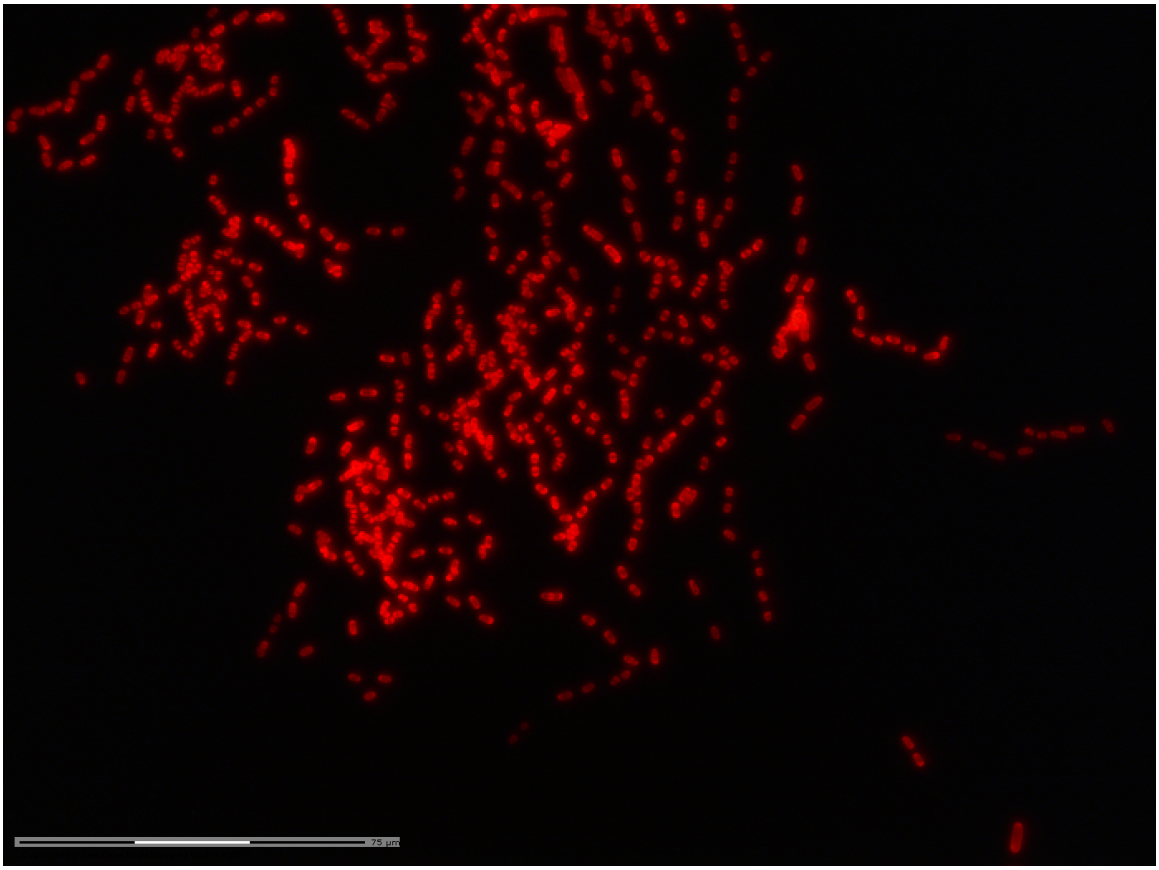
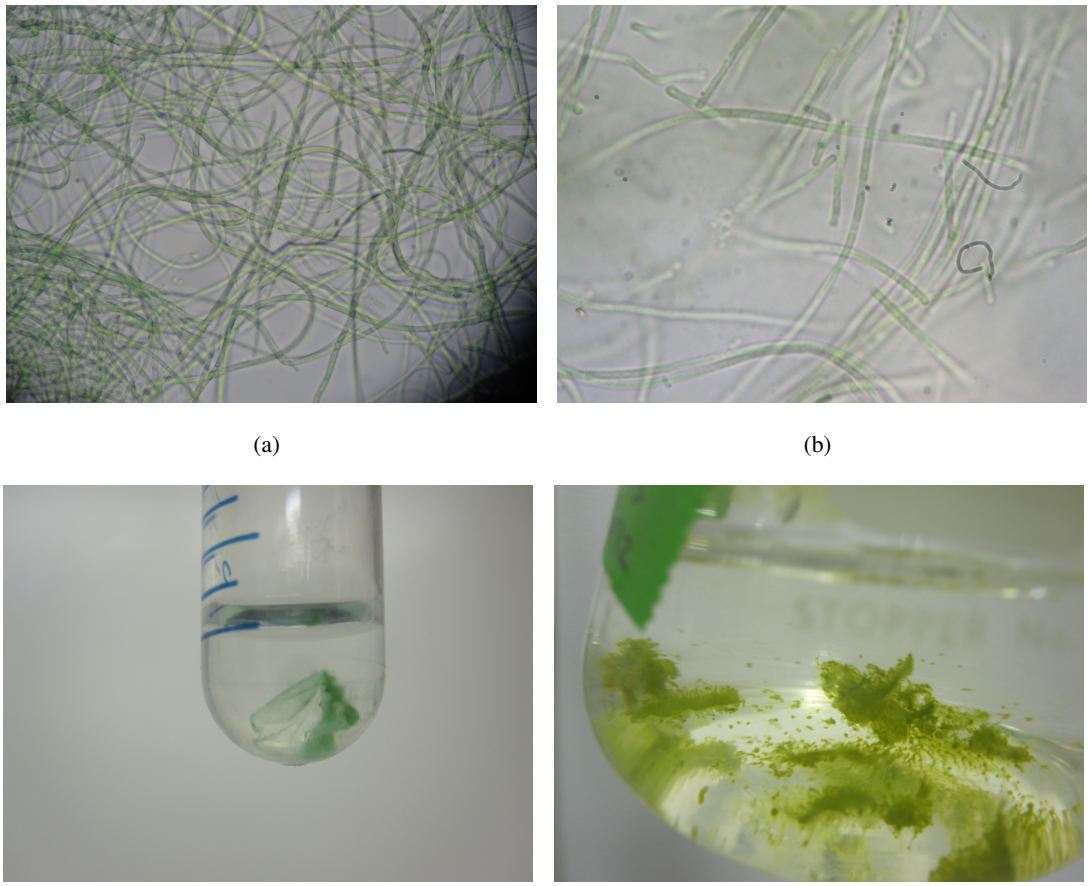


Figure 3.4. Autofluorescent *Gloeobacter* sp. JS1 cells. Dividing cells are brighter at their polar regions. Scale bar is 75 μm long.

3.4.1.2 Cultivation of *Leptolyngbya* sp. JS2

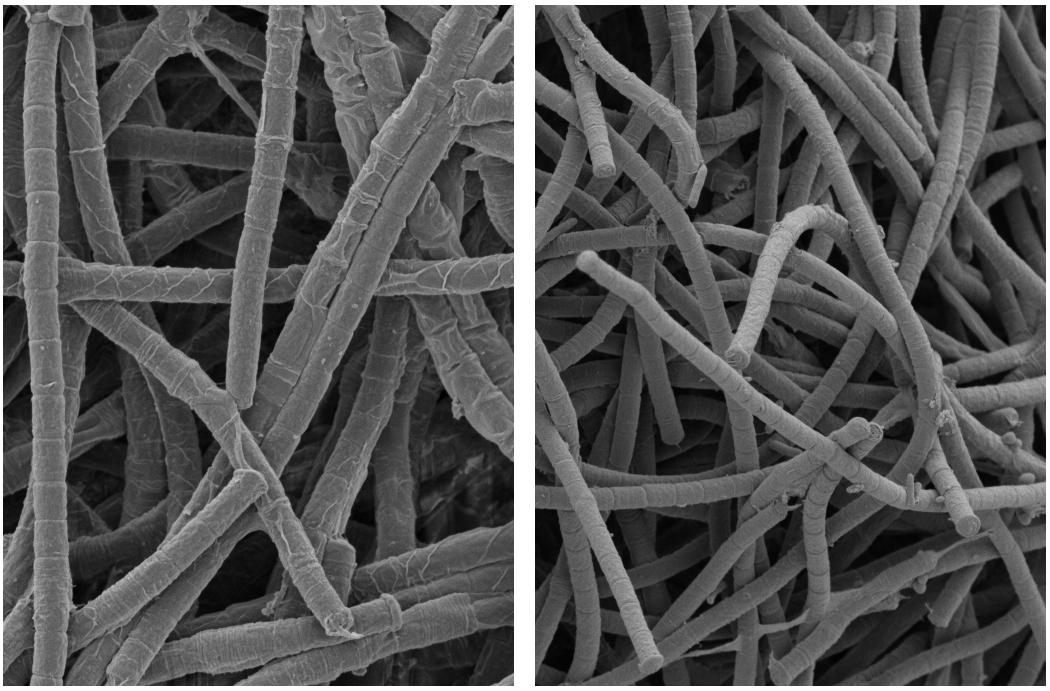
Leptolyngbya sp. filaments formed green biofilms when incubated and shaken in a liquid medium (Figure 3.5(c)). These turned yellowish-green after prolonged shaking and incubation (Figure 3.5(d)), perhaps indicative of an inability to fix nitrogen given this coloration in such cultures is usually associated with nitrogen starvation, a condition known as chlorosis [117, 118]. *Leptolyngbya* cells were filamentous, and appeared by light microscopy to form continuous cells without clear lines of division between cells (Figures 3.5(a) and 3.5(b)). However, distinct cells were visible when viewed by scanning electron microscopy (Figure 3.6).



(a) (b)

(c) (d)

Figure 3.5. Light micrograph and photographs of *Leptolyngbya* sp. JS2. (a) *Leptolyngbya* sp. JS2 filaments. (b) Close-up of *Leptolyngbya* sp. JS2 filaments. (c) *Leptolyngbya* sp. JS2 cells after incubation in a liquid medium, showing biofilm. (d) *Leptolyngbya* sp. JS2 after incubation in a liquid medium, showing biofilm. The yellowish-green coloration (chlorosis) arose after prolonged incubation.



(a) (b)

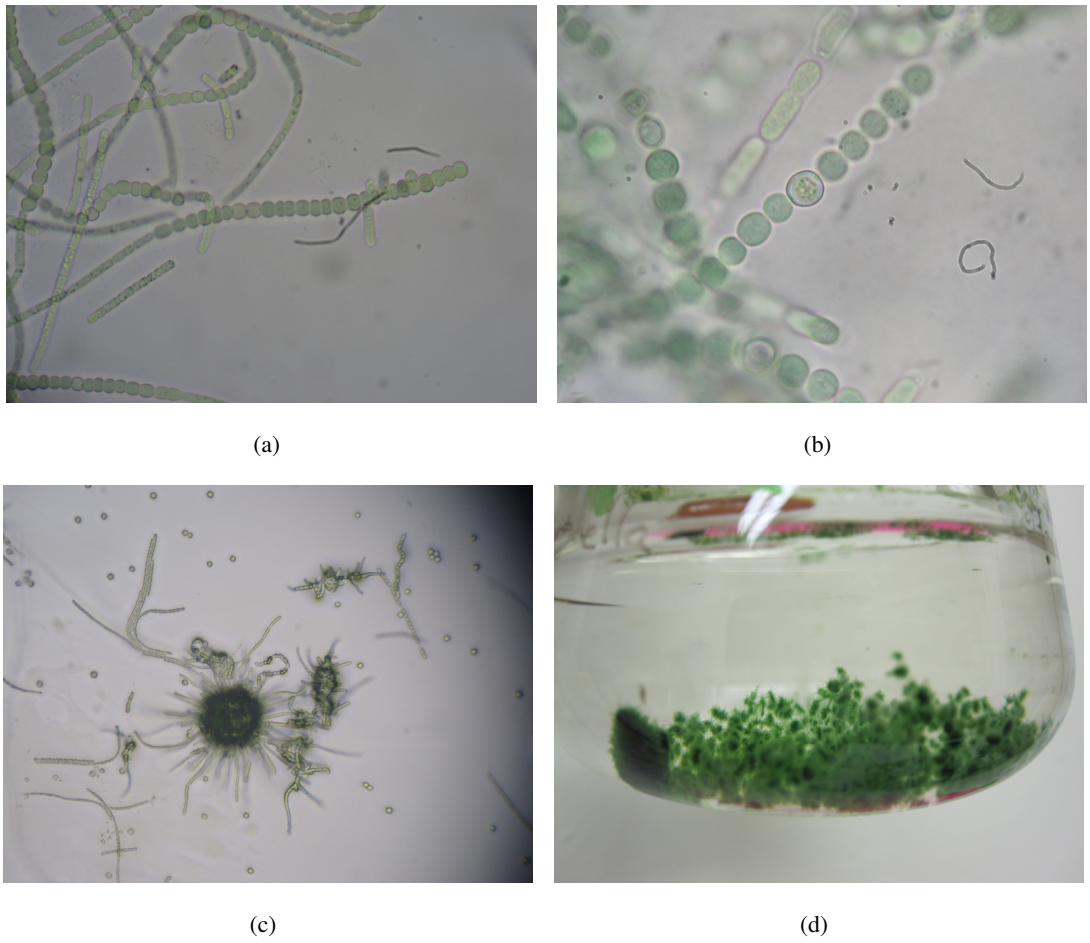
Figure 3.6. Scanning electron micrographs of *Leptolyngbya* sp. JS2 cells. (a) SEM of *Leptolyngbya* sp. JS2 filaments in a biofilm. (b) SEM of *Leptolyngbya* sp. JS2 filaments. Individual cells are visible.

3.4.1.3 Cultivation of a *Fischerella* sp. JS3

Fischerella belongs in the order Stigonematales, a family of true branching cyanobacteria broadly classified into three major categories: T, V, or Y-branching [119]. They are highly differentiated cyanobacteria capable of nitrogen fixation and heterocyst formation [119]. No matches to the 16S rRNA gene sequence of *Fischerella* spp. were detected in the clone library. A BLASTn search against the popset data showed a handful of the top hits shared no more than 93% identity with *Fischerella* spp., specifically those under accession numbers EF032787 (92.5%), EF032783 (91.9%), EF032781 (91.5%), EF032788 (91.4%), and EF032782 (89.8%). It does appear that distant relatives of Stigonematales are in the HAVO biofilm sample (as determined by Lamprinou et al. (2011) [120]), but they are not exact matches as the putative *Fischerella* sp. that was cultivated here.

The *Fischerella* sp. cultivated here was labeled JS3 (Hereafter referred to as *Fischerella* sp. JS3). The strain forms filaments that branch extensively (true-branching) (Figures 3.7 and

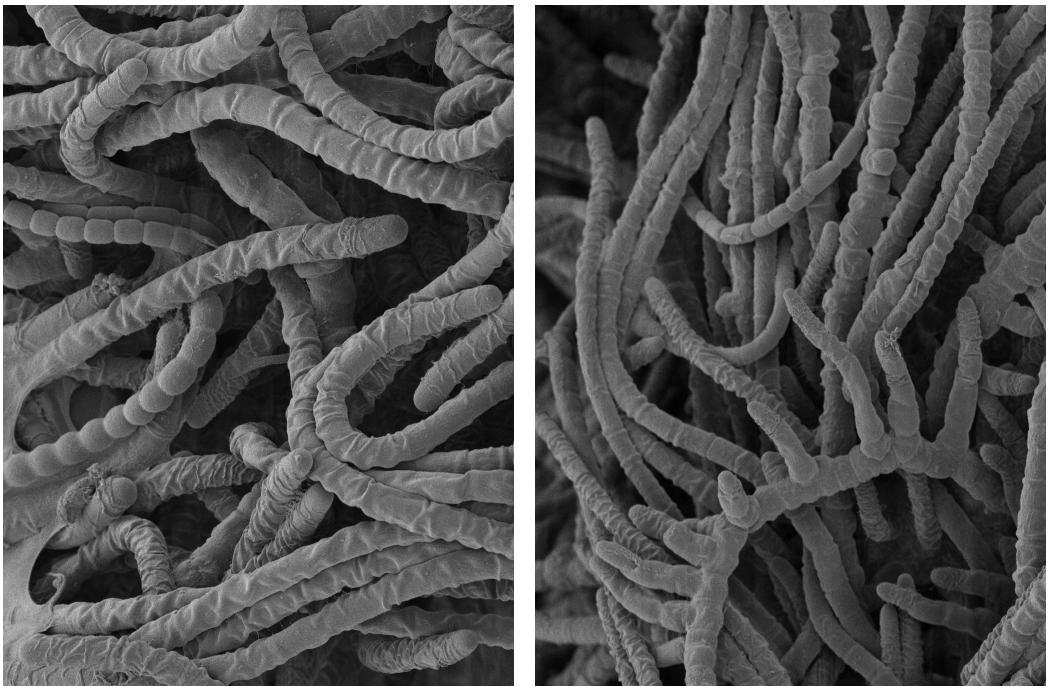
3.8) and hormogonia (an important feature in survival and gliding motility [121]) can be seen near termini of cell filaments (Figure 3.7(a)). The strain also tolerates temperatures above 45°C; in fact, optimal growth in *Fischerella* and *Mastigocladus* has been noted at 45°C [121]. Thermophilic Stigonematales have been isolated from hot springs [119], so the isolation from a cave in Kīlauea caldera of a thermophilic putative Stigonematales supports observations of this cyanobacterium being adapted to geothermal environments.



(a) (b)

(c) (d)

Figure 3.7. Light micrographs and photographs of cultivated *Fischerella* sp. JS3 cells. (a) *Fischerella* filaments showing braching cells that are growing outward perpendicular to the main filament and hormogonia can be seen near termini of cell filaments. (b) Heterocyst near the center of a *Fischerella* sp. JS3 filament. (c) Extensively branched *Fischerella* sp. JS3 filaments on an agar plate, under a dissecting microscope. (d) *Fischerella* sp. JS3 cell clumps in a conical flask.



(a) (b)

Figure 3.8. Scanning electron micrograph of *Fischerella* sp. JS3 cells. (a) Scanning electron micrograph of *Fischerella* sp. JS3 filaments encased in a thick sheath. (b) Scanning electron micrograph of *Fischerella* sp. JS3 filaments, showing true branching patterns.

3.4.2 Pigment analysis

HPLC detected chlorophyll *a* and β-carotene in non-axenic cultures of *Gloeobacter* sp. JS1 (Figure 3.9). The culture contains very few heterotrophic bacteria, bacteria that actually lack these pigments. The method is somewhat limited in that it confirmed the presence of these two pigments in the *Gloeobacter* sp. JS1 culture, but it does not rule out the presence of other pigments. However, while this analysis may not provide a complete profile of the pigments in JS1, the method is widely used to determine water-soluble pigments in bacteria.

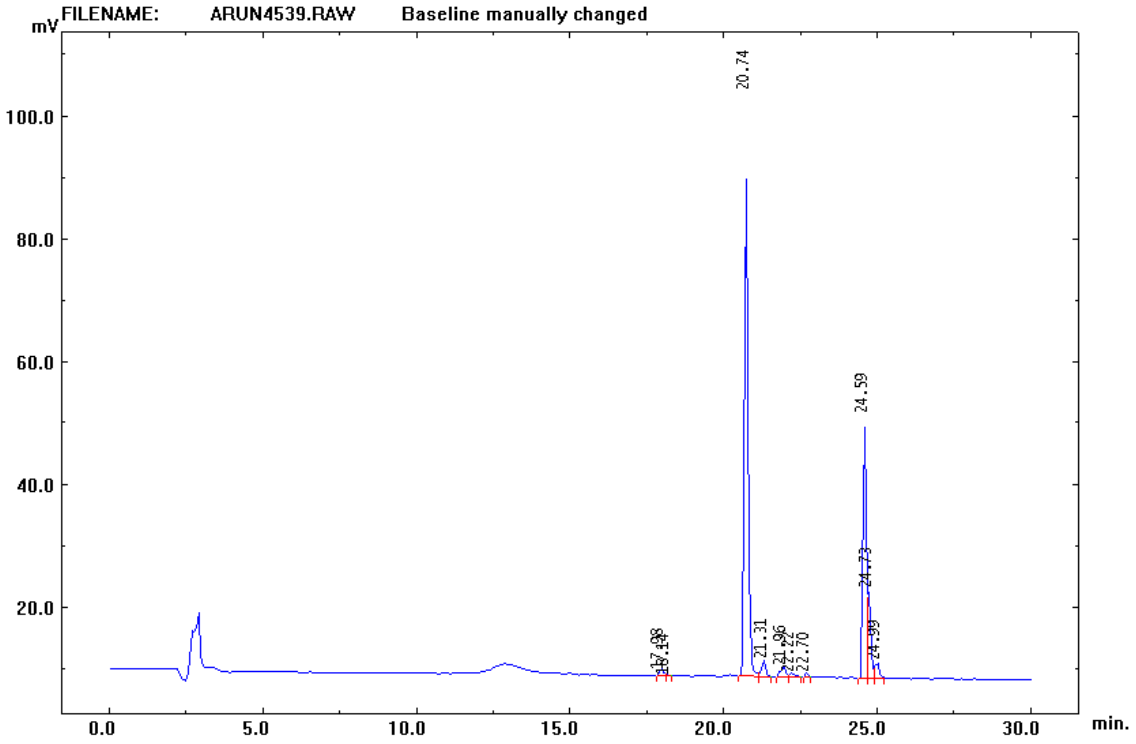


Figure 3.9. HPLC absorbance spectra for pigment analysis. The figure shows retention times characteristic of chlorophyll *a* at 20.74 min and β-carotene at 24.59 min.

3.4.3 Phylogenetic analysis of cultivated cyanobacteria and comparison with cloned 16S rRNA genes

The closest relatives of the cyanobacteria cultivated here were determined by BLAST searches of the amplified 16S rRNA genes from each culture. The complete genome sequence of *Gloeobacter* sp. JS1 is described in Chapter 4 (a proposal to give it a new species name is also described in detail in Chapter 4); the genome has been deposited in GenBank, and will be available once the manuscript describing it is submitted. The 16S rDNA sequences of *Leptolyngbya* and *Fischerella* have been deposited in GenBank under accession numbers JX524204 (*Leptolyngbya* sp. JS2), and JX524205 (*Fischerella* sp. JS3).

The phylogenetic positions of the three potentially novel cyanobacteria were viewed in maximum likelihood trees (Figs. 3.10, 3.11, 3.12). A more detailed maximum likelihood phylogenetic tree of cultivated *Gloeobacter* spp. was also prepared (Fig. 4.22). *Leptolyngbya* sp. JS2 shares very low sequence identity with any known cyanobacteria in publicly available databases;

the closest described species are *Leptolyngbya tenuis* PMC304.07 with which it shares 95.0% nucleotide identity [122], *Pseudanabaena tremula* UTCC 471 (94.9%) [123], and *Leptolyngbya frigida* ANT.L53B.2 (94.2%) [124]. *Leptolyngbya* sp. JS2 was placed deep within the *Leptolyngbya* clade (Figure 3.10).

Fischerella sp. JS2 shares 98.2% 16S rRNA gene sequence nucleotide identity with *Mastigocladus laminosus* Greenland_8 isolate 8 (Accession number: DQ431003.1, as of 06/14/12). The closest Type strain is *Fischerella muscicola* PCC 7414 (Figure 3.11). *Leptolyngbya* sp. JS shares 94.7% 16S rDNA sequence identity with Cf. Leptolyngbya sp. Greenland_10 (Accession number: DQ431005.1, as of 06/14/12) [125].

Relatives of clones in the popset by Brown et al. have been described as belonging in the Stigonematales (Lamprinou et al. [120]), yet they share very low sequence identity with *Fischerella* sp. JS2 cultivated from the biofilm, HAVOmat106 (91.2%) and HAVOmat34 (92.5%). The phylogenetic tree clearly shows that *Fischerella* sp. JS2 clearly is not a match to HAVOmat106 and HAVOmat34 (Figure 3.11). Moreover, the *Fischerella* sp. JS3 16S rDNA sequence was not identified in either the clone library or the metagenomic data, while the other two cultivated cyanobacteria were. Potential explanations for the absence of *Fischerella* sp. JS3 from clone library or metagenomic sequences is that they resist lysis during DNA extraction because of their sheaths [126, 127]. Based on clone sequences from the popset data, it seems other Stigonematales are yet to be cultivated from the HAVO biofilm.

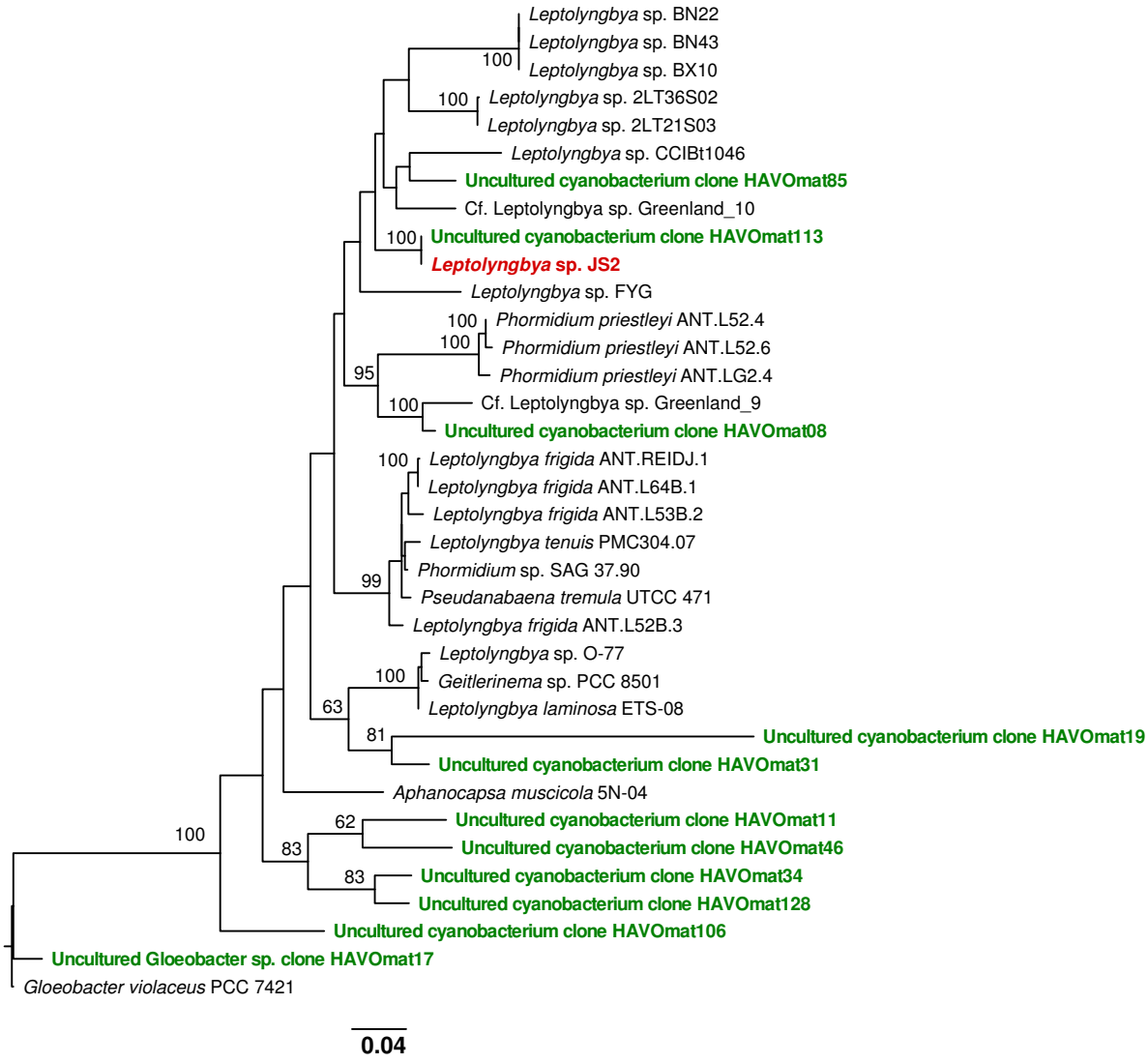


Figure 3.10. Maximum likelihood phylogenetic tree based on 16S rRNA gene sequences of the newly cultivated *Leptolyngbya* sp. and select hits from a BLASTn search. *Leptolyngbya* sp. JS2 is highlighted in red. Only bootstrap values higher than 60 are shown. Note that the 16S rRNA sequence 'Uncultured cyanobacterium clone HAVOmat113' is from a HAVO clone library constructed previously, and matches the cultivated cyanobacterium. Cyanobacteria clones from the clone library are highlighted in green.

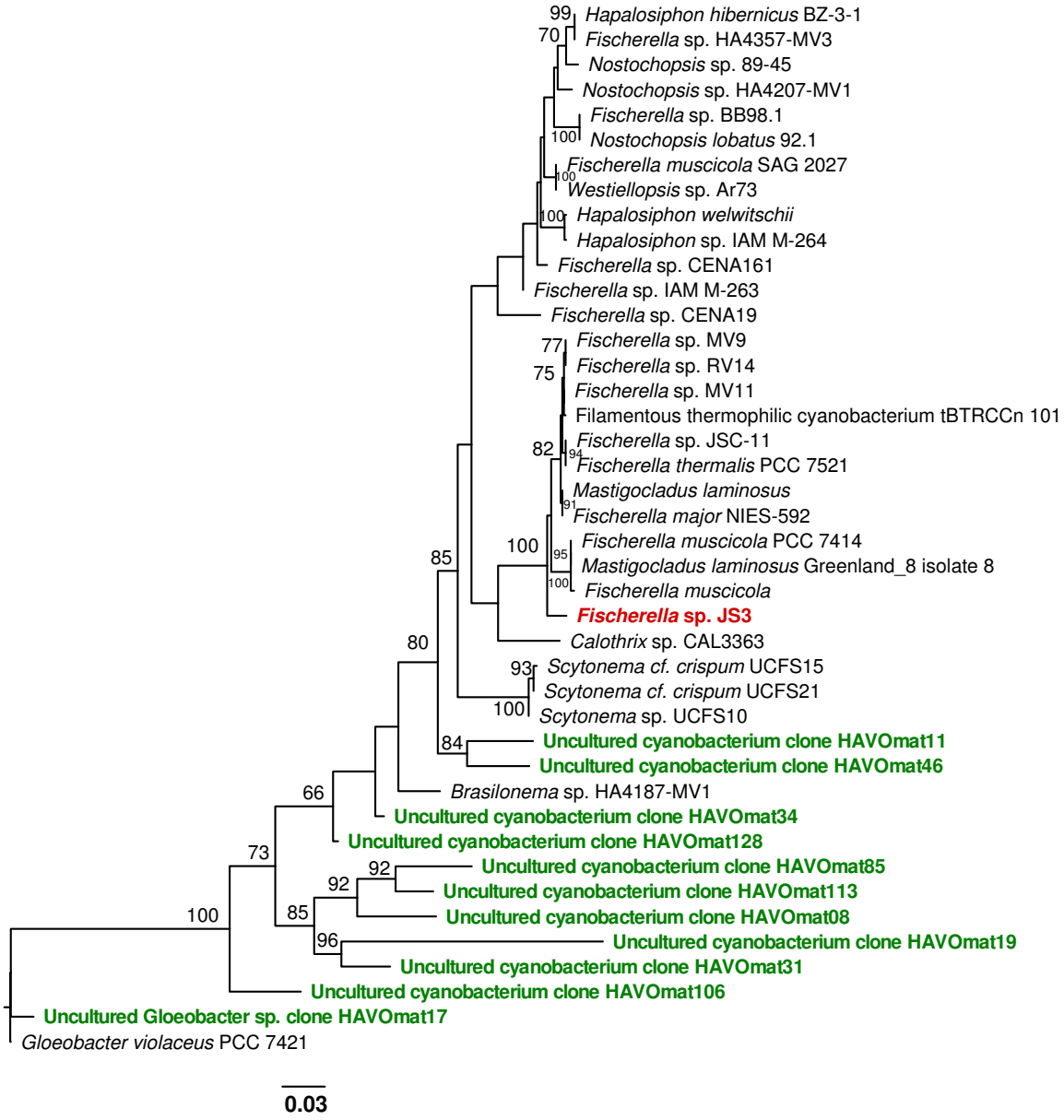


Figure 3.11. Maximum likelihood phylogenetic tree based on 16S rDNA sequences of the newly cultivated *Fischerella* sp. JS3 and selected neighbors from a BLASTn search. The position of *Fischerella* sp. JS3 is shown in red. Only bootstrap values higher than 60 are shown. Cyanobacteria clones from the clone library are highlighted in green; note that this culture has no matching clone in the clone library.

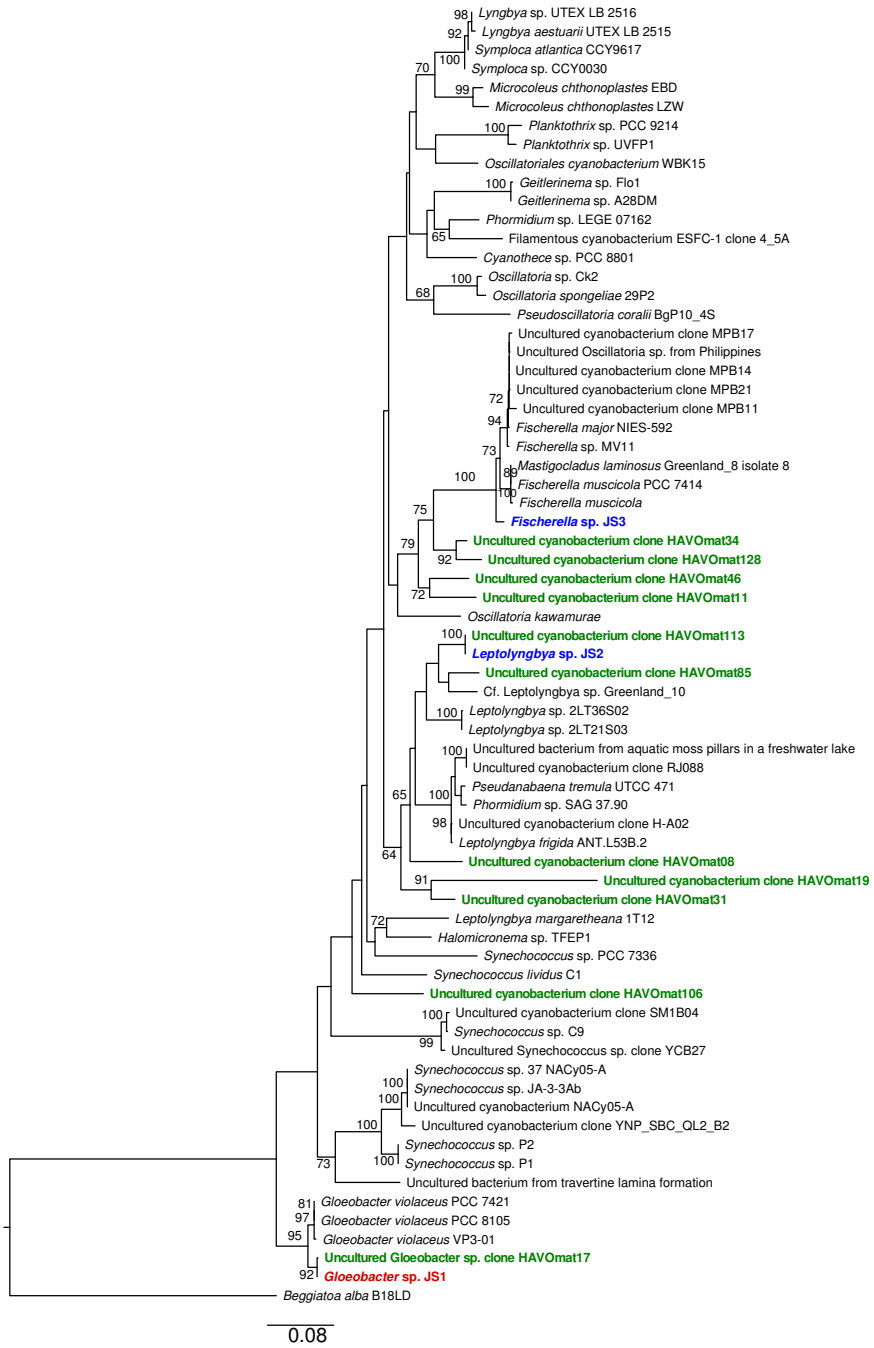


Figure 3.12. Maximum likelihood phylogenetic tree based on 16S rDNA sequences of the three cultivated cyanobacteria and their nearest neighbors from BLAST searches. *Gloeobacter* sp. JS11 is highlighted in red, and the two other cultivated cyanobacteria are in blue. Sequences from the clone library are highlighted in green. 'Uncultured Gloeobacter sp. clone HAVOmat17' is a matching clone from a previous clone library.

3.5 Conclusions

A cyanobacterium with a long evolutionary history, that belonging to the genus *Gloeobacter*, was cultivated. Only one *Gloeobacter violaceus* species, strain PCC 7421, has been cultivated and deposited in an international culture collection. The work described here cultivated only the second known species in the genus. Two previously undescribed filamentous cyanobacteria were also brought into culture. One was initially detected in a clone library, and now represents a likely new species (or genus) with closest formally described relatives in the *Leptolyngbya*. The other filamentous strain affiliates with the order Stigonematales, and is likely a relative of *Fischerella* or *Mastigocladus*.

The *Leptolyngbya* sp. JS2 forms biofilms and may *in situ* be a contributor to the HAVO biofilm's formation or structure. This particular taxonomic assignment is rather tentative because the percentage of nucleotide identity with the nearest *Leptolyngbya* suggests the strain from the HAVO mat may actually constitute a new genus. *Leptolyngbya* spp. are of potential value in biotechnological applications because they have higher lipid and monosaturated fatty acid contents than the *Arthrospira* species which are often used in the industry [128].

The *Fisherella* sp. JS3 cultivated from the HAVO biofilm is a true-branching cyanobacterium, which shares 98.0% 16S rRNA gene sequence identity with a cyanobacterium (*Fischerella* JSC11) whose genome is currently being sequenced. It is certainly feasible that the HAVO *Fischerella* sp. is a novel species, but it is currently referred to only as *Fisherella* sp. JS3. It is worth noting that this *Fischerella* was only cultivated from the HAVO biofilm, and was not detected in the clone library prepared previously from the same mat. This suuports the contention that cultivation approaches should not be abandoned, but should rather be practiced together with molecular approaches in studies of microbial diversity [17].

The lava cave in which the HAVO epilithic biofilm is located is less than 100 years old. The microbial community on the rock surface in the cave entrance likely entered the cave from the surrounding volcanic soil, or from the rhizosphere of nearby plants, especially those that penetrate the cave ceiling. Others must surely have been transported by the wind. *Gloeobacter* is an early-branching cyanobacteria that is rarely cultivated, but which is reported in clone libraries from time to time. Quite how a novel *Gloeobacter* species came to form such a visually conspicuous part of an epilithic biofilm in a cave in a relatively young lava flow, and in an active volcano in the middle of the Pacific Ocean, will surely attract questions, given it is it is so far from the rock wall in Switzerland from which the only other known Type strain was isolated almost forty years ago.

Chapter 4

Complete genome sequence of *Candidatus* Gloeobacter kilaueaensis from Kīlauea Caldera

Jimmy Saw, Michael Schatz, Mark Brown, Jamie Foster, Shaobin Hou, Dennis Kunkel, Maqsudul Alam, and Stuart Donachie. *In preparation. To be submitted to the PNAS journal.*

4.1 Abstract

Gloeobacter belongs to an ancient lineage of early diverging cyanobacteria usually associated with rock surfaces. Divergence of *Gloeobacter* from its sister cyanobacteria occured before that of the plant plastids and other cyanobacteria. Due to the deep divergence of *Gloeobacter* within cyanobacterial lineages and its lack of thylakoid membranes, *Gloeobacter* is an interesting organism in which to study the evolution of cyanobacteria, particularly as it retains many ancestral features of early oxygenic phototrophs. Only a handful of *Gloeobacter* have been detected in 16S ribosomal gene clone libraries, and only one Type strain exists in the enire order. The complete genome of this Type strain has been sequenced. Now, however, a second *Gloeobacter* sp., termed JS1, has been cultivated, and its complete genome sequenced. *Gloeobacter* sp. JS1^T was isolated from an epilithic biofilm found in a lava cave entrance in volcanically active Kīlauea caldera, Hawai'i. Due to difficulties in obtaining an axenic culture, the genome was sequenced from an enriched culture resembling a low-complexity metagenomic sample. Combined 9 kb paired-end 454 pyrosequences and Illumina short reads enabled assembly of the complete genome. Comparison of the assembled genome with that of the closely related *Gloeobacter violaceus* PCC 7421 confirmed PCC7421

and JS1 are not the same species. Very little gene synteny exists between these two *Gloeobacter* genomes, despite their sharing 2842 orthologous genes. Based on differences in the genome and calculated distance, '*Candidatus* Gloeobacter kilaueaensis' is proposed to accommodate strain JS1. The complete genome sequence of '*Candidatus* Gloeobacter kilaueaensis' should lead to a better understanding of cyanobacteria evolution, and the transition from anoxygenic to oxygenic photosynthesis.

4.2 Introduction

The *Cyanobacteria* phylum hosts some of the most diverse microbes to have evolved on Earth. Pioneers of oxygenic photosynthesis, their production of free oxygen permanently changed the gas composition of Earth's atmosphere, paving the way for the evolution of aerobic respiration [129, 130, 7]. *Gloeobacter* is known as an early branching cyanobacterium that diverged before the emergence of plant plastids from other cyanobacteria [131, 132, 133]. It is thus believed to be one of the earliest cyanobacteria capable of oxygenic photosynthesis, that is, an intermediary organism because of its primordial characteristics [134]. Only one species in the Class *Gloeobacter* has been described [107]. Here, the isolation and complete genome sequence of the second member of the *Gloeobacter* is described, and compared with that of the Type strain of the Class, genus and species, *Gloeobacter violaceus* PCC 7421. *Gloeobacter* is unique among cyanobacteria due to its lack of the thylakoid membranes found in all other cyanobacteria [107]. Thylakoid membranes are crucial in other cyanobacteria and plant plastids, as the light-dependent reaction centers of photosynthesis. The lack of thylakoid membranes in *Gloeobacter* has led to speculation it may be one of the earliest ancestors in the cyanobacteria lineage [134, 135].

Photosynthesis was thought to have evolved once on Earth [136, 137, 138, 139, 140, 141]. The origin of photosynthesis began in the domain *Bacteria* and the ability of eukaryotes to photosynthesize resulted from symbiosis events [142, 143]. The evolution of anoxygenic photosynthesis first took place in anaerobic phototrophs when Earth's atmosphere was strongly reducing [144, 145]. Oxygenic photosynthesis was estimated to have taken place nearly 2.8 billion years ago [146] and the rise of molecular oxygen in Earth's atmosphere correlates with the rise of oxygenic photoautotrophs, such as modern-day cyanobacteria and plants [147]. Thus, it would be of enormous significance if additional links between anoxygenic and oxygenic photorophs are found, and their evolutionary paths mapped. Such discoveries can only advance our understanding of the evolution of photosynthesis.

Only three strains of *Gloeobacter* have been cultivated, and the genome of only one of these has been completely sequenced, specifically *Gloeobacter violaceus* PCC 7421, the Type strain of the species, genus and Class. *Gloeobacter violaceus* PCC 7421 was isolated from the surface of a limestone rock in Oberwald, Switzerland in 1974 [107] and its complete genome was sequenced in 2003 [108]. Two other strains exist (PCC 8105 [132] and VP3-01 [148]) but these have never been considered as different species and generally categorized as strains of *Gloeobacter violaceus*.

During this research, a new species of *Gloeobacter* was cultivated from an epilithic biofilm on the wall of a cave entrance in Kīlauea caldera on Hawai‘i. Initial pyrotag and metagenomic sequencing of the community revealed high diversity, and a community hosting phyla from both *Bacteria* and *Archaea* (See Chapter 2). Numerous taxa with no cultivated representatives were detected in the community, including a potentially novel *Gloeobacter*. Using a modified growth medium and low-light conditions, the putative *Gloeobacter* sp. JS1 was brought into culture. However, an axenic culture was difficult to obtain due to the presence of heterotrophic bacteria that tended to outgrow the *Gloeobacter*. Sequencing the entire *Gloeobacter* sp. genome from this mixed culture was deemed feasible, since the sequence pool resembled a low-complexity metagenome, where most of the sequences would come from the most abundant organism, and few from contaminating organisms; *de novo* assembly of the sequences enabled construction of the complete genome.

To obtain insights into the divergence and evolution of *Gloeobacter* from other cyanobacteria, the genome of this newly isolated *Gloeobacter* was compared with that of *Gloeobacter violaceus* PCC 7421T. Comparison of gene and sequence conservation, synteny, and genome-to-genome distances calculated between the two organisms, confirmed JS1 belongs to a different species, for which the name *Candidatus* Gloeobacter kilaueaensis was proposed. The complete genome sequence of *Candidatus* Gloeobacter kilaueaensis JS1^T from the deeply divergent *Gloeobacter* clade is described here.

4.3 Materials and Methods

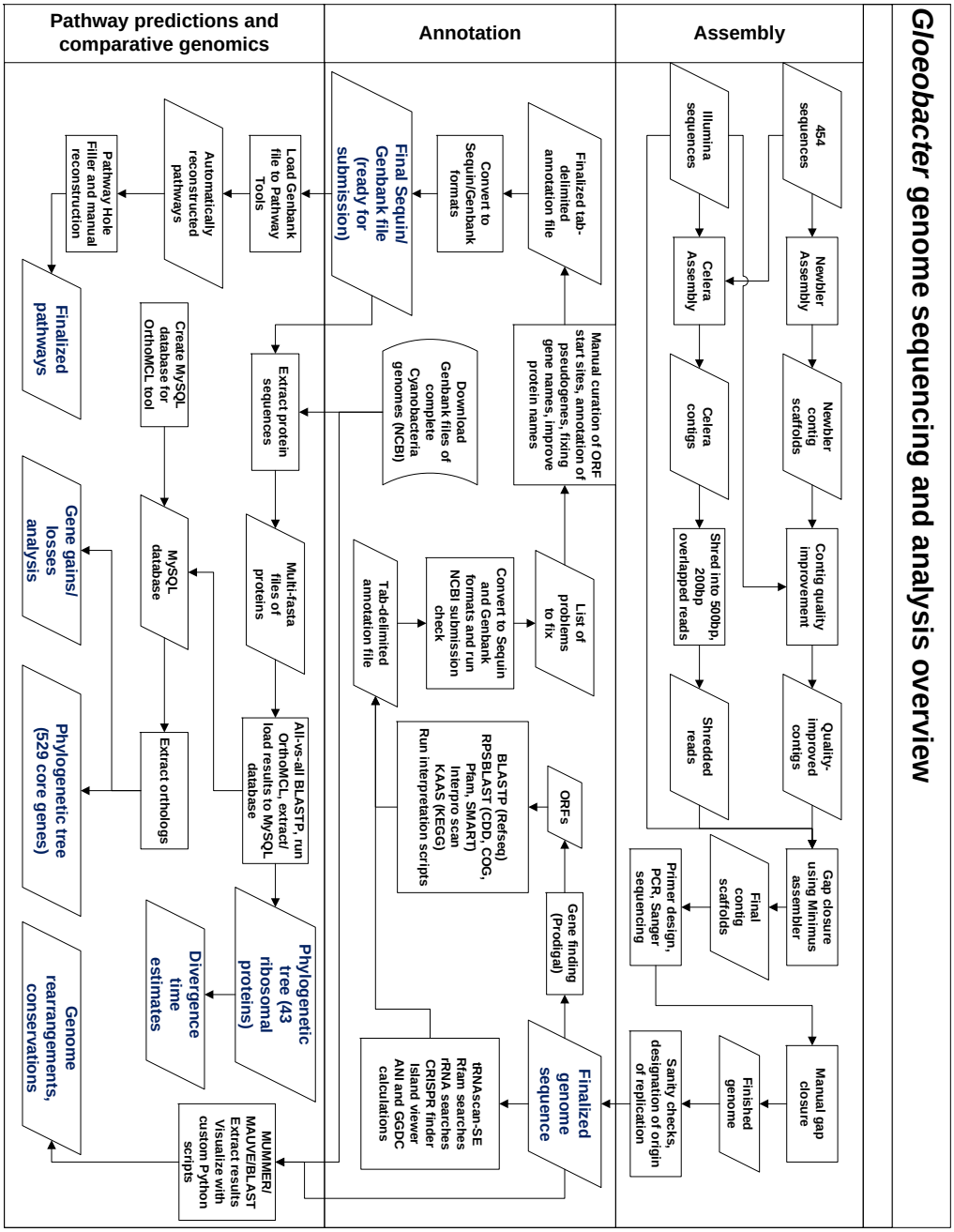


Figure 4.1. Flowchart of sequencing and analysis of the *Candidatus* Gloeobacter kilaueaensis JS1^T genome. Flowchart comprises three sections: Assembly, annotation, and comparative genomics. A number of custom scripts were written for some steps shown in the flowchart.

The sequencing, assembly, annotation, and comparative genomic analyses of *Candidatus* G. kilaueaensis JS1 genome is summarized in terms of the flow of information across different stages (Figure 4.1).

4.3.1 Sampling and cultivation

Punch cores of 5 mm diamter were aseptically removed from an epilithic biofilm on a lava cave wall in Kīlauea Caldera. Within eight hours of collection the cores were dissected with a sterile scalpel and transferred to a modified liquid BG11 medium with reduced phosphate, wrapped in a white paper towel to mimic the low intensity of light in the cave entrance, and shaken at 200 rpm in a light incubator with 2% CO_2. The modified BG11 medium (BG11M) contained, in grams per liter: $NaNO_3$ (1.5 gl^{-1}), $CaCl_2 2H_2O$ (0.036 gl^{-1}), $FeNH_4$ Citrate (0.012 gl^{-1}), Na_2EDTA (0.001 gl^{-1}), K_2HPO_4 (0.02 gl^{-1}), $MgSO_4 7H_2O$ (0.075gl^{-1}), Na_2CO_3 (0.02 gl^{-1}), and a solution of micronutrients (Appendix B.2). After shaking at 29°C for about two weeks, a mass of purple flocs were visible near the bottom of the culture tube. Sub-samples (10 μl) of the cell suspension were spread on a solid BG11 medium. After incubation for one week, a dense purple 'slime' appeared on the medium's surface. Using a Pasteur pipette drawn to a fine point, cells from the purple biofilm were 'spotted' to another BG11M plate. Purple, non-axenic colonies arose after two weeks of incubation. One such colony was transferred to liquid BG11M and shaken for two weeks, after which 10μl of purple floc was spread on solid BG11M and incubated for another two weeks. This cycle was repeated until cultures appeared uniform by light microscopy, although non-*Gloeobacter* cells remained in very low numbers. Cells were prepared for scanning electron microscopy (Section 3.3.3).

4.3.2 Genomic DNA extraction and quality control

An axenic *Gloeobacter* culture was not available for complete sequencing, so genomic DNA was extracted from a culture determined visually to be predominantly of *Gloeobacter* cells. DNA was extracted from $\sim$1 g wet weight of cells using the MO BIO Ultraclean® Soil DNA isolation kit. Bacterial primers (27F and 1492R) were used in PCR amplification of a fragment of the 16S rRNA gene in this DNA, in PCRs containing 5 μl of 10X *Pfu* buffer, 1 μl of 10 μM dNTP mixture, 5 μl of *Pfu* DNA polymerase, 1 μl of 10 mM primer, 1 μl of DNA template, and nuclease-free water for a total of 50 μl. The conditions for PCR were 95°C (5 min), followed by 35 cycles of 95°C (30 sec), 52°C (30 sec), 72°C (30 sec), and a final extension of 72°C (7 min). PCR

products were cleaned with the MO BIO UltraClean® PCR Clean-Up Kit. Purified PCR products were cloned into pCR®-Blunt II-TOPO vector (Life Technologies, Carlsbad, CA) and transformed into chemically competent One Shot® TOP10 *E. coli* cells. Transformed cells were plated on LB + Kanamycin agar plates, isolated and grown in Circle Grow® (Q-BIOgene, Carlsbad, CA). Cloned inserts were amplified using M13F and M13R primers and sequenced using 27F primer. A total of 42 clones were sequenced to assess the level of contaminant DNA from heterotrophs.

4.3.3 Sequencing, genome assembly, and finishing

Only two of the 42 clones sequenced were not *Gloeobacter* (see results section 4.4.1). Upon determining that the level of contamination by heterotrophic bacteria in the *Candidatus* G. kilaueaensis JS1 culture would not preclude aseembly of the *Gloeobacter* sp. JS1 genome, the genomic DNA extracted above was used to prepare an 8-9 kb paired-end 454 library in the University of Hawaii's 'Advanced Studies in Genomics, Proteomics, and Bioinformatics Center' (ASGPB), according to the Roche protocol, and sequenced in a 454 GS-FLX Titanium DNA sequencer (454 Life Sciences, Branford, CT). A total of 222,335 pyrosequences were generated, representing 155,068 paired-end sequences and 66,513 singletons, for a total of 221,581 usable sequences. The remainder were discarded because of poor sequence quality. A total of 4,792,504 Illumina sequences (2,396,252 paired-end sequences) were generated in an Illumina Genome Analyzer *IIx* sequencer (Illumina Inc, San Diego, CA). After trimming for quality, 4,756,989 original sequences remained for assembly or for read recruitment.

Assembly and finishing followed the procedure shown (Figure 4.1). Custom utility scripts written for certain steps along the pipeline are listed (Chapter 5). Raw sequences produced by the Roche 454 GS FLX sequencer were first assembled using Newbler version 2.6. The MUMMER sequence alignment tool [149] was used to recruit sequences produced by the Illumina Genome Analyzer *IIx* to the assembled Newbler contig scaffolds. Each Newbler contig scaffold was then assembled with quality-trimmed Illumina reads using the Minimus assembler (AMOS) package. Often, coverage of Illumina reads was found to be more than required for quality improvement, so a Python script was written to recruit reads to only $\sim$15x coverage. This procedure improved and corrected the sequence quality of the Newbler assembled contigs that initially contained only 454 pyrosequences. Illumina sequences also helped correct ambiguous sequence regions caused by homopolymers usually present in 454 reads.

Pyrosequences and Illumina sequences were also assembled together using the Celera Assembler to compare Celera contigs with Newbler contigs. Celera contigs were shredded into

500bp fragments with 200bp overlapping regions and used in Minimus assemblies to help close gaps. Final gaps between quality-improved and Minimus-assembled contigs were then manually closed using the Seqman program (DNAstar Inc, Madison, WI). To close gaps, Illumina reads were used first, but where gap persisted, specific primers helped amplify the gap regions with products then sequenced by capillary sequencing (ABI3730xl, Life Technologies, Carlsbad, CA). The error rate of the final assembled genome is less than 1 nt in 100,000. Illumina and 454 sequences provided roughly 93x coverage of the genome, *i.e.*, 440,800,613 bases.

4.3.4 Verification of genome assembly

Trimmed 454 pyrosequences were taxonomically assigned using the PhymmBL binning tool [70]. Mate pairs with at least one read belonging to the genus *Gloeobacter* were aligned against the assembled genome with the MUMMER alignment tool, and overlapping paired-ended reads binned as *Gloeobacter* were graphically represented in tiling paths using a custom Python script (see section 5.1.17). The algorithm to select mate pairs was such that mate pairs spanning a given segment of the genome were searched for *Gloeobacter*-binned reads, and where found, only those pairs fitting the expected insert size range (5000 - 12000 bp) were reported (see section 5.1.20).

Contig scaffolds produced by the Newbler and Celera assemblers were also aligned against the assembled genome to determine if the genome may have been misassembled. PCR amplification of suspicious boundaries were performed in regions where G+C content significantly varies from the rest of the genome, *i.e.*, less than the mean of 60.5%, and where coverage of reads binned as *Gloeobacter* fell, but reads from other organisms dominated. Primers were designed by a custom Python script using the Primer3 program [150] to pick primers meeting the criteria needed for long-range PCR (Table 4.1). Primers were designed to amplify $\sim$15 kb fragments. The Qiagen® LongRange PCR kit was used to amplify suspicious genome segments from the genomic DNA isolated from the *Gloeobacter* culture described above.

Table 4.1. Primers to check questionable regions

Primer name	Primer sequence
SR4-F	5'-GTCTTGCCCTTGCTGATGATCAAG-3'
SR4-R	5'-ATAGTCGCGGGTATCTTGCAGATC-3'
SR5-F	5'-TCCTGGCTTGAGTACCTGATCAAC-3'
SR5-R	5'-CCTGCTTTGATAGAGCCTCACTCA-3'
SR6-F	5'-CGATTACCCGAGCCAGAAATTTCG-3'
SR6-R	5'-GGCAGATGGTAGAGCTTGATCACA-3'
SR10-F	5'-CAAGGGGCAGTGACTTTCTTTGAC-3'
SR10-R	5'-GTTGCTCACCAACCAGCTTTAGAG-3'
IR1-F	5'-GCAACTGTCGCCACCTGATTTATG-3'
IR1-R	5'-AGGTAGATAGCAGCCGACGATTTC-3'
IR2-F	5'-GCACCAGACTCGACCTTCTATTTC-3'
IR2-R	5'-CTCGCTTCGATGTATCTGGGAACT-3'
IR3-F	5'-CGTCGCCGGTAGTTTTCATACTCT-3'
IR3-R	5'-TGGTTGGCTCATCCCAATCTACTG-3'
CR1-F	5'-ATCAGCGATCTTACCGAGCAGATC-3'
CR1-R	5'-TTAAAGAGCGTCTCGGAGGTAAGG-3'

SR denotes 'suspicious/questionable region', IR is 'important region', and CR is 'control region'.

4.3.5 Genome annotation

The genome annotation procedure followed a defined protocol (Figure 4.1). Putative coding regions in the genome were identified using the Prodigal gene finder program [151], and submitted to the NCBI submission check tool to curate ORF start sites, and to identify frameshifts and gene fragments. ORFs with partially conserved domains were inspected individually to determine if the products are functionally inactive, and assigned as pseudogenes where necessary. ORFs were searched against the NCBI Refseq database using BLASTp [82], and top hits were checked against the Protein Clusters database from NCBI to assign names to ORFs. Intergenic regions were extracted and searched against the Refseq database using BLASTx to identify potential coding regions missed by gene finders, and manually assigned. RPS-BLAST searches were performed against Conserved Domain Database (CDD) to identify protein domains, and the resulting XML

output files were parsed in a custom Python script (see section 5.1.10) to check protein domain arrangement and counts.

4.3.6 Phylogenetic analyses

A phylogenetic tree was constructed using 16S rRNA gene sequences of the top 35 hits of the *Candidatus* G. kilaueaensis JS1 16S rRNA gene sequence, and others in Cordeau et al. [43], aligned in Muscle [114] and edited with Gblocks [115]; a maximum likelihood tree was built using the RAxML program [116]. For the ribosomal protein tree, all ribosomal proteins identified in *Gloeobacter* sp. JS1 1 were searched against the 40 cyanobacterial and *Beggiatoa* sp. PS genomes. A total of 43 ribosomal proteins were found to occur in all these genomes, and each was aligned in Muscle, edited with Gblocks, and concatenated (Appendix 5.1.25). The maximum likelihood phylogenetic tree was inferred from 5357 aligned and concatenated amino acid characters using RAxML and the Γ+WAG model of amino acid substitution, and 100 bootstrap replicates. The divergence time between the cyanobacteria was calculated on the basis of 43 concatenated ribosomal proteins from 41 cyanobacterial genomes and *Beggiatoa* sp. PS, aligned, edited, and analyzed in MCMC using the CODEML and MCMCTREE programs in PAML [152]; the tree was visualized in FigTree. Gene gains and losses along the cyanobacteria lineage were calculated, and phyletic patterns constructed on the basis of 13,655 orthologous groups identified among the 41 cyanobacteria, with events calculated through the GLOOME web server with default parameters [153].

4.3.7 Metagenome recruitment

Metagenome reads were recruited using MUMMER, with parameters set as "-minmatch 10", and BLASTn with parameters set as "-F mL -U T -e 1e-4 -r 8 -q -8 -z 95386202 -X 150 -v 1000000 -b 1000000 -m 8 -a 20"

4.3.8 Resolving the *Gloeobacter* lineage by genome-to-genome distance and average nucleotide identities.

To determine if *Candidatus* G. kilaueaensis JS1 from the HAVO epilithic biofilm qualifies as a new species, an *in silico* DNA-DNA Hybridization (DDH) using Genome-To-Genome sequence comparison [154] and Average Nucleotide Identities (ANI) [155] was conducted. For genome-to-genome distances, the complete genome sequence was uploaded with the reference *G. violaceus* genome sequence to the GGDC website (`http://ggdc.gbdp.org/`). The ANI between *Can-*

didatus G. kilaueaensis JS1 and *G. violaceus* PCC 7421 were calculated using the Jspecies program [156].

4.3.9 Comparative genomics analyses

Complete sequenced genomes of 40 *Cyanobacteria* (as of March 3, 2012) used in comparative genomic analyses were downloaded from the NCBI website (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Several cyanobacteria compared had multiple amplicons, and these were pooled into a single data set for each genome. Local BLAST databases of amino acid sequences for each genome were created by the 'formatdb' command (BLAST package) and all-*vs*-all BLASTp searches were used to create all possible combinations of relationships between all the amino acid sequences. BLAST results were loaded to a custom MySQL database, and orthologous groups in 41 cyanobacteria genomes were identified using scripts provided in the OrthoMCL program [157, 158]. To compare metabolic pathways in *Candidatus* G. kilaueaensis JS1 and *G. violaceus* PCC 7421, amino acid sequences were annotated using the KEGG automatic annotation server (KAAS) to assign KO numbers, and submitted to the iPath2.0 server to create metabolic pathway maps. Whole genome comparison plots were generated by custom Python scripts that parses MUMMER alignment output files to draw custom plots (see sections 5.1.12, 5.1.13, and 5.1.22).

4.4 Results and Discussions

4.4.1 Sampling, cultivation, and sequencing

Previously, a 16S rRNA gene sequence sharing 98.6% nucleotide identity with that of *G. violaceus* PCC 7421 was detected (accession number EF032784) in a 16S rRNA gene clone library prepared from community DNA extracted from a purple-pigmented epilithic biofilm on the wall of a lava cave in Kilauea caldera. Samples collected in October 2009 provided material for attempts to cultivate this *Gloeobacter*. During collection, steam rose from the cave entrance, and the temperature several meters into the cave ranged from 35 to 40°C. The cave floor was hot to the touch, but close to the entrance the air temperature was 30-35°C, and condensation flowed steadily over and dripped from the purple biofilm on the wall and ceiling.

Several 'plugs' of ∼5 mm diameter were taken directly from purple sections of the epilithic biofilm into 2 mL cryovials and returned at ambient temperature to the laboratory at the University of Hawai'i at Manoa. One sample was transferred to and shaken in a modified BG11 liquid medium,

with reduced phosphate (0.02 instead of 0.04 gl^{-1} K_2HPO_4) and incubated at 28°C under 500±20 lux (∼6.5 μE m^{-2} s^{-1}) light in a continuous light cycle. Subsamples were also streaked or diluted prior to plating on BG11 and incubated under the same conditions. Purple colonies arose after a month on agar plates, while purple clumps were visible in liquid BG11 at the same time. Repeated streaking for isolation also maintained the cultures, but axenic cultures were not attained because heterotrophic bateria tended to outgrow the purple, presumed *Gloeobacter* sp. cells.

Candidatus G. kilaueaensis JS1 cells form raised purple colonies that tend to become elevated (Figure 3.2) and large after repeated transfers, but they lose color on agar when not transferred for several weeks. *Candidatus* G. kilaueaensis JS1 cells are ovoid and autofluoresce (Figure 3.4). They form copious amounts of a mucilaginous material, and the cells often appear surrounded by such material (Figure 4.2). Cells are approximately 1 × 1.5 μm in size. Prior to constructing paired-end 454 and Illumina libraries, steps were taken to make sure that the DNA used for sequencing contained as little DNA from other organisms as was practically possible. Cells were initially observed by light microscopy (wet mount, Gram stain) to gauge semi-quantitatively the diversity and abundance of cell types in the culture. Cells observed by light and fluorescence microscopy predominantly matched the the characteritics (∼ 3.5 x 1.5 μm) of *Gloeobacter*, and very few other cells were observed (Figure 3.4). A clone library containing 16S rRNA gene amplicons from the genomic DNA extracted from the ∼1 g wet wt, of cells contained 40 inserts that affiliated with 'Uncultured Gloeobacter sp. clone HAVOmat17' (from the popset clones), and 1 sequence each from a (*Bradyrhizobium* and a *Burkholderia*). Laboratory cultures of *Candidatus* G. kilaueaensis JS1 are non-axenic, but they have been deposited in the Pasteur Culture Collection (PCC) and the PCC number is pending attainment of axenic cultures at PCC.

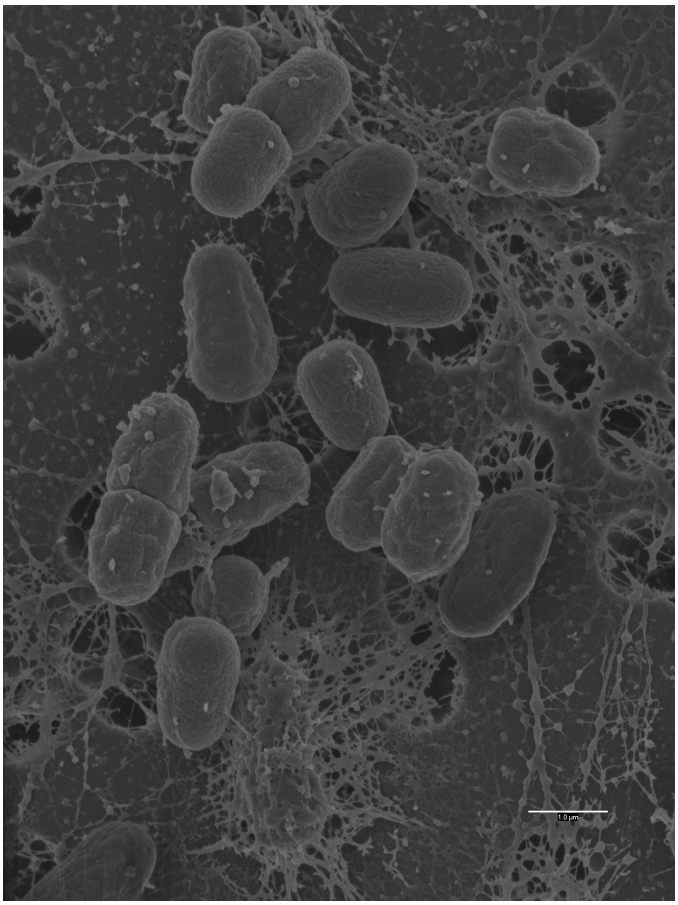


Figure 4.2. Scanning electron micrograph of *Candidatus* G. kilaueaensis JS1. Scale bar is 1μm. SEM of *Candidatus* G. kilaueaensis JS1 in modified BG11 liquid medium. Dividing cells are evident near the top and left of the field. Cells surrounded by mucilaginous material are near the bottom of the field. Scale bar is 1μm.

4.4.2 Genome assembly and verification

Candidatus G. kilaueaensis JS1 genome sequence fragments were generated from a single pool of genomic DNA extracted from a non-axenic culture deemed to contain few other cells, as described above. A total of 376,649 pyrosequences (310,136 paired-end and 66,513 singleton reads) and 4,792,504 Illumina reads were generated. Average read length of pyrosequences was 199.1 bp after splitting to left and right segments. Average length of singletons was 281.6 bp. Illumina sequences were generated from 400 bp paired-end fragments and comprise 2,396,252 paired-end reads (total of 4,792,504 reads). The total number of sequences generated and assembly statistics are in Table 4.2; the Newbler assembly metric file is in Appendix C.

Table 4.2. Assembly statistics

Total number of 454 reads	376,649
Total number of Illumina reads	4,792,504
Newbler contigs	145
Newbler scaffolds	1
Celera contigs	83
Celera scaffolds	66
Velvet contigs	3,157
Total sequence coverage	93x

Due to the non-axenic nature of the culture used for sequencing, there was a possibility of sequence misassembly from contaminant organisms. To prevent co-assembly of sequences from other organisms with *Gloeobacter*-specific sequences, a paired-end 9kb library was constructed and a paired-end constraint was applied by Newbler assembler to prevent misassemblies. Illumina sequences were also paired-end sequences with insert sizes of approximately 400bp. Pyrosequences were assembled using Newbler assembler version 2.6, resulting in a single scaffold of 146 contigs (Figure 4.3 and Table 4.2). All the contigs assembled had 9kb mate pairs linking the contigs (Figure 4.3). Each contig produced by Newbler also had paired-end 454 reads spanning the whole contig (Figure 4.4). Hybrid assembly using Celera assembler utilizing both 454 and Illumina reads produced a total of 66 contig scaffolds with 83 contigs. Total bases in the scaffolds totalled 4,799,862 bp. Contigs from the Celera assembly were aligned against Newbler contigs to checks discrepancies between assemblies (Figure 4.3). Generally, contigs produced by both assembly methods were comparable and complemented each other, although contig breaks occurred at different positions along the length of the genome.

Table 4.3. Questionable regions within the genome

Region	Start	Stop	Size (bp)	G+C%
1	415517	431636	16120	59.47
2	903207	946727	43521	59.81
3	1004060	1015740	11681	54.28
4	1732580	1828970	96391	55.52
5	4262380	4279810	17431	53.42

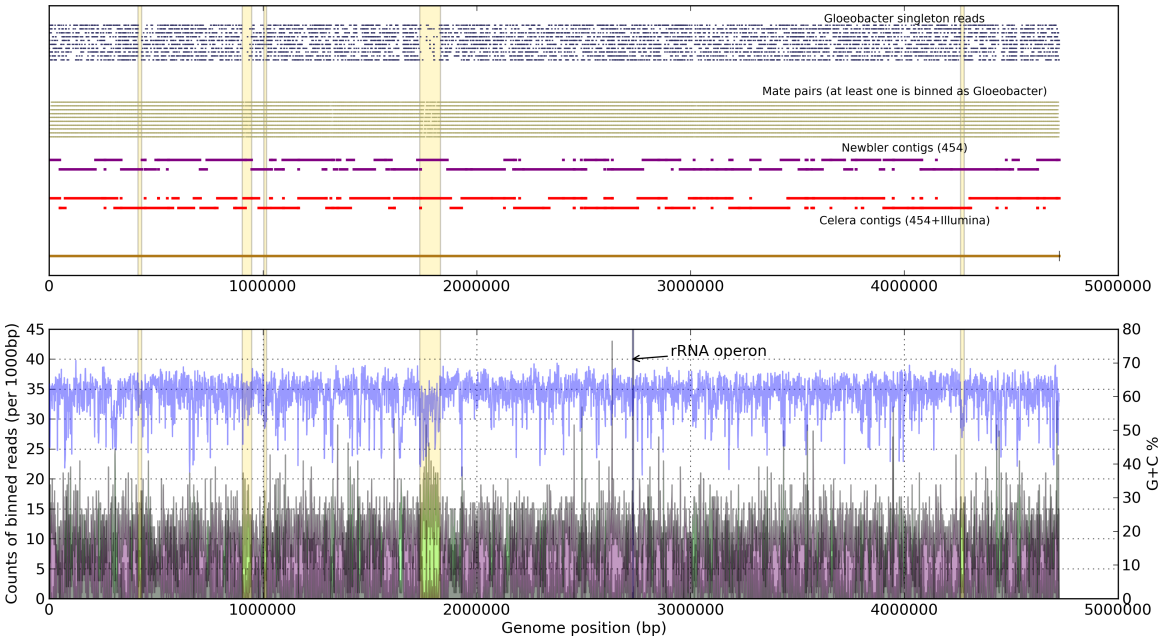


Figure 4.3. *Candidatus* G. kilaueaensis JS1 assembly verification plot. Top panel: Contigs produced by Celera and Newbler assemblers aligned against the finished genome represented as a gold line near the bottom of top panel. Consistent (∼9kb) mate pairs identified as *Gloeobacter* in origin and aligned against the finished genome are plotted as black line segments (appear here as continuous black lines across the genome because of the close proximity between mate pairs). Singleton reads binned as *Gloeobacter* are shown as blue line segments. Suspicious regions with low G+C% are highlighted as beige rectangles. Bottom panel: G+C% for a given 1000 bp region along the genome, as blue lines. Also shows coverage of reads binned as either *Gloeobacter* in origin or not. Reads binned as *Gloeobacter* are purple, while others are green. The plot was produced by a custom Python script (See section 5.1.17).

Newbler contig scaffolds were used as a framework to orient the contigs and to close the remaining gaps between contigs. To aid closure of these gaps, Illumina sequences were independently assembled using the Velvet assembler [72], producing 3,157 contigs with an average contig size of 1,741 bp (Figure 5.1.17). The largest contig was 56 kbp. Velvet contigs were shredded to 500 bp fragments with 250 bp overlapping regions, and manually assembled with Newbler contigs by MINIMUS and SeqmanII (DNASTAR Inc, Madison, WI). PCR amplification of the remaining gap regions followed by capillary sequencing of the PCR products closed all gaps.

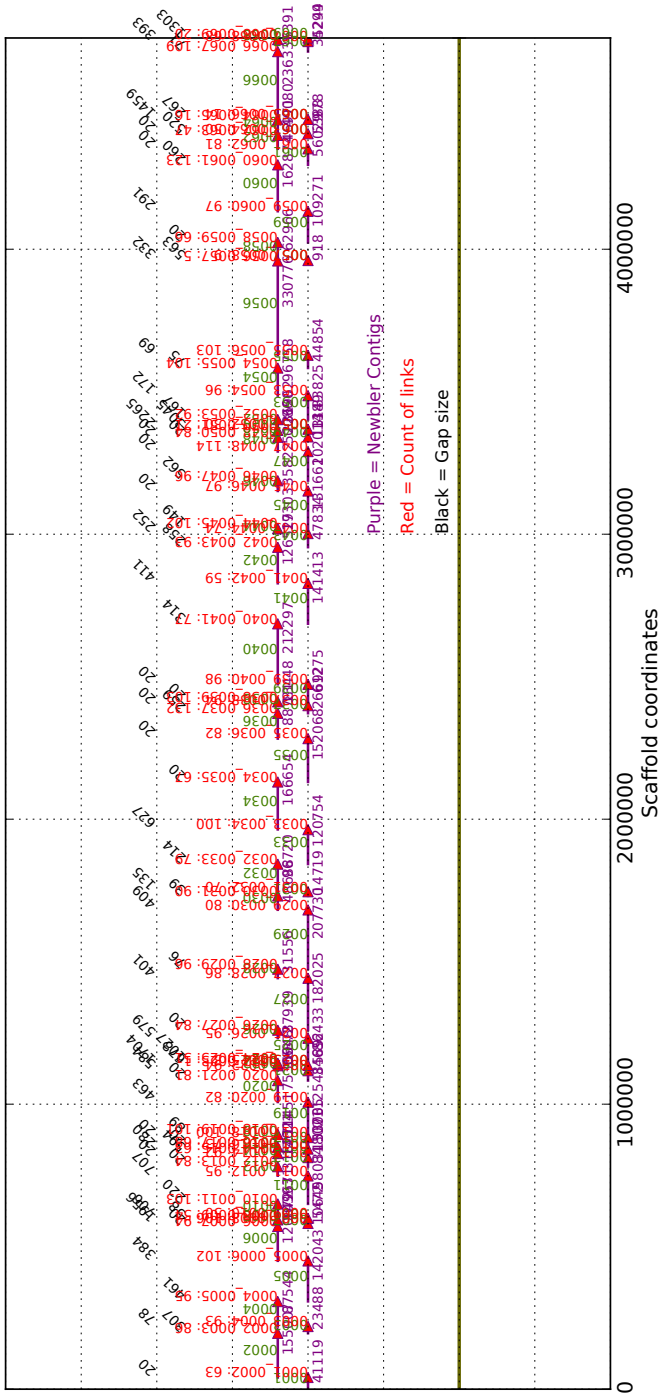
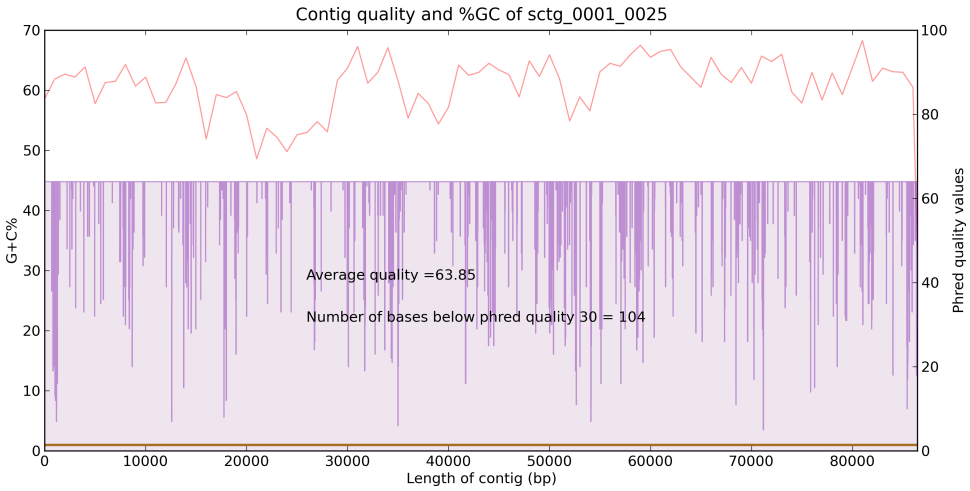
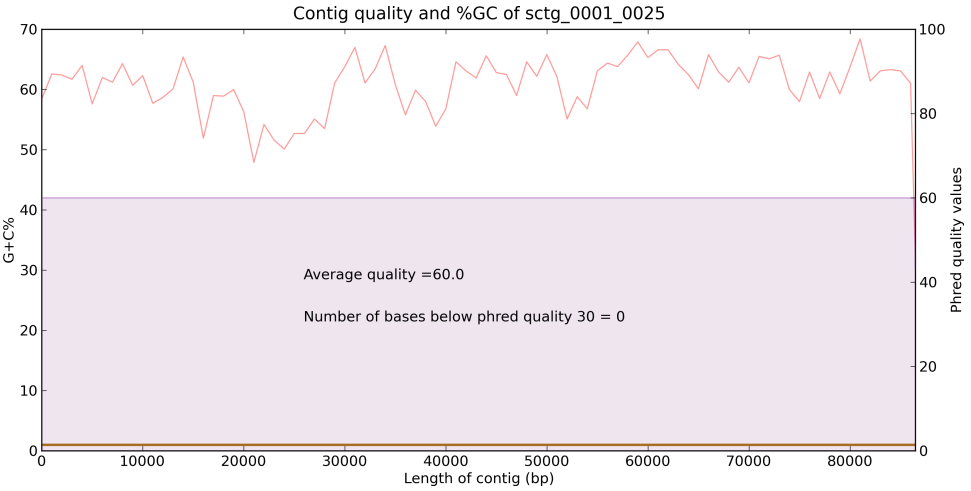


Figure 4.4. Newbler scaffolds visualized by a custom Python script (see 5.1.19). Scaffold contigs are shown as 'sticks' with arrows, and purple text indicating contig size in bp. Vertical red text shows number of consistent mate pairs between any two given scaffolds. Slanted black text indicates hypothetical gap size as determined by Newbler assembler.

To improve the sequence quality of final assembled contig, Illumina reads were recruited using MUMMER [149], and recruited reads were assembled with the final contig using MINIMUS. This step fixed ambiguous bases introduced by homopolymers present in 454 pyrosequences, and improved the overall quality of the assembled genome sequence (Figure 4.5).



(a)



(b)

Figure 4.5. Contig quality improvement. (a) Newbler contig before quality was improved using Illumina reads, (b) after polishing with Illumina reads. To improve the contig quality, Newbler contigs were re-assembled by the Minimo tool from theAMOS package utilizing Illumina reads recruited with MUMMER.

To verify correct assembly and to identify potentially misassembled regions in the genome, Newbler and Celera contigs and pyrosequences were aligned against the finished genome using

MUMMER. Mate pairs and singletons produced by 454 sequencing were binned in PhymmBL [69, 70] to assign taxonomic ranks to the reads. The intention was to bin *Gloeobacter* reads from non-*Gloeobacter* reads to visualize sequence coverage along the genome. Where sequence coverage for *Gloeobacter*-specific reads dropped below average coverage (for example, below the average count of ∼15 per 1000 bp in the bottom panel of Figure 4.3), those regions were manually checked and amplified in long-range PCRs to confirm their presence and their sequence in the genome. Five such regions seemed questionable in this respect due to the low coverage of *Gloeobacter* sequences, and because their G+C% dropped below 60% (G+C% of the *Candidatus* G. kilaueaensis JS1 genome is 60.5%). ORFs and their taxonomic ranks are recorded (Tables A.1 to A.8). Taxonomic ranks of these ORFs were estimated by taking the consensus of the top 10 (or fewer) BLASTp hits. It is important to note that the PhymmBL program has an accuracy of 78.4% in assigning taxonomic ranks at the genus level, and the algorithm involves comparison with known sequences from Genbank [69]. Since there is only one representative *Gloeobacter* genome in GenBank, it is possible that binning may have false positives or negatives due to the low representation of *Gloeobacter*-specific sequences in the database.

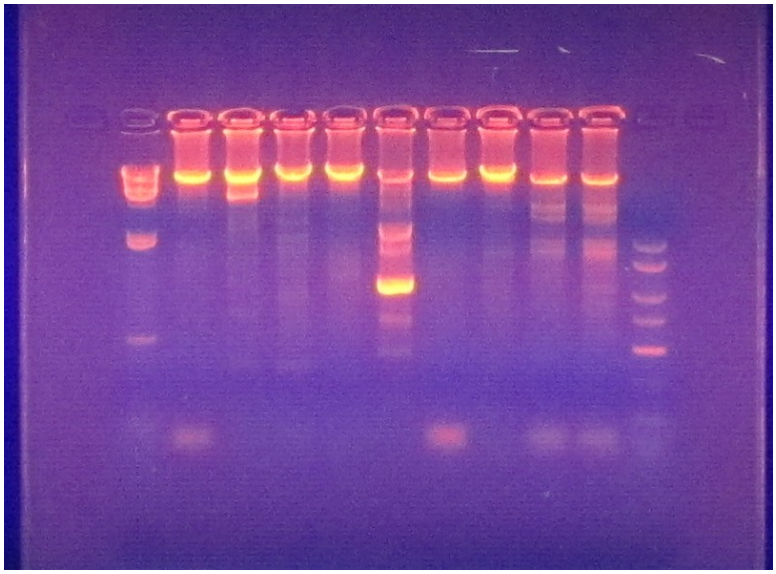


Figure 4.6. Long PCR gel. Gel picture showing ∼15kb bands. Two outermost lanes are DNA markers (left: λ marker, right: 1 kb marker).

Some parts of ther assembled genome did appear to contain genes from other organisms, i.e., top BLAST hits from organisms outside of *Gloeobacter* genus (Tables A.1 to A.8) (Figure 4.3).

To confirm the presence of such regions in *Candidatus* G. kilaueaensis JS1, primers were designed to amplify approximately 15 kb fragments from these parts of the genome (Table 4.1). A total of 9 long-range PCR reactions confirmed these regions are in fact part of the genome, and that they are not derived from other bacteria or artifacts of sequence assembly (Figure 4.6).

The complete genome sequence of *Candidatus* G. kilaueaensis JS1 has been deposited in GenBank with an accession number of CP003587.

4.4.3 Genome characteristics and features

The *Candidatus* G. kilaueaensis JS1 (hereafter referred to as JS1) genome comprises 4,724,791 bp, with a G+C% of 60.5 (Table 4.4). The G+C content of the genome is 1.5% lower than that of *G. violaceus* PCC 7421 (hereafter referred to as PCC 7421) which has a G+C content of 62%. G+C content variations within the chromosome derive from several regions, *i.e.*, the suspect regions checked by long PCRs that appeared to contain phage-related genes or mobile genetic elements such as transposons. The genes in and top BLAST hits of these low G+C regions are provided in Tables A.1 to A.8.

A total of 49 tRNAs, 1 rRNA operon, and 4,508 protein coding genes were identified in the genome. Functions were predicted for 2862(63.5%) of the 4508 proteins; 1655(36.7%) were annotated as hypothetical proteins, and 313(6.9%) had no BLAST hit in the Refseq database at an E-value cutoff of $1e^{-5}$. About 34% of the proteome has no hits to COGs (Cluster of Orthologous Groups). Protein-coding genes were compared by COG functional categories with those in PCC 7421 (Figure 4.8). The top three COG functional categories are cell wall/membrane/envelope biogenesis (5.9%), transcription (4.7%), and amino acid transport and metabolism (4.4%). Generally, the distribution of COG categories between the two *Gloeobacter* species is similar, indicating some conservation of their functional potential (Figure 4.8).

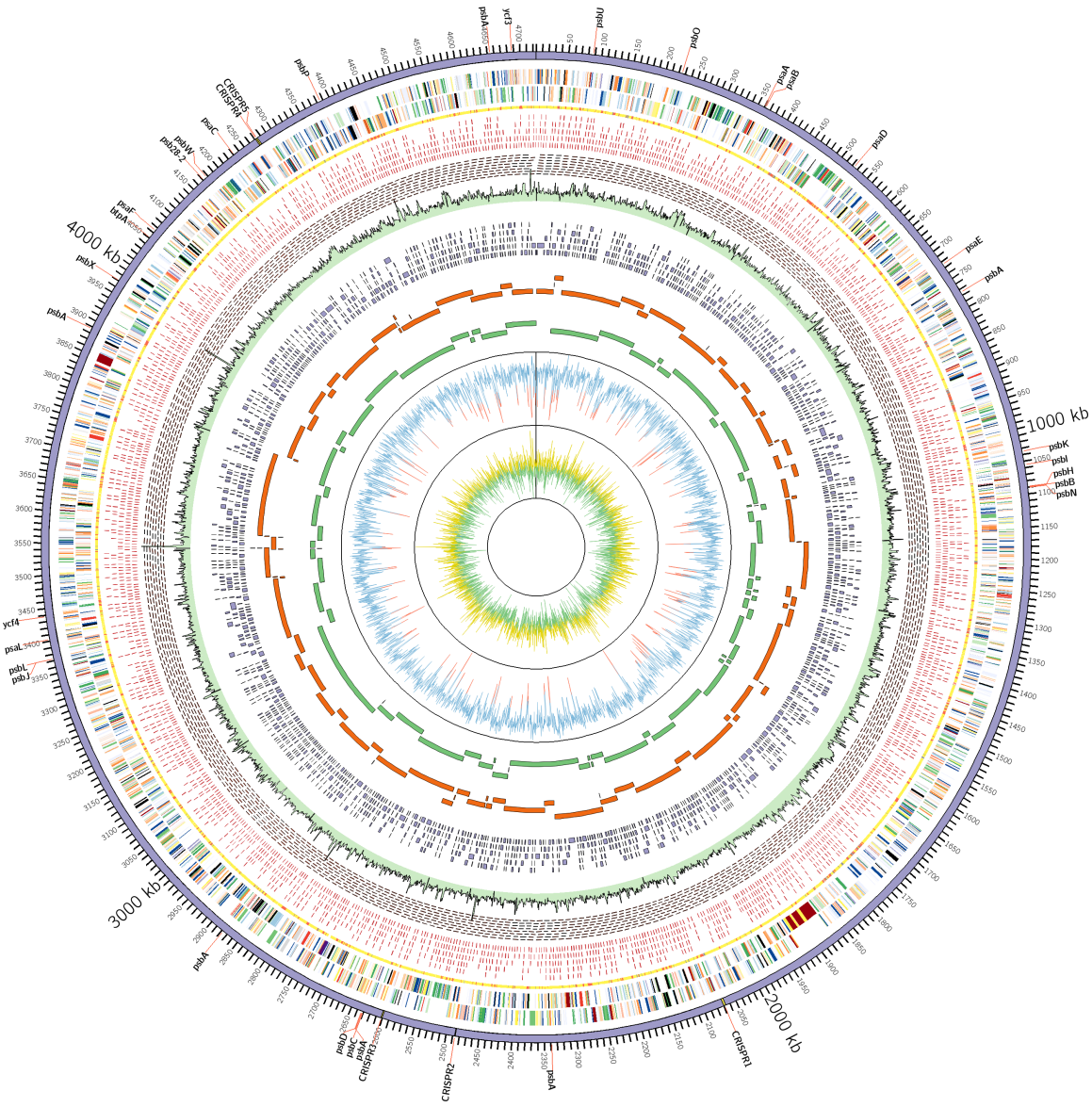


Figure 4.7. Representation of the *Candidatus* G. kilaueaensis JS1 genome. From inside out: GC skew (Yellow >0, Green <0), GC percent (Blue >50%, Red <50%), Newbler scaffold contigs, Celera contigs, Velvet contigs (Illumina reads only), read coverage (Combined 454 and Illumina reads sampled for 1000bp window. Highest coverage is 368x), minimal tiling clone pairs (shown in red), recruited reads from metagenome, taxonomic rank of top BLAST hit (yellow = Cyanobacteria, Red = others, Grey = no BLAST hit), coding regions in minus and plus strands (colored by COG functional categories). CRISPR repeat regions are highlighted in yellow in the outermost circle. Locations of genes involved in photosystems are labeled in the outermost circle.

Table 4.4. General features of the *Candidatus* G. kilaueaensis JS1 genome and comparison with *G. violaceus* PCC 7421

Organism	*Candidatus* G. kilaueaensis JS1	*G. violaceus* PCC 7421
Size (bp)	4,724,791	4,659,019
G+C%	60.5	62.0
Total number of ORFs	4,508	4,430
Protein coding (%)	90.4	89.4
Proteins with known functions	2,245	1,788
Hypothetical proteins	1,642	2,642
Total number of rRNA operons	1	1
Total number of tRNA genes	49	45
Other RNA	8	4
CRISPR repeat regions	5	0

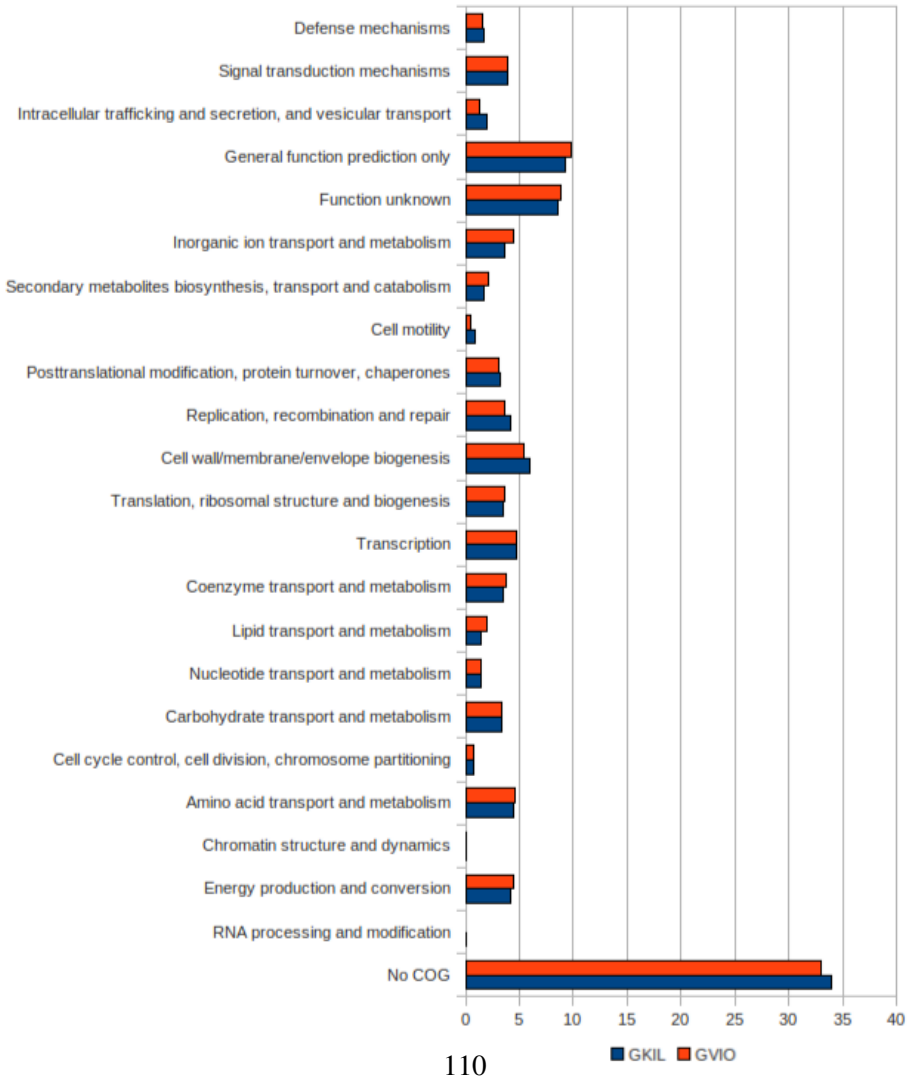


Figure 4.8. Comparison of COG functional categories in *G. violaceus* PCC 7421 and *Candidatus* G. kilaueaensis JS1. Numbers on x-axis represent percent of total protein coding genes.

Phage are important agents in genetic exchange between bacteria, and phage-related regions constitute genomic hotspots in cyanobacteia such as *Prochlorococcus* [159]. These hotspots or 'genomic islands' can contribute as much as 10-30% of the diversity between different strains of bacteria [160]. The genome of JS1 contains regions that seem to have been acquired from other organisms. These occur mostly in the suspect regions mentioned above (Table 4.3); genes in these regions have either no BLAST hits, are mostly from other bacteria, or are of viral origin (Tables A.1 to A.8). Of 196 ORFs identified in these regions, 75 have no BLAST hits and 136 have no known function and are annotated as hypothetical proteins. Among genes of viral origins, taxonomic affiliations suggest they are derived from Caudovirales, double-stranded DNA viruses with no RNA stage.

Using the CRISPR Finder tool, 5 CRISPR repeats were detected in the JS1 genome (Table 4.5). There are no CRISPR repeat regions in the PCC 7421 genome. In addition to CRISPR repeats, CRISPR-associated proteins (Cas1, Cas2, Cas4, and Csc2) were located in the genome. Cas1 (GKIL_1965), Cas2 (GKIL_1966), Cas4 (GKIL_1964), and Csc2 (GKIL_1961) were found close to CRISPR repeat region 1 (2066878-2070197). Additional copies of CRISPR-associated proteins - Cas1 (GKIL_4060) and Cas2 (GKIL_4059) were found close to CRISPR region 5 (4273038-4274931). A CRISPR-associated protein from the APE2256 family (GKIL_2360) was found close to CRISPR region 2 (2486198-2486962). CRISPR repeats are components of a type of bacterial immune system that helps them defend against viruses [161, 162]. The presence of phage genes and CRISPR regions in the JS1 genome suggests the strain may be in an environment in which viruses and bacteriophages are threats. This is a noteworthy observation because CRISPR regions have been reported in hotspring phototrophic mats in volcanically active Yellowstone National Park [163], suggesting that viruses may be quite common in geothermal areas.

Table 4.5. CRISPR regions in the *Candidatus* G. kilaueaensis JS1 genome

Region	Direct Repeat	Number of spacers
2066878-2070197	ATCGAAACGACCACCATCCCTGCAAAGGGATTGAAAC	45
2486198-2486962	GTTTCCGTCCCCTCGCGGGGATTAGGTCCACTCGAAC	9
2600618-2602373	GCGATTCAATCAGTGACTCCTTTCGGAGTTGAGCAC	24
4271404-4272947	GTTTCCAATCTAATCGTCCGCTGAGGGACGTCGAAC	19
4273038-4274931	GTTTCCAATCTAATCGTCCGCTGAGGGACGTCGAAC	22

4.4.4 Metabolic pathway analysis

A total of 212 pathways were identified in the JS1 genome by the Pathway Tools program [164]. These pathways are considered complete because all the enzymes required are present. Pathway prediction was done mostly automatically, but some pathways were manually inspected to verify whether or not they were complete. KEGG orthologous (KO) numbers were submitted to the iPath2.0 program to create customized pathway atlases for both the JS1 and PCC 7421 genomes so an overall assessment of the pathways present in the two genomes could be performed visually (Figures 4.15 to 4.18). Due to the large amount of space required to display each of these pathways legibly, just a few representative pathways considered important to JS1 are presented here.

4.4.4.1 Pathways involved in photosynthesis

Candidatus G. kilaueaensis JS1 has a complete set of enzymes needed for photosynthesis, except those for formation of thylakoid membranes (*e.g.,* thylakoidal processing peptidase). Genes involved in the Calvin-Benson-Bassham cycle, oxygenic photosynthesis, photorespiration, and photosynthesis light reaction are shown (Figures 4.9 to 4.12).

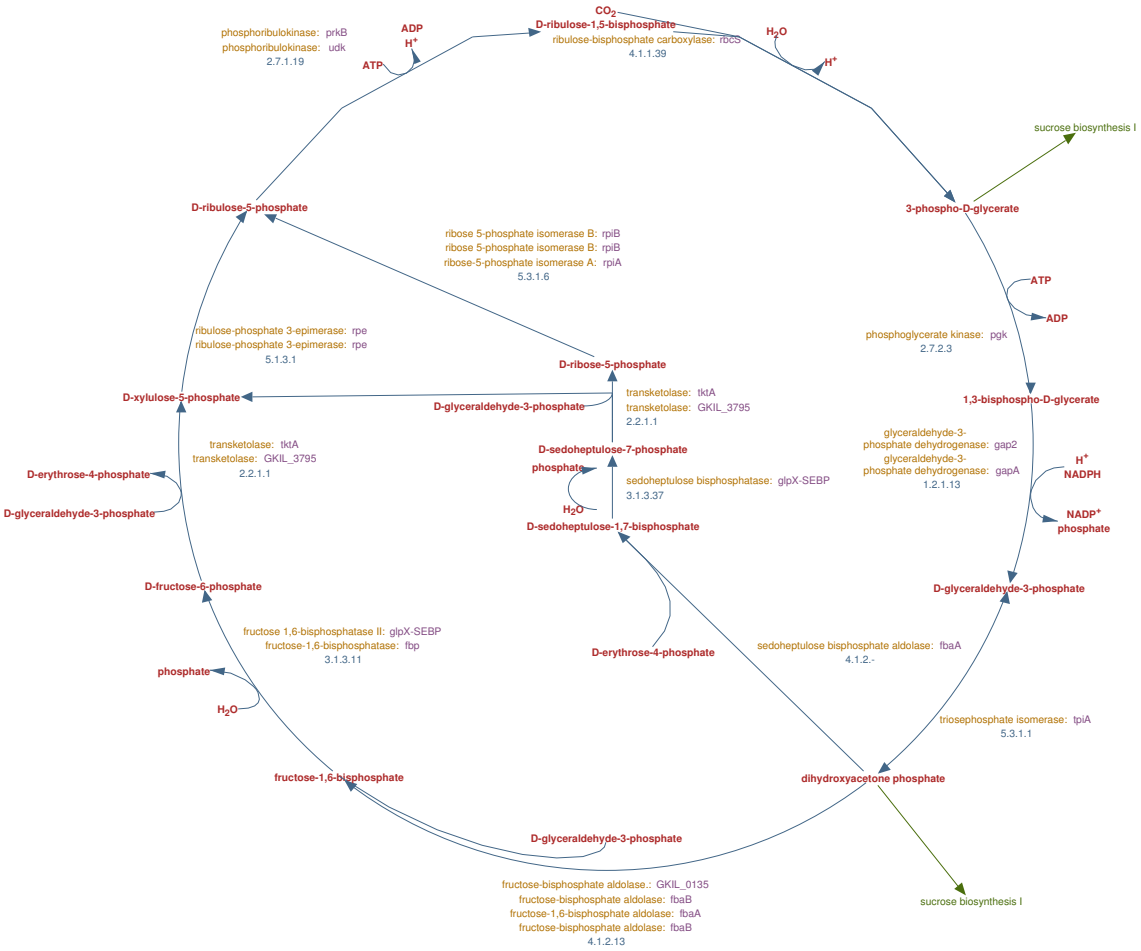


Figure 4.9. Calvin-Benson-Bassham cycle. Enzymes present in JS1 genome are shown in purple text.

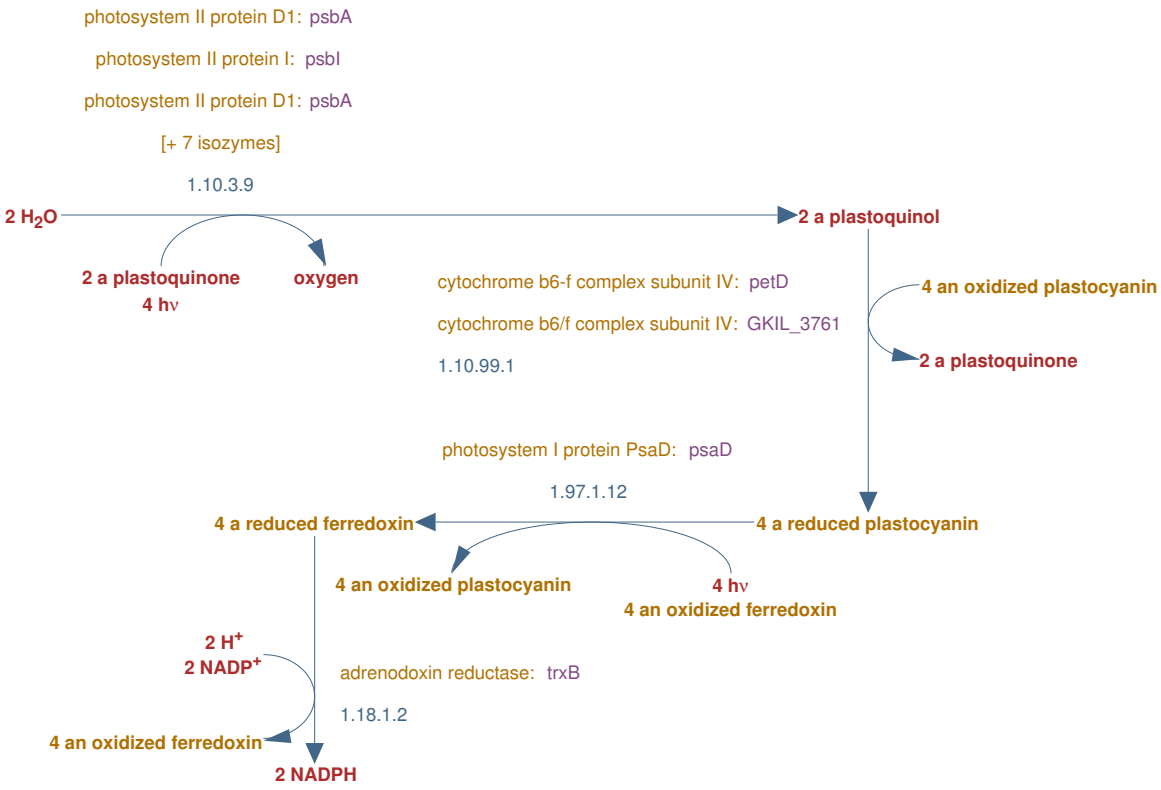


Figure 4.10. Photosynthesis light reactions pathway. Enzymes present in JS1 genome are shown in purple text.

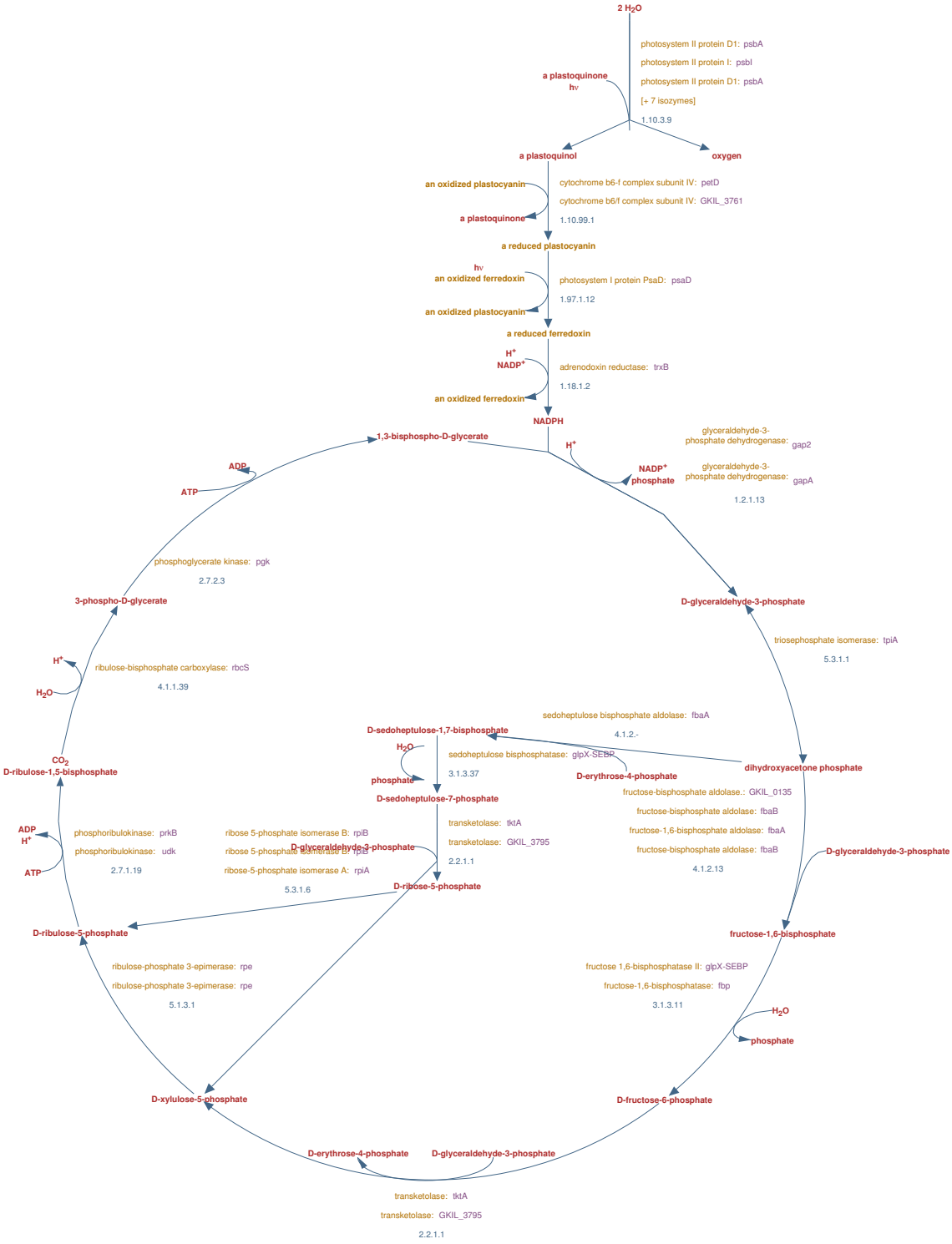


Figure 4.11. Oxygenic photosynthesis pathway. Enzymes present in JS1 genome are shown in purple text.

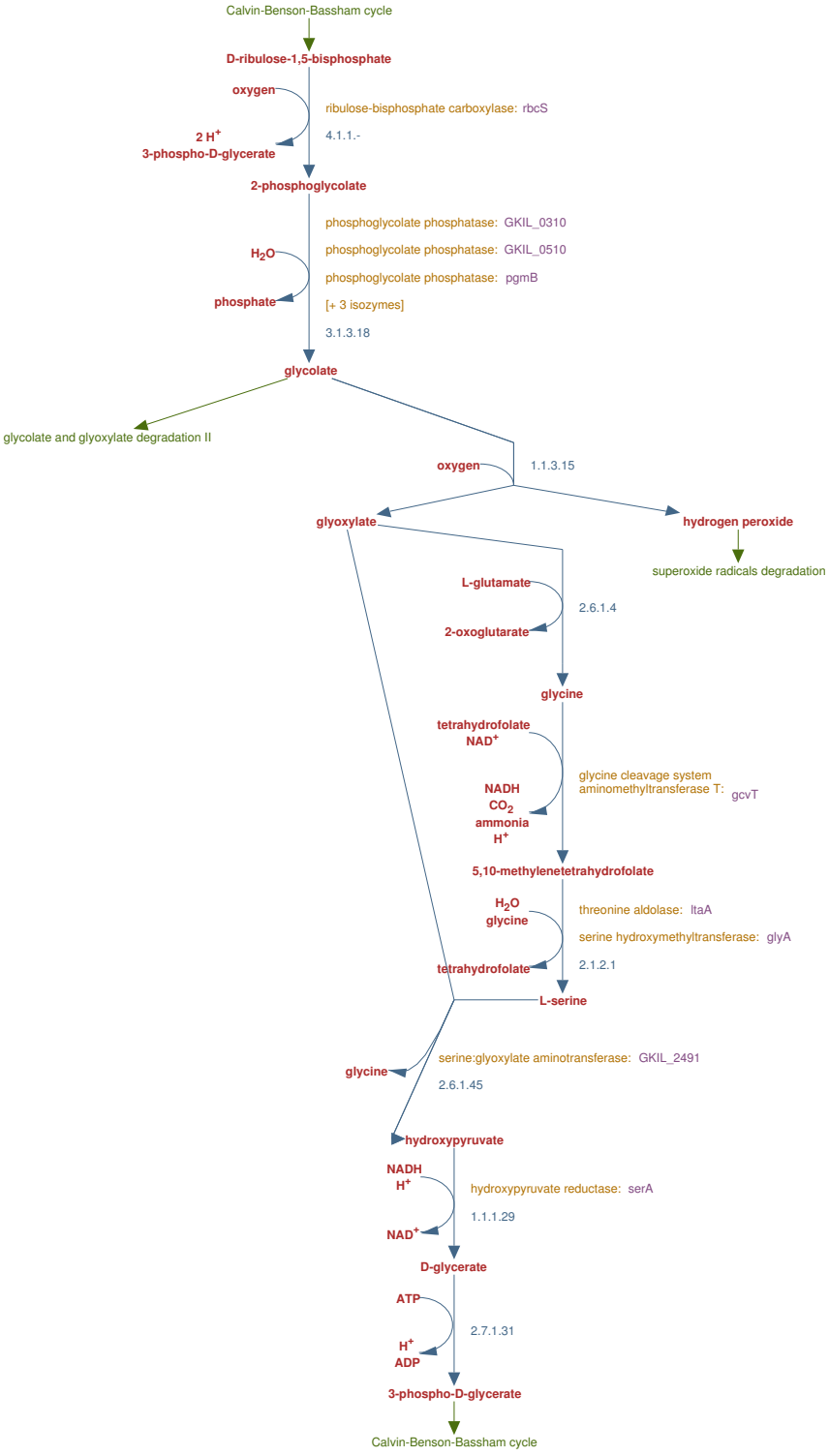


Figure 4.12. Photorespiration pathway. Enzymes present in JS1 genome are shown in purple text.

4.4.4.2 Secondary metabolite biosynthesis pathways

Gloeobacter is known to produce different pigments such as β-carotene, oscillol diglycoside, and echinenone [165, 166, 167, 168, 169]. It unusual purple coloration is hypothesized to have been due to the result of low chlorophyll content in the cells [107]. HPLC analysis detected chlorophyll *a* and β-carotene pigments in JS1, but it is not known if other pigments are present (Figure 3.9). Pathway Tools was used to check and identify metabolic pathways involved in pigment synthesis. This revealed a biosynthesis pathway for neurosporene, a subclass of *trans*-lycopene biosynthesis I in *Bacteria* (Figure 4.13).

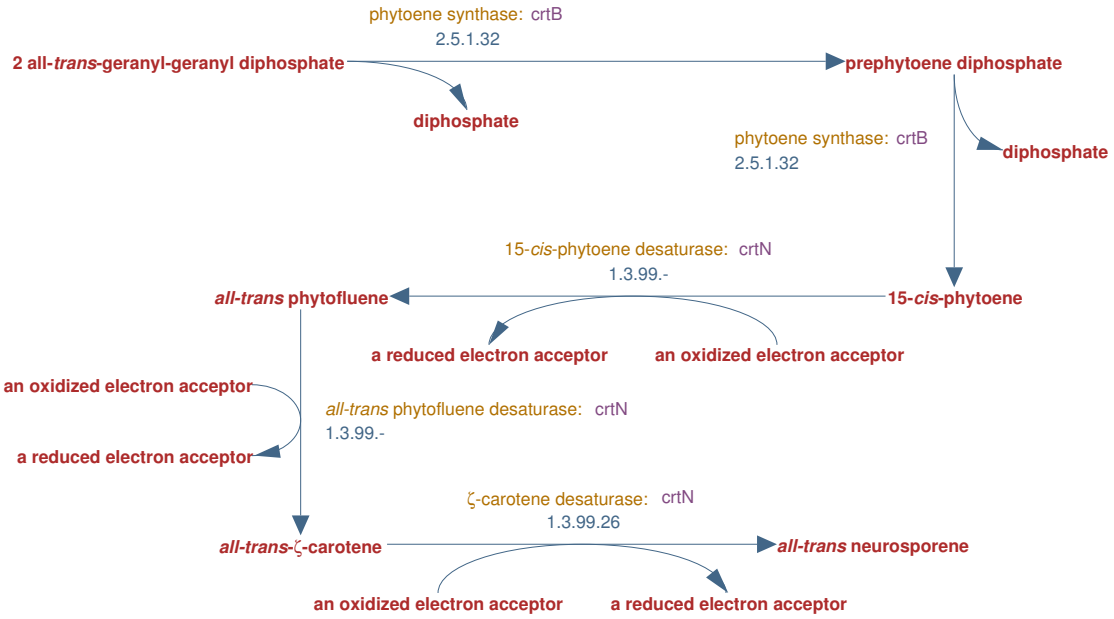


Figure 4.13. Neurosporene biosynthesis pathway in *Candidatus* G. kilaueaensis JS1.

The *trans*-lycopene biosynthesis I pathway synthesizes *all-trans*-lycopene, a bright red carotenoid pigment usually found in photosynthetic organisms and a precursor to other pigments. *Gloeobacter violaceus* is known to use bacterial-type phytoene desaturase from this pathway to synthesize major pigments such as β-carotene and (2S,2'S)-oscillol 2,2'-di(α-L-fucoside), and a minor pigment known as echinenone [165]. Phytoene synthase (*crtB*) and phytoene desaturatase (*crtN*) were identified in the JS1 genome and it is expected that JS1 is able to synthesize these carotenoid pigments. The neurosporene biosynthesis pathway, a sub-class of the *trans*-lycopene biosynthesis I

pathway that converts *all-trans*-phytoene to *all-trans*-neurosporene is utilized by purple non-sulfur bacteria such as *Rhodobacter capsulatus* and *Rhodobacter sphaeroides* to produce a pigment known as Spheroidene required in the photoreaction centers of these bacteria [170]. This is interesting because *Rhodobacter* species are the focus of studies involving anoxygenic photosynthesis due to their ability to function in both aerobic and anaerobic conditions [171]. It would be interesting to see if JS1 can synthesize Spheroidene as well. So far, HPLC analysis only revealed chlorophyll *a* and β-carotene pigments in JS1. An alternative and more sensitive test will be needed to determine if Spheroidene is present in JS1.

4.4.4.3 Vancomycin resistance genes

Pathway Tools predicted that *Candidatus* G. kilaueaensis JS1 has a pathway for vancomycin resistance, revealed by the presence of *vanB* (GKIL_3597), *vanX* (GKIL_1509 and GKIL_1879), and *serA* (GKIL_0932) (Figure 4.14). In contrast, *G. violaceus* PCC 7421 only has a copy of *vanX* (gll1805) and *serA* (gvip294), while missing the *vanB* found in JS1. Five essential gene products are required for a high-level of vancomycin resistance: VanR, VanS, VanH, VanX, and either VanA, VanB or VanD [172, 173, 174]. The vancomycin resistance pathway in JS1 only comprises genes for VanX and VanB, so the strain may not be as resistant to vancomycin as the Pathway Tool predicted. Experiments using susceptibility test discs showed *Candidatus* G. kilaueaensis JS1 is not resistant to vancomycin (results not shown), although susceptibility should be tested again with different concentrations of vancomycin to determine the level, if any, of resistance to vancomycin.

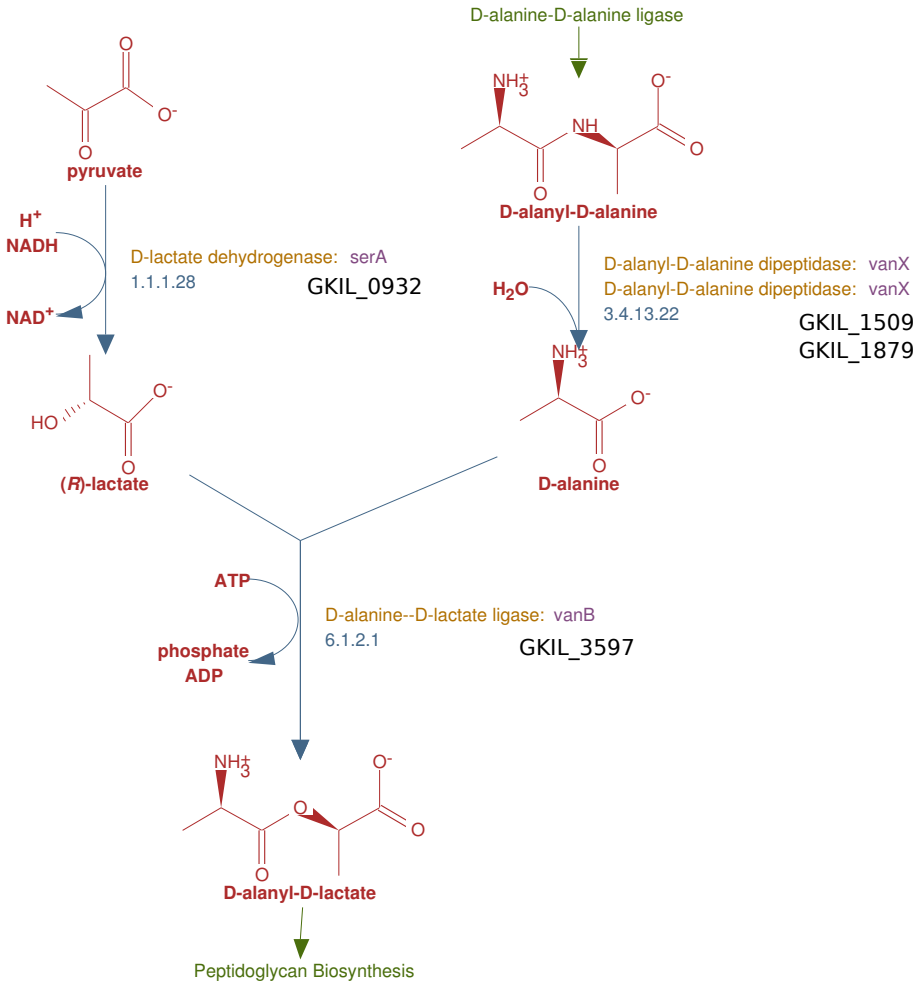


Figure 4.14. Vancomycin resistance pathway in *Candidatus* G. kilaueaensis JS1.

4.4.4.4 Comparison of metabolic pathways in *Candidatus* G. kilaueaensis JS1 and *G. violaceus* PCC 7421

The metabolic capabilities of *Candidatus* G. kilaueaensis JS1 and *G. violaceus* PCC 7421 were compared on the basis of KEGG ortholog (KO) groups in the two genomes. In JS1, 168 KO groups were identified, compared to 182 in PCC 7421, and 1138 in both. To visualize the overall metabolic potentials of these organisms, the KO groups present in each organism were submitted to the iPath2.0 program to generate pathway atlases (Figures 4.15 to 4.18). iPath2.0 metabolic pathway atlases for main metabolic pathways, plus secondary metabolite biosynthesis pathways in

PCC 7421 and JS1 were generated (Figures 4.16 and 4.18). Pathway components were highlighted in these figures by colors depending on the metabolic pathway category.

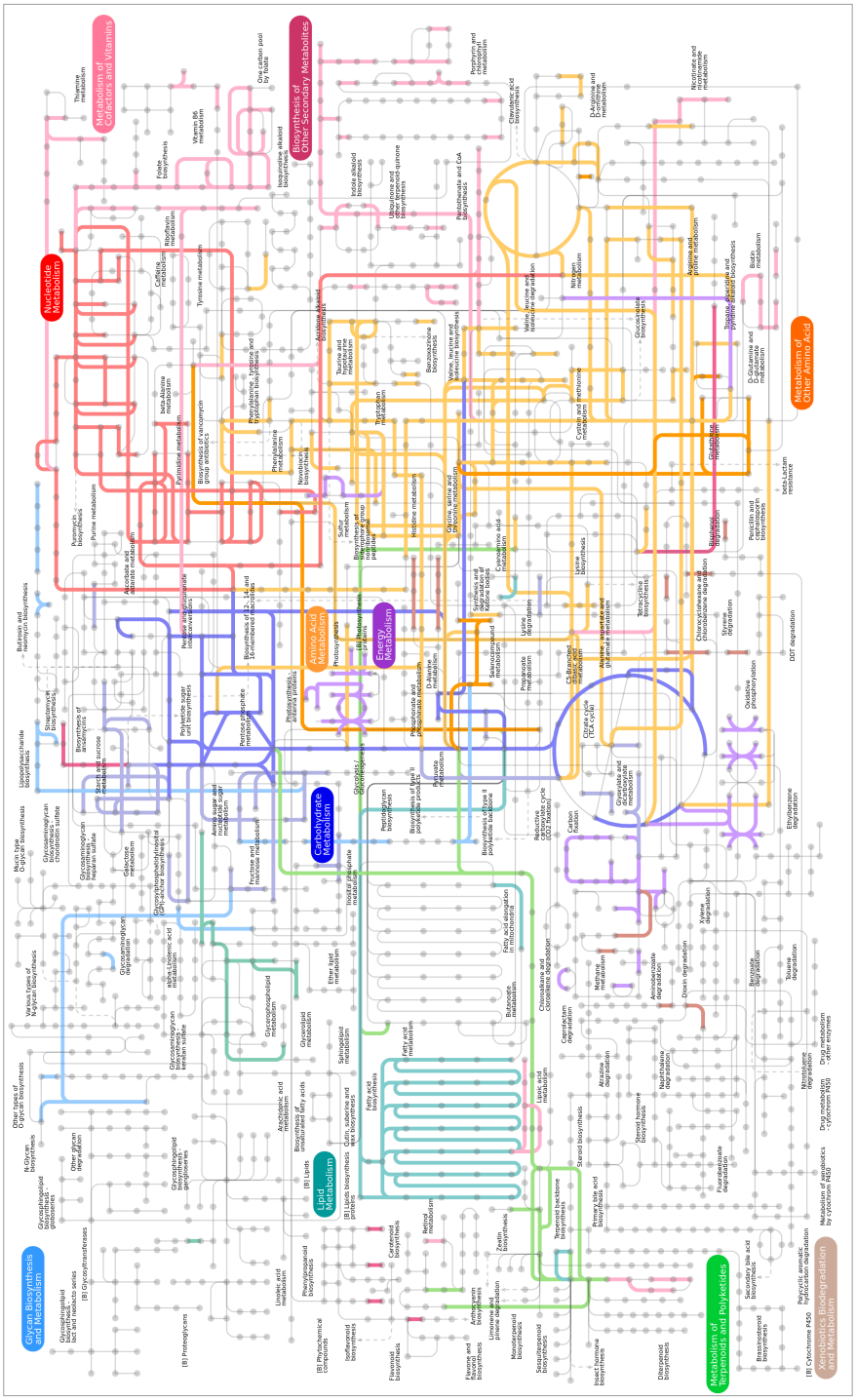


Figure 4.15. Pathway atlas of *Candidatus* G. kilaueaensis JS1 based on KEGG orthologous groups (KO). Pathway modules for which enzymes are present in the organism are highlighted in different colors according to the pathway.

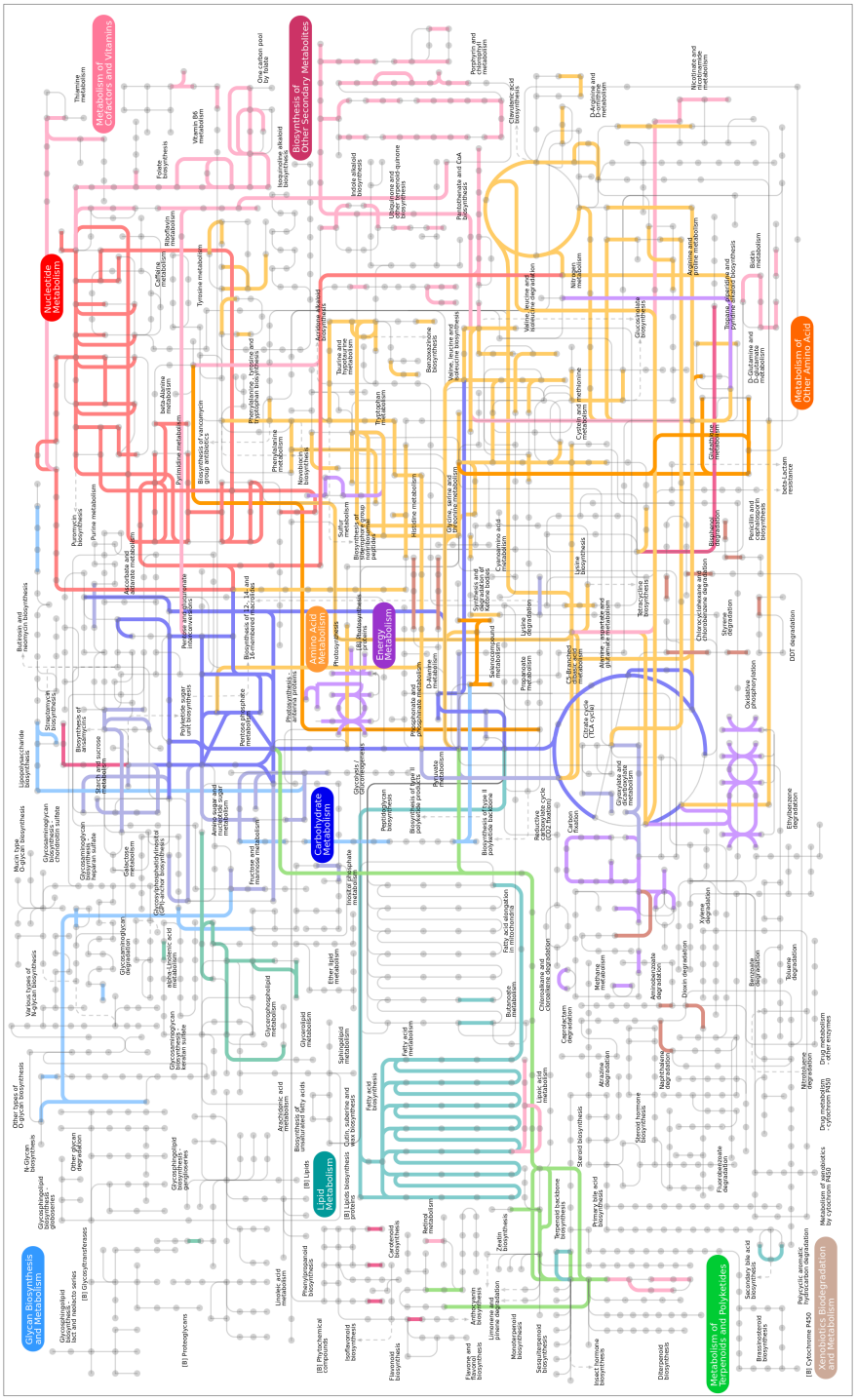


Figure 4.16. Pathway atlas of metabolic pathways in *G. violaceus* PCC 7421 based on KEGG orthologous groups (KO). Pathway modules for which enzymes are present are highlighted in different colors according to the pathway.

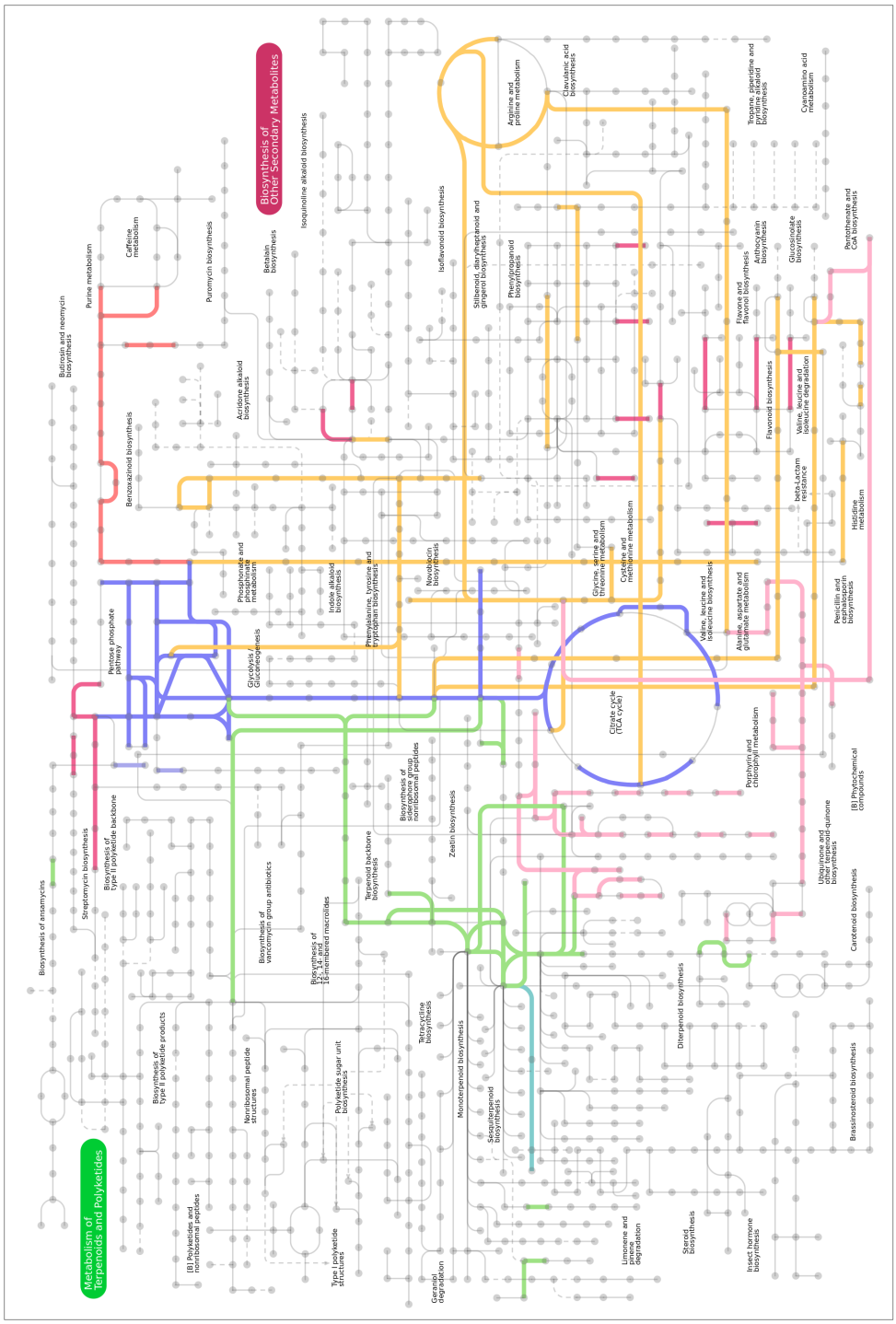


Figure 4.17. Pathway atlas of secondary metabolites in *Candidatus* G. kilaueaensis JS1 based on KEGG orthologous groups (KO). Pathway modules for which enzymes are present are highlighted in different colors according to the pathway.

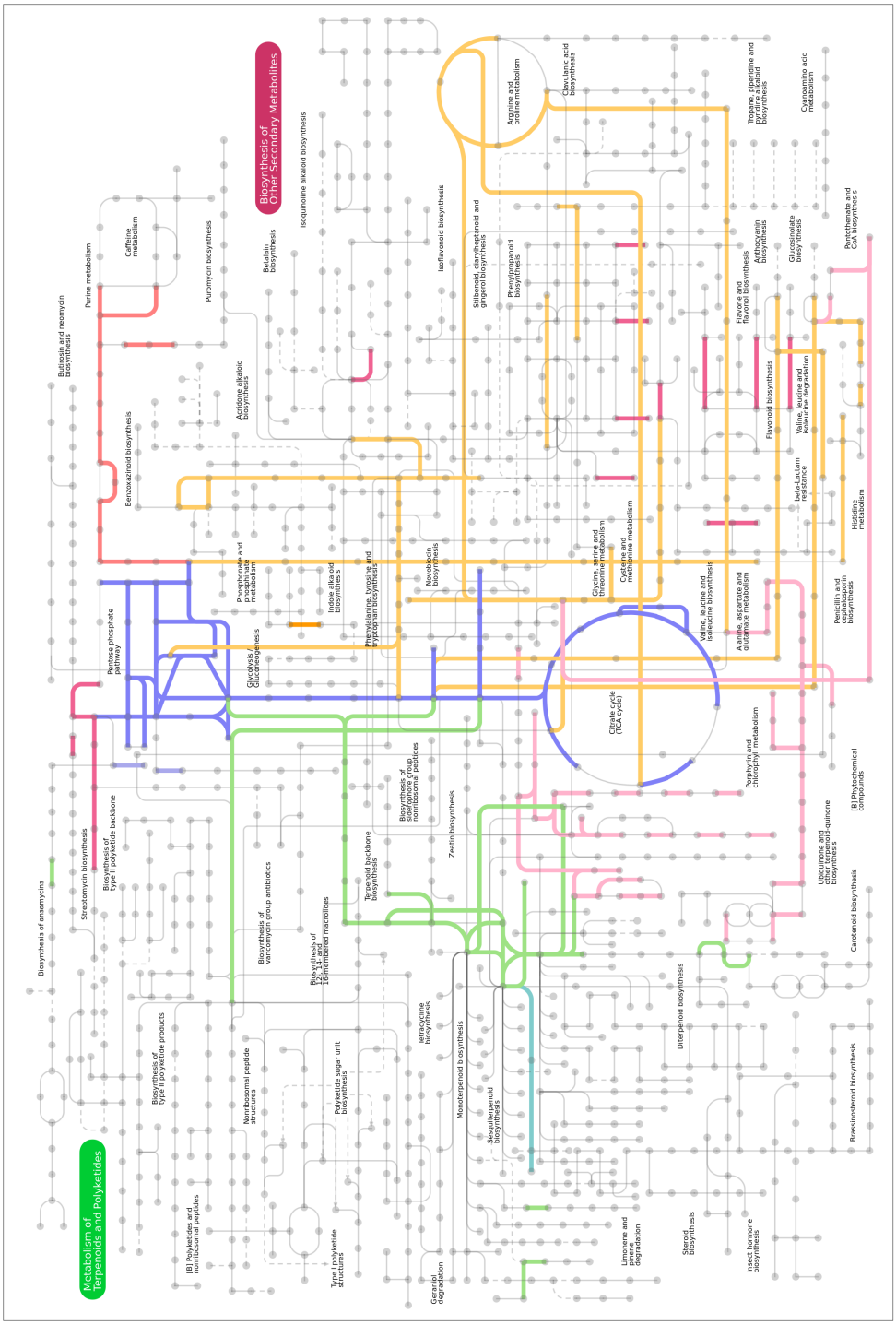


Figure 4.18. Pathway atlas of secondary metabolites in *G. violaceus* PCC 7421 based on KEGG orthologous groups (KO). Pathway modules for which enzymes are present are highlighted in different colors according to the pathway.

Most of the pathway components in the two organism are similar, for both major metabolic and secondary metabolites biosynthesis pathways. Differences in these two organisms will be presented as supplementary tables when the manuscript is submitted to a peer-reviewed journal.

4.4.5 *In silico* DNA-DNA hybridization and determination of species rank

Species definition and delineation in bacteria (*Archaea* and *Bacteria*) is not a trivial task, but DNA-DNA hybridization (DDH) values ranging from 60 to 70% have traditionally been used, with different species sharing less than the 'cut-off' value [175]. However, with more complete genomes becoming readily available, there have been efforts to replace laboratory DDH experiments with *in silico* genome comparisons, such as the Average Nucleotide Identities (ANI), or Genome-to-Genome distances [176, 155, 156]. Assignment of species/strain names among closely related *Bacillus* species was recently demonstrated to be possible on the basis of genome comparisons since they correlated very closely with actual DDH results [177].

Based on 16S rDNA sequence identity alone, *Candidatus* G. kilaueaensis JS1 and *G. violaceus* PCC 7421 would be considered to belong to the same species because their 16S rRNA genes share 98.7% (1465/1485 bp match) nucleotide identity [178]. However, the complete genome sequence of *Candidatus* G. kilaueaensis JS1 and *G. violaceus* PCC 7421 revealed major differences between the two organisms, and an *in silico* DDH equivalent to ∼11%, well below the 70% threshold recommended for distinguishing species [178]. With both slow growth and a non-axenic culture, it would have been challenging and perhaps unreliable to perform DDH experiments *in vitro* with *Candidatus* G. kilaueaensis JS1.

Genomes of *Candidatus* G. kilaueaensis JS1 and *G. violaceus* PCC 7421 were used to calculate percent identities between their respective genomic DNA fragments ([155]). Using JSpecies [156] with default parameters, the ANI between JS1 and PCC 7421 genomes was found to be 73.75% (with BLAST) and 83.11% (with MUMMER), well below the cut-off value of 90% for species delineation used in this approach. The *in silico* DDH values were calculated using the Genome-to-Genome Distance Calculator (GGDC) with three formulae [154]. GGDC calculations using these formulae with BLAST revealed DDH values of 11.3%, 13.5%, and 8.72%. Using MUMMER, DDH value between JS1 and PCC 7421 was 14.96%. Both methods give DDH values well below the cutoff of 60% for delineation of species by this method. Moreover, there was little synteny between the genomes (Figure 4.27).

Using MUMMER with default parameters, JS1 and PCC 7421 genomes were aligned; alignment plots were visualized with a custom Python script (see section 5.1.12). Matching regions

were visualized by connecting lines between the two genomes, with lines in different colors representing different sequence identities based on MUMMER results. Sequence identities (MUMs or Maximal Unique Matches) between the two genomes averaged 83.4% at the DNA level (light green lines). Matching segments are very small (average 1 kbp, largest 6.1 kbp) and scattered througout the genome (Figure 4.27), as opposed to in large conserved syntenous blocks often found in closely related bacteria species (Figure 5.4).

Based on these results and taxonomic criteria, especially DDH, *Candidatus* Gloeobacter kilaueaensis JS1 does not belong in the same spexcies as *Gloeobacter violaceus* CPP 7421.

4.4.6 Analysis of individual genes of interest

Genes associated with thylakoid membranes: *G. violaceus* PCC 7421 is known to lack thylakoid membranes, and it is this lack of thylakoid membrane that led to intense investigation of this species on the grounds it may be the missing link in anoxygenic to oxygenic photosynthesis [134, 135]. The presence or absence of thylakoid membranes in *Candidatus* G. kilaueaensis JS1 was not tested by transmission electron microscopy (TEM), but the genome annotation did not detect some of the genes involved in thylakoid membrane formation. The genes *sqdB* (encoding sulfolipid biosynthesis protein) and *sqdX* (encoding UDP-sulfoquinovose:DAG sulfoquinovosyltransferase) are known to be required for synthesis of sulfoquinovosyl diacylglycerol (SQDG) which is required for photosystem stabilization (the product SQDB is usually found in thylakoid membranes) in other cyanobacteria but is absent from *G. violaceus* PCC 7421 [179, 180]. Both *sqdB* and *sqdX* are also absent from the JS1 genome, *i.e.*, BLASTp searched yielded only weak homologs with less than 30% sequence identities at the amino acid sequence level.

The Vipp1 protein is known to be essential for the formation of thylakoid membrane in *Synechocystis* [181] and *Arabidopsis thaliana* [182] and has been detected in *G. violaceus* PCC 7421, although the ortholog in PCC 7421 (which is annotated as phage shock protein, PspA) seems to be missing the conserved C-terminal region in its amino acid sequence and is not expected to function the same way as *Synechocystis* or plant Vipp1 protein [108]. A copy of the Vipp1 homolog was also detected in the JS1 genome (GKIL_4366 - phage shock protein A, PspA) but is nearly identical to PCC 7421 PspA protein, and also lacks the conserved C-terminal mentioned in [108]. Thus, it can be reasonably deduced from the genome information that *Candidatus* G. kilaueaensis JS1 lacks thylakoid membranes, although this would be best confirmed by TEM.

Squalene hopene cyclase: Hopanoids are important biomarkers that have been used to date divergence and appearence of certain bacterial phylotypes in fossil records. The squalene

hopene cyclase gene was identified in *Candidatus* Gloeobacter kilaueaensis JS1, and its sequence was compared with those in other cyanobacteria to determine the phylogenetic affiliation of this important gene. Squalene hopene cyclase (*shc*; GKIL_2413) was first searched against the nr database in NCBI to retrieve the top 250 hits. These hits were then aligned with GKIL_2413 using Muscle, edited with Gblocks, and a maximum likelihood analysis performed using RAxML. The maximum likelihood phylogenetic tree built using RAxML indicates that the *Candidatus* Gloeobacter kilaueaensis JS1 *shc* gene clustered closely with that in *Gloeobacter violaceus* PCC 7421 *shc* and in the *Cyanobacteria* clade. Surprisingly, the *shc* gene from *Candidatus* Chloracidobacterium thermophilum B, the only known photoheterotrophic bacterium in the phylum *Acidobacteria*, was located near the root of the *Cyanobacteria* clade, suggesting the gene may have been horizontally transferred from the ancestors of photosynthetic bacteria.

Photosynthetic genes: A search for copies of the *psbA* gene, an important gene for photosystem II, revealed 6 copies in the *Candidatus* G. kilaueaensis JS1 genome. These were used as a query to search against the 40 extant complete cyanobacteria genomes to see if the *psbA* gene lineage could be traced back to an earliest ancestor, and if some pattern of clustering exists for this essential gene in cyanobacteria. A total of 190 matching *psbA* gene orthologs were found in the genomes after their alignment with the 6 copies in *Candidatus* G. kilaueaensis JS1 to build a maximum likelihood phylogenetic tree. Procedures for alignment and phylogenetic analysis were the same as those for the *shc* genes. Five copies in *Candidatus* G. kilaueaensis JS1 clustered closely with the 5 copies found in *G. violaceus* PCC 7421 (Figure 4.20). The *psbA* gene is not strictly conserved and is most likely transferred horizontally among different species of cyanobacteria (and plant plastids) [183, 184]. Building this particular gene tree is informative in many ways. First, one might reveal if each of these paralogs are vertically transferred, and thus indicate which could have been acquired from a distant organism. Next, pinpointing locations of genes potentially transferred from other organisms into these genomes might reveal genetic hotspots amenable to more detailed scrutiny.

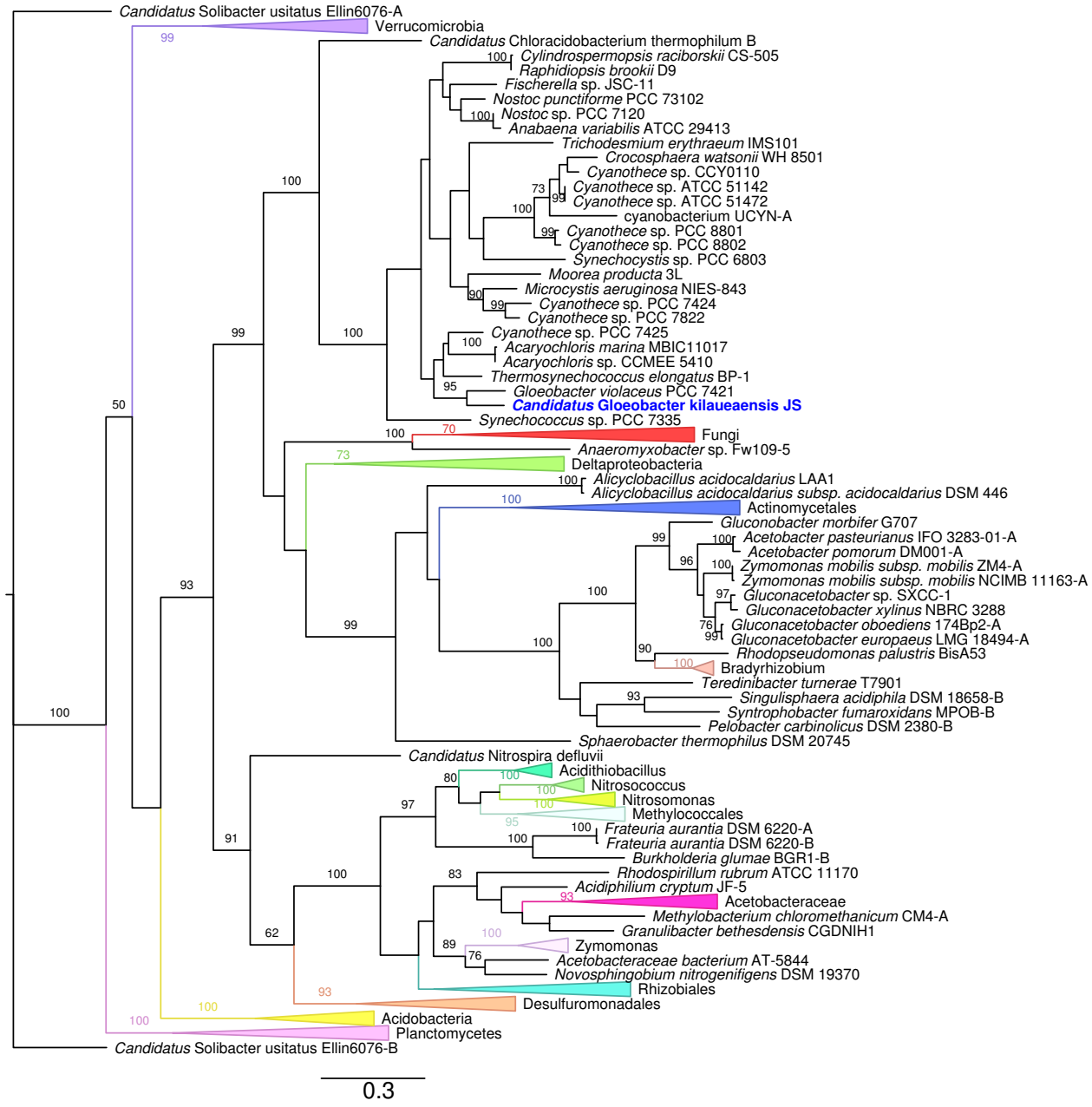


Figure 4.19. Unrooted maximum likelihood phylogenetic tree of Shc proteins from top hit organisms. *Candidatus* G. kilaueaensis JS1 is highlighted in blue. Several taxa identified as belonging to a specific taxonomic group are collapsed to make the tree easier to visualize. Only bootstrap values higher than 60 are shown.

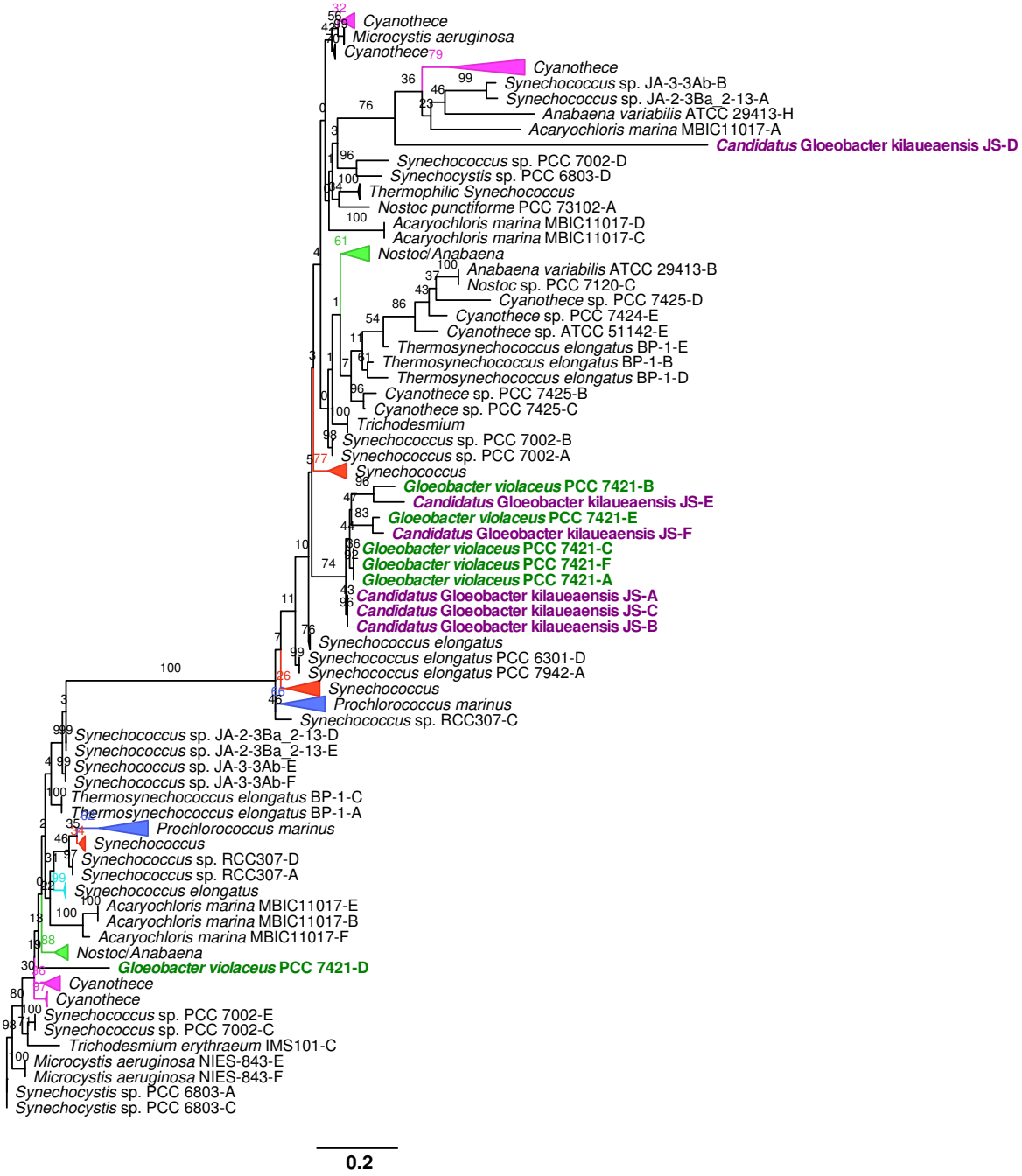


Figure 4.20. Unrooted maximum likelihood phylogenetic tree of PsbA copies in 40 completely sequenced cyanobacteria genomes and *Candidatus* G. kilaueaensis JS1. *Candidatus* G. kilaueaensis JS1 is highlighted in purple and *G. violaceus* PCC 7421 in green.

Bacteriorhodopsin: *Candidatus* G. kilaueaensis JS1 lacks the rhodopsin (gll0198) present in *G. violaceus* PCC 7421. In PCC 7421, the bacteriorhodopsin gene seems likely to have been horizontally acquired because it is contained in a region flanked by tRNA gene and a transposon. A gene neighborhood comparison between JS1 and PCC 7421 is not straightforward because of the lack of conservation in gene synteny between the two species. Since JS1 lacks orthologous genes for the rhodopsin gene in PCC 7421, orthologs of neighboring genes that flank the PCC 7421 rhodopsin gene (gll0198) were sought in the JS1 genome. The gene flanking immediately to the left of the PCC 7421 rhodopsin gene is a tRNA encoding gene (gvit003). Further left flanking genes (gll0194, gll0195, glr0196, and gll0197) have orthologs in JS1, but the gene order is reversed (GKIL_2504, GKIL_2503, GKIL_2502, and GKIL_2501) (Figure 4.21). Immediate right-flanking genes (gsl0199, glr0200, and glr0201) in PCC 7421 have no orthologs in JS1. gll0202 in PCC 7421 is annotated as 'pilin gene inverting protein', and it has several orthologs in JS1, all annotated as transposases. Thus, this region in *G. violaceus* PCC 7421 seems to be a horizontally transferred region due to the presence of a tRNA gene and transposase, hallmarks of genetic hotspots in bacteria [185, 186]. A BLAST search of gll0198 against the nr database yielded top BLAST hits from diverse range of taxa with no clear representation of one phylum, thus further indicating the promiscuous nature of this gene.

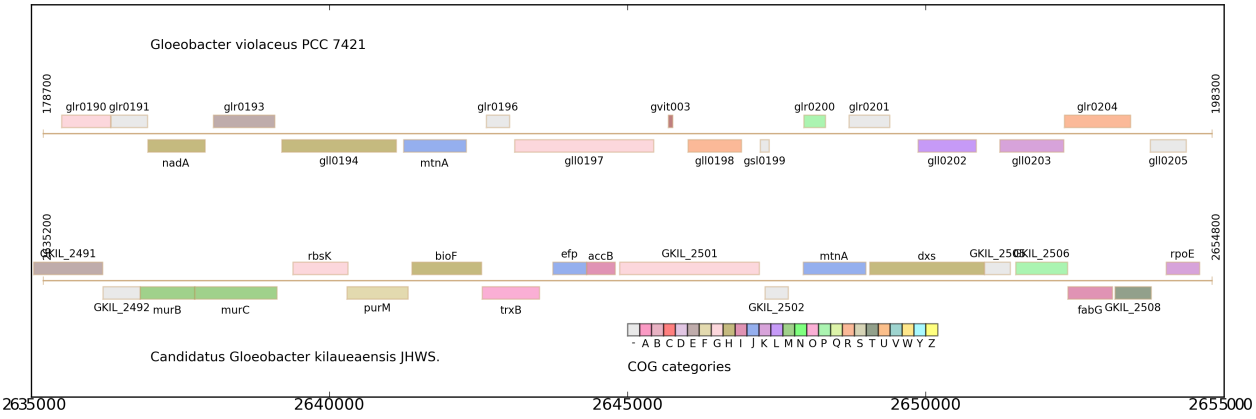


Figure 4.21. Comparison of the rhodopsin gene neighborhoods in *G. violaceus* PCC 7421 and *Candidatus* G. kilaueaensis JS1

.

4.4.7 *Cyanobacteria* and *Gloeobacter* phylogeny and evolution

4.4.7.1 Placement of *Candidatus* Gloeobacter kilaueaensis JS1 in the cyanobacteria lineage

Candidatus G. kilaueaensis JS1 shares 98.7%, 98.6%, and 98.6% 16S rDNA sequence identity respectively with *G. violaceus* PCC 7421, VP3-01, and PCC 8105. A maximum likelihood phylogenetic tree based on these 16S rRNA gene sequences and with that of *Beggiatoa alba* B18LD as an outgroup, revealed that *G. violaceus* places deeper along the cyanobacterial lineage than *Candidatus* G. kilaueaensis JS1 (Figure 4.22). Outgroup selection is known to affect the topology of phylogenetic trees, and *Beggiatoa* was used here as an outgroup because it has the shortest distance to the cyanobacteria clade, and gives a more accurate tree topology than other outgroups, which also results in the *Gloeobacter* being near the root of the cyanobacterial lineage [49]. The 16S rRNA gene tree was constructed with the intent of identifying cyanobacteria closely related to *Candidatus* G. kilaueaensis JS1. Though limited in sequence variability, the availability of 16S rRNA gene sequences from a vast number of cyanobacteria allows us to trace the evolutionary lineage of the *Gloeobacter* clade. Some sequences mentioned in Couradeau et al. [43] were also included in the tree to determine if the intra-cellular carbonate forming cyanobacteria clade (*Candidatus* Gloeomargarita lithophora) branches more deeply than the *Gloeobacter* in this newly proposed order Gloeobacterales. The 16S rDNA phylogenetic tree shows that the clade including *Gloeomargarita* actually forms its own group distinct from the *Gloeobacter*, and closer to thermophilic *Synechococcus* than to *Gloeobacter* (Figure 4.22).

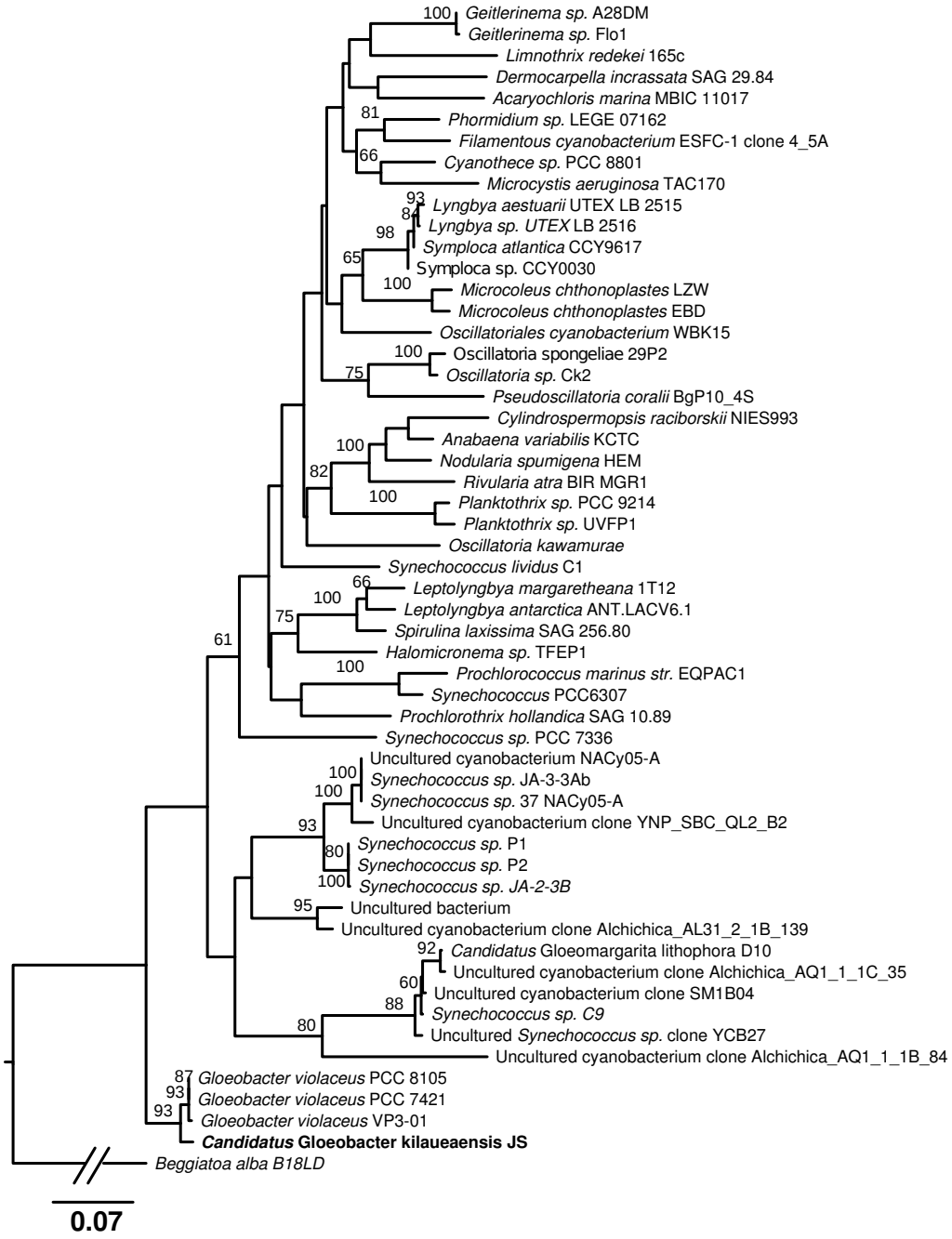


Figure 4.22. Maximum likelihood phylogenetic tree based on 16S rRNA gene sequences. Sequences were aligned in Muscle, edited with Gblocks, and the maximum likelihood tree inferred using the RAxML program with 100 bootstrap replicates and the GTR + Γ model of rate substitution. The root of the tree was shortened to fit the figure. Sequences in the tree were the top 50 BLASTn hits to the *Candidatus* G. kilaueaensis JS1 16S rRNA gene sequence, plus clones from Couradeau et al. [43] that were recently included in a new proposed order Gloeobacterales.

4.4.7.2 Phylogeny of *Candidatus* G. kilaueaensis JS1 with respect to completely sequenced cyanobacteria genomes

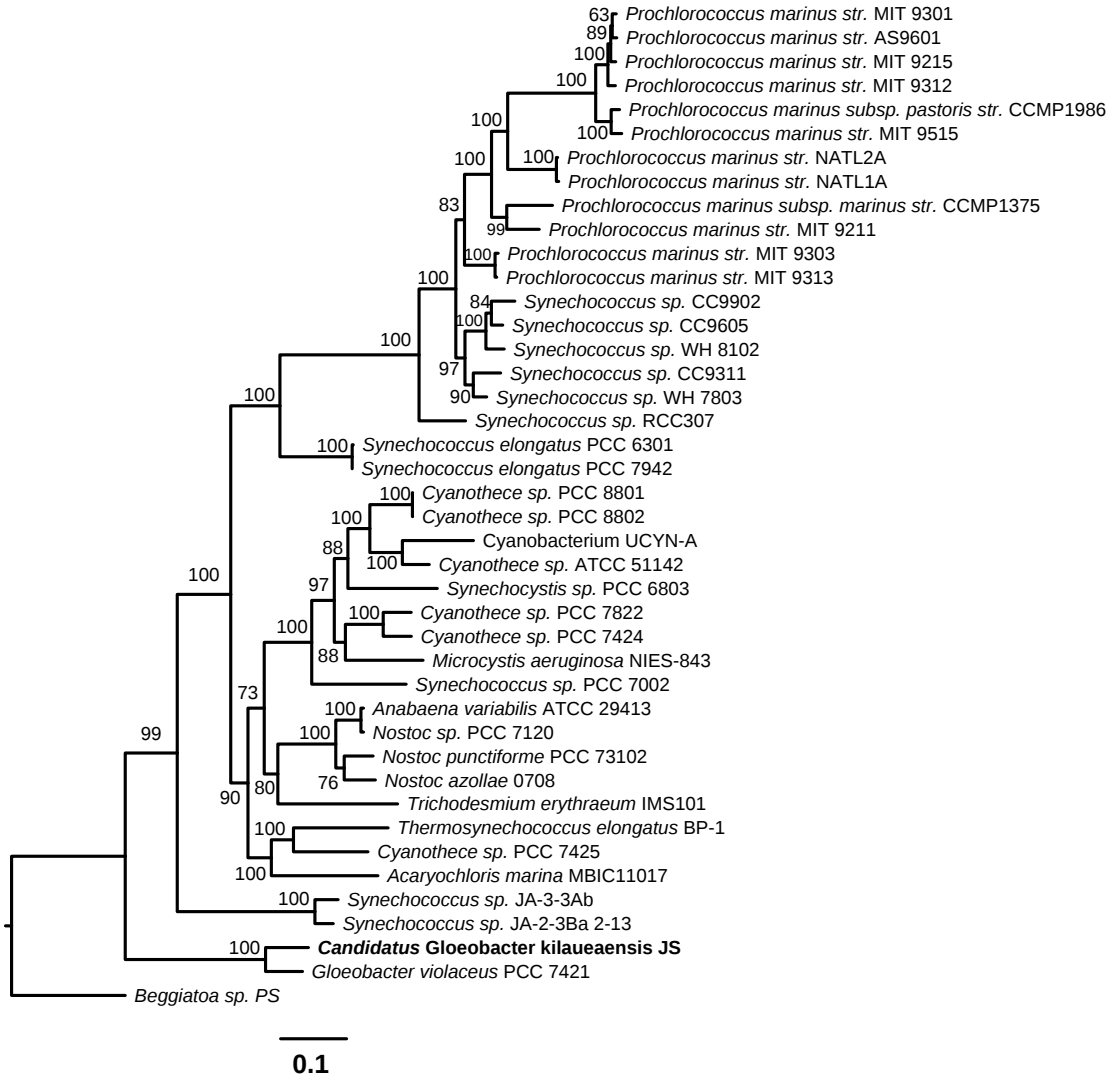


Figure 4.23. Phylogenetic tree based on 43 concatenated ribosomal proteins found in 41 cyanobacteria and the *Beggiatoa* outgroup. List of 43 ribosomal proteins identified in each genome, concatenated, aligned using Muscle, and edited with Gblocks. Based on 5359 aligned characters, maximum likelihood phylogenetic tree constructed using the RAxML program, Γ+WAG model of amino acid substitution, and 100 bootstrap replicates.

In order to better resolve the lineage of *Candidatus* G. kilaueaensis JS1 in the cyanobacteria clade, ribosomal proteins present in the 41 cyanobacteria genomes and the (*Beggiatoa*) outgroup were aligned, concatencated, and the maximum likelihood phylogenetic tree was constructed. A phylogenetic tree based on 43 ribosomal proteins identified in the 41 completely sequenced cyanobacteria (including *Candidatus* G. kilaueaensis JS1) and *Beggiatoa* sp. PS also places *G. violaceus* PCC 7421 closer to the root than *Candidatus* G. kilaueaensis JS1 (Figure 4.23). Whole-genome phylogenetic trees of cyanobacteria genomes in previous studies largely agree with the tree topology here [187, 49, 188].

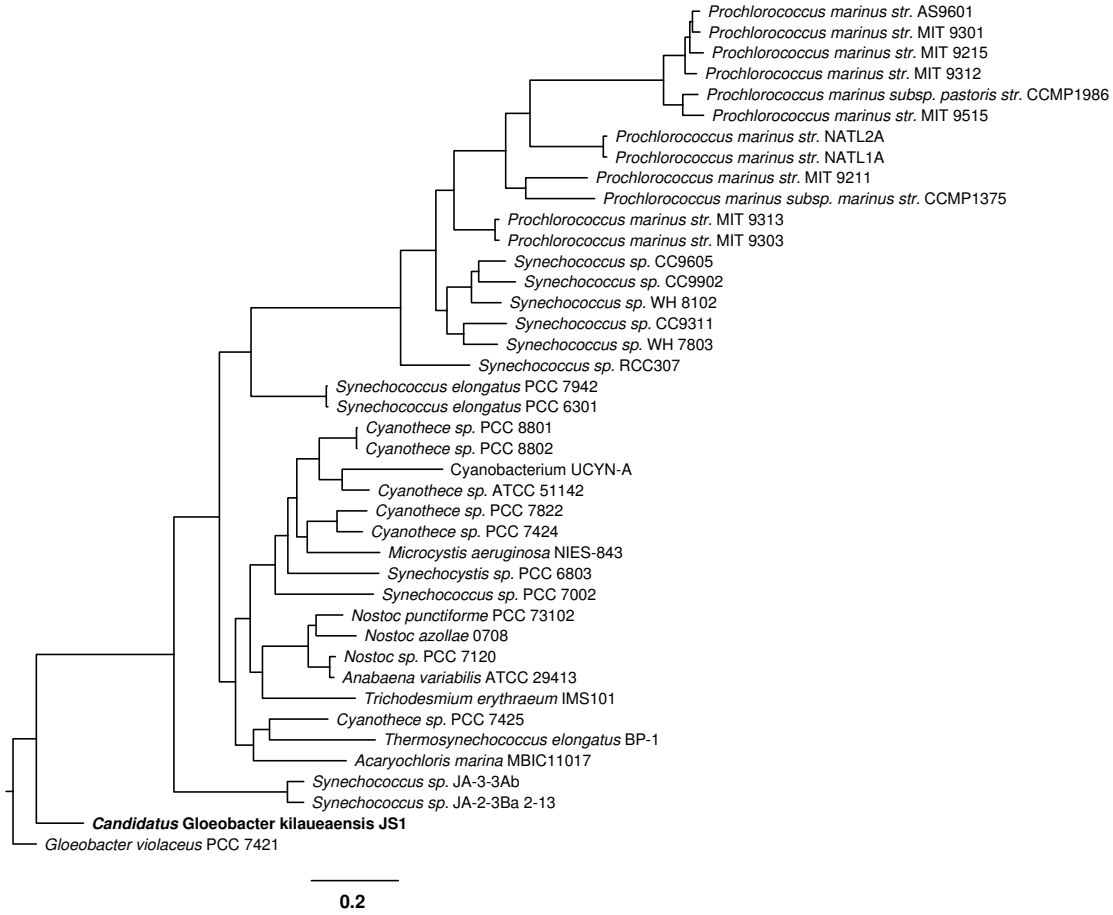


Figure 4.24. Phylogenetic tree based on 529 orthologous genes identified among 41 *Cyanobacteria*. Single representative copies of each of 529 genes were concatenated, aligned using MAFFT [189], and edited with Gblocks. A maximum likelihood phylogenetic tree was constructed in the RAxML program using the Γ+WAG model of amino acid substitution and 100 bootstrap replicates. All nodes are supported by 100% bootstrap values.

An additional tree was built using 529 aligned and concatenated amino acid sequences of orthologous genes identified as all present among the 41 cyanobacteria genomes compared (Figure 4.24). For this tree, *G. violaceus* PCC 7421 was used as an outgroup based on the assumption it is the most basal in the cyanobacteria lineage as determined by 16S rDNA and ribosomal protein phylogenetic trees. The purpose of this tree is to further test the placement of JS1 among all the completely sequenced genomes of cyanobacteria, and to see if better resolution of species lineage is achieved by comparing all shared genes present in the genomes. The tree topology and placement of different cyanobacteria within this tree is almost identical to that of the ribosomal protein tree, although variations in branch lengths were observed (Figure 4.24). This proves that ribosomal proteins are good indicators for species deliniation and could be used to trace the evolutionary history of a certain lineage of bacteria.

4.4.7.3 Divergence time of *Candidatus* Gloeobacter kilaueaensis JS1 and *Gloeobacter violaceus* PCC 7421 from their last common ancestor

Ribosomal proteins tend to be present in single copies, are usually conserved enough to be good gene markers, and are less likely to be horizontally transferred. To identify evolution and divergence to *Candidatus* G. kilaueaensis JS1 from the last ancestor of the Gloeobacterales, the Monte Carlo Markov Chain (MCMC) analysis was used to calculate the divergence times of 41 completely sequenced cyanobacteria from the *Beggiatoa* outgroup. Species divergence time was calculated using the PAML package [152], and the tree shows *Candidatus* G. kilaueaensis JS1 diverged from *G. violaceus* 153 million years ago (MYA) (Figure 4.25). The divergence time between the ancestor of *Gloeobacter* and *Beggiatoa* sp. PS was calculated to be $\sim$658 MYA.

In addition to the MCMC tree calculation using the PAML package, the PATHd8 program [190] was used to calculate divergence times among the cyanobacteria. MCMC calculation from the PAML package uses Bayesian statistics to calculate divergence time, but PATHd8 uses a different algorithm that calculates note ages locally, and allows for calculation of larger trees [190]. PATHd8 calculations showed the divergence time between *Candidatus* G. kilaueaensis JS1 and *G. violaceus* PCC 7421 to be 324 $\pm$ 24.9 MYA. This is $\sim$150 million years earlier than that computed in the MCMC tree calculation in the PAML package (153 MYA).

To obtain consistent and reliable estimates of divergence time between these two organisms, three other programs will be used to provide other time frames, specifically BEAST [191], Phylobayes [192], and MrBayes [193]; the aim was to at least find some consistent estimates among

different methods. The results from these three programs will be presented as soon as they become available.

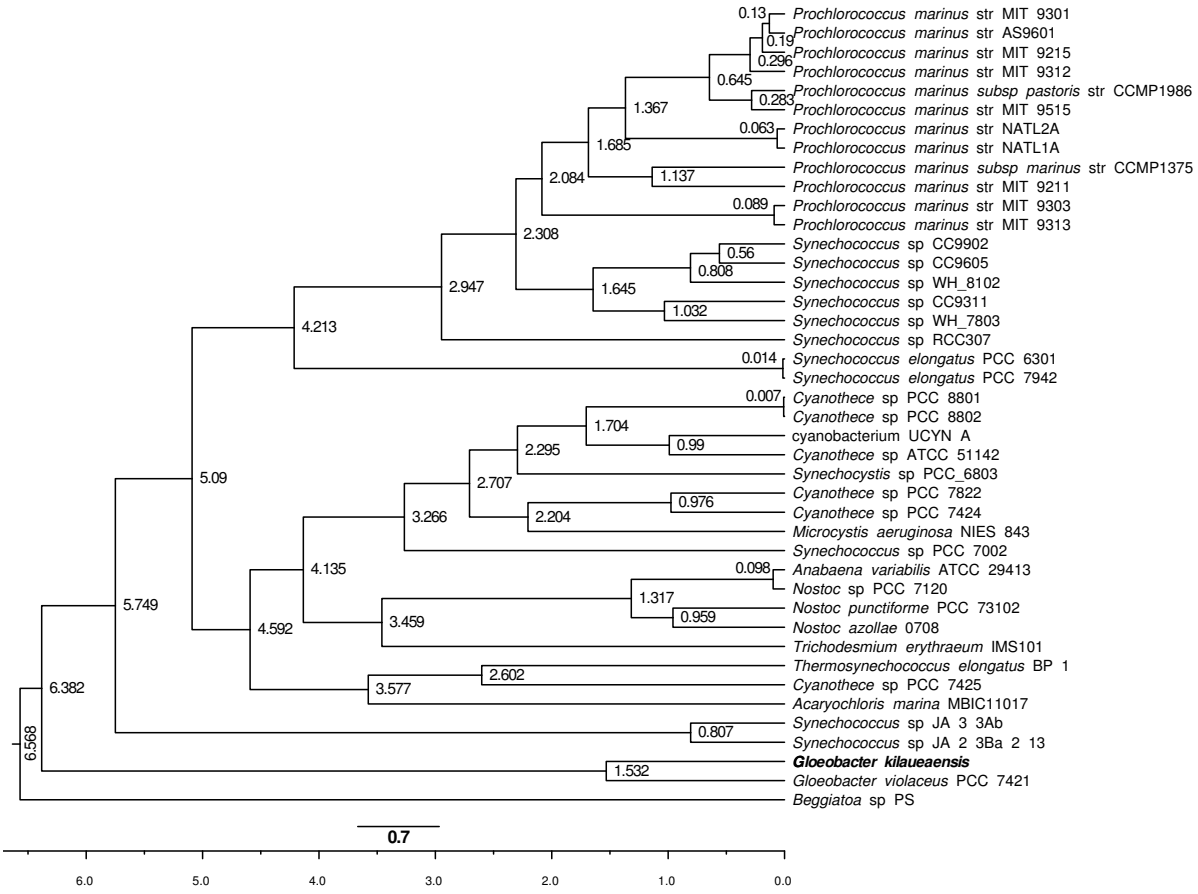


Figure 4.25. MCMC tree showing divergence times in the cyanobacteria lineage. The tree was built using 43 concatenated ribosomal proteins aligned with Muscle, edited with Gblocks, and divergence time calculated using CODEML and MCMCTREE from the PAML package. MCMC runs were repeated until results showed consistent divergence time with each iteration. Numbers near the nodes specify approximate divergence time in hundreds of million of years. The tree was calibrated using a previously calculated divergence time between *Prochlorococcus* and *Synechococcus* of 150 million years [194].

4.4.7.4 Gene gains and losses along the *Cyanobacteria* phylum

To map genes gained and lost along the cyanobacterial lineage, phyletic patterns were first compiled based on presence or absence of 13,655 orthologous genes identified among the 41

cyanobacteria compared. These phyletic patterns were then uploaded to the Gain Loss Mapping Engine (GLOOME) server (http://gloome.tau.ac.il/) [153] to calculate gene gain and loss events using a stochastic mapping approach [195]. The goal of this analysis was to detect genes gained or lost during the evolutionary history of the last common ancestor of two *Gloeobacter* spp. that led to the emergence of two *Gloeobacter* species. The analysis revealed that *Candidatus* G. Kilaueaensis JS1 gained 493 and lost 363 genes from the node branching from *G. violaceus* PCC 7421 (Figure 4.26). The genes gained or lost are not shown in this work but will be included as supplemental tables as part of the manuscript submitted to a peer-reviewed journal.

Figure 4.26. Gene gain/loss events in the cyanobacteria lineage. Phylogenetic tree built by stochastic mapping of phyletic patterns representing gene gains or losses. Scale bar represents the number of gain events, and branch length represents gain events. Numbers in blue indicate gene gains, those in red indicate gene losses.

4.4.8 Comparative genomic analyses

4.4.8.1 Gene synteny and genomic rearrangements

There was a surprising lack of synteny between the *Candidatus* G. kilaueaensis JS1 and *G. violaceus* PCC 7421 genomes (see DDH comparison in Section 4.4.5). Comparison of gene synteny and genome rearrangements between *Candidatus* G. kilaueaensis JS1 and *G. violaceus* are shown in Figures 4.27 and 4.28. Despite a 16S rRNA gene sequence identity of 98.7%, very little in the respective genomes was conserved (Figure 4.27). One would usually expect to see conserved gene synteny and a large block of colinear genomic regions in closely related bacterial species or strains (*e.g.*, Figure 5.4). This was not the case with *Candidatus* G. kilaueaensis JS1 and *G. violaceus* PCC 7421; JS1 appearing to have gone through a considerable genome rearrangements.

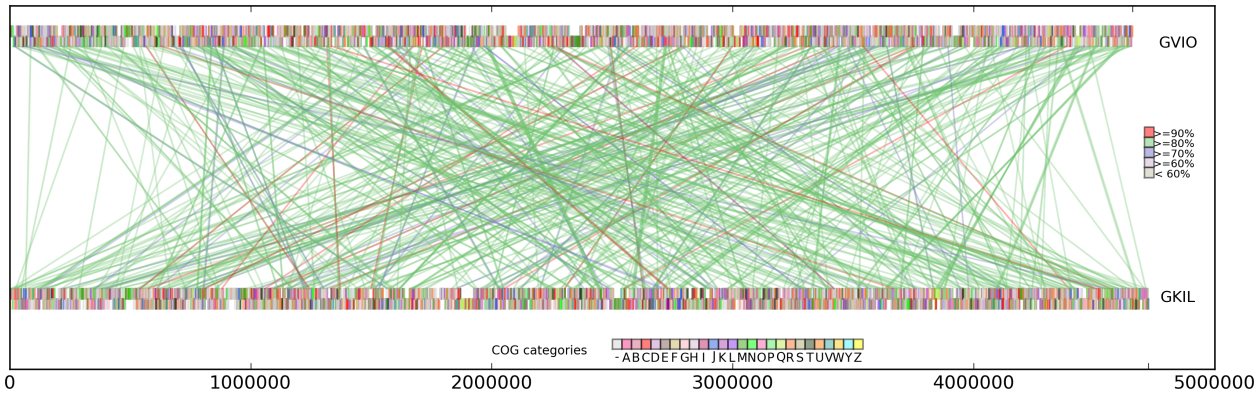


Figure 4.27. MUMMER alignment between JS1 and PCC 7421. Colored rectangular blocks represent protein coding sequences according to COG functional categories. Lines represent matching DNA segments between the two genomes. Colors of connecting line segments are categorized according to % identities. This plot was generated by a custom Python script (See section 5.1.12).

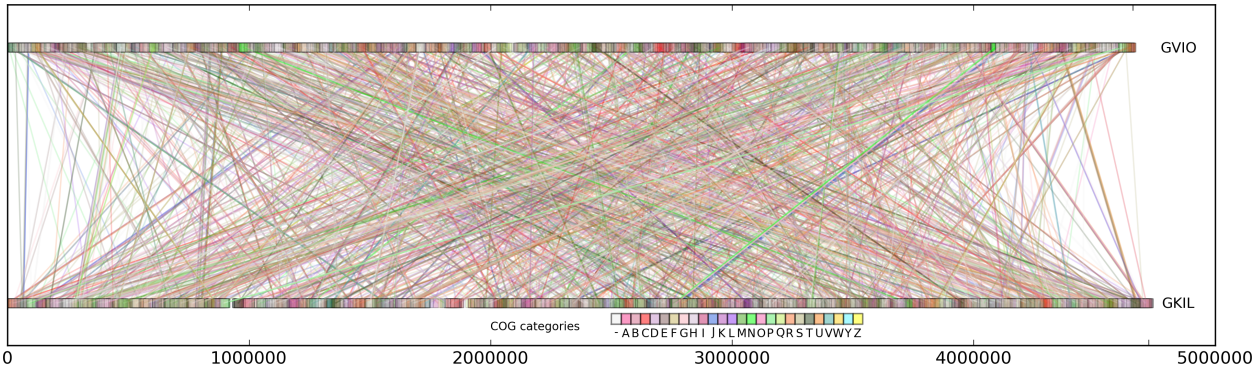


Figure 4.28. Shared orthologs identified between JS1 and PCC 7421 and their locations in the genomes. Lines connect orthologous genes and are colored according to COG functional categories. BLASTp E-values between bi-directional best hits are less than or equal to $1e^{-5}$. This plot was generated by a custom Python script (See section 5.1.22).

4.4.8.2 Ecophysiological roles of different cyanobacteria

Using orthologous groups identified by the OrthoMCL program, the presence or absence of the same orthologous groups was counted in each of the 42 compared cyanobacteria genomes. A 13655×42 matrix of '1's and '0's representing presence or absence was constructed, and the Pearson correlation coefficient calculated in the R statistical analysis package. Results were visualized as a clustered heatmap (Figure 4.29). This approach has been shown to be useful in understanding niche specialization in studies where complete genomes of *Bacteroidetes* were compared, and showed strong correlation and clustering of bacteria adapted to different lifestyles, *e.g.*, anaerobic oral pathogens, endosymbionts of insects, or nearshore decomposers [196]. Using this approach the distinct clusters formed by different strains of (marine) *Prochlorococcus* and *Synechococcus*, and freshwater *Synechococcus* are clear (Figure 4.29). *Candidatus* G. kilaueaensis JS1 and *G. violaceus* PCC 7421 grouped tightly in a cluster and separate from other cyanobacteria. Despite the two genomes having gone through large-scale rearrangements, they still share a large number of orthologous groups, and seem to perform similar functions based on a comparison of orthologous groups of genes.

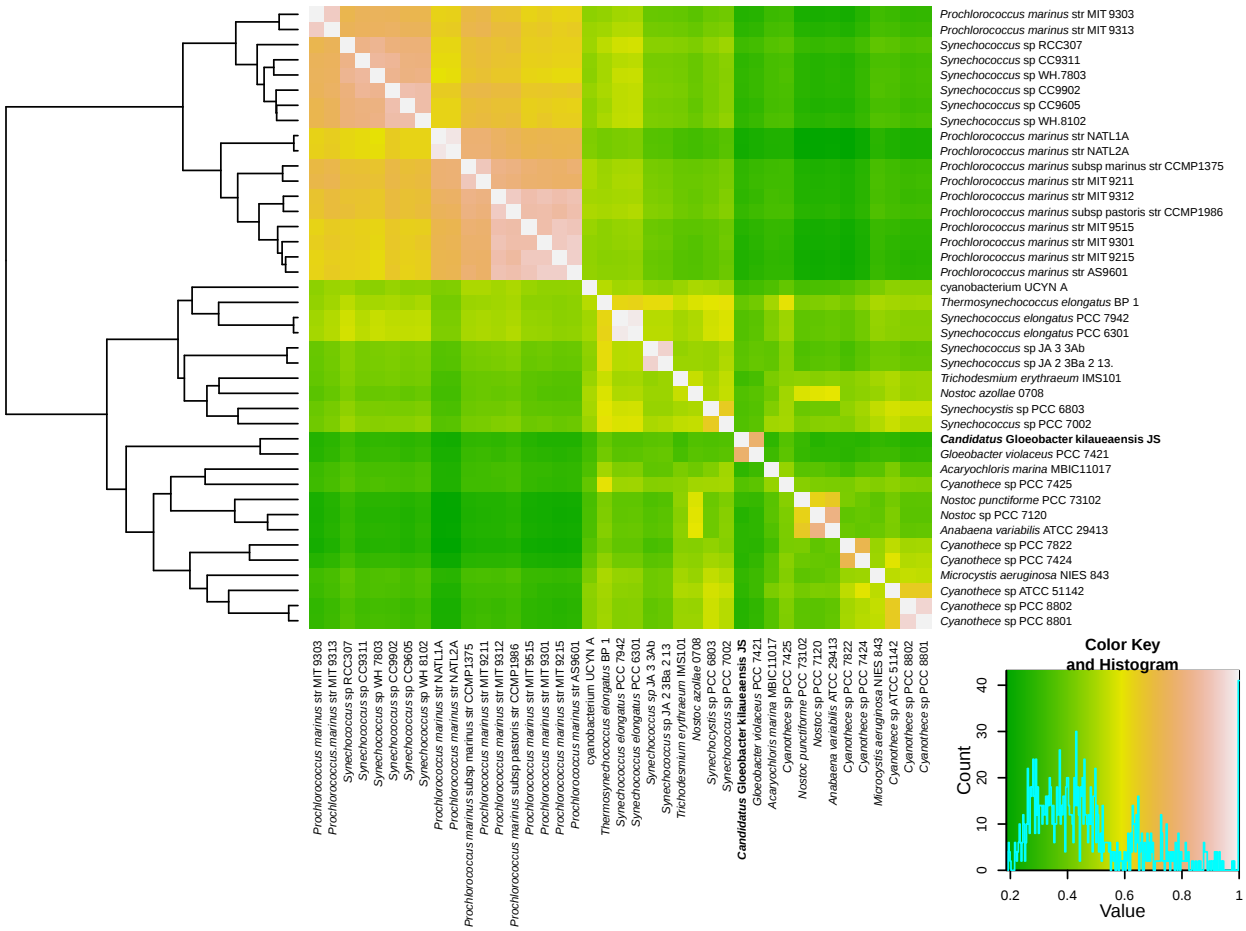


Figure 4.29. Hierarchical clustered heatmap portraying a comparison of completely sequenced cyanobacterial genomes. A correlation matrix based on presence or absence of 13,655 orthologous groups identified in 41 cyanobacteria was first created. The 13,655 x 41 matrix was imported into the R program, and Pearson correlation coefficients were calculated using the 'gplots' package.

4.4.9 Recruitment of *Gloeobacter* reads from the cave biofilm metagenome

The final assembled *Candidatus* G. kilaueaensis JS1 genome was used to 'recruit' *Gloeobacter*-specific reads from the HAVO epilithic biofilm metagenome described above (Chapter 2). Recruitment using the NUCMER script from the MUMMER aligner identified 3474 unique metagenomic reads (20,433 unique reads with BLASTn using relaxed parameters); only 596 unique reads (19,101 reads with BLASTn using relaxed parameters) were recruited using *G. violaceus* PCC 7421 as the reference genome. Note that the BLASTn parameters used were relaxed to recruit reads that had sequence identities as low as ~60% in order to recover reads from distantly related organisms.

The two recruitment plots showed recruitment using the *G. violaceus* PC 7421 genome yielded mostly reads that matched with less than 90% sequence identity (Figures 4.30 and 4.31).

This result indicates the need for more reference genomes in public databases because even at 98.7% 16S rRNA sequence identity, *G. violaceus* PCC 7421 is not the perfect organism with which to extract all sequences belonging to the genus *Gloeobacter* from metagenomic sequences. This also highlights a need for more reference genome sequences of rare but still important organisms, and especially for their genomes to be completely sequenced.

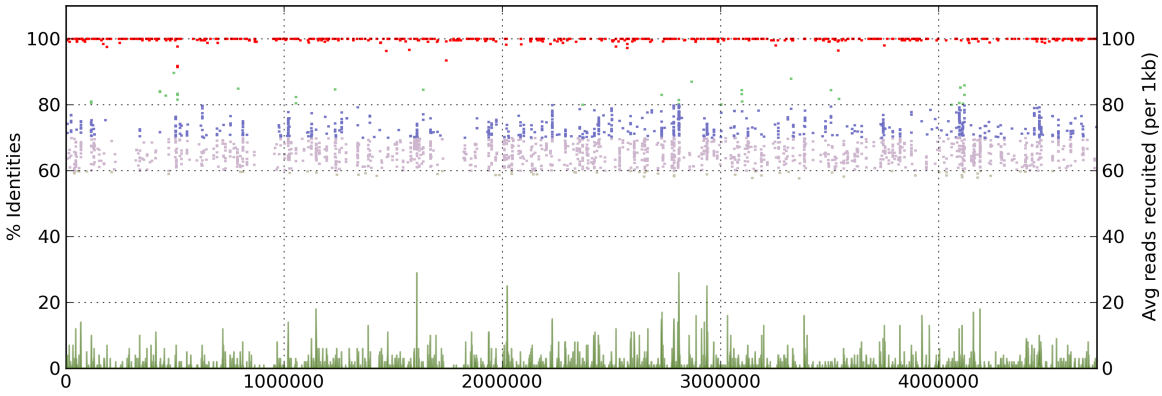


Figure 4.30. Fragment recruitment plot of *Candidatus* G. kilaueaensis JS1 against the cave epilithic biofilm metagenome. The *Candidatus* G. kilaueaensis JS1 genome was searched against the metagenome data set of the HAVO biofilm sample using BLASTn to recruit reads that may be of *Gloeobacter* in origin. Identities are color-coded: Red (≥90%), green (≥80%), blue (≥70%), lavendar (≥60%), grey (<60%). BLASTn parameters were relaxed to recruit reads with identity as low as 60%. Plot generated by a custom Python script (See Section 5.1.24)

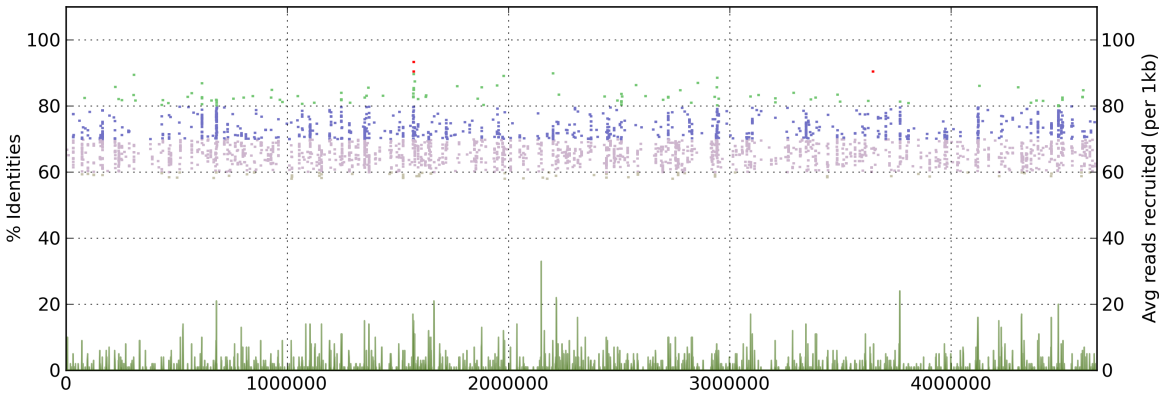


Figure 4.31. Fragment recruitment plot of *G. violaceus* PCC 7421 against the cave epilithic biofilm metagenome. The *G. violaceus* PCC 7421 genome was searched against the metagenome data set of the HAVO biofilm using BLASTn to recruit reads that may be of *Gloeobacter* in origin. Note the recruitment of reads falling mostly within 60 to 80% nucleotide sequence identity.

4.5 Conclusions

A new *Gloeobacter* sp. was isolated from an epilithic biofilm in the indirectly illuminated entance of a lava cave in Kilauea Caldera, Hawai‘i. This is only the second *Gloeobacter* species to be isolated, 38 years after *G. violaceus* PCC 7421, the Type strain of the species, genus, family and class was published [107]. Assembly of the complete genome from DNA extracted from an enriched but non-axenic culture resembling a low-complexity metagenome was completed, allowing comparison of the *Candidatus* Gloeobacter kilaeuaensis JS1 complete genome with the only known reference genome of a *Gloeobacter*, that of *G. violaceus* PCC 7421. *In silico* DDH analyses revealed the *Gloeobacter* sp. isolated to be a new species distinct from *Gloeobacter violaceus* PCC 7421 and for which the name *Candidatus* Gloeobacter kilaeuaensis JS1 is proposed.

Despite the genome of *Candidatus* G. kilaeuaensis JS1 showing little synteny with that of *G. violaceus* PCC 7421, the gene contents of the two organisms are comparable, and they share the largest number of orthologous genes between them rather than with other cyanobacteria. Phylogenetic trees (16S rRNA gene, 43 concatenated ribosomal proteins, and 529 concatenated shared orthologous genes) placed *Candidatus* G. kilaeuaensis JS1 in the same deep branching, monophyletic clade as *G. violaceus* PCC 7421, but the latter seems to be more deeply-branching than *Candidatus* G. kilaeuaensis JS1. Using amino acid sequences from 43 concatenated ribosomal proteins, the divergence time between different cyanobacterial species was calculated, providing an estimate of

153 to 324 million years for the divergence of *G. violaceus* and *G. kilaueaensis*. Metabolic pathway analysis revealed minor differences between the two *Gloeobacter* genomes.

Chapter 5

Bioinformatics Work

Custom tools were needed for the data analyses used here, and to visualize the results of the metagenomic and genomic analyses presented. Considerable time has been expended in learning how to use programming languages in such work. Although writing these scripts is mostly technical, they represent considerable time in actually formulating the problem, developing approaches to analyze specific types of data, testing approaches, and then applying those approaches to actual data. Analyzing the sometimes specifically formatted outcomes also requires a steep learning curve. A chapter devoted to explaining the rationale for the development of these tools is appropriate.

Scripts I developed and utilized are presented here since they may be of use to other researchers in the field. Some tools/scripts are highly specific to the context of data analysis, while others are more general in that the tasks they perform can be applied to common biological data analyses (such as format conversion), and they may thus be of use for someone in search of a quick solution to a bioinformatic data analysis problem. Most of the tools developed here were written in the Python programming language, while some were in Ruby, or Bash shell scripting languages, depending on the task at hand.

A detailed list of scripts written is included (Appendix B). It is important to describe in detail the scripts/codes written for data analysis but because of space limitations it is most feasible to deposit them in an online repository and direct readers to a particualr URL. Therefore, all the scripts are written on the Google Code repository at: http://code.google.com/p/jimmysawdissertation/source/browse/trunk/dissertation (Full source codes can be navigated through the directories listed on the left of the page). Some scripts written are described in detail in this dissertation. It is important generally in this field to produce figures of the type used in this dissertation, especially for publication. Some of these scripts will likely be developed further into graphical interfaces that will be user-friendly for biologists. This would

permit creation of publication-quality figures that are often difficult to produce without knowing specialized bioinformatics or graphical tools.

5.1 Scripts for analysis of the *Candidatus* Gloeobacter kilaueaensis JS1 genome

Custom Python, Ruby, or Bash scripts written for the analysis of *Candidatus* Gloeobacter kilaueaensis JS1 genome, including scripts that are both complete and incomplete (work in progress) are shown in (Table 5.3). Some important scripts and examples of their use are explained in detail in the sections below.

Table 5.1. Bioinformatic scripts used in the analysis of the *Candidatus* Gloeobacter kilaueaensis JS1 genome

Number	Script name	Language
1	dissertation_BlastnRetrieveTopHits.py	Python
2	dissertation_BLASTPLineageVotes.py	Python
3	dissertation_CheckBLASTPLength.py	Python
4	dissertation_CheckGenes.py	Python
5	dissertation_CheckPathwayModules.py	Python
6	dissertation_CogSummary.py	Python
7	dissertation_CompareGenes.py	Python
8	dissertation_ConcatConvertMSA.py	Python
9	dissertation_ConvertAlignment.py	Python
10	dissertation_CountPhymmBL_Phyla.sh	Bash
11	dissertation_CountSharedOrthologs.py	Python
12	dissertation_CreateOrthologMatrix.py	Python
13	dissertation_CyanoOrthologsMSA.py	Python
14	dissertation_DigitalPCR.py	Python
15	dissertation_DomainParser.py	Python
16	dissertation_DownloadGenomes.py	Python
17	dissertation_DrawGenesArrows.py	Python
18	dissertation_DrawGenes.py	Python
19	dissertation_DrawGeneswithPtt.py	Python
20	dissertation_DrawMUMMER.py	Python
21	dissertation_DrawMUMMERwithPtt.py	Python
22	dissertation_DrawMUMMERwithPttZoomRegion.py	Python
23	dissertation_ECfromKEGG.py	Python
24	dissertation_GapCloserMinimo.py	Python
25	dissertation_GCskew.py	Python
26	dissertation_GeneNamesfromKEGG2.py	Python
27	dissertation_GeneNamesfromKEGG.py	Python
28	dissertation_GenerateCircosTracks.py	Python
29	dissertation_GenerateCircosTracksReadsCoverage.py	Python
30	dissertation_GloeoAsmVerification.py	Python
31	dissertation_IgsBlast.py	Python
32	dissertation_IlluminaCoverage.py	Python
33	dissertation_IndividualGeneTrees.sh	Bash
34	dissertation_KeggModule.rb	Ruby
35	dissertation_KeggOrthologInfo.py	Python
36	dissertation_NewblerContigScaffold.py	Python
37	dissertation_NewblerFilledScaffolds.py	Python
38	dissertation_OrthologMatrix.sh	Bash
39	dissertation_OrthologsTreeIndividual.sh	Bash
40	dissertation_ParseOverlappingMatePairs.py	Python
41	dissertation_PhymmBLParser.py	Python
42	dissertation_PlotContigQuality.py	Python
43	dissertation_PrimerPicker.py	Python
44	dissertation_ReciprocalBestHitPlot.py	Python
45	dissertation_ReciprocalBestHitPlotWithPtt.py	Python
46	dissertation_RecruitmentPlotBlast.py	Python
47	dissertation_RecruitmentPlot.py	Python
48	dissertation_RenameOrthomclCompliant.py	Python
49	dissertation_RibosomalGenesIndividual.sh	Bash
50	dissertation_RibosomalGenes.sh	Bash
51	dissertation_SingleCopyGenes.sh	Bash
52	dissertation_TopBlastRank.py	Python

5.1.1 dissertation_BlastnRetrieveTopHits.py

This script was written to automatically perform BLAST and retrieve top n hit organisms from a given fasta sequence file. See Appendix D.1 for full code.

```
Example usage:
dissertation_BlastnRetrieveTopHits.py test.fasta 10
```

5.1.2 dissertation_CheckBLASTPLength.py

This script was written to check the length of a protein (to see if it is within an expected size range) by comparing its length to hits from the BLASTp search. This is necessary because if the length of the ORF predicted is much shorter or longer than of the top hits, the ORF predicted may be truncated (too short to be functional or should be annotated as pseudogene) or bifunctional (two functional domains fused together in a single ORF). The start site may need to be adjusted as the ORF may be a non-functional pseudogene.

```
Example usage:
python dissertation_CheckBLASTPLength.py GKIL_3100.refseq.blastp.tbl

This prints:
GKIL_3100 95.614
```

5.1.3 dissertation_CheckGenes.py

This script was written to check any number of genes contained within a given start and stop coordinates in a genome. The script takes Genbank file, taxonomic classification file, and integers as input. The goal of this script is to quickly identify genes within a genomic region and to determine their taxonomic affiliation, *i.e.*, are they cyanobacterial in origin or not? The script relies on output from another script that parses BLAST results of ORFs to assign taxonomic affiliation to each ORF.

```
Example usage:
python dissertation_CheckGenes.py GKIL.v6.gbf GKIL.v6.tophits_class.txt 10000 20000
```

```
This prints:
GKIL_0010 Gloeobacteria          glutathione synthetase
GKIL_0011 Gloeobacteria          hypothetical protein
GKIL_0012 Gloeobacteria          carboxylate-amine ligase
GKIL_0013 Gloeobacteria          benzoyl-CoA oxygenase/reductase, BoxA protein
GKIL_0014 Gloeobacteria          phosphoribulokinase
GKIL_0015 Gloeobacteria          transketolase
GKIL_0016 Gloeobacteria          single-stranded DNA-binding protein
GKIL_0017 Gloeobacteria          nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase
GKIL_0018 no BLAST hit           hypothetical protein
GKIL_0019 Gloeobacteria          succinate dehydrogenase iron-sulfur subunit
```

5.1.4 dissertation_CombineFastq.py

This script was written to combine paired end Fastq files that came with Illumina sequencing technology. It takes 2 files as input and combines them into 1 resulting Fastq file and removes trailing 'B's that represent low quality sequence towards the end of each read.

```
Example usage:
dissertation_CombineFastq2BtrimmedFastaQual.py file1.fastq file2.fastq new.fastq
```

5.1.5 dissertation_CompareGenes.py

This script was written to compare the gene neighborhood between two genomes. It takes GenBank files of two organisms and a COG categories file. See Figure 5.1. This example shows genes next to the *psbA* gene involved in photosynthesis between *G. violaceus* PCC 7421 and *Candidatus* G. kilaueaensis JS1. See Appendix D.2 for full code.

```
Example usage:
dissertation_CompareGenes.py ../annotation/GKIL.v6.gbf NC_005125.1.gbk NC_005125.ptt
orthologs/orthomcl/cogs.t.list 773000 783000  2814800 2824800
```

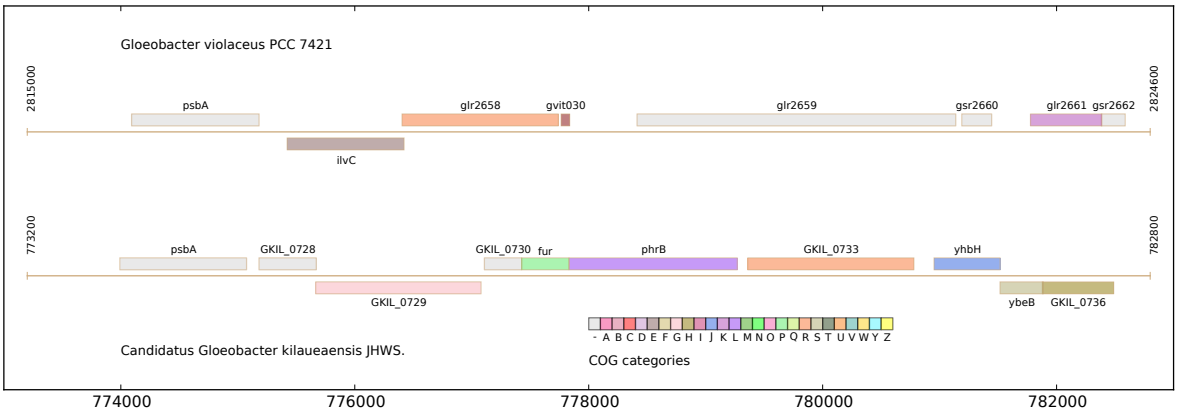


Figure 5.1. Example figure showing comparison of gene clusters between *G. violaceus* PCC 7421 and *Candidatus* G. kilaueaensis JS1.

5.1.6 dissertation_ConcatConvertMSA.py

This script was written to concatenate and convert individual Gblocks-edited multiple sequence gene alignment files to Phylip format.

```
Example usage:
dissertation_ConcatConvertMSA.py r43.list cyano.list
```

5.1.7 dissertation_ContigQualityPlot.py

This script was written to draw a plot of quality of a given contig and to see how many bases fall below a given threshold. An example of such a drawing is in Figure 5.2

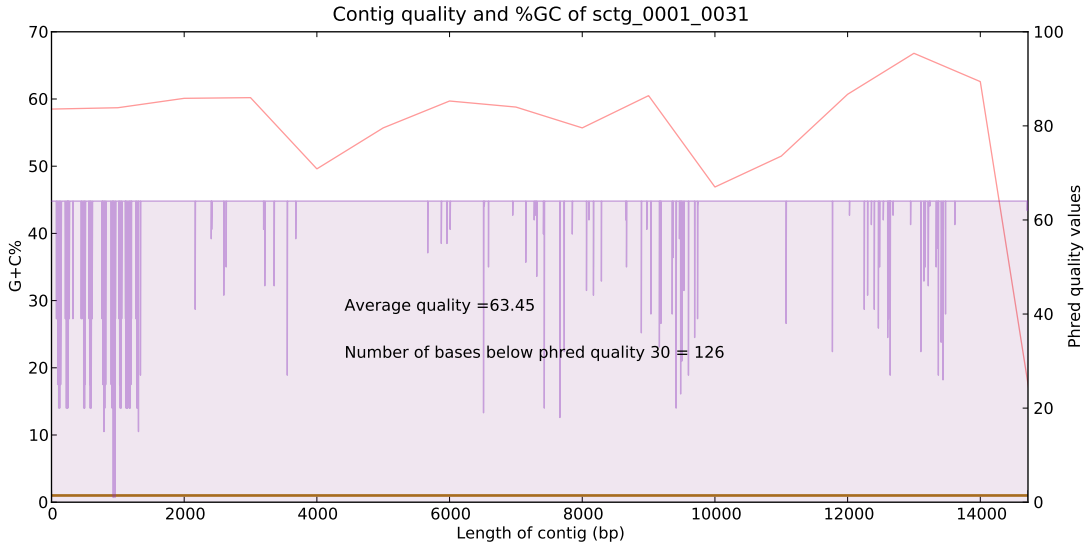


Figure 5.2. Example figure produced by dissertation_DrawGenes.py script, showing a contig produced by Newbler assembler and the average quality of the contig. Also shown is how many bases within the contig are below a given threshold of 30. Purple lines represent the contig quality at a given position within the contig (Phred quality values are on Y axis on the right). G+C% of 1000bp sequence window is plotted and shown as a connected x-y scatter plot in red and G+C% values are shown on Y axis on the left.

Example usage:

```
dissertation_ContigQualityPlot.py sctg_0001_0031.fasta sctg_0001_0031.qual 30
```

5.1.8 dissertation_ConvertAlignment.py

This script was written to quickly convert multiple sequence alignment formats.

Example usage:

```
dissertation_ConvertAlignment.py msa.fasta fasta msa.phy phylip
```

5.1.9 dissertation_CountSharedOrthologs.py

This script was written to count shared orthologs between two genomes. This script expects a file produced by the OrthoMCL program to calculate the orthologs.

```
Example usage:
dissertation_CountSharedOrthologs.py CYANO.orthologs.txt GKIL 58011
```

5.1.10 dissertation_DomainParser.py

This script was written to identify Pfam domains after a given ORF (amino acid sequence) has been searched against the Pfam database through RPS-BLAST. Need to run RPS-BLAST first with option to produce XML files, as the script expects XML format to parse results.

```
Example usage:
Go to /host/Users/JS/UH-work/gloeobacter/final_work/annotation/gkil_rpsblast
dissertation_DomainParser.py GKIL_4101.pfam.rpsblast.xml

This prints:
GKIL_4101 GST_N + GST_C
```

5.1.11 dissertation_DrawGenes.py

This script is needed to create custom gene diagrams such as the one shown in 5.3. It can be improved upon to create better gene diagrams for publication quality images. See Appendix D.3 for full code.

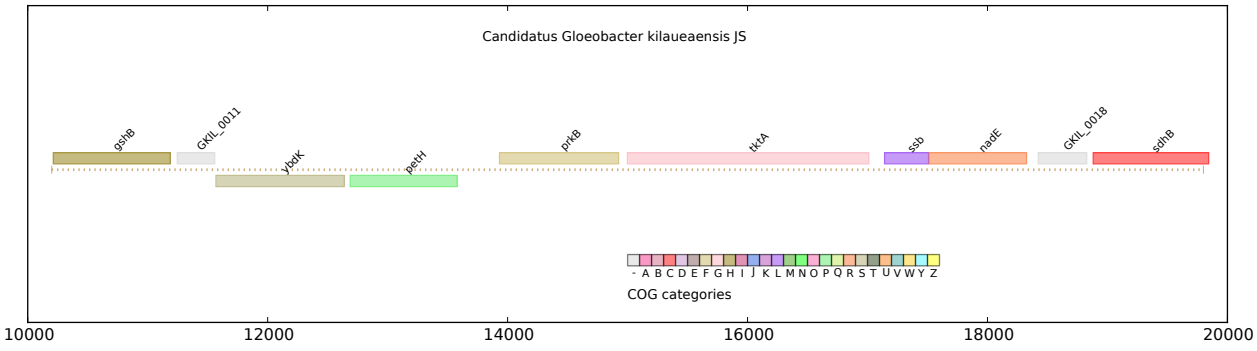


Figure 5.3. Example gene figure showing 10000 to 20000 region from *Gloeobacter* genome.

5.1.12 dissertation_DrawMUMMER.py

This script is needed to create custom MUMMER plots such as one shown in Figure 4.27. This script is special (improved visualization compared to native MUMMER visualization option using Gnuplot) because I have modified it to show protein coding regions (ORFs) color-coded by COG categories, and to show where DNA alignment between the genomes take place. See Appendix D.4 for full code.

```
Example usage:
dissertation_DrawMUMMER.py ../annotation/GKIL.v6.gbf NC_005125.1.gbk NC_005125.ptt
  orthologs/orthomcl/cogs.t.list GKIL_vs_GVIO.coords
```

5.1.13 dissertation_DrawMUMMERwithPtt.py

This script plots similar figures as 4.27 but the script was improved to utilize Genbank and Ptt files from NCBI to parse COG information. See Appendix D.5 for full code.

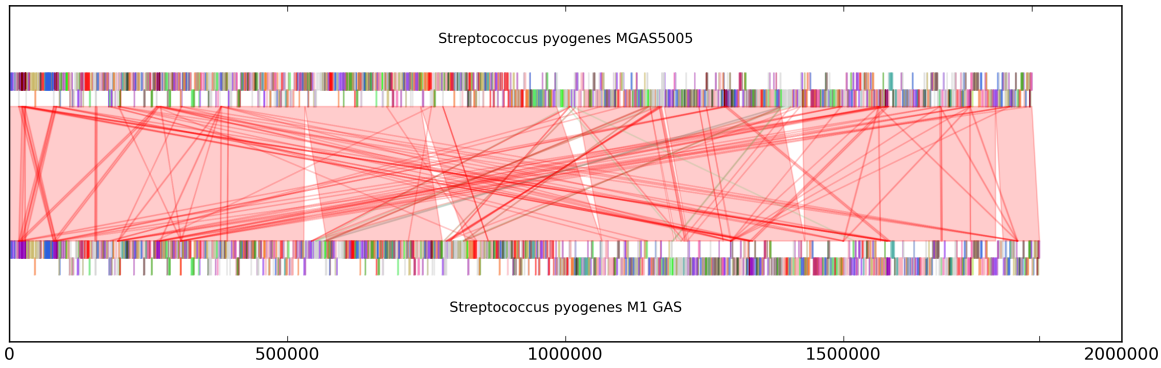


Figure 5.4. Genome alignment between two *Streptococcus pyogenes* strains showing conserved genomic blocks.

5.1.14 dissertation_GapCloserMinimo.py

This script can generate a list of reads generated from shreds of contigs (from Celera or other assemblers) spanning two contigs scaffolds to help close gaps between these scaffolds. This script only lists the gap-spanning reads and they need to be manually assembled with the Minimo program from the AMOS package. See Appendix D.6 for full code.

```
Example usage:
dissertation_GapCloserMinimo.py sctgs.list
```

5.1.15 dissertation_GCskew.py

This script is needed to create data points to draw GC skew plots in a genome circle diagram drawn with the Circos program. See Appendix D.7 for full code.

```
Example usage:
dissertation_GCskew.py GKIL.v6.gbf percent
dissertation_GCskew.py GKIL.v6.gbf skew
```

5.1.16 dissertation_GeneNamesfromKEGG.py

This script remotely retrieves gene names based on KEGG ortholog (KO) IDs. The idea is to automate discovery of gene names from annotated metagenomic data. Note that this script requests web services provided by KEGG database and could be slow if more than a few thousand sequences need to be processed.

```
Example usage:
dissertation_GeneNamesfromKEGG.py meta.ko.txt > meta.genes.txt
```

5.1.17 dissertation_GloeoAsmVerification.py

While the genome assembly produced by Newbler was fairly intuitive to navigate, no integrated visualization tool exists to check contig scaffolds and matepair distribution between the contig scaffolds. To solve this problem, a custom python script 'dissertation_GloeoAsmVerification.py' was written to show the problematic regions in the assembled genome. See Figure 4.3 for the plot produced by this script and Appendix D.8 for the full code.

```
Example usage:
dissertation_GloeoAsmVerification.py kl_vs_AtleastOneAndSingletons.coords
    glbkl_vs_non-gloeo.coords glbkl_vs_454scaffolds.coords
    glbkl_vs_celeractgs.coords binned.gloeo.pairs.txt ../annotation/fixed.final_assembly_noCN4.fasta
```

5.1.18 dissertation_IgsBlast.py

This script parses intergenic regions between ORFs, BLASTs them automatically, and saves the results. See Appendix D.9 for full code.

```
Example usage:
dissertation_IgsBlast.py annotation.tab sequence.fasta
```

5.1.19 dissertation_NewblerFilledScaffolds.py

This script was written to visualize Newbler assembly scaffolds to estimate number of mate pairs between contig scaffolds and to estimate gap sizes between the contigs. Consistent mate pairs are necessary to unambiguously link contigs to close gaps.

```
Example usage:
dissertation_NewblerFilledScaffolds.py 454Scaffolds.txt newbler_scaffolds_vs_mates.coords
```

5.1.20 dissertation_ParseOverlappingMatePairs.py

This program parses overlapping mate pairs between contig scaffolds. First, the paired-end 454 sequence fasta file needs to be aligned against contig scaffolds using MUMMER and a coordinate file needs to be produced before this script can be run.

```
Example usage:
dissertation_ParseOverlappingMatePairs.py file.coords > file.matepairs.txt
```

5.1.21 dissertation_PhymmBLParser.py

This program parses mate pairs with consistent taxonomic assignments from phymmBL and extracts the sequences that match the criteria, *i.e.*, reads binned as being *Gloeobacter* for both mate pairs or at least one of the pair.

```
Example usage:
dissertation_PhymmBLParser.py results.03.xx.txt reads.fasta > taxa_pairs.fasta
```

5.1.22 dissertation_ReciprocalBestHitPlot.py

This script plots locations of reciprocal BLAST hits along the genome coordinates between two given organisms. The script expects GenBank file, ortholog pair file (parsed from OrthoMCL output), and COG functional categories file, and produces figures similar to Figure 4.28. Currently, it is hard-coded for alignment between *G. violaceus* PCC 7421 and *Candidatus* G. kilaueaensis JS1, but it can be modified to compare other genomes.

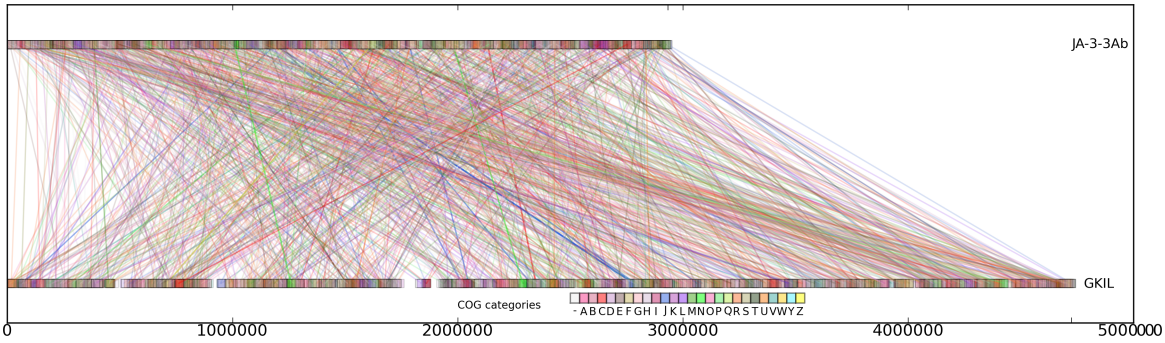


Figure 5.5. Reciprocal Best BLAST hit plot comparison between *Candidatus* G. kilaueaensis JS1 and *Synechococcus* sp. JA-3-3Ab.

```
Example usage:
dissertation_ReciprocalBestHitPlot.py ../../../annotation/GKIL.v6.gbf
  ../../NC_005125.1.gbk pair_GKIL_58011.txt cogs.t.list
dissertation_ReciprocalBestHitPlot.py ../cyano/61581.refseq.gbk
  ../cyano/61607.refseq.gbk pair_61581_61607.txt cogs.t.list
```

5.1.23 dissertation_ReciprocalBestHitPlotWithPtt.py

This script plots a reciprocal BLAST hit plot similar to one plotted in Figure 5.5 but uses just the Ptt file that comes with NCBI genomes, instead of the GenBank file. This makes it easier and faster to parse files and display results quicker than the previous script (dissertation_ReciprocalBestHitPlot.py). An example output file is shown in Figure 5.6.

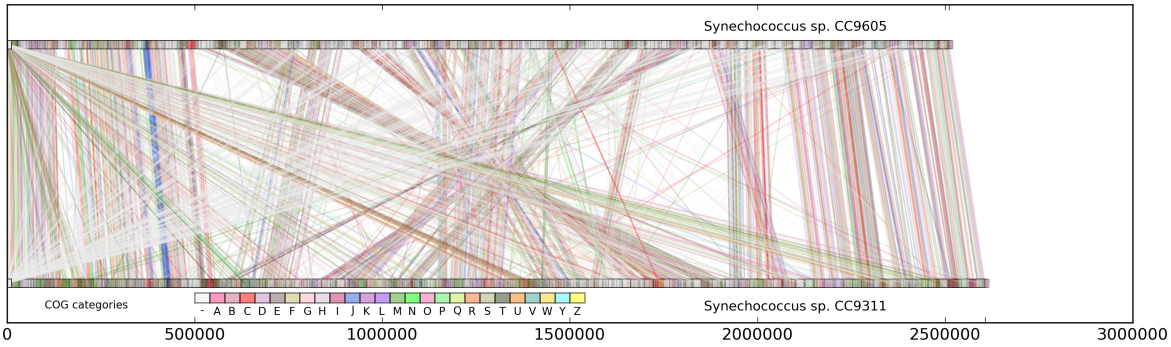


Figure 5.6. Reciprocal Best BLAST hit plot comparison between *Synechococcus* sp. CC9311 and *Synechococcus* sp. CC9605.

```
Example usage:
dissertation_ReciprocalBestHitPlotWithPtt.py 58123 58319 NC_008319.ptt
  NC_007516.ptt orthologs/orthomcl/CYANO.orthologs.txt orthologs/orthomcl/cogs.t.list
```

5.1.24 dissertation_RecruitmentPlotBlast.py

This script displays a custom recruitment plot as in Figures 4.30 and 4.31 for coordinate file produced with BLASTn -m 8 option. Full code is listed in Google Code URL given above.

```
Example usage:
dissertation_RecruitmentPlot.py genome.blastn genome.fasta
```

5.1.25 dissertation_RibosomalGenesIndividual.sh

This bash script Blasts and extracts individual ribosomal genes (using *Gloeobacter* as query) from other cyanobacteria, aligns them using Muscle, automatically trims the gaps or non-conserved blocks using Gblocks, then concatenates them to prepare them for analysis using RAxML. See Appendix D.13 for full code. Full code is listed online in Google Code URL given above.

```
Example usage:
dissertation_RibosomalGenesIndividual.sh r43.list cyano
```

5.2 Scripts used to analyze the epilithic biofilm metagenome

Table 5.2. Bioinformatic scripts used in the analysis of the epilithic biofilm metagenome

Number	Script name	Language
1	dissertation_DownloadPopset.py	Python
2	dissertation_TetraNTCalculatorImproved.py	Python
3	dissertation_TetraNTCalculator.py	Python

5.2.1 dissertation_DownloadPopset.py

This script downloads nucleotide sequences from popsets (usually 16S rDNA sequences) from NCBI. Given a list of popset IDs, it can automatically download Fasta files and save them locally. An internet connection is needed for it to work.

```
Example usage:
dissertation_DownloadPopset.py popset.list
```

5.2.2 dissertation_TetraNTCalculatorImproved.py

The aim is to use this script to bin metagenomic reads by tetranucleotide frequency, among other components such as G+C%. I attempted to use Z score because it is a normalized score instead of a raw score which can change based on length of the sequence. It is important to take into account the differences between sequence lengths in metagenomic reads. Calculation of Z score uses the following formula:

$$z = \frac{x - \mu}{\sigma} \tag{5.2.1}$$

where x is the raw tetranucleotide count, μ is the mean, and σ is the standard deviation for each metagenomic sequence read. Since there are 256 combinations, each sequence read produces Z scores for each tetranucleotide combination. Although this script was intended for metagenomic binning, it can also be used to calculate tetranucleotide frequencies in any given genome or genes. See Appendix D.10 for full code.

```
Example usage:
dissertation_TetraNTCalculatorImproved.py test-multi.fasta tetra.list
```

5.2.3 dissertation_KeggModule.rb

This script queries the KEGG database given a module name. The idea is to query the KO and COG groups (or EC numbers) in a given pathway module to extract a list of KO and COGs when analyzing a given pathway module. This can be automated when working on a complete genome to annotate pathways as being complete or incomplete in a genome. See Appendix D.11 for full code.

```
Example usage:
dissertation_KeggModule.rb M00001

This prints:
Module info:
MD: M00001 Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate

KO info:
KO: K01689 enolase [EC:4.2.1.11] [RN:R00658]
KO: K01623,K01624 fructose-bisphosphate aldolase [EC:4.1.2.13] [RN:R01070]
KO: K00844,K12407,K00845 hexokinase/glucokinase [EC:2.7.1.1 2.7.1.2] [RN:R01786]
KO: K01834 phosphoglycerate mutase [EC:5.4.2.1] [RN:R01518]
KO: K01803 triosephosphate isomerase [EC:5.3.1.1] [RN:R01015]
KO: K00927 phosphoglycerate kinase [EC:2.7.2.3] [RN:R01512]
KO: K00134,K00150 glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12 1.2.1.59] [RN:R01061 R01063]
KO: K00850 6-phosphofructokinase [EC:2.7.1.11] [RN:R04779]
KO: K00873 pyruvate kinase [EC:2.7.1.40] [RN:R00200]
KO: K01810,K06859,K13810,K15916 glucose-6-phosphate isomerase [EC:5.3.1.9] [RN:R02740]
Total KOs found: 10

COG info:
COG: K00134 COG0057 glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12]
COG: K00150 COG0057 glyceraldehyde-3-phosphate dehydrogenase (NAD(P)) [EC:1.2.1.59]
COG: K00845 COG0837 glucokinase [EC:2.7.1.2]
COG: K00850 COG0205 6-phosphofructokinase [EC:2.7.1.11]
COG: K00850 COG1105 6-phosphofructokinase [EC:2.7.1.11]
COG: K00873 COG0469 pyruvate kinase [EC:2.7.1.40]
COG: K00927 COG0126 phosphoglycerate kinase [EC:2.7.2.3]
COG: K01623 COG1830 fructose-bisphosphate aldolase, class I [EC:4.1.2.13]
COG: K01623 COG3588 fructose-bisphosphate aldolase, class I [EC:4.1.2.13]
COG: K01624 COG0191 fructose-bisphosphate aldolase, class II [EC:4.1.2.13]
COG: K01689 COG0148 enolase [EC:4.2.1.11]
COG: K01803 COG0149 triosephosphate isomerase (TIM) [EC:5.3.1.1]
```

```
COG: K01810 COG0166 glucose-6-phosphate isomerase [EC:5.3.1.9]
COG: K01834 COG0588 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase [EC:5.4.2.1]
COG: K06859 COG2140 glucose-6-phosphate isomerase, archaeal [EC:5.3.1.9]
COG: K15916 COG0166 glucose/mannose-6-phosphate isomerase [EC:5.3.1.9 5.3.1.8]
Total COGs found: 14
```

5.3 Other general utility scripts

Table 5.3. General utility scripts

Number	Script name	Language
1	dissertation_BibTeX.rb	Ruby
2	dissertation_CalculateGC.py	Python
3	dissertation_CombineFastq.py	Python
4	dissertation_ConvertFastq2FastaQual.py	Python
5	dissertation_SanityCheckDNA.py	Python
6	dissertation_SplitMultiFastaInBatches.py	Python
7	dissertation_SplitMultiFasta.py	Python

5.3.1 dissertation_BibTeX.rb

This script retrieves BibTeX files to be used in dissertation writing. BibTeX files are necessary for use with LaTeX documents as this dissertation was written in LaTeX language. See Appendix D.12 for full code.

```
Example usage:
dissertation_BibTeX.rb 20200567
dissertation_BibTeX.rb 20200567 > 20200567.bib

This prints:

@article{PMID:20200567,
  author       = {Falcon, L. I. and Magallon, S. and Castillo, A.},
  title        = {Dating the cyanobacterial ancestor of the chloroplast.},
  journal      = {ISME J},
  year         = {2010},
  volume       = {4},
  number       = {6},
  pages        = {777--783},
  url          = {http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=
                  PubMed&dopt=Citation&list_uids=20200567},
}
```

Chapter 6

Summary and Conclusions

6.1 Summary of accomplishments and findings

This section summarizes the specific aims proposed in this dissertation, and whether or not these aims were met. Summaries of the main findings for each aim are also presented here.

Three main aims were proposed in this dissertation:

1. **Aim 1:** To describe phylogenetic diversity and metabolic potential of microorganisms in the epilithic biofilm through 16S rDNA variable sequence (pyrotag) and metagenomic data.
2. **Aim 2:** To target for cultivation potentially novel microbes identified in molecular data from the epilithic biofilm.
3. **Aim 3:** To isolate and sequence a novel *Gloeobacter* sp. identified in preliminary studies of the epilithic biofilm.

In the following sub-sections, I have listed the findings of each aim, and significance of each.

Findings and conclusions for Aim 1: To describe phylogenetic diversity and metabolic potential of microorganisms in the epilithic biofilm through 16S rDNA variable sequence (pyrotag) and metagenomic data

Pyrotag and metagenomic sequences revealed the HAVO epilithic biofilm microbial community is dominated by *Proteobacteria*, *Acidobacteria*, and *Chloroflexi*. These taxa are not represented by just a few members each, but the community comprises a diverse assemblage of bacteria;

the community is very complex and requires more sampling depth than was conducted in this work. The effective genome size was determined to be about 4.2 Mbp, indicating a complexity level approaching soil microbial communities. Analysis of metabolic functional categories revealed the most abundant gene category is for amino acid transport and metabolism. Comparative metagenomic analyses revealed the HAVO epilithic biofilm is more closely related to soil microbial communities than to known microbial mat communities, suggesting perhaps that this community may have originated with microorganisms from nearby soils. Fragment recruitment analysis identified several unexpected *Bacteria* and *Archaea* taxa that may be amenable to cultivation and genome sequencing.

Significances of findings for Aim 1:

No previous work has characterized microbial diversity in this particular community. Somewhat similar work has been conducted, but only of microbial communities inside a handful of older lava caves inside and outside Hawai‘i. This is the first metagenomic characterization of such a microbial community in Kīlauea, however, and may lead to further characterizations of rare individuals in the community. Novel species and lineages detected in this metagenome could also be subject to further targeted cultivation and genome sequencing projects.

Findings and conclusions for Aim 2: To target for cultivation potentially novel microbes identified in molecular data from the epilithic biofilm

Three novel cyanobacteria were cultivated from the HAVO epilithic biofilm. One is a new *Gloeobacter* species later confirmed to be markedly different from the only known species in the Class, *Gloeobacter violaceus* PCC 7421. Another cyanobacterium belongs in the genus *Leptolyngbya*, but no close species has yet been identified; it is likely a new species, and is pending further genomic and physiological characterization. A third cultivated cyanobacterium belongs in the order Stigonematales, and had close relatives in the *Fischerella* and *Mastigocladus*. This culture, too, may prove to be a new species after further genomic and physiological characterization.

Significances of findings for Aim 2:

Candidatus Gloeobacter kilaueaensis JS1 is only the second *Gloeobacter* species to be isolated, since the genus, Family and Class was established 38 years ago. This is a significant finding in terms of species discovery, and may well advance our knowledge of ancient lineages of cyanobacteria and the evolution of photosynthesis. Cultivation of *Candidatus* Gloeobacter kilaueaensis JS

also provided material from which its complete genome was sequenced. Both the *Leptolyngbya* sp. JS2 and *Fischerella* sp. JS3 cultivated from the epilithic biofilm are potential biofilm formers given their filamentous natures; sequencing their genomes may reveal genes involved in biofilm formation on rock surfaces.

Findings and conclusions for Aim 3: To isolate and sequence a novel *Gloeobacter* sp. identified in preliminary studies of the epilithic biofilm

In silico DDH experiments revealed this organism differs from the only known Type strain *G. violaceus* PCC 7421. Little gene synteny exists between these two *Gloeobacter* species; *Candidatus* Gloeobacter kilaueaensis JS1 may have gone through extensive genome rearrangements. Despite these rearrangements, the two species still share the largest number of orthologs when compared to other cyanobacteria. The divergence time of the two *Gloeobacter* species is currently estimated to be between 153 and 324 million years ago.

Significances of findings for Aim 3:

Candidatus Gloeobacter kilaueaensis JS1 is only the second *Gloeobacter* species to be recognized since 1974. Its relation to other early diverging cyanobacteria will be of enormous use to evolutionary biologists interested in the early evolution of oxygenic photosynthesis. Attempting to cultivate this strain only began after it was detected in a 16S rRNA gene clone library. This shows that cultivation approaches remain important and able to yield important results and material for subsequent experimental work. It should also be noted that the complete genome was sequenced and assembled from a non-axenic culture resembling a low-complexity metagenome, and not from a pure culture. *Candidatus* Gloeobacter kilaueaensis JS1 is not the deepest-branching cyanobacterium. Analyses of gene gains/losses and evolution rates of genes identified here may help shed light on *Gloeobacter* evolution.

Limitations and improvements to be made for further research

Although metagenomic sequencing and analysis revealed a vast number of novel microbes in the HAVO eopilithic biofilm, the functional importance of these microbes is unknown. Transcriptomic experiments would likely reveal specific roles and *in situ* activities. For example, extraction of mRNA from the biofilm during the day and night might reveal metabolic processes more active at specific times of the day. Another interesting experiment to understand the HAVO epilithic biofilm's formation and maintenance would be to conduct spatial and temporal sampling; the biofilm was first

sampled in 2006 when it looked very different from 2009. In 2006, a larger area of the same biofilm was colored purple, almost certainly from the *Gloeobacter*, but the areal extent of the purple component had shrunk markedly by 2009. How this *Gloeobacter* sp. reached this cave and if it is even endemic to Hawai‘i remains to be seen. The same or even other unique *Gloeobacter* sp. and other cyanobacteria may well occupy other marginally lit zones in caves throughout Hawaii.

Cultivation approaches led to the isolation of three novel cyanobacteria from the epilithic biofilm, but other bacteria from diverse lineages surely await cultivation. Thermophilic bacteria such as *Deinococcus* and *Meiothermus* were detected in metagenomic sequences, and also in screening of early cultures (data not shown). Such rare taxa deserve to be further investigated and cultivation should be attempted in the future. Complete genome sequencing revealed that the *Gloeobacter* identified first in a 16S rRNA gene clone library is in fact a new species, and that it differs significantly from its nearest cultivated neighbor, *G. violaceus* PCC 7421 in gene synteny, showing large scale rearrangements. However, specific physiological activities of this novel *Gloeobacter* remain to be determined.

6.2 Final conclusions

The work presented here provides insights into microbial community composition and genetic diversity in a highly complex microbial community on the indirectly illuminated wall of a warm, wet lava cave in an active volcano. Combined pyrotag and metagenomic sequences detected similar species diversities and abundances. However, the unique characterisitics of both the site and the biofilm itself mean finding analogous biofilms for comparative study could be extfremely difficult. This work also cultivated three novel cyanobacteria from the biofilm, the genome of one of which was sequenced and annotated.

Past studies of Kīlauea Caldera's microbial communities in terms of their diversity and activities have focused mostly on ecology and biogeochemical surveys; no cultivation or genomic work has been reported. Here, a microbial community survey comprising cultivation approaches, and molecular methods (clone libraries and metagenomics), culminated in the complete genome sequencing of an ancient cyanobacterium from a lava cave in Kīlauea Caldera, just hundreds of meters from the Halema‘uma‘u pit crater. The approach described here combined both traditional cultivation and more recent genomic methods. The findings of this work should prove useful in directing further sampling and characterization of microbes from volcanic habitats.

Appendix A

Supplemental Tables

A.1 Tables of questionable regions in *Candidatus* Gloeobacter kilaueaensis JS1

Table A.1. Genes within questionable region 1

Region	locus tag	product	taxonomic class
415517-431636	GKIL_0383	hypothetical protein	Chroococcales
415517-431636	GKIL_0384	RecA-family ATPase	Clostridia
415517-431636	GKIL_0385	hypothetical protein	no BLAST hit
415517-431636	GKIL_0386	hypothetical protein	no BLAST hit
415517-431636	GKIL_0387	hypothetical protein	no BLAST hit
415517-431636	GKIL_0388	hypothetical protein	no BLAST hit
415517-431636	GKIL_0389	hypothetical protein	no BLAST hit
415517-431636	GKIL_0390	phage terminase large subunit	Caudovirales
415517-431636	GKIL_0391	phage tail tape measure protein, TP901 family	Negativicutes
415517-431636	GKIL_0392	hypothetical protein	no BLAST hit
415517-431636	GKIL_0393	hypothetical protein	no BLAST hit
415517-431636	GKIL_0394	hypothetical protein	Gloeobacteria
415517-431636	GKIL_0395	transcriptional regulator	Chroococcales
415517-431636	GKIL_0396	swim zinc finger domain protein	Halobacteria
415517-431636	GKIL_0397	hypothetical protein	no BLAST hit
415517-431636	GKIL_0398	hypothetical protein	no BLAST hit
415517-431636	GKIL_0399	hypothetical protein	no BLAST hit
415517-431636	GKIL_0400	hypothetical protein	no BLAST hit
415517-431636	GKIL_0401	hypothetical protein	no BLAST hit

Table A.2. Genes within questionable region 2

Region	locus tag	product	taxonomic class
903207-946727	GKIL_0863	phage integrase family protein	Oscillatoriales
903207-946727	GKIL_0864	hypothetical protein	no BLAST hit
903207-946727	GKIL_0865	hypothetical protein	no BLAST hit
903207-946727	GKIL_0866	hypothetical protein	Chroococcales
903207-946727	GKIL_0867	hypothetical protein	no BLAST hit
903207-946727	GKIL_0868	hypothetical protein	Deltaproteobacteria
903207-946727	GKIL_0869	hypothetical protein	no BLAST hit
903207-946727	GKIL_0870	pentapeptide repeat-containing protein	Caudovirales
903207-946727	GKIL_0871	hypothetical protein	no BLAST hit
903207-946727	GKIL_0872	hypothetical protein	Chroococcales
903207-946727	GKIL_0873	hypothetical protein	Negativicutes
903207-946727	GKIL_0874	hypothetical protein	no BLAST hit
903207-946727	GKIL_0875	hypothetical protein	no BLAST hit
903207-946727	GKIL_0876	hypothetical protein	no BLAST hit
903207-946727	GKIL_0877	hypothetical protein	no BLAST hit
903207-946727	GKIL_0878	hypothetical protein	no BLAST hit
903207-946727	GKIL_0879	hypothetical protein	no BLAST hit
903207-946727	GKIL_0880	DNA-cytosine methyltransferase	unclassified phages
903207-946727	GKIL_0881	D12 class N6 adenine-specific DNA methyltransferase	Bacillales
903207-946727	GKIL_0882	hypothetical protein	no BLAST hit
903207-946727	GKIL_0883	hypothetical protein	Chroococcales
903207-946727	GKIL_0884	hypothetical protein	Alphaproteobacteria
903207-946727	GKIL_0885	hypothetical protein	Betaproteobacteria
903207-946727	GKIL_0886	hypothetical protein	no BLAST hit
903207-946727	GKIL_0887	hypothetical protein	no BLAST hit
903207-946727	GKIL_0888	hypothetical protein	no BLAST hit
903207-946727	GKIL_0889	hypothetical protein	Bacillales
903207-946727	GKIL_0890	phage terminase GpA	Alphaproteobacteria
903207-946727	GKIL_0891	hypothetical protein	Caudovirales
903207-946727	GKIL_0892	phage portal protein, lambda family	Caudovirales
903207-946727	GKIL_0893	conserved hypothetical protein	Betaproteobacteria

Table A.3. Genes within questionable region 2 - continued

Region	locus tag	product	taxonomic class
903207-946727	GKIL_0894	hypothetical protein	no BLAST hit
903207-946727	GKIL_0895	hypothetical protein	no BLAST hit
903207-946727	GKIL_0896	hypothetical protein	no BLAST hit
903207-946727	GKIL_0897	hypothetical protein	no BLAST hit
903207-946727	GKIL_0898	hypothetical protein	no BLAST hit
903207-946727	GKIL_0899	poly(3-hydroxybutyrate) depolymerase	Gloeobacteria
903207-946727	GKIL_0900	hypothetical protein	no BLAST hit
903207-946727	GKIL_0901	hypothetical protein	Betaproteobacteria
903207-946727	GKIL_0902	hypothetical protein	no BLAST hit
903207-946727	GKIL_0903	hypothetical protein	no BLAST hit
903207-946727	GKIL_0904	hypothetical protein	no BLAST hit
903207-946727	GKIL_0905	hypothetical protein	Gloeobacteria
903207-946727	GKIL_0906	hypothetical protein	no BLAST hit
903207-946727	GKIL_0907	hypothetical protein	no BLAST hit
903207-946727	GKIL_0908	Mu-like prophage protein	Gammaproteobacteria
903207-946727	GKIL_0909	conserved hypothetical protein	Alphaproteobacteria
903207-946727	GKIL_0910	conserved hypothetical protein	Alphaproteobacteria
903207-946727	GKIL_0911	cell wall-associated hydrolases (invasion-associated proteins)	Betaproteobacteria
903207-946727	GKIL_0912	hypothetical protein	Alphaproteobacteria
903207-946727	GKIL_0913	hypothetical protein	Alphaproteobacteria
903207-946727	GKIL_0914	hypothetical protein	no BLAST hit
903207-946727	GKIL_0915	hypothetical protein	no BLAST hit
903207-946727	GKIL_0916	multi-sensor signal transduction histidine kinase	Actinobacteridae
903207-946727	GKIL_0917	conserved hypothetical protein	Alphaproteobacteria
903207-946727	GKIL_0918	hypothetical protein	no BLAST hit
903207-946727	GKIL_0919	hypothetical protein	no BLAST hit
903207-946727	GKIL_0920	XRE family transcriptional regulator	Chroococcales
903207-946727	GKIL_0921	hypothetical protein	no BLAST hit
903207-946727	GKIL_0922	conserved hypothetical protein	Bacillales
903207-946727	GKIL_0923	hypothetical protein	no BLAST hit
903207-946727	GKIL_0924	hypothetical protein	Rubrobacteridae

Table A.4. Genes within questionable region 3

Region	locus tag	product	taxonomic class
1004060-1015740	GKIL_0974	alcohol dehydrogenase	environmental samples
1004060-1015740	GKIL_0975	2,3-dihydroxybenzoate decarboxylase	Betaproteobacteria
1004060-1015740	GKIL_0976	conserved hypothetical protein	Gammaproteobacteria
1004060-1015740	GKIL_0977	aldehyde dehydrogenase	Gammaproteobacteria
1004060-1015740	GKIL_0978	hypothetical protein	Deinococci
1004060-1015740	GKIL_0979	hypothetical protein	no BLAST hit
1004060-1015740	GKIL_0980	conserved hypothetical protein	Gammaproteobacteria
1004060-1015740	GKIL_0981	hypothetical protein	no BLAST hit
1004060-1015740	GKIL_0982	LysR family transcriptional regulator	Alphaproteobacteria
1004060-1015740	GKIL_0983	TetR family transcriptional regulator	Alphaproteobacteria
1004060-1015740	GKIL_0984	aldo/keto reductase	Alphaproteobacteria
1004060-1015740	GKIL_0985	D,D-heptose 1,7-bisphosphate phosphatase	Gammaproteobacteria

Table A.5. Genes within questionable region 4

Region	locus tag	product	taxonomic class
1732580-1828970	GKIL_1689	hypothetical protein	Verrucomicrobiae
1732580-1828970	GKIL_1690	transposase	Oscillatoriales
1732580-1828970	GKIL_1691	hypothetical protein	Verrucomicrobiae
1732580-1828970	GKIL_1692	hypothetical protein	Verrucomicrobiae
1732580-1828970	GKIL_1693	type IV secretory pathway, VirD4 components	Alphaproteobacteria
1732580-1828970	GKIL_1694	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1695	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1696	permease	Chloroflexales
1732580-1828970	GKIL_1697	conserved hypothetical protein	Streptophyta
1732580-1828970	GKIL_1698	MerR family transcriptional regulator	Betaproteobacteria
1732580-1828970	GKIL_1699	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1700	hypothetical protein	Betaproteobacteria
1732580-1828970	GKIL_1701	CaCA family Na(+)/Ca(+) antiporter	Stigonematales
1732580-1828970	GKIL_1702	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1703	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1704	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1705	ATP-dependent metalloprotease FtsH	Dikarya
1732580-1828970	GKIL_1706	hypothetical protein	Betaproteobacteria
1732580-1828970	GKIL_1707	conserved hypothetical protein	Chroococcales
1732580-1828970	GKIL_1708	recombination and DNA strand exchange inhibitor protein	Spirochaetales
1732580-1828970	GKIL_1709	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1710	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1711	small-conductance mechanosensitive channel	Hexamitidae
1732580-1828970	GKIL_1712	conserved hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1713	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1714	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1715	pullulanase, type I	Alphaproteobacteria
1732580-1828970	GKIL_1716	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1717	transposase	Chroococcales
1732580-1828970	GKIL_1718	signal peptidase I	Negativicutes
1732580-1828970	GKIL_1719	conserved hypothetical protein	Gloeobacteria
1732580-1828970	GKIL_1720	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1721	plasmid segregation protein ParM	Nostocales
1732580-1828970	GKIL_1722	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1723	hypothetical protein	Deltaproteobacteria
1732580-1828970	GKIL_1724	hypothetical protein	Deltaproteobacteria
1732580-1828970	GKIL_1725	hypothetical protein	Deltaproteobacteria
1732580-1828970	GKIL_1726	prolipoprotein diacylglyceryl transferase	Chroococcales
1732580-1828970	GKIL_1727	DNA-directed RNA polymerase subunit beta	Chroococcales
1732580-1828970	GKIL_1728	endonuclease/exonuclease/phosphatase	Gloeobacteria

Table A.6. Genes within questionable region 4 - continued

Region	locus tag	product	taxonomic class
1732580-1828970	GKIL_1729	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1730	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1731	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1732	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1733	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1734	hypothetical protein	Chroococcales
1732580-1828970	GKIL_1735	Ycf35	Nostocales
1732580-1828970	GKIL_1736	conserved hypothetical protein	Clostridia
1732580-1828970	GKIL_1737	ATP-dependent metalloprotease	Chroococcales
1732580-1828970	GKIL_1738	hypothetical protein	Gloeobacteria
1732580-1828970	GKIL_1739	pentapeptide repeat-containing protein	Chroococcales
1732580-1828970	GKIL_1740	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1741	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1742	DNA polymerase III subunit beta	Gloeobacteria
1732580-1828970	GKIL_1743	ribonuclease E	Oscillatoriales
1732580-1828970	GKIL_1744	hypothetical protein	Oscillatoriales
1732580-1828970	GKIL_1745	conserved hypothetical protein	Nostocales
1732580-1828970	GKIL_1746	hypothetical protein	Oscillatoriales
1732580-1828970	GKIL_1747	hypothetical protein	Chroococcales
1732580-1828970	GKIL_1748	hypothetical protein	Chroococcales
1732580-1828970	GKIL_1749	hypothetical protein	Nostocales
1732580-1828970	GKIL_1750	thiamine biosynthesis protein ThiF	Oscillatoriales
1732580-1828970	GKIL_1751	type I restriction-modification system methyltransferase subunit	Chroococcales
1732580-1828970	GKIL_1752	ATP-dependent DNA helicase Rep	Coriobacteridae
1732580-1828970	GKIL_1753	cob(I)alamin adenosyltransferase	Chroococcales
1732580-1828970	GKIL_1754	DNA primase	Deltaproteobacteria
1732580-1828970	GKIL_1755	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1756	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1757	YD repeat-containing protein	Gloeobacteria
1732580-1828970	GKIL_1758	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1759	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1760	type IV secretory pathway, VirB4 components	Deltaproteobacteria
1732580-1828970	GKIL_1761	apolipoprotein N-acyltransferase	Clostridia
1732580-1828970	GKIL_1762	conserved hypothetical protein	Chroococcales
1732580-1828970	GKIL_1763	hypothetical protein	Gloeobacteria
1732580-1828970	GKIL_1764	conserved hypothetical protein	Alphaproteobacteria
1732580-1828970	GKIL_1765	transcriptional regulator	Alphaproteobacteria
1732580-1828970	GKIL_1766	protein-S-isoprenylcysteine methyltransferase	Oscillatoriales
1732580-1828970	GKIL_1767	hypothetical protein	no BLAST hit

Table A.7. Genes within questionable region 4 - continued

Region	locus tag	product	taxonomic class
1732580-1828970	GKIL_1768	conserved hypothetical protein	Chroococcales
1732580-1828970	GKIL_1769	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1770	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1771	transposase	Deltaproteobacteria
1732580-1828970	GKIL_1772	transposase IS3/IS911 family protein	Gammaproteobacteria
1732580-1828970	GKIL_1773	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1774	GCN5-related N-acetyltransferase	Gammaproteobacteria
1732580-1828970	GKIL_1775	conserved hypothetical protein	Alphaproteobacteria
1732580-1828970	GKIL_1776	conserved hypothetical protein	Gloeobacteria
1732580-1828970	GKIL_1777	hypothetical protein	no BLAST hit
1732580-1828970	GKIL_1778	conserved hypothetical protein	Gammaproteobacteria

Table A.8. Genes within questionable region 5

Region	locus tag	product	taxonomic class
4262380-4279810	GKIL_4053	conserved hypothetical protein	Oscillatoriales
4262380-4279810	GKIL_4054	conserved hypothetical protein	Oscillatoriales
4262380-4279810	GKIL_4055	conserved hypothetical protein	Oscillatoriales
4262380-4279810	GKIL_4056	hypothetical protein	Oscillatoriales
4262380-4279810	GKIL_4057	conserved hypothetical protein	Oscillatoriales
4262380-4279810	GKIL_4058	hypothetical protein	Oscillatoriales
4262380-4279810	GKIL_4059	CRISPR-associated protein Cas2	Chroococcales
4262380-4279810	GKIL_4060	CRISPR-associated protein Cas1	Chroococcales
4262380-4279810	GKIL_4061	hypothetical protein	Chroococcales
4262380-4279810	GKIL_4062	conserved hypothetical protein	Chroococcales
4262380-4279810	GKIL_4063	plasmid stabilization system protein	Gloeobacteria
4262380-4279810	GKIL_4064	conserved hypothetical protein	Oscillatoriales
4262380-4279810	GKIL_4065	HNH endonuclease	Chroococcales

Appendix B

Media and recipes

B.1 Ammonia-oxidizing *Archaea* medium

Synthetic *Crenarchaeota* Media (1L)	
NaCl	26g
$MgCl_2 \cdot 6\,H_2O$	5g
$MgSO_4 \cdot 7\,H_2O$	5g
$CaCl_2$	1.5g
KBr	0.1g

- Note: Add less salt for non-marine cultures
- Autoclave
- Add aseptically:
  - 1ml non-chelated trace element mixture
  - 1ml vitamin solution
  - 10ml KH_2PO_4 (Potassium phosphate) solution (4g/L) → 0.4g KH_2PO_4 in 100mL
  - 1ml Selenite-tungstate ($Na_2WO_4 \cdot 2\,H_2O$) medium
  - 1ml bicarbonate solution (1M) → 8.4g $NaHCO_3$ in 100mL
  - 0.5-1ml ammonium chloride (1M) → 5.35g NH_4Cl in 100mL
- Adjust pH to 7.0 - 7.2 using NaOH

Trace Element Solution SL-10 (per liter)

$FeCl_2 \cdot 4\,H_2O$	1.5g
$CoCl_2 \cdot 6\,H_2O$	190mg (0.19g)
$MnCl_2 \cdot 4\,H_2O$	100mg (0.10g)
$ZnCl_2$	70mg (0.07g)
$Na_2MoO_4 \cdot 2\,H_2O$	36mg (0.036g)
$NiCl_2 \cdot 6\,H_2O$	24mg (0.024g)
H_3BO_3	6mg (0.006g)
$CuCl_2 \cdot 2\,H_2O$	2mg (0.002g)
HCl (25% solution)	10ml

- Add $FeCl_2 \cdot 4\,H_2O$ to 10.0ml of HCl. Mix thoroughly. Add distilled/deionized water and bring volume to 1.0L.
- Add remaining components. Mix thoroughly.
- Sparge with 80% N_2 + 20% CO_2.
- Autoclave for 15min at 15psi pressure - 121 °C.

Trace Element Solution (Drews, 1974)

$MnCl_2 \cdot 4\,H_2O$	100mg (0.1g)
$CoCl_2$	20mg (0.02g)
$CuSO_4$	10mg (0.01g)
$Na_2MoO_4 \cdot 2\,H_2O$	10mg (0.01g)
$ZnCl_2$	20mg (0.02g)
LiCl	5mg (0.005g)
$SnCl_2 \cdot 2\,H_2O$	5mg (0.005g)
H_3BO_3	10mg (0.01g)
KBr	20mg (0.02g)
KI	20mg (0.02g)
EDTA, $Na-Fe^{3+}$ salt (trihydrate)	8g

- Dissolve in 1L water, filter sterilize.

Selenite-Tungstate solution (per liter)

NaOH	0.5g
$Na_2WO_4 \cdot 2\,H_2O$	4mg (0.004g)
$Na_2SeO_3 \cdot 5\,H_2O$	3mg (0.003g)

- Add components to distilled/deionized water and bring volume to 1.0L. Mix thoroughly. Sparge with 100% N_2. Filter sterilize.

Vitamin solution (per liter)	
Pyridoxine.HCl	10mg (0.01g)
Thiamine.HCl $\cdot 2\,H_2O$	5mg (0.005g)
Riboflavin	5mg (0.005g)
Nicotinic Acid	5mg (0.005g)
Calcium D-(+)-pantothenate	5mg (0.005g)
p-Aminobenzoic acid	5mg (0.005g)
Lipoic Acid	5mg (0.005g)
Biotin	2mg (0.002g)
Folic Acid	2mg (0.002g)
Vitamin B_{12}	0.1mg (0.0001g)

- Add components to distilled/deionized water and bring volume to 1.0L. Mix thoroughly. Sparge with 80% H_2 + 20% CO_2. Filter sterilize.

B.2 Cyanobacteria medium

Modified BG 11 Agar	
Agar	10.0g
$NaNO_3$	1.5g
$MgSO_4 \cdot 7\,H_2O$	0.075g
K_2HPO_4	0.04g
$CaCl_2 \cdot 2\,H_2O$	0.036g
Na_2CO_3	0.02g
Citric Acid	6.0mg (0.006g)
Ferric ammonium citrate	6.0mg (0.006g)
Disodium EDTA	1.0mg (0.001g)
Trace metal mix A5	1.0mL

- Add components to distilled/deionized water and bring volume to 1.0L. Mix thoroughly. Heat gently to boiling. Distribute into tubes or flasks.
- Autoclave for 15min at 15psi pressure - 121°C.
- For solid medium, pour into sterile Petri dishes or leave in tubes.

Trace metal mix A5	
H_3BO_3	2.86g
$MnCl_2 \cdot 4\,H_2O$	1.81g
$Na_2MoO_4 \cdot 2\,H_2O$	0.39g
$ZnSO_4 \cdot 7\,H_2O$	0.222g
$CuSO_4 \cdot 5\,H_2O$	0.079g
$Co(NO_3)_2 \cdot 6\,H_2O$	0.049g

- Add components to distilled/deionized water and bring volume to 1.0L. Mix thoroughly.

B.3 ATCC Medium (1473 LPBM acido-thermophile medium)

NH_4Cl	1.0g
KH_2PO_4	1.0g
$Na_2HPO_4 \cdot 7\,H_2O$	0.1g
$MgSO_4 \cdot 7\,H_2O$	0.2g
$CaCl_2 \cdot 2\,H_2O$	0.02g
Yeast extract	1.0g
Sigmacell alpha Type 50 (Sigma S5504)	5.0g
Cellobiose	0.5g
Agar (for plates)	20.0g
Distilled water	1.0L

- Adjust medium to pH 5.2 with H_3PO_4 prior to addition of carbon sources. Autoclave at 121°C for 15 minutes.

B.4 FS1 and FS2 Media

FS1 (per liter)

NH_4Cl	0.2g
KH_2PO_4	0.05g
$MgSO_4 \cdot 7\,H_2O$	0.02g
$CaCl_2 \cdot 6\,H_2O$	0.01g
Yeast Extract	10mg
FeEDTA solution	3ml
Trace element solution 1	3ml
Phyagel (gellan)	15g

- Adjust pH to 4.5 - 5.5
- Autoclave
- Add filter-sterilized vitamin solution if needed

FS2 (per liter)

KNO_3	0.4g
KH_2PO_4	0.05g
$MgSO_4 \cdot 7\,H_2O$	0.02g
$CaCl_2 \cdot 6\,H_2O$	0.01g
Yeast Extract	10mg
FeEDTA solution	3ml
Trace element solution 1	3ml
Phyagel (gellan)	15g

- Adjust pH to 4.5 - 5.5
- Autoclave
- Add filter-sterilized vitamin solution if needed

FeEDTA solution (per liter)

$FeSO_4 \cdot 7\,H_2O$	1.54g
Na_2EDTA	2.06g

Trace elements solution 1 (per liter)

$ZnSO_4 \cdot 7\,H_2O$	0.44g
$CuSO_4 \cdot 5\,H_2O$	0.20g
$MnCl_4 \cdot H_2O$	0.19g
$Na_2MoO_4 \cdot 2\,H_2O$	0.06g
H_3BO_3	0.10g
$CoCl_2 \cdot 6\,H_2O$	0.08g

Trace elements solution 2 (per liter)

Nitrilotriacetic acid	1.5g
$Fe(NH_4)_2(SO_4)_2 \cdot 6\,H_2O$	0.2g
Na_2SeO_3	0.2g
$CoCl_2 \cdot 6\,H_2O$	0.1g
$MnSO_4 \cdot 2\,H_2O$	0.1g
$Na_2MoO_4 \cdot 2\,H_2O$	0.1g
$Na_2WO_4 \cdot 2\,H_2O$	0.1g
$ZnSO_4 \cdot 7\,H_2O$	0.1g
$AlCl_3 \cdot 6\,H_2O$	0.04g
$NiCl_2 \cdot 6\,H_2O$	0.025g
H_3BO_3	0.01g
$CuSO_4 \cdot 5\,H_2O$	0.01g

- Adjust pH to 7.

Vitamin solution (100mg)

Folic acid	0.8mg
Vitamin B1	8mg
Vitamin B2	4mg
Niacin	1mg
Niacinamide	10mg
Pantothenate	15mg
Pyridoxine	15mg
Cobalamin	5mg
Biotin	5mg
Choline	15mg
Inositol	15mg
Para-amino benzoic acid	7mg

Appendix C

Newbler assembly metrics

The following shows Newbler assembly metrics file produced by an assembly using Newber version 2.6.

```
/*****************************************************************************
**
**      454 Life Sciences Corporation
**         Newbler Metrics Results
**
**      Date of Assembly: 2011/10/19 19:51:43
**      Project Directory: /host/Users/JS/UH-work/gloeobacter/final_assembly/newbler/454GapSeqsConsed
**      Software Release: 2.6 (20110517_1502)
**
*****************************************************************************/

/*
**  Input information.
*/

runData
{
        file
        {
                path = "/host/Users/JS/UH-work/gloeobacter/final_assembly/042811.clean.fasta";

                numberOfReads = 37, 37;
                numberOfBases = 25681, 25247;
        }

        file
        {
                path = "/host/Users/JS/UH-work/gloeobacter/final_assembly/042911.clean.fasta";

                numberOfReads = 18, 18;
                numberOfBases = 9065, 8861;
        }

}

pairedReadData
{
        file
        {
                path = "/host/Users/JS/UH-work/gloeobacter/final_assembly/GM6SIKE01.sff";

                numberOfReads = 222335, 376380;
                numberOfBases = 83921268, 75347400;
                numWithPairedRead = 155047;
        }

}

/*
**  Operation metrics.
*/

```

```
runMetrics
{
        inputFileNumReads  = 222390;
        inputFileNumBases  = 83956014;

        totalNumberOfReads = 376435;
        totalNumberOfBases = 75381508;

        numberSearches   = 96152;
        seedHitsFound    = 5838236, 60.72;
        overlapsFound    = 904179, 9.40, 15.49%;
        overlapsReported = 844213, 8.78, 93.37%;
        overlapsUsed     = 614115, 6.39, 72.74%;
}

readAlignmentResults
{
        file
        {
                path = "/host/Users/JS/UH-work/gloeobacter/final_assembly/042811.clean.fasta";

                numAlignedReads     = 34, 91.89%;
                numAlignedBases     = 22198, 87.92%;
                inferredReadError  = 0.60%, 134;
        }

        file
        {
                path = "/host/Users/JS/UH-work/gloeobacter/final_assembly/042911.clean.fasta";

                numAlignedReads     = 18, 100.00%;
                numAlignedBases     = 8780, 99.09%;
                inferredReadError  = 0.77%, 68;
        }

}

pairedReadResults
{
        file
        {
                path = "/host/Users/JS/UH-work/gloeobacter/final_assembly/GM6SIKE01.sff";

                numAlignedReads     = 364733, 96.91%;
                numAlignedBases     = 73436810, 97.46%;
                inferredReadError  = 0.81%, 592987;

                numberWithBothMapped  = 149185;
                numWithOneUnmapped    = 858;
                numWithMultiplyMapped = 1815;
                numWithBothUnmapped   = 3189;
        }

}

/*
** Consensus distribution information.
*/
consensusDistribution
{
        fullDistribution
        {
                signalBin =  0.0, 135695;
                signalBin =  0.5, 2;
                signalBin =  0.6, 31;
                signalBin =  0.7, 1322;
                signalBin =  0.8, 116239;
                signalBin =  0.9, 1647809;
                signalBin =  1.0, 822310;
                signalBin =  1.1, 10355;
                signalBin =  1.2, 130;
                signalBin =  1.3, 12;
                signalBin =  1.4, 4;
                signalBin =  1.5, 15;
                signalBin =  1.6, 178;
                signalBin =  1.7, 3807;
                signalBin =  1.8, 93081;
                signalBin =  1.9, 439696;
                signalBin =  2.0, 135627;
                signalBin =  2.1, 3139;
                signalBin =  2.2, 65;
                signalBin =  2.3, 6;
                signalBin =  2.4, 1;
                signalBin =  2.5, 18;
                signalBin =  2.6, 192;
                signalBin =  2.7, 2879;
                signalBin =  2.8, 33937;
```

```
                signalBin =  2.9, 94386;
                signalBin =  3.0, 33084;
                signalBin =  3.1, 1522;
                signalBin =  3.2, 46;
                signalBin =  3.3, 4;
                signalBin =  3.4, 5;
                signalBin =  3.5, 32;
                signalBin =  3.6, 190;
                signalBin =  3.7, 1458;
                signalBin =  3.8, 7530;
                signalBin =  3.9, 16500;
                signalBin =  4.0, 12251;
                signalBin =  4.1, 3319;
                signalBin =  4.2, 461;
                signalBin =  4.3, 44;
                signalBin =  4.4, 2;
                signalBin =  4.5, 15;
                signalBin =  4.6, 23;
                signalBin =  4.7, 185;
                signalBin =  4.8, 1227;
                signalBin =  4.9, 5307;
                signalBin =  5.0, 5811;
                signalBin =  5.1, 1581;
                signalBin =  5.2, 161;
                signalBin =  5.3, 10;
                signalBin =  5.4, 2;
                signalBin =  5.5, 12;
                signalBin =  5.6, 30;
                signalBin =  5.7, 97;
                signalBin =  5.8, 355;
                signalBin =  5.9, 1031;
                signalBin =  6.0, 1469;
                signalBin =  6.1, 774;
                signalBin =  6.2, 177;
                signalBin =  6.3, 25;
                signalBin =  6.4, 7;
                signalBin =  6.5, 6;
                signalBin =  6.6, 14;
                signalBin =  6.7, 43;
                signalBin =  6.8, 105;
                signalBin =  6.9, 264;
                signalBin =  7.0, 273;
                signalBin =  7.1, 141;
                signalBin =  7.2, 48;
                signalBin =  7.3, 18;
                signalBin =  7.4, 2;
                signalBin =  7.5, 6;
                signalBin =  7.6, 8;
                signalBin =  7.7, 19;
                signalBin =  7.8, 27;
                signalBin =  7.9, 49;
                signalBin =  8.0, 34;
                signalBin =  8.1, 20;
                signalBin =  8.2, 6;
                signalBin =  8.3, 7;
                signalBin =  8.4, 1;
                signalBin =  8.5, 1;
                signalBin =  8.6, 1;
                signalBin =  8.7, 2;
                signalBin =  8.8, 3;
                signalBin =  8.9, 6;
                signalBin =  9.0, 1;
                signalBin =  9.3, 1;
        }

        distributionPeaks
        {
                signalPeak = 1, 0.98;
                signalPeak = 2, 1.94;
                signalPeak = 3, 2.94;
                signalPeak = 4, 3.96;
                signalPeak = 5, 5.00;
                signalPeak = 6, 6.02;
                signalPeak = 7, 7.00;
        }

        thresholdsUsed
        {
                threshold = 0, 1, 0.48;
                threshold = 1, 2, 1.42;
                threshold = 2, 3, 2.42;
                threshold = 3, 4, 3.40;
                threshold = 4, 5, 4.44;
                threshold = 5, 6, 5.42;
                threshold = 6, 7, 6.56;

                interpolationAmount = 1.01;
```

```
        }
}


/*
**  Alignment depths.
*/
alignmentDepths
{
              1 = 2035;
              2 = 4150;
            3-4 = 35369;
            5-6 = 129277;
            7-8 = 294126;
           9-10 = 486919;
          11-13 = 991916;
          14-16 = 1023127;
          17-19 = 812777;
          20-22 = 524497;
          23-25 = 302125;
          26-28 = 162714;
          29-31 = 81204;
          32-34 = 37476;
          35-38 = 20449;
          39-42 = 7374;
          43-46 = 2974;
          47-50 = 1831;
          51-55 = 1441;
          56-60 = 1724;
          61-70 = 2867;
          71-80 = 1539;
          81-90 = 635;
         91-100 = 155;
        101-140 = 391;
        141-180 = 39;
        181-240 = 102;
        241-300 = 693;
        301-400 = 1374;
        401-500 = 6;
        501-600 = 0;
        601-700 = 0;
        701-850 = 0;
        851-1000 = 0;
          1001+ = 0;

        peakDepth            = 14.0;
        estimatedGenomeSize  = "5.4 MB";
}

/*
**  Consensus results.
*/
consensusResults
{
        readStatus
        {
                numAlignedReads    = 364785, 96.91%;
                numAlignedBases    = 73467788, 97.46%;
                inferredReadError = 0.81%, 593189;

                numberAssembled = 362615;
                numberPartial   = 2170;
                numberSingleton = 9597;
                numberRepeat    = 1951;
                numberOutlier   = 102;
                numberTooShort  = 0;
        }

        pairedReadStatus
        {
                numberWithBothMapped   = 149185;
                numberWithOneUnmapped  = 858;
                numberMultiplyMapped   = 1815;
                numberWithBothUnmapped = 3189;

                library
                {
                        libraryName       = "GM6SIKE01.sff";
                        libraryNumPairs   = 155047;
                        numInSameScaffold = 141167, 91.0%;

                        pairDistanceRangeUsed      = 4724..14173;
                        computedPairDistanceAvg    = 9449.2;
                        computedPairDistanceDev    = 2362.3;
                }
        }

```

```
        scaffoldMetrics
        {
                numberOfScaffolds   = 1;
                numberOfBases       = 4741261;

                avgScaffoldSize     = 4741261;
                N50ScaffoldSize     = 4741261, 1;
                largestScaffoldSize = 4741261;

                numberOfScaffoldContigs     = 69;
                numberOfScaffoldContigBases = 4714628;

                avgScaffoldContigSize       = 68327;
                N50ScaffoldContigSize       = 120754, 14;
                largestScaffoldContigSize   = 330776;

                scaffoldEndMetrics
                {
                        NoEdges   = 2, 100.0%;
                        OneEdge   = 0, 0.0%;
                        TwoEdges  = 0, 0.0%;
                        ManyEdges = 0, 0.0%;
                }

                scaffoldGapMetrics
                {
                        BothNoEdges   = 42, 61.8%;
                        OneNoEdges    = 17, 25.0%;
                        BothOneEdge   = 5, 7.4%;
                        MultiEdges    = 4, 5.9%;
                }
        }

        largeContigMetrics
        {
                numberOfContigs   = 129;
                numberOfBases     = 4670559;

                avgContigSize     = 36205;
                N50ContigSize     = 73468;
                largestContigSize = 165340;

                Q40PlusBases      = 4656603, 99.70%;
                Q39MinusBases     = 13956, 0.30%;

                largeContigEndMetrics
                {
                        NoEdges   = 104, 40.3%;
                        OneEdge   = 122, 47.3%;
                        TwoEdges  = 21, 8.1%;
                        ManyEdges = 11, 4.3%;
                }
        }

        allContigMetrics
        {
                numberOfContigs = 146;
                numberOfBases   = 4675190;
        }
}
```

Appendix D

Full source codes of selected scripts written for bioinformatic analyses

This section lists the full source code for selected scripts deemed significant for analysis of data in this dissertation. Complete list of scripts written for this dissertation can be viewed and downloaded at the following URL:
http://code.google.com/p/jimmysawdissertation/source/browse/trunk/dissertation

D.1 dissertation_BlastnRetrieveTopHits.py

```
#!/usr/bin/python
"""
Author: Jimmy Saw
Last update: 08-06-2012
Description: This script runs BLAST of sequences then retrieves top n hits as instructed.
Usage example: dissertation_BlastnRetrieveTopHits.py test.fasta 10 blastp 1e-5
"""

import sys
import re
from Bio import SeqIO
from Bio import Entrez
from Bio.Blast import NCBIWWW
from Bio.Blast import NCBIXML

Entrez.email = 'jimmy@hawaii.edu'

def blast(sequences, blastprog, database, maxevalue):
    for s in sequences:
        result_handle = NCBIWWW.qblast(blastprog, database, s.seq,
          expect=maxevalue, filter=None)
        save_file = open(s.id + ".xml", "w")
        save_file.write(result_handle.read())
        save_file.close()
```

```
        result_handle.close()
        print "Done with BLAST for: ", s.id

def parseblastn(seqlist):
    accessions = []
    for seq in seqlist:
        xml_file = seq + ".xml"
        xf = open(xml_file, "rU")
        r = NCBIXML.parse(xf)
        results_rec = r.next()
        for hit in results_rec.alignments[0:maxhits]:
#            print seq, hit.accession, hit.hit_def
            accessions.append(hit.accession)
    return accessions

def retrieveseqs(acclist):
    sequences = []
    outfile = "top_" + str(maxhits) + "_neighbors.fasta"
    for acc in acclist:
        handle = Entrez.efetch(db="nucleotide", id=acc, rettype="fasta",
          retmode="text")
        tmpseq = SeqIO.read(handle, "fasta")
        sequences.append(tmpseq)
    SeqIO.write(sequences, outfile, "fasta")

if __name__ == "__main__":
    seqslist = []
    seqids = []
    seqs = SeqIO.parse(sys.argv[1], "fasta")
    maxhits = int(sys.argv[2])
    blasttype = sys.argv[3]
    maxeval = sys.argv[4]
    for seq in seqs:
        seqslist.append(seq)
        seqids.append(seq.id)
    if blasttype == "blastn":
        blast(seqslist, "blastn", "nt", maxeval)
    elif blasttype == "blastp":
        blast(seqslist, "blastp", "nr", maxeval)
    elif blasttype == "blastx":
        blast(seqslist, "blastx", "nr", maxeval)
    else:
        print "Choose either: blastn, blastp, or blastx"
    accs = parseblastn(seqids)
    retrieveseqs(accs)
```

D.2 dissertation_CompareGenes.py

```
#!/usr/bin/python

"""
This program draws gene clusters to make publication quality figures. It takes in
Genbank file, COG category file, and expects start and stop coordinates of region
to inspect.

Usage: dissertation_CompareGenes.py seq1.gbk seq2.gbk seq2.ptt cogs.t.list
    <seq1start> <seq1stop> <seq2start> <seq2stop>
Examples:

dissertation_CompareGenes.py ../annotation/GKIL.v6.gbf NC_005125.1.gbk
```

```
    NC_005125.ptt orthologs/orthomcl/cogs.t.list 10000 20000 10000 20000
dissertation_CompareGenes.py ../annotation/GKIL.v6.gbf NC_005125.1.gbk
    NC_005125.ptt orthologs/orthomcl/cogs.t.list 773000 783000  2814800 2824800

dissertation_CompareGenes.py ../annotation/GKIL.v6.gbf NC_005125.1.gbk
    NC_005125.ptt orthologs/orthomcl/cogs.t.list 2635000 2655000
    178500 198500 (This one is comparing rhodopsin gene cluster between
    GVIO and GKIL)
dissertation_CompareGenes.py ../annotation/GKIL.v6.gbf NC_005125.1.gbk
    NC_005125.ptt orthologs/orthomcl/cogs.t.list 2625000 2665000 168500 208500

Note: Resolution is best if the segment in view is less than 10000bp.

Author: Jimmy Saw
Date of last update: 05-01-2012
"""

import sys
import re
import matplotlib.pyplot as plt
import pylab
import matplotlib
from matplotlib import mpl
from matplotlib.patches import Rectangle
from matplotlib.transforms import Bbox
from Bio import SeqIO
from Bio.SeqUtils import GC
import matplotlib.patches as mpatch

#Regex and other stuffs
cogcat = re.compile('\[(.*)\]\t(\w+)\t.*')
pttcog = re.compile('(COG\d{4})(\w).*')

cogdict = {
    'J' : '#2B60DE', 'A' : '#F6358A', 'K' : '#B048B5', 'L' : '#8E35EF', 'B' : '#D16587',
    'D' : '#C38EC7', 'Y' : '#52F3FF', 'V' : '#3EA99F', 'T' : '#254117', 'M' : '#41A317',
    'N' : '#00FF00', 'Z' : '#FFFF00', 'W' : '#FDD017', 'U' : '#F88017', 'O' : '#F660AB',
    'C' : '#FF0000', 'G' : '#FAAFBA', 'E' : '#7F5A58', 'F' : '#C8B560', 'H' : '#8B7500',
    'I' : '#C12869', 'P' : '#57E964', 'Q' : '#BCE954', 'R' : '#F87431', 'S' : '#ADA96E',
    '-' : '#D3D3D3'
    }

g1starts = []
g1stops = []
g2starts = []
g2stops = []
toplot = []

g1spanx1 = int(sys.argv[5])
g1spanx2 = int(sys.argv[6])
g2spanx1 = int(sys.argv[7])
g2spanx2 = int(sys.argv[8])

##Genome one Genbank file
g1seq = SeqIO.read(sys.argv[1], "gb")
g1length = len(g1seq.seq)
g1featdict = {}
for feat in g1seq.features:
    g1start = feat.location._start.position
    g1stop = feat.location._end.position
    g1starts.append(g1start)
    g1stops.append(g1stop)
    if g1spanx1 <= g1start <= g1spanx2 and g1spanx1 <= g1stop <= g1spanx2:
        if feat.type == 'CDS':
```

```
            g1featdict[feat.qualifiers['locus_tag'][0]] = feat
        if feat.type == 'tRNA':
            g1featdict[feat.qualifiers['locus_tag'][0]] = feat
        if feat.type == 'rRNA':
            g1featdict[feat.qualifiers['locus_tag'][0]] = feat

##Genome 2 Genbank file
g2seq = SeqIO.read(sys.argv[2], "gb")
g2length = len(g2seq.seq)
g2featdict = {}
for feat in g2seq.features:
    g2start = feat.location._start.position
    g2stop = feat.location._end.position
    g2starts.append(g2start)
    g2stops.append(g2stop)
    if g2spanx1 <= g2start <= g2spanx2 and g2spanx1 <= g2stop <= g2spanx2:
        if feat.type == 'CDS':
            g2featdict[feat.qualifiers['locus_tag'][0]] = feat
        if feat.type == 'tRNA':
            g2featdict[feat.qualifiers['locus_tag'][0]] = feat
        if feat.type == 'rRNA':
            g2featdict[feat.qualifiers['locus_tag'][0]] = feat

g1starts.sort()
g1stops.sort()
g2starts.sort()
g2stops.sort()

##Genome 2 ptt file
g2cogs = {}
g2pttfile = open(sys.argv[3], "rU")
g2ptt = g2pttfile.readlines()
for line in g2ptt[3:]:
    c = line.split('\t')
    g2ltag = c[5]
    g2gene = c[4]
    g2cog = '-'
    if pttcog.match(c[7]):
        p = pttcog.match(c[7])
        g2cog = p.group(1)
    g2cogs[g2ltag] = g2cog

cogcatfile = open(sys.argv[4], "rU")
cfl = cogcatfile.readlines()

cogcatdict = {}

for line in cfl:
    tmp = line.strip()
    if cogcat.match(tmp):
        pattern = cogcat.match(tmp)
        cogcatdict[pattern.group(2)] = pattern.group(1)[0]

cogcatfile.close()

glist = []

genome1x = [g1spanx1, g1spanx2]
genome1y = [2, 2]
genome2x = [g1spanx1, g1spanx2]
genome2y = [10, 10]
mid1x = [g1spanx1+200, g1spanx2-200]
mid1y = [10.75, 10.75]
mid2x = [g1spanx1+200, g1spanx2-200]
```

```
mid2y = [4.75, 4.75]

xdiff = 0

if g1spanx1 > g2spanx1:
    xdiff = g1spanx1 - g2spanx1
    padding = xdiff - g1starts[0]
else:
    xdiff = g2spanx1 - g1spanx1
    padding = g2starts[0] - xdiff

##Start plotting
fig = plt.figure(1, figsize=(16,5))
#ax1 = fig.add_subplot(211) #makes the subplot and squeezes the figure to half panel
ax1 = fig.add_subplot(111) #makes the full figure plot. larger.

ax1.plot(mid1x, mid1y, color='#CDAA7D', marker='|', mec='#CDAA7D', ls ='-', lw=1.0)
ax1.text(g1spanx1+200, 5.6, g1spanx1+200, fontsize=8, color='black', rotation=90)
ax1.text(g1spanx2-200, 5.6, g1spanx2-200, fontsize=8, color='black', rotation=90)
ax1.plot(mid2x, mid2y, color='#CDAA7D', marker='|', mec='#CDAA7D', ls ='-', lw=1.0)
ax1.text(g1spanx1+200, 11.6, g2spanx1+200, fontsize=8, color='black', rotation=90)
ax1.text(g1spanx2-200, 11.6, g2spanx2-200, fontsize=8, color='black', rotation=90)

#ax1.axis([0, g1length, 0, 14])
ax1.axis([g1spanx1, g1spanx2, 0, 16])

for k, v in g1featdict.iteritems():
    g1feat = v
    g1start = g1feat.location._start.position
    g1stop = g1feat.location._end.position
    g1size = g1stop - g1start + 1
    g1mid = g1start + ((g1stop - g1start) / 2.0)
    g1desc = g1feat.qualifiers['product'][0]
    g1gene = ""
    if g1feat.qualifiers.has_key('gene'):
        g1gene = g1feat.qualifiers['gene'][0] #displays gene name
        #g1gene = g1desc #displays product description
    else:
        g1gene = g1feat.qualifiers['locus_tag'][0] #displays locus tag
        #g1gene = g1desc #displays product description
    cogcolor = "#D3D3D3" #base color
    if g1feat.qualifiers.has_key('note'):
        cog = g1feat.qualifiers['note'][0]
        if cog in cogcatdict:
            cogcolor = cogdict[cogcatdict[cog]]
    if g1feat.type == 'tRNA':
        cogcolor = '#800000'
    if g1feat.type == 'rRNA':
        cogcolor = '#9400D3'
        g1gene = g1desc
    if g1feat.strand == -1:
        if g1spanx1 <= g1start <= g1spanx2 and g1spanx1 <= g1stop <= g1spanx2:
            rect = Rectangle((g1start, 4.0), g1size, 0.5, fc=cogcolor,
                ec='#CDAA7D', alpha=0.5)
            plt.gca().add_patch(rect)
            #ax1.text(g1mid, 4.5, g1gene, fontsize=8, color='black', rotation=45)
            ax1.text(g1mid, 3.4, g1gene, fontsize=8, color='black',
                horizontalalignment='center')
    else:
        if g1spanx1 <= g1start <= g1spanx2 and g1spanx1 <= g1stop <= g1spanx2:
            rect = Rectangle((g1start, 5.0), g1size, 0.5, fc=cogcolor,
                ec='#CDAA7D', alpha=0.5)
            plt.gca().add_patch(rect)
            #ax1.text(g1mid, 5.5, g1gene, fontsize=8, color='black', rotation=45)
```

```
            ax1.text(g1mid, 5.6, g1gene, fontsize=8, color='black',
                horizontalalignment='center')

#Genome 2
for k, v in g2featdict.iteritems():
    g2feat = v
    g2locustag = g2feat.qualifiers['locus_tag'][0]
    g2start = g2feat.location._start.position
    g2stop = g2feat.location._end.position
    g2size = g2stop - g2start + 1
    g2mid = g2start + ((g2stop - g2start) / 2.0)
    g2desc = g2feat.qualifiers['product'][0]
    g2gene = ""
    if g2feat.qualifiers.has_key('gene'):
        g2gene = g2feat.qualifiers['gene'][0] #displays gene name
        #g1gene = g1desc #displays product description
    else:
        g2gene = g2feat.qualifiers['locus_tag'][0] #displays locus tag
        #g1gene = g1desc #displays product description
    cogcolor = "#D3D3D3" #base color
    if g2locustag in g2cogs:
        if g2cogs[g2locustag] != '-':
            #cogcolor = cogdict[g2cogs[g2locustag]]
            cogcolor = cogdict[cogcatdict[g2cogs[g2locustag]]]
    if g2feat.type == 'tRNA':
        cogcolor = '#800000'
    if g2feat.type == 'rRNA':
        cogcolor = '#9400D3'
        g2gene = g2desc
    if g2feat.strand == -1:
        #if g1spanx1 <= g1start <= g1spanx2 and g1spanx1 <= g1stop <= g1spanx2:
        #if g2start >= g1spanx1 and g2stop <= g1spanx2:
        newg2start = g2start + padding
        newg2stop = g2stop + padding
        newg2mid = newg2start + ((newg2stop - newg2start) / 2.0)
        rect = Rectangle((newg2start, 10.0), g2size, 0.5, fc=cogcolor,
            ec='#CDAA7D', alpha=0.5)
        plt.gca().add_patch(rect)
        #ax1.text(g2mid, 10.5, g2gene, fontsize=8, color='black', rotation=45)
        ax1.text(newg2mid, 9.4, g2gene, fontsize=8, color='black',
            horizontalalignment='center')
        #print "g2start, newg2start", g2start, newg2start
    else:
        #if spanx1 <= g2start <= spanx2 and spanx1 <= g2stop <= spanx2:
        #if g2start >= g1spanx1 and g2stop <= g1spanx2:
        newg2start = g2start + padding
        newg2stop = g2stop + padding
        newg2mid = newg2start + ((newg2stop - newg2start) / 2.0)
        rect = Rectangle((newg2start, 11.0), g2size, 0.5, fc=cogcolor,
            ec='#CDAA7D', alpha=0.5)
        plt.gca().add_patch(rect)
        #ax1.text(newg2mid, 11.5, g2gene, fontsize=8, color='black', rotation=45)
        ax1.text(newg2mid, 11.6, g2gene, fontsize=8, color='black',
            horizontalalignment='center')
        #print "g2start, newg2start", g2start, newg2start

#Draw legend box for COG categories
coglist = []
for k, v in cogdict.iteritems():
    coglist.append((k,v))

coglist.sort()

ccounts = len(coglist)
```

```
a = 0
spansize = g1spanx2 - g1spanx1
spanmid = g1spanx1 + ((g1spanx2 - g1spanx1) / 2.0)
xstart = spanmid
#increment = 20000 #for synecoccus
increment = spansize * 0.01 #for gloeobacter
while a < ccounts:
    fc = coglist[a][1]
    tx = coglist[a][0]
    #rect = Rectangle((xstart, 2.5), 20000, 0.25, facecolor=fc,
    #  alpha=0.5) #for synecococcus
    rect = Rectangle((xstart, 2.5), increment, 0.5, fc=fc, alpha=0.5) #for gloeobacter
    plt.gca().add_patch(rect)
    ax1.annotate(tx, xy=(xstart+(increment/2.0), 2.2),
        horizontalalignment='center', verticalalignment='center', fontsize=8)
    a += 1
    xstart = xstart + increment

ax1.annotate('COG categories', xy=(spanmid, 1.2), horizontalalignment='left',
    verticalalignment='center', fontsize=10)
ax1.annotate(g1seq.annotations['organism'], xy=(0.1, 0.1),
    xycoords='axes fraction', horizontalalignment='left',
    verticalalignment='center', fontsize=10)
ax1.annotate(g2seq.annotations['organism'], xy=(0.1, 0.9),
    xycoords='axes fraction', horizontalalignment='left',
    verticalalignment='center', fontsize=10)

frame1 = plt.gca()
for tick in frame1.axes.get_yticklines():
    tick.set_visible(False)
for y in frame1.axes.get_yticklabels():
    y.set_visible(False)
ax1.grid(False)

plt.show()
```

D.3 dissertation_DrawGenes.py

```
#!/usr/bin/python

"""
This program draws gene clusters to make publication quality figures. It takes in
Genbank file, COG category file, and expects start and stop coordinates of region
to inspect.

Usage: dissertation_DrawGenes.py seq.gbk cogs.t.list 10000 20000
Examples:
dissertation_DrawGenes.py ../../../annotation/GKIL.v6.gbf cogs.t.list 10000 20000

Note: Resolution is best if the segment in view is less than 50000bp.

Author: Jimmy Saw
Date of last update: 04-23-2012

"""

import sys
import re
import matplotlib.pyplot as plt
import pylab
```

```
import matplotlib
from matplotlib import mpl
from matplotlib.patches import Rectangle
from matplotlib.transforms import Bbox
from Bio import SeqIO
from Bio.SeqUtils import GC
import matplotlib.patches as mpatch

#Regex and other stuffs
cogcat = re.compile('\[(.*)\]\t(\w+)\t.*')

cogdict = {
    'J' : '#2B60DE', 'A' : '#F6358A', 'K' : '#B048B5', 'L' : '#8E35EF', 'B' : '#D16587',
    'D' : '#C38EC7', 'Y' : '#52F3FF', 'V' : '#3EA99F', 'T' : '#254117', 'M' : '#41A317',
    'N' : '#00FF00', 'Z' : '#FFFF00', 'W' : '#FDD017', 'U' : '#F88017', 'O' : '#F660AB',
    'C' : '#FF0000', 'G' : '#FAAFBA', 'E' : '#7F5A58', 'F' : '#C8B560', 'H' : '#8B7500',
    'I' : '#C12869', 'P' : '#57E964', 'Q' : '#BCE954', 'R' : '#F87431', 'S' : '#ADA96E',
    '-' : '#D3D3D3'
    }

##Genome one Genbank file
g1seq = SeqIO.read(sys.argv[1], "gb")
g1length = len(g1seq.seq)
g1featdict = {}
for feat in g1seq.features:
    if feat.type == 'CDS':
        g1featdict[feat.qualifiers['locus_tag'][0]] = feat
    if feat.type == 'tRNA':
        g1featdict[feat.qualifiers['locus_tag'][0]] = feat
    if feat.type == 'rRNA':
        g1featdict[feat.qualifiers['locus_tag'][0]] = feat

cogcatfile = open(sys.argv[2], "rU")
cfl = cogcatfile.readlines()

cogcatdict = {}

for line in cfl:
    tmp = line.strip()
    if cogcat.match(tmp):
        pattern = cogcat.match(tmp)
        cogcatdict[pattern.group(2)] = pattern.group(1)[0]

cogcatfile.close()

spanx1 = int(sys.argv[3])
spanx2 = int(sys.argv[4])

glist = []

genome1x = [spanx1, spanx2]
genome1y = [2, 2]
genome2x = [spanx1, spanx2]
genome2y = [10, 10]
mid1x = [spanx1+200, spanx2-200]
mid1y = [6.75, 6.75]

##Start plotting
fig = plt.figure(1, figsize=(16,4))
#ax1 = fig.add_subplot(211) #makes the subplot and squeezes the figure to half panel
ax1 = fig.add_subplot(111) #makes the full figure plot. larger.
#ax1.plot(genome1x, genome1y, color='#FFFFFF', marker='|', markersize=8.0,
#    mec='black', ls='-', lw=2.0)
#ax1.plot(genome2x, genome2y, color='#FFFFFF', marker='|', markersize=8.0,
```

```
#    mec='black', ls='-', lw=2.0)
ax1.plot(mid1x, mid1y, color='#CDAA7D', marker='|',
    mec='#CDAA7D', ls =':', lw=2.0)

#ax1.axis([0, g1length, 0, 14])
ax1.axis([spanx1, spanx2, 0, 14])

for k, v in g1featdict.iteritems():
    g1feat = v
    g1start = g1feat.location._start.position
    g1stop = g1feat.location._end.position
    g1size = g1stop - g1start + 1
    g1mid = g1start + ((g1stop - g1start) / 2.0)
    g1desc = g1feat.qualifiers['product'][0]
    g1gene = ""
    if g1feat.qualifiers.has_key('gene'):
        g1gene = g1feat.qualifiers['gene'][0] #displays gene name
        #g1gene = g1desc #displays product description
    else:
        g1gene = g1feat.qualifiers['locus_tag'][0] #displays locus tag
        #g1gene = g1desc #displays product description
    cogcolor = "#D3D3D3" #base color
    if g1feat.qualifiers.has_key('note'):
        cog = g1feat.qualifiers['note'][0]
        if cog in cogcatdict:
            cogcolor = cogdict[cogcatdict[cog]]
    if g1feat.type == 'tRNA':
        cogcolor = '#800000'
    if g1feat.type == 'rRNA':
        cogcolor = '#9400D3'
        g1gene = g1desc
    if g1feat.strand == -1:
        if spanx1 <= g1start <= spanx2 and spanx1 <= g1stop <= spanx2:
            rect = Rectangle((g1start, 6.0), g1size, 0.5,
                fc=cogcolor, ec=cogcolor, alpha=0.5)
            plt.gca().add_patch(rect)
            ax1.text(g1mid, 6.5, g1gene, fontsize=8, color='black', rotation=45)
            #ax1.plot(g1start, 6.0, 'r<', mec='red')
    else:
        if spanx1 <= g1start <= spanx2 and spanx1 <= g1stop <= spanx2:
            rect = Rectangle((g1start, 7.0), g1size, 0.5, fc=cogcolor,
                ec=cogcolor, alpha=0.5)
            plt.gca().add_patch(rect)
            ax1.text(g1mid, 7.5, g1gene, fontsize=8, color='black', rotation=45)
            #ax1.plot(g1stop, 7.0, 'r>', mec='red')

#Draw legend box for COG categories
coglist = []
for k, v in cogdict.iteritems():
    coglist.append((k,v))

coglist.sort()

ccounts = len(coglist)
a = 0
spansize = spanx2 - spanx1
spanmid = spanx1 + ((spanx2 - spanx1) / 2.0)
xstart = spanmid
#increment = 20000 #for synecoccus
increment = spansize * 0.01 #for gloeobacter
while a < ccounts:
    fc = coglist[a][1]
    tx = coglist[a][0]
    #rect = Rectangle((xstart, 2.5), 20000, 0.25, facecolor=fc,
```

```
    #  alpha=0.5) #for synecococcus
    rect = Rectangle((xstart, 2.5), increment, 0.5,
        fc=fc, alpha=0.5) #for gloeobacter
    plt.gca().add_patch(rect)
    ax1.annotate(tx, xy=(xstart+(increment/2.0), 2.2),
        horizontalalignment='center', verticalalignment='center', fontsize=8)
    a += 1
    xstart = xstart + increment

ax1.annotate('COG categories', xy=(spanmid, 1.2), horizontalalignment='left',
    verticalalignment='center', fontsize=10)
ax1.annotate(g1seq.annotations['organism'], xy=(0.5, 0.9),
    xycoords='axes fraction', horizontalalignment='center',
    verticalalignment='center', fontsize=10)

frame1 = plt.gca()
for tick in frame1.axes.get_yticklines():
    tick.set_visible(False)
for y in frame1.axes.get_yticklabels():
    y.set_visible(False)
ax1.grid(False)

plt.show()
```

D.4 dissertation_DrawMUMMER.py

```
#!/usr/bin/python

"""
This program draws MUMMER alignment results and shows connecting segments
based on % identity.

Usage: dissertation_DrawMUMMER.py g1.gbk g2.gbk g2.ptt cogs.t.list mummer.coord
Examples:
Go to this directory:
/host/Users/JS/UH-work/gloeobacter/final_work/comparisons

dissertation_DrawMUMMER.py ../annotation/GKIL.v6.gbf NC_005125.1.gbk
    NC_005125.ptt orthologs/orthomcl/cogs.t.list GKIL_vs_GVIO.coords

Author: Jimmy Saw
Date of last update: 04-23-2012

"""

import sys
import re
import matplotlib.pyplot as plt
import pylab
import matplotlib
from matplotlib import mpl
from matplotlib.patches import Rectangle
from matplotlib.transforms import Bbox
from Bio import SeqIO
from Bio.SeqUtils import GC
import matplotlib.patches as mpatch

#Regex and other stuffs
cogcat = re.compile('\[(.*)\]\t(\w+)\t.*')
pttcog = re.compile('(COG\d{4})(\w).*')
```

```
35
36  cogdict = {
37      'J' : '#2B60DE', 'A' : '#F6358A', 'K' : '#B048B5', 'L' : '#8E35EF', 'B' : '#D16587',
38      'D' : '#C38EC7', 'Y' : '#52F3FF', 'V' : '#3EA99F', 'T' : '#254117', 'M' : '#41A317',
39      'N' : '#00FF00', 'Z' : '#FFFF00', 'W' : '#FDD017', 'U' : '#F88017', 'O' : '#F660AB',
40      'C' : '#FF0000', 'G' : '#FAAFBA', 'E' : '#7F5A58', 'F' : '#C8B560', 'H' : '#8B7500',
41      'I' : '#C12869', 'P' : '#57E964', 'Q' : '#BCE954', 'R' : '#F87431', 'S' : '#ADA96E',
42      '-' : '#D3D3D3'
43      }
44
45  ##Genome one Genbank file
46  g1seq = SeqIO.read(sys.argv[1], "gb")
47  g1length = len(g1seq.seq)
48  g1featdict = {}
49  for feat in g1seq.features:
50      if feat.type == 'CDS':
51          g1featdict[feat.qualifiers['locus_tag'][0]] = feat
52      if feat.type == 'tRNA':
53          g1featdict[feat.qualifiers['locus_tag'][0]] = feat
54      if feat.type == 'rRNA':
55          g1featdict[feat.qualifiers['locus_tag'][0]] = feat
56
57  ##Genome two Genbank file
58  g2seq = SeqIO.read(sys.argv[2], "gb")
59  g2length = len(g2seq.seq)
60  g2featdict = {}
61  for feat in g2seq.features:
62      if feat.type == 'CDS':
63          g2featdict[feat.qualifiers['locus_tag'][0]] = feat
64      if feat.type == 'tRNA':
65          g2featdict[feat.qualifiers['locus_tag'][0]] = feat
66      if feat.type == 'rRNA':
67          g2featdict[feat.qualifiers['locus_tag'][0]] = feat
68
69  ##Genome 2 ptt file
70  g2cogs = {}
71  g2pttfile = open(sys.argv[3], "rU")
72  g2ptt = g2pttfile.readlines()
73  for line in g2ptt[3:]:
74      c = line.split('\t')
75      g2ltag = c[5]
76      g2gene = c[4]
77      #g2cogcat = '-'
78      g2cog = '-'
79      if pttcog.match(c[7]):
80          p = pttcog.match(c[7])
81          #g2cogcat = p.group(2)
82          g2cog = p.group(1)
83      #g2cogs[g2ltag] = g2cogcat
84      g2cogs[g2ltag] = g2cog
85
86  cogcatfile = open(sys.argv[4], "rU")
87  cfl = cogcatfile.readlines()
88
89  cogcatdict = {}
90
91  for line in cfl:
92      tmp = line.strip()
93      if cogcat.match(tmp):
94          pattern = cogcat.match(tmp)
95          cogcatdict[pattern.group(2)] = pattern.group(1)[0] #slices the first letter
96
97  cogcatfile.close()
98
```

```
#spanx1 = int(sys.argv[3])
#spanx2 = int(sys.argv[4])

glist = []

genome1x = [0, g1length]
genome1y = [2, 2]
genome2x = [0, g2length]
genome2y = [11, 11]

largergenome = 0

if g1length > g2length:
    largergenome = g1length
else:
    largergenome = g2length

##Start plotting
fig = plt.figure(1, figsize=(14,4))
#ax1 = fig.add_subplot(211) #makes the subplot and squeezes the figure to half panel
ax1 = fig.add_subplot(111) #makes the full figure plot. larger.
ax1.plot(genome1x, genome1y, color='#FFFFFF', marker='|', markersize=8.0,
    mec='black', ls='-', lw=2.0)
ax1.plot(genome2x, genome2y, color='#FFFFFF', marker='|', markersize=8.0,
    mec='black', ls='-', lw=2.0)

ax1.axis([0, largergenome, 0, 14])
#ax1.axis([spanx1, spanx2, 0, 14])

for k, v in g1featdict.iteritems():
    g1feat = v
    g1start = g1feat.location._start.position
    g1stop = g1feat.location._end.position
    g1size = g1stop - g1start + 1
    g1mid = g1start + ((g1stop - g1start) / 2.0)
    g1desc = g1feat.qualifiers['product'][0]
    g1gene = ""
    if g1feat.qualifiers.has_key('gene'):
        g1gene = g1feat.qualifiers['gene'][0] #displays gene name
        #g1gene = g1desc #displays product description
    else:
        g1gene = g1feat.qualifiers['locus_tag'][0] #displays locus tag
        #g1gene = g1desc #displays product description
    cogcolor = '#D3D3D3' #base color
    if g1feat.qualifiers.has_key('note'):
        cog = g1feat.qualifiers['note'][0]
        if cog in cogcatdict:
            cogcolor = cogdict[cogcatdict[cog]]
    if g1feat.type == 'tRNA':
        cogcolor = '#800000'
    if g1feat.type == 'rRNA':
        cogcolor = '#9400D3'
        g1gene = g1desc
    if g1feat.strand == -1:
        rect = Rectangle((g1start, 3.5), g1size, 0.25, fc=cogcolor,
            ec=cogcolor, alpha=0.5)
        plt.gca().add_patch(rect)
        #ax1.text(g1mid, 4.5, g1gene, fontsize=8, color='black', rotation=45)
    else:
        rect = Rectangle((g1start, 3.75), g1size, 0.25, fc=cogcolor,
            ec=cogcolor, alpha=0.5)
        plt.gca().add_patch(rect)
        #ax1.text(g1mid, 5.5, g1gene, fontsize=8, color='black', rotation=45)

```

```
for k, v in g2featdict.iteritems():
    g2feat = v
    g2locustag = g2feat.qualifiers['locus_tag'][0]
    g2start = g2feat.location._start.position
    g2stop = g2feat.location._end.position
    g2size = g2stop - g2start + 1
    g2mid = g2start + ((g2stop - g2start) / 2.0)
    g2desc = g2feat.qualifiers['product'][0]
    g2gene = ""
    if g2feat.qualifiers.has_key('gene'):
        g2gene = g2feat.qualifiers['gene'][0] #displays gene name
        #g1gene = g1desc #displays product description
    else:
        g2gene = g2feat.qualifiers['locus_tag'][0] #displays locus tag
        #g1gene = g1desc #displays product description
    cogcolor = '#D3D3D3' #base color
    if g2locustag in g2cogs:
        if g2cogs[g2locustag] != '-':
            #cogcolor = cogdict[g2cogs[g2locustag]]
            cogcolor = cogdict[cogcatdict[g2cogs[g2locustag]]] #check this out! :)
    if g2feat.type == 'tRNA':
        cogcolor = '#800000'
    if g2feat.type == 'rRNA':
        cogcolor = '#9400D3'
        g1gene = g1desc
    if g2feat.strand == -1:
        rect = Rectangle((g2start, 10.0), g2size, 0.25, fc=cogcolor,
            ec=cogcolor, alpha=0.5)
        plt.gca().add_patch(rect)
        #ax1.text(g2mid, 9.5, g2gene, fontsize=8, color='black', rotation=45)
    else:
        rect = Rectangle((g2start, 10.25), g2size, 0.25, fc=cogcolor,
            ec=cogcolor, alpha=0.5)
        plt.gca().add_patch(rect)
        #ax1.text(g2mid, 10.5, g2gene, fontsize=8, color='black', rotation=45)

ax1.annotate('GKIL', xy=(0.97, 0.2), xycoords='axes fraction',
    horizontalalignment='center', verticalalignment='center', fontsize=10)
ax1.annotate('GVIO', xy=(0.97, 0.9), xycoords='axes fraction',
    horizontalalignment='center', verticalalignment='center', fontsize=10)

##Parse MUMMER alignment file

mummerfile = open(sys.argv[5], "rU")
mfl = mummerfile.readlines()
for line in mfl[4:]:
    c = line.split('\t')
    g1start = int(c[0])
    g1stop = int(c[1])
    g2start = int(c[2])
    g2stop = int(c[3])
    ident = float(c[6])
    fillcolor = '#AAAAAA'
    if ident >= 90:
        fillcolor = '#FF0000'
    elif ident >= 80:
        fillcolor = '#71C671'
    elif ident >= 70:
        fillcolor = '#7171C6'
    elif ident >= 60:
        fillcolor = '#CDB5CD'
    else:
        fillcolor = '#C5C1AA'
    x = [g1start, g2start, g2stop, g1stop]
```

```
    y = [4, 10, 10, 4]
    ax1.fill(x, y, color=fillcolor, alpha=0.3)

#draw legend for % identities
pcx = 4750000
pctlgndx = [pcx, pcx, pcx, pcx, pcx]
pctlgndy = [7.75, 7.5, 7.25, 7.0, 6.75]
pctfc = ['#FF0000', '#71C671', '#7171C6', '#CDB5CD', '#C5C1AA']
pcttx = ['>=90%', '>=80%', '>=70%', '>=60%', '< 60%']
for a, b, c, d in zip(pctlgndx, pctlgndy, pctfc, pcttx):
    prect = Rectangle((a-42000, b), 40000, 0.25, facecolor=c, alpha=0.5)
    plt.gca().add_patch(prect)
    ax1.annotate(d, xy=(a, b), fontsize=7)

#Draw legend box for COG categories

coglist = []
for k, v in cogdict.iteritems():
    coglist.append((k,v))

coglist.sort()

ccounts = len(coglist)
a = 0
xstart = 2500000
#increment = 20000 #for synecoccus
increment = 40000 #for gloeobacter
while a < ccounts:
    fc = coglist[a][1]
    tx = coglist[a][0]
    #rect = Rectangle((xstart, 2.5), 20000, 0.25, facecolor=fc,
    #    alpha=0.5) #for synecococcus
    rect = Rectangle((xstart, 2.5), 40000, 0.25, facecolor=fc, alpha=0.5) #for gloeobacter
    plt.gca().add_patch(rect)
    ax1.annotate(tx, xy=(xstart+20000, 2.3), horizontalalignment='center',
        verticalalignment='center', fontsize=8)
    a += 1
    xstart = xstart + increment
ax1.annotate('COG categories', xy=(2000000, 2.3), fontsize=8)

frame1 = plt.gca()
for tick in frame1.axes.get_yticklines():
    tick.set_visible(False)
for y in frame1.axes.get_yticklabels():
    y.set_visible(False)
ax1.grid(False)

plt.show()
```

D.5 dissertation_DrawMUMMERwithPtt.py

```
#!/usr/bin/python

"""
This program draws MUMMER alignment results and shows connecting segments
based on % identity.

Usage: dissertation_DrawMUMMER.py g1.gbk g2.gbk g1.ptt g2.ptt cogs.t.list mummer.coord
Examples:
Go to this directory:
```

```
/host/Users/JS/UH-work/gloeobacter/final_work/comparisons

dissertation_DrawMUMMERwithPtt.py NC_007776.gbk NC_007775.gbk NC_007776.ptt
    NC_007775.ptt orthologs/orthomcl/cogs.t.list NC_007776_vs_NC_007775.coords
dissertation_DrawMUMMERwithPtt.py NC_007516.gbk NC_007513.gbk NC_007516.ptt
    NC_007513.ptt orthologs/orthomcl/cogs.t.list NC_007516_vs_NC_007513.coords
dissertation_DrawMUMMERwithPtt.py NC_011748.gbk NC_008253.gbk NC_011748.ptt
    NC_008253.ptt orthologs/orthomcl/cogs.t.list NC_011748_vs_NC_008253.coords
dissertation_DrawMUMMERwithPtt.py NC_002737.gbk NC_007297.gbk NC_002737.ptt
    NC_007297.ptt orthologs/orthomcl/cogs.t.list NC_002737_vs_NC_007297.coords

Steps:
1. Download fna, gbk, and ptt files
2. Run MUMMER
3. Run this program

Author: Jimmy Saw
Date of last update: 04-23-2012

"""

import sys
import re
import matplotlib.pyplot as plt
import pylab
import matplotlib
from matplotlib import mpl
from matplotlib.patches import Rectangle
from matplotlib.transforms import Bbox
from Bio import SeqIO
from Bio.SeqUtils import GC
import matplotlib.patches as mpatch

#Regex and other stuffs
cogcat = re.compile('\[(.*)\]\t(\w+)\t.*')
#pttcog = re.compile('(COG\d+)(\w).*')
pttcog = re.compile('(COG\d{4})(\w).*')

cogdict = {
    'J' : '#2B60DE', 'A' : '#F6358A', 'K' : '#B048B5', 'L' : '#8E35EF', 'B' : '#D16587',
    'D' : '#C38EC7', 'Y' : '#52F3FF', 'V' : '#3EA99F', 'T' : '#254117', 'M' : '#41A317',
    'N' : '#00FF00', 'Z' : '#FFFF00', 'W' : '#FDD017', 'U' : '#F88017', 'O' : '#F660AB',
    'C' : '#FF0000', 'G' : '#FAAFBA', 'E' : '#7F5A58', 'F' : '#C8B560', 'H' : '#8B7500',
    'I' : '#C12869', 'P' : '#57E964', 'Q' : '#BCE954', 'R' : '#F87431', 'S' : '#ADA96E',
    '-' : '#D3D3D3'
    }

##Genome 1 Genbank file
g1seq = SeqIO.read(sys.argv[1], "gb")
g1length = len(g1seq.seq)
g1featdict = {}
for feat in g1seq.features:
    if feat.type == 'CDS':
        g1featdict[feat.qualifiers['locus_tag'][0]] = feat
    if feat.type == 'tRNA':
        g1featdict[feat.qualifiers['locus_tag'][0]] = feat
    if feat.type == 'rRNA':
        g1featdict[feat.qualifiers['locus_tag'][0]] = feat

##Genome 2 Genbank file
g2seq = SeqIO.read(sys.argv[2], "gb")
g2length = len(g2seq.seq)
g2featdict = {}
for feat in g2seq.features:
```

```
    if feat.type == 'CDS':
        g2featdict[feat.qualifiers['locus_tag'][0]] = feat
    if feat.type == 'tRNA':
        g2featdict[feat.qualifiers['locus_tag'][0]] = feat
    if feat.type == 'rRNA':
        g2featdict[feat.qualifiers['locus_tag'][0]] = feat

##Genome 1 ptt file
g1cogs = {}
g1pttfile = open(sys.argv[3], "rU")
g1ptt = g1pttfile.readlines()
for line in g1ptt[3:]:
    c = line.split('\t')
    g1ltag = c[5]
    g1gene = c[4]
    #g1cogcat = '-'
    g1cog = '-'
    if pttcog.match(c[7]):
        p = pttcog.match(c[7])
        #g1cogcat = p.group(2)
        g1cog = p.group(1)
    #g1cogs[g1ltag] = g1cogcat
    g1cogs[g1ltag] = g1cog

##Genome 2 ptt file
g2cogs = {}
g2pttfile = open(sys.argv[4], "rU")
g2ptt = g2pttfile.readlines()
for line in g2ptt[3:]:
    c = line.split('\t')
    g2ltag = c[5]
    g2gene = c[4]
    #g2cogcat = '-'
    g2cog = '-'
    if pttcog.match(c[7]):
        p = pttcog.match(c[7])
        #g2cogcat = p.group(2)
        g2cog = p.group(1)
    #g2cogs[g2ltag] = g2cogcat
    g2cogs[g2ltag] = g2cog

cogcatfile = open(sys.argv[5], "rU")
cfl = cogcatfile.readlines()

cogcatdict = {}

for line in cfl:
    tmp = line.strip()
    if cogcat.match(tmp):
        pattern = cogcat.match(tmp)
        cogcatdict[pattern.group(2)] = pattern.group(1)[0]

cogcatfile.close()

#spanx1 = int(sys.argv[7])
#spanx2 = int(sys.argv[8])

glist = []

genome1x = [0, g1length]
genome1y = [2, 2]
genome2x = [0, g2length]
genome2y = [12, 12]

```

```
largergenome = 0

if g1length > g2length:
    largergenome = g1length
else:
    largergenome = g2length

##Start plotting
fig = plt.figure(1, figsize=(14,4))
#ax1 = fig.add_subplot(211) #makes the subplot and squeezes the figure to half panel
ax1 = fig.add_subplot(111) #makes the full figure plot. larger.
ax1.plot(genome1x, genome1y, color='#FFFFFF', marker='|', markersize=8.0,
    mec='black', ls='-', lw=2.0)
ax1.plot(genome2x, genome2y, color='#FFFFFF', marker='|', markersize=8.0,
    mec='black', ls='-', lw=2.0)

ax1.axis([0, largergenome, 0, 14])
#ax1.axis([spanx1, spanx2, 0, 14])

for k, v in g1featdict.iteritems():
    g1feat = v
    g1locustag = g1feat.qualifiers['locus_tag'][0]
    g1start = g1feat.location._start.position
    g1stop = g1feat.location._end.position
    g1size = g1stop - g1start + 1
    g1mid = g1start + ((g1stop - g1start) / 2.0)
    g1desc = g1feat.qualifiers['product'][0]
    g1gene = ""
    if g1feat.qualifiers.has_key('gene'):
        g1gene = g1feat.qualifiers['gene'][0] #displays gene name
        #g1gene = g1desc #displays product description
    else:
        g1gene = g1feat.qualifiers['locus_tag'][0] #displays locus tag
        #g1gene = g1desc #displays product description
    cogcolor = '#D3D3D3' #base color
    if g1locustag in g1cogs:
        if g1cogs[g1locustag] != '-':
            #cogcolor = cogdict[g1cogs[g1locustag]]
            cogcolor = cogdict[cogcatdict[g1cogs[g1locustag]]]
    if g1feat.type == 'tRNA':
        cogcolor = '#800000'
    if g1feat.type == 'rRNA':
        cogcolor = '#9400D3'
        g1gene = g1desc
    if g1feat.strand == -1:
        rect = Rectangle((g1start, 4.0), g1size, 0.5, fc=cogcolor,
            ec=cogcolor, alpha=0.5)
        plt.gca().add_patch(rect)
        #ax1.text(g1mid, 4.5, g1gene, fontsize=8, color='black', rotation=45)
    else:
        rect = Rectangle((g1start, 4.5), g1size, 0.5, fc=cogcolor,
            ec=cogcolor, alpha=0.5)
        plt.gca().add_patch(rect)
        #ax1.text(g1mid, 5.5, g1gene, fontsize=8, color='black', rotation=45)

for k, v in g2featdict.iteritems():
    g2feat = v
    g2locustag = g2feat.qualifiers['locus_tag'][0]
    g2start = g2feat.location._start.position
    g2stop = g2feat.location._end.position
    g2size = g2stop - g2start + 1
    g2mid = g2start + ((g2stop - g2start) / 2.0)
    g2desc = g2feat.qualifiers['product'][0]
    g2gene = ""
```

```
    if g2feat.qualifiers.has_key('gene'):
        g2gene = g2feat.qualifiers['gene'][0] #displays gene name
        #g1gene = g1desc #displays product description
    else:
        g2gene = g2feat.qualifiers['locus_tag'][0] #displays locus tag
        #g1gene = g1desc #displays product description
    cogcolor = '#D3D3D3' #base color
    if g2locustag in g2cogs:
        if g2cogs[g2locustag] != '-':
            #cogcolor = cogdict[g2cogs[g2locustag]]
            cogcolor = cogdict[cogcatdict[g2cogs[g2locustag]]]
    if g2feat.type == 'tRNA':
        cogcolor = '#800000'
    if g2feat.type == 'rRNA':
        cogcolor = '#9400D3'
        g1gene = g1desc
    if g2feat.strand == -1:
        rect = Rectangle((g2start, 9.0), g2size, 0.5, fc=cogcolor,
            ec=cogcolor, alpha=0.5)
        plt.gca().add_patch(rect)
        #ax1.text(g2mid, 9.5, g2gene, fontsize=8, color='black', rotation=45)
    else:
        rect = Rectangle((g2start, 9.5), g2size, 0.5, fc=cogcolor,
            ec=cogcolor, alpha=0.5)
        plt.gca().add_patch(rect)
        #ax1.text(g2mid, 10.5, g2gene, fontsize=8, color='black', rotation=45)

ax1.annotate(g1seq.annotations['organism'], xy=(0.5, 0.1),
    xycoords='axes fraction', horizontalalignment='center', verticalalignment='center',
      fontsize=10)
ax1.annotate(g2seq.annotations['organism'], xy=(0.5, 0.9),
    xycoords='axes fraction', horizontalalignment='center', verticalalignment='center',
      fontsize=10)

##Parse MUMMER alignment file

mummerfile = open(sys.argv[6], "rU")
mfl = mummerfile.readlines()
for line in mfl[4:]:
    c = line.split('\t')
    g1start = int(c[0])
    g1stop = int(c[1])
    g2start = int(c[2])
    g2stop = int(c[3])
    ident = float(c[6])
    fillcolor = '#AAAAAA'
    if ident >= 90:
        fillcolor = '#FF0000'
    elif ident >= 80:
        fillcolor = '#71C671'
    elif ident >= 70:
        fillcolor = '#7171C6'
    elif ident >= 60:
        fillcolor = '#CDB5CD'
    else:
        fillcolor = '#C5C1AA'
    x = [g1start, g2start, g2stop, g1stop]
    y = [5, 9, 9, 5]
    ax1.fill(x, y, color=fillcolor, alpha=0.2)

frame1 = plt.gca()
for tick in frame1.axes.get_yticklines():
    tick.set_visible(False)
for y in frame1.axes.get_yticklabels():
```

```
    y.set_visible(False)
ax1.grid(False)

plt.show()
```

D.6 dissertation_GapCloserMinimo.py

```
#!/usr/bin/python
"""
Author:                         Jimmy Saw
Date modified:          10-22-2011

Description:    This script can generate a list of reads generated from shreds
                of contigs (from Celera assembly) spanning two contigs scaffolds
                to help close gaps between these scaffolds. It prints something like this:

Between  sctg_0001_0001  and  sctg_0001_0002 ctg220003834103_1200_1700 68.2
Between  sctg_0001_0002  and  sctg_0001_0003 ctg220003834132_45000_45500 36.2
Between  sctg_0001_0003  and  sctg_0001_0004 ctg220003834132_21600_22100 45.6
Between  sctg_0001_0004  and  sctg_0001_0005 ctg220003834123_127200_127700 23.6
Between  sctg_0001_0006  and  sctg_0001_0007 ctg220003834091_14400_14900 17.6
Between  sctg_0001_0006  and  sctg_0001_0007 ctg220003834091_14100_14600 77.6
Between  sctg_0001_0008  and  sctg_0001_0009 ctg220003834092_11100_11600 39.2
Between  sctg_0001_0009  and  sctg_0001_0010 ctg220003834092_16500_17000 49.4
Between  sctg_0001_0010  and  sctg_0001_0011 ctg220003834092_66300_66800 24.0
Between  sctg_0001_0011  and  sctg_0001_0012 ctg220003834092_164700_165200 19.8
Between  sctg_0001_0012  and  sctg_0001_0013 ctg220003834128_31500_32000 33.6
Between  sctg_0001_0014  and  sctg_0001_0015 ctg220003834094_2400_2900 50.0

                Need to work further to call Minimo (AMOS package) to run
                assembly automatically. Currently, need to extract these reads
                and combine with original contig scaffold in a fasta and
                manually and run Minimo with the following command (example):

Minimo testgap_0069_0001.fasta -D QUAL_IN=testgap_0069_0001.qual -D MIN_LEN=30
  -D FASTA_EXP=1 -D ACE_EXP=1 -D OUT_PREFIX=tmp_69_1

Usage:          dissertation_GapCloserMinimo.py <list of 454 contig scaffolds>

Example:        dissertation_GapCloserMinimo.py sctgs.list

Note:           Run from this folder:
/host/Users/JS/UH-work/gloeobacter/final_assembly/newbler/454GapSeqsConsed/assembly/
contig_scaffs
Needs to follow these steps:

1. Shred Celera Contigs into smaller chunks
2. Run MUMMER of Newbler assembly scaffolds vs. these shreds and generate .coord files
3. Create a list of Newbler assembly scaffold files
4. Run this script
"""
import sys
import os
import re
from Bio import SeqIO
from subprocess import call

def fmt(f):
    st = '{0:.4}'.format(f)
    return st
```

```
54
55  sctgfile = open(sys.argv[1], "rU")
56  sctgs = sctgfile.readlines()
57
58  mummerlist = []
59
60  for index, i in enumerate(sctgs):
61      scaflist = []
62      celera = open(i.strip()+".celerashreds.coords", "rU") #check Celera shreds
63  #     velvet = open(i.strip()+".velvetshreds.coords", "rU") #check Velvet shreds
64      scaflist.append(i)
65      cl = celera.readlines()
66      tmp = cl[4]
67      t = tmp.split('\t')
68      ctgsize = int(t[7])
69      ctgbegin = {}
70      ctgend = {}
71      for line in cl[4:]:
72          l = line.split('\t')
73          scafstart = int(l[0])
74          scafstop = int(l[1])
75          readname = l[-1].rstrip()
76          scafname = l[-2]
77          qcoverage = float(l[-3])
78          if qcoverage < 100: #if alignment coverage is < 100, the rest is in another
79                              #contig
80              if scafstop < 1000: #work on beginning of contig
81                  ctgbegin[readname] = ((scafstart, scafstop, qcoverage)) #dictionary
82                                                                          #of tuples
83              if scafstart > ctgsize - 1000: #work on end of contig
84                  ctgend[readname] = ((scafstart, scafstop, qcoverage))
85      nctgbegin = {}
86      nctgend = {}
87      if index < len(sctgs)-1:
88          nextscaf = sctgs[index+1]
89          nc = open(nextscaf.strip()+".celerashreds.coords", "rU")
90  #        nv = open(nextscaf.strip()+".celerashreds.coords", "rU")
91          ncl = nc.readlines()
92          tmp2 = ncl[4]
93          t2 = tmp2.split('\t')
94          nctgsize = int(t2[7])
95          for line in ncl[4:]:
96              l = line.split('\t')
97              nscafstart = int(l[0])
98              nscafstop = int(l[1])
99              nreadname = l[-1].rstrip()
100             nscafname = l[-2]
101             nqcoverage = float(l[-3])
102             if nqcoverage < 100:
103                 if nscafstop < 1000:
104                     nctgbegin[nreadname] = ((nscafstart, nscafstop, nqcoverage))
105                 if nscafstart > nctgsize - 1000:
106                     nctgend[nreadname] = ((nscafstart, nscafstop, nqcoverage))
107
108     for k, v in ctgend.iteritems():
109         if k in nctgbegin:
110             print "Between ", i.strip(), " and ", sctgs[index+1].strip(), k, fmt(v[2])
111 #            tmpfasta = ""
112 #            call(["exfasta", k, "celera.shredded.ctgs.fasta"])
113
114     celera.close()
115     nc.close()
116 #generate list of gap-spanning shreds
117
```

```
#minimocmd = os.system("Minimo testgap_0069_0001.fasta -D QUAL_IN=testgap_0069_0001.qual\
#     -D MIN_LEN=30 -D FASTA_EXP=1 -D ACE_EXP=1 -D OUT_PREFIX=tmp_69_1")
```

D.7 dissertation_GCskew.py

```
#!/usr/bin/python
"""
This program makes data points for GC skew plot

Usage:
Examples:
dissertation_GCskew.py GKIL.v6.gbf percent
dissertation_GCskew.py GKIL.v6.gbf skew

Note:

Author: Jimmy Saw
Date: 04-28-2012
"""

import sys
import re
import matplotlib.pyplot as plt
import pylab
import random
from Bio import SeqIO
from Bio.SeqUtils import GC

genome = SeqIO.read(sys.argv[1], "gb")
genome_size = len(genome.seq)

choice = str(sys.argv[2])

def calAvg(countlist):
    total = 0
    count = len(countlist)
    for i in countlist:
        total += i
    avg = float(total/count)
    return avg

def skew(seq):
    g = 0
    c = 0
    a = 0
    t = 0
    for i in seq:
        if i == 'G':
            g += 1
        elif i == 'C':
            c += 1
        elif i == 'A':
            a += 1
        elif i == 'T':
            t += 1
    gc = g + c
    at = a + t
    total = gc + at
    percentgc = (float(gc) / float(total)) * 100
    gcskew = float(g-c)/float(g+c) * 100
```

```
    #return gcskew
    if choice == 'skew':
        return gcskew
    elif choice == 'percent':
        return percentgc

#calculate skew
skewpoints = []

x = 1000
while x < len(genome.seq):
    start = x - 1000
    stop = x
    gc_skew = skew(genome.seq[start:stop])
    skewpoints.append((start, stop, gc_skew))
    x += 1000 #increment by 1kb

for i, j in enumerate(skewpoints):
    print "chr1", j[0], j[1], j[2]
```

D.8 dissertation_GloeoAsmVerification.py

```
#!/usr/bin/python
"""
Author: Jimmy Saw
Last updated: 06-20-2012
Usage:
dissertation_GloeoAsmVerification.py glbkl_vs_AtleastOneAndSingletons.coords
  glbkl_vs_non-gloeo.coords glbkl_vs_454scaffolds.coords glbkl_vs_celeractgs.coords
  binned.gloeo.pairs.txt ../annotation/fixed.final_assembly_noCN4.fasta

Run the above command in this folder:
/host/Users/JS/UH-work/gloeobacter/final_work/assembly_verification/

Usage: python gloeoAssemblyVerificationFigure2.py 1 2 3 4 5
1: mummer coordinate file of assembled contig/genome vs.
   phymmBL-binned mate pairs and singletons
2: mummer coordinate file of assembled contig/genome vs.
   phymmBL-binnned non-gloeobacter reads
3: mummer coordinate file of assembled contig/genome vs.
   Newbler contigs
4: mummer coordinate file of assembled contig/genome vs.
   Contigs from other assemblers (Velvet or Celera)
5: mummer coordinate file of assembled contig/genome vs.
   mate pairs list (binned gloeobacter reads)
6: Fasta file of assembled contig(only one) or assembled genome

mate pair list looks like this:
GM6SIKE01AULG1  L->R    1 -> 9218       9093
GM6SIKE01B095V  L->R    1080 -> 11672   10261
GM6SIKE01ARQG1  L->R    2009 -> 13492   11328
GM6SIKE01AW9LY  L->R    3003 -> 13365   10120
GM6SIKE01BWLSW  L->R    4031 -> 14075   9896
GM6SIKE01BNO9P  R->L    5081 -> 13130   7966
GM6SIKE01CF48K  L->R    6078 -> 16829   10665
GM6SIKE01AWV4O  L->R    7036 -> 16355   9232

"""

import sys
```

```
import matplotlib.pyplot as plt
import pylab
import re
import random
from Bio import SeqIO
from Bio.SeqUtils import GC

##Functions
def fmt(f):
    st = '{0:.4}'.format(f)
    return st

def calAvg(countlist):
    total = 0
    count = len(countlist)
    for i in countlist:
        total += i
    avg = float(total/count)
    return avg

##regular expressions
m0 = re.compile('.*0$')
m1 = re.compile('.*1$')
m2 = re.compile('.*2$')
m3 = re.compile('.*3$')
m4 = re.compile('.*4$')
m5 = re.compile('.*5$')
m6 = re.compile('.*6$')
m7 = re.compile('.*7$')
m8 = re.compile('.*8$')
m9 = re.compile('.*9$')
left = re.compile('\w+_L')
right = re.compile('\w+_R')
#ctg = re.compile('sctg_\d+_\d+')

##Gloeobacter-specific reads
file1 = sys.argv[1]
f1 = open(file1, "rU")
fl1 = f1.readlines()

##Other reads
file2 = sys.argv[2]
f2 = open(file2, "rU")
fl2 = f2.readlines()

##454 contigs alignment
file3 = sys.argv[3]
f3 = open(file3, "rU")
fl3 = f3.readlines()

##Solexa contigs alignment
file4 = sys.argv[4]
f4 = open(file4, "rU")
fl4 = f4.readlines()

##phymmBL binning results

#Gloeobacter clone pairs
file5 = sys.argv[5]
f5 = open(file5, "rU")
fl5 = f5.readlines()

##Sequence file
seqfile = sys.argv[6]
```

```
sf = SeqIO.read(seqfile, "fasta")
gc_values = []
gcx = []

i = 0

while i < len(sf.seq):
    gc = GC(sf.seq[i:i+1000])
    gc_values.append(gc)
    gcx.append(i)
    i += 1000

##Parsing the file1 contents
line = fl1[4]
l = line.split('\t')
genome_size = int(l[7])

fl_coords = {}

for i in range(1, genome_size+1):
    fl_coords[i] = 0

pairsx1 = []
pairsx2 = []
pairsy = []

singletonsx1 = []
singletonsx2 = []
singletonsy = []

for i, line in enumerate(fl1[4:]):
    l = line.split('\t')
    start = int(l[0])
    stop = int(l[1])
    readname = l[12].rstrip()
    identity = float(l[6])
    for a in range(start, stop):
        fl_coords[a] = fl_coords[a] + 1

    if left.match(readname) or right.match(readname):
        pairsx1.append(start)
        pairsx2.append(stop)
        if m0.match(str(start)):
            pairsy.append(11.2)
        elif m1.match(str(start)):
            pairsy.append(11.4)
        elif m2.match(str(start)):
            pairsy.append(11.6)
        elif m3.match(str(start)):
            pairsy.append(11.8)
        elif m4.match(str(start)):
            pairsy.append(12)
        elif m5.match(str(start)):
            pairsy.append(12.2)
        elif m6.match(str(start)):
            pairsy.append(12.4)
        elif m7.match(str(start)):
            pairsy.append(12.6)
        elif m8.match(str(start)):
            pairsy.append(12.8)
        elif m9.match(str(start)):
            pairsy.append(13)
    else:
        singletonsx1.append(start)
```

```
        singletonsx2.append(stop)
        if m0.match(str(start)):
            singletonsy.append(11.2)
        elif m1.match(str(start)):
            singletonsy.append(11.4)
        elif m2.match(str(start)):
            singletonsy.append(11.6)
        elif m3.match(str(start)):
            singletonsy.append(11.8)
        elif m4.match(str(start)):
            singletonsy.append(12)
        elif m5.match(str(start)):
            singletonsy.append(12.2)
        elif m6.match(str(start)):
            singletonsy.append(12.4)
        elif m7.match(str(start)):
            singletonsy.append(12.6)
        elif m8.match(str(start)):
            singletonsy.append(12.8)
        elif m9.match(str(start)):
            singletonsy.append(13)

sxa = [singletonsx1, singletonsx2]
sxb = [singletonsy, singletonsy]

cov1 = []

for k, v in f1_coords.iteritems():
    cov1.append(v)

w = 1000
depthx1 = []
depthy1 = []
while w < len(cov1):
    average = calAvg(cov1[w-1000:w])
    depthx1.append(w-1000+500)
    depthy1.append(average)
    w += 1000

##Parsing the file2 contents

f2_coords = {}
for i in range(1, genome_size+1):
    f2_coords[i] = 0

for j, line in enumerate(fl2[4:]):
    l = line.split('\t')
    start = int(l[0])
    stop = int(l[1])
    identity = float(l[6])
    for a in range(start, stop):
        f2_coords[a] = f2_coords[a] + 1

cov2 = []

for k, v in f2_coords.iteritems():
    cov2.append(v)

x = 1000
depthx2 = []
depthy2 = []
while x < len(cov2):
    average = calAvg(cov2[x-1000:x])
    depthx2.append(x+500)
```

```
    depthy2.append(average)
    x += 1000

##For line representing the assembled genome
gx = [1,genome_size]
gy = [1,1]

xcoords = range(1, genome_size+1)
#ycoords = cov

##Parsing the file3 contents
##For 454 contigs assembled with Newbler
x1 = []
x2 = []
y454 = []

for i, line in enumerate(fl3[4:]):
    l = line.split('\t')
    start = int(l[0])
    stop = int(l[1])
    x1.append(start)
    x2.append(stop)
    if (i+1-1)%2 == 0:
        y454.append(6)
    else:
        y454.append(5.5)
a = [x1, x2]
b = [y454, y454]

##Parsing the file4 contents
##For Solex contigs assembled with Velvet
x3 = []
x4 = []
ys = []

for i, line in enumerate(fl4[4:]):
    l = line.split('\t')
    start = int(l[0])
    stop = int(l[1])
    x3.append(start)
    x4.append(stop)
    if (i+1-1)%2 == 0:
        ys.append(4)
    else:
        ys.append(3.5)
c = [x3, x4]
d = [ys, ys]

##Parsing the file5 contents
##Gloeobacter clone pairs
gpx1 = []
gpx2 = []
gpy = []

for i, line in enumerate(fl5):
    l = line.split('\t')
    start = int(l[2].split(' ')[0])
    stop = int(l[2].split(' ')[2])
    gpx1.append(start)
    gpx2.append(stop)
    if m0.match(str(i)):
        gpy.append(7.2)
    elif m1.match(str(i)):
        gpy.append(7.4)
```

```
    elif m2.match(str(i)):
        gpy.append(7.6)
    elif m3.match(str(i)):
        gpy.append(7.8)
    elif m4.match(str(i)):
        gpy.append(8)
    elif m5.match(str(i)):
        gpy.append(8.2)
    elif m6.match(str(i)):
        gpy.append(8.4)
    elif m7.match(str(i)):
        gpy.append(8.6)
    elif m8.match(str(i)):
        gpy.append(8.8)
    elif m9.match(str(i)):
        gpy.append(9)

gpa = [gpx1, gpx2]
gpb = [gpy, gpy]

##Close the file handlers
f1.close()
f2.close()
f3.close()
f4.close()

##Print notice..
print "# of records in xcoords: ", len(xcoords)
print "# of records in cov1: ", len(cov1)

##Start plotting
fig = plt.figure(1, figsize=(15,8))
#plt.subplots_adjust(wspace=4.0)

##First subplot
ax1 = fig.add_subplot(211)
#plt.subplot(311)
#plt.title('Assembly verification')
ax1.plot(gx, gy, color='#AF7817', marker='|', markersize=8.0,
    mec='black', ls='-', lw=2.0)
ax1.plot(sxa, sxb, color='#333366', linestyle='-')
ax1.plot(a, b, color='purple', linestyle='-', lw=2.0)
ax1.plot(c, d, color='red', linestyle='-', lw=2.0)
ax1.plot(gpa, gpb, color='#ADA96E', linestyle='-')
ax1.annotate('Gloeobacter singleton reads', xy=(0.8, 0.95),
    xycoords='axes fraction', horizontalalignment='center',
    verticalalignment='center', fontsize=9)
ax1.annotate('Mate pairs (at least one is binned as Gloeobacter)',
    xy=(0.8, 0.68), xycoords='axes fraction', horizontalalignment='center',
    verticalalignment='center', fontsize=9)
ax1.annotate('Newbler contigs (454)', xy=(0.8, 0.47), xycoords='axes fraction',
    horizontalalignment='center', verticalalignment='center', fontsize=9)
ax1.annotate('Celera contigs (454+Illumina)', xy=(0.8, 0.20),
    xycoords='axes fraction', horizontalalignment='center',
    verticalalignment='center', fontsize=9)

ax1.axis([0, genome_size, 0, 14])

frame1 = plt.gca()
for tick in frame1.axes.get_yticklines():
    tick.set_visible(False)
for y in frame1.axes.get_yticklabels():
    y.set_visible(False)
ax1.grid(False)
```

```
359
360 ##Second subplot
361 ax2 = fig.add_subplot(212)
362 #plt.subplot(312)
363 ax2.fill_between(depthx2, depthy2, facecolor='#33FF33', alpha=0.4)
364 ax2.fill_between(depthx1, depthy1, facecolor='#660066', alpha=0.4)
365 #ax2.annotate('Green = other organisms', xy=(0.2, 0.80),
366 #    xycoords='axes fraction', horizontalalignment='center',
367 #    verticalalignment='center')
368 #ax2.annotate('Purple = Gloeobacter', xy=(0.2, 0.60), xycoords='axes fraction',
369 #    horizontalalignment='center', verticalalignment='center')
370 ax2.axvspan(2728368, 2733167, facecolor='blue', alpha=0.4)
371 ax2.annotate('rRNA operon', xy=(2728368, 40), xycoords='data', xytext=(30, 0),
372     textcoords='offset points', arrowprops=dict(arrowstyle="->"))
373
374 ax2.axis([0, genome_size, 0, max(depthy1)+2])
375
376 ax2.set_xlabel('Genome position (bp)')
377 ax2.set_ylabel('Counts of binned reads (per 1000bp)')
378
379 ax3 = ax2.twinx()
380 ax3.plot(gcx, gc_values, color='blue', linestyle='-', alpha=0.4)
381 ax3.axis([0, genome_size, 0, 80])
382
383 ax3.set_ylabel('G+C %')
384
385 ax1.axvspan(415517, 431636, facecolor='#FFD700', alpha=0.2)
386 ax1.axvspan(903207, 946727, facecolor='#FFD700', alpha=0.2)
387 ax1.axvspan(1004060, 1015740, facecolor='#FFD700', alpha=0.2)
388 ax1.axvspan(1732580, 1828970, facecolor='#FFD700', alpha=0.2)
389 ax1.axvspan(4262380, 4279810, facecolor='#FFD700', alpha=0.2)
390
391 ax2.axvspan(415517, 431636, facecolor='#FFD700', alpha=0.2)
392 ax2.axvspan(903207, 946727, facecolor='#FFD700', alpha=0.2)
393 ax2.axvspan(1004060, 1015740, facecolor='#FFD700', alpha=0.2)
394 ax2.axvspan(1732580, 1828970, facecolor='#FFD700', alpha=0.2)
395 ax2.axvspan(4262380, 4279810, facecolor='#FFD700', alpha=0.2)
396
397 ax2.grid(True)
398
399 plt.show()
400
401 #plt.savefig(outfile, format='pdf')
```

D.9 dissertation_IgsBlast.py

```
1  #!/usr/bin/python
2  """
3  Usage: python auto_anno.py annofile.txt seqfile.fasta
4  Author:                          Jimmy Saw
5  Date modified:          04-24-2011
6
7  Description:     This script can extract intergenic regions and BLAST them to
8                   get hits to known protein sequences.
9
10 Usage:           dissertation_IgsBlast.py <cog count file>
11
12 Example:         dissertation_IgsBlast.py annotation.tab seq.fasta
13 Note:
14 """
```

```
15
16  import sys
17  from Bio import SeqIO
18  from Bio.Blast import NCBIWWW
19
20  annofile = sys.argv[1]
21  seqfile = sys.argv[2]
22  prefix = sys.argv[3]
23  af = open(annofile, "rU")
24  sf = open(seqfile, "rU")
25  rec = SeqIO.read(sf, "fasta")
26  lines = af.readlines()
27  num = len(lines)
28
29  i = 0
30
31  while i < num:
32      if i == 0:
33          curr_line = lines[i].split('\t')
34          curr_id = curr_line[0]
35          curr_start = int(curr_line[3])
36          substop = curr_start - 1
37          igs_id = prefix + "_IGS_" + str(i).zfill(4) + "_" + "0-" + str(substop)
38          subseq = rec.seq[0:substop]
39          if len(subseq) > 90:
40              print "Running BLASTx of " + igs_id
41              result_handle = NCBIWWW.qblast("blastx", "nr", subseq,
42                  expect=0.00001, filter=None)
43              save_file = open(igs_id + ".xml", "w")
44              save_file.write(result_handle.read())
45              save_file.close()
46              result_handle.close()
47              print "Done BLASTx"
48
49      if i == num - 1:
50          curr_line = lines[i].split('\t')
51          curr_stop = int(curr_line[4])
52          substart = curr_stop + 1
53          substop = len(rec.seq)
54          igs_id = prefix + "_IGS_" + str(i).zfill(4) + "_" + str(substart) \
55              + "-" + str(substop)
56          subseq = rec.seq[substart:substop]
57          if len(subseq) > 90:
58              print "Running BLASTx of " + igs_id
59              result_handle = NCBIWWW.qblast("blastx", "nr", subseq,
60                  expect=0.00001, filter=None)
61              save_file = open(igs_id + ".xml", "w")
62              save_file.write(result_handle.read())
63              save_file.close()
64              result_handle.close()
65              print "Done BLASTx"
66
67      if i != num - 1 and i != 0:
68          curr_line = lines[i].split('\t')
69          curr_id = curr_line[0]
70          curr_locus_tag = curr_line[1]
71          curr_feat_type = curr_line[2]
72          curr_start = int(curr_line[3])
73          curr_stop = int(curr_line[4])
74          curr_frame = curr_line[5]
75
76          next_line = lines[i + 1].split('\t')
77          next_id = next_line[0]
78          next_locus_tag = next_line[1]
```

```
        next_feat_type = next_line[2]
        next_start = int(next_line[3])
        next_stop = int(next_line[4])
        next_frame = next_line[5]

        prev_line = lines[i - 1].split('\t')
        prev_id = prev_line[0]
        prev_locus_tag = prev_line[1]
        prev_feat_type = prev_line[2]
        prev_start = int(prev_line[3])
        prev_stop = int(prev_line[4])
        prev_frame = prev_line[5]

        if curr_feat_type == "CDS":
            substart = prev_stop + 1
            substop = curr_start - 1
            if substop > substart:
                igs_id = prefix + "_IGS_" + str(i).zfill(4) + "_" \
                    + str(substart) + "-" + str(substop)
                subseq = rec.seq[substart:substop]
                if len(subseq) > 90:
                    print "Running BLASTx of " + igs_id
                    result_handle = NCBIWWW.qblast("blastx", "nr", subseq,
                        expect=0.00001, filter=None)
                    save_file = open(igs_id + ".xml", "w")
                    save_file.write(result_handle.read())
                    save_file.close()
                    result_handle.close()
                    print "Done BLASTx"
    i += 1

af.close()
sf.close()
```

D.10 dissertation_TetraNTCalculatorImproved.py

```
#!/usr/bin/python
"""
Author:                         Jimmy Saw
Date modified:          06-19-2012

Description:    This program calculates tetranucleotide frequencies from a given
                multi-fasta file and reports the z score.

Usage:          dissertation_TetraNTCalculatorImproved.py <multi-fasta file>
                <tetra list>

Example:        dissertation_TetraNTCalculatorImproved.py test-multi.fasta
                tetra.list

Note:

z = (x - mu) / rho

tetra.list file should contain a list of tetranucleotide combinations like this:
AAAA
AAAC
AAAG
AAAT
AACA
```

```
AACC
AACG
AACT
AAGA
AAGC
.
.
.
TTCG
TTCT
TTGA
TTGC
TTGG
TTGT
TTTA
TTTC
TTTG
TTTT
And should contain a total of 256 combinations.

Note: currently it prints z score in a tab-delimited format like this:
EM7JFSU01D21YQ        -0.471361238918        -0.471361238918
  -0.471361238918        3.91658338519
The idea is to use this script to bin metagenomic reads by tetra-nt freq among
other components such as G+C% and other things. I attempt to use z score because
it is a normalized score instead of a raw score which can change based on length
of the sequence.

"""

import re
import sys
import numpy
from Bio import SeqIO

def zscore(x, u, r):
    z = (x - u) / float(r)
    return z

#genome = SeqIO.parse(sys.argv[1], "fasta") #in generator
genome = SeqIO.index(sys.argv[1], "fasta") #in dictionary

tetrafile = open(sys.argv[2], "rU")
tetras = tetrafile.readlines()

tetradict = {}

#for so in genome: #to use with generator
for i, so in genome.iteritems(): #to use with dictionary iteritems
    for t in tetras:
        x = t.strip()
        #tetradict[x] = so.seq.count(x)
        #this statement below counts overlapping tetranucleotides
        #such as GGGGG
        #see this: http://stackoverflow.com/questions/6844005/
        #how-can-i-find-the-number-of-overlapping-sequences-in-a-string-with-python
        tetradict[x] = len(re.findall(r'(?=(%s))' % re.escape(x), so.seq.tostring()))
        #print "Done with", x, tetradict[x]
    tetralist = tetradict.items()
    tetralist.sort()
    total = 0
    numbers = []

    for i in tetralist:
```

```
        numbers.append(i[1])
        total = numpy.sum(numbers)
        average = numpy.average(numbers)
        stdev = numpy.std(numbers)
        toprint = so.id

    for j in tetralist:
        z = zscore(j[1], average, stdev)
        toprint += "\t" + '{0:.8}'.format(str(z))
    print toprint

tetrafile.close()
```

D.11 dissertation_KeggModule.rb

```
#!/usr/bin/ruby
# Author: Jimmy Saw
# Last updated: 12-29-2010
# Usage: dissertation_KeggModule.rb <KEGG module name>
# Example: dissertation_KeggModule.rb M00001

require 'bio'
require 'soap/wsdlDriver'

wsdl = "http://soap.genome.jp/KEGG.wsdl"
serv = SOAP::WSDLDriverFactory.new(wsdl).create_rpc_driver
serv.generate_explicit_type = true

entry = ARGV[0]

#Get module entry from KEGG
et = serv.bget("md:#{entry}")

#Set up MODULE object
o = Bio::KEGG::MODULE.new(et)

#Read ORTHOLOGY info from MODULE entry
orthology = o.orthologs_as_array

#Print MODULE entry and name
print "Module info:\n"
print "MD:", "\t", o.entry_id, "\t", o.name, "\n"

print "\n"

#Set up ORTHOLOGs from MODULE as hash
orth = o.orthologs_as_hash
key = orth.keys
val = orth.values

#Print a list of KOs
print "KO info:\n"
for i in 0..key.length
  if key[i] =~ /^K0/
    print "KO:", "\t", key[i], "\t", val[i], "\n"
  end
end

print "Total KOs found: #{key.length}\n"

```

```
print "\n"

#Print COGs found for each KO
print "COG info:\n"

cogcount = 0

orthology.each do |oo|
  f = serv.bget("orthology:#{oo}")
  ot = Bio::KEGG::ORTHOLOGY.new(f)
  oid = ot.entry_id
  odf = ot.definition

  cs = ot.dblinks_as_strings
  for s in cs
    if s =~ /COG/
      cogcount += 1
      slice = s.split(" ")
      if slice.length == 2
        print "COG:", "\t", oid, "\t", slice[1], "\t", odf, "\n"
      else
        for j in 1..slice.length
          if slice[j] != nil
            print "COG:", "\t", oid, "\t", slice[j], "\t", odf, "\n"
          end
        end
      end
    end
  end
end

print "Total COGs found: #{cogcount}\n"
```

D.12 dissertation_BibTeX.rb

```
#!/usr/bin/ruby
# This program fetches bibtex file for a given PMID
# Usage example: ruby get.bibtex.rb 20176788 >  20176788.bib

require 'bio'

pmid = ARGV[0]

Bio::NCBI.default_email = "jimmy@hawaii.edu"

entries = Bio::PubMed.esearch(pmid)

Bio::PubMed.efetch(entries).each do |entry|
  medline = Bio::MEDLINE.new(entry)
  reference = medline.reference
  puts reference.bibtex
end
```

D.13 dissertation_RibosomalGenesIndividual.sh

```
#!/bin/sh
#Usage: directories need to be named with phylum names, such as cyano or chlorobi
#Note: This script runs alignment on each gene, then concatenates them
#      (instead of concat first, then aligning)
#Eg: dissertation_SingleCopyGenes.sh genes.list cyano
#    dissertation_RibosomalGenes.sh r53.list cyano

phylum=$1
genes=$2
count=`wc -l $genes | awk '{print $1}'`

#New. Run alignment for each gene instead of concatenating them and aligning.
#BLAST and extract sequences

cd /host/Users/JS/UH-work/gloeobacter/final_work/comparisons/ribosomal_genes/

for i in `$genes`;do
    cp $i.fasta /host/Users/JS/UH-work/gloeobacter/final_work/comparisons/orthologs/
      $phylum/$i.GKIL.fasta
done

cd /host/Users/JS/UH-work/gloeobacter/final_work/comparisons/orthologs/$phylum

for i in `cat $phylum.list`;do
    for j in `cat $genes`;do
        echo "Working on $i and $j"
        blastall -p blastp -i /host/Users/JS/UH-work/gloeobacter/final_work/comparisons/
          ribosomal_genes/$j.fasta -d $i.refseq.faa -F F -e 0.01 -m 8 -o $j.$i.blastp
        head -1 $j.$i.blastp | awk '{print $2}' > tmp.out
        exfasta `cat tmp.out` $i.refseq.faa > $j.$i.fasta
        sed -i 's/>.*/>'$i'/g' $j.$i.fasta
    done
done

for i in `cat $genes`;do
    echo "Doing alignments and Gblocks $i"
    #Get each gene ready for alignment
    cat $i*.fasta > $i.align.fasta
    #Run alignment using Muscle
    muscle -in $i.align.fasta -out $i.align.muscle
    #Trim sequences using Gblocks
    Gblocks $i.align.muscle -t=p -e=-gb -b4=2 #output $i.all.muscle-gb
    cp $i.align.muscle-gb /host/Users/JS/UH-work/gloeobacter/final_work/comparisons/
      ribosomal_genes/
    rm *.htm
    echo "Done!"
done

#Convert/concatenate alignments
dissertation_ConcatConvertMSA.py $genes $phylum.list
cat *.concat.faa > ribo43.each.fasta
dissertation_ConvertAlignment.py ribo43.each.fasta fasta ribo43.each.phy phylip

#Now run RAxML on a Cluster with the following command:
#raxmlHPC-PTHREADS-SSE3 -T 20 -f a -m PROTGAMMAWAG -x 12345 -# 100 -p 11386 -s
#  rc43.muscle-gb.phy -n Ribo43
```

Bibliography

[1] Chivian D, Brodie EL, Alm EJ, Culley DE, Dehal PS, et al. (2008) Environmental genomics reveals a single-species ecosystem deep within Earth. Science 322: 275–278.

[2] Borgonie G, Garcia-Moyano A, Litthauer D, Bert W, Bester A, et al. (2011) Nematoda from the terrestrial deep subsurface of South Africa. Nature 474: 79–82.

[3] Stivaletta N, Barbieri R, Billi D (2012) Microbial colonization of the salt deposits in the driest place of the Atacama Desert (Chile). Orig Life Evol Biosph 42: 187–200.

[4] Neilson JW, Quade J, Ortiz M, Nelson WM, Legatzki A, et al. (2012) Life at the hyperarid margin: novel bacterial diversity in arid soils of the Atacama Desert, Chile. Extremophiles 16: 553–566.

[5] Parro V, de Diego-Castilla G, Moreno-Paz M, Blanco Y, Cruz-Gil P, et al. (2011) A microbial oasis in the hypersaline Atacama subsurface discovered by a life detector chip: implications for the search for life on Mars. Astrobiology 11: 969–996.

[6] Lacap DC, Warren-Rhodes KA, McKay CP, Pointing SB (2011) Cyanobacteria and chloroflexi-dominated hypolithic colonization of quartz at the hyper-arid core of the Atacama Desert, Chile. Extremophiles 15: 31–38.

[7] Johnston DT, Wolfe-Simon F, Pearson A, Knoll AH (2009) Anoxygenic photosynthesis modulated Proterozoic oxygen and sustained Earth's middle age. Proc Natl Acad Sci U S A 106: 16925–16929.

[8] Pomeroy L, Wiebe W (1988) Energetics of microbial food webs. Hydrobiologia 159: 7–18.

[9] Pomeroy L, Wiebe W (1993) Energy sources for microbial food webs. Mar Microb Food Webs 7: 101–118.

[10] Pomeroy L, Wiebe W (1993) Seasonal uncoupling of the microbial loop and its potential significance for the global carbon cycle. Trends in Microbial Ecology : 407–9.

[11] Azam F, Worden AZ (2004) Oceanography. Microbes, molecules, and marine ecosystems. Science 303: 1622–1624.

[12] Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276: 734–740.

[13] Head IM, Saunders JR, Pickup RW (1998) Microbial Evolution, Diversity, and Ecology: A Decade of Ribosomal RNA Analysis of Uncultivated Microorganisms. Microb Ecol 35: 1–21.

[14] Rappe MS, Connon SA, Vergin KL, Giovannoni SJ (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 418: 630–633.

[15] Britschgi TB, Giovannoni SJ (1991) Phylogenetic analysis of a natural marine bacterioplankton population by rRNA gene cloning and sequencing. Appl Environ Microbiol 57: 1707–1713.

[16] Donachie SP, Hou S, Lee KS, Riley CW, Pikina A, et al. (2004) The Hawaiian Archipelago: a microbial diversity hotspot. Microb Ecol 48: 509–520.

[17] Donachie SP, Foster JS, Brown MV (2007) Culture clash: challenging the dogma of microbial diversity. ISME J 1: 97–99.

[18] Curtis TP, Sloan WT, Scannell JW (2002) Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci U S A 99: 10494–10499.

[19] King GM (2003) Contributions of atmospheric CO and hydrogen uptake to microbial dynamics on recent Hawaiian volcanic deposits. Appl Environ Microbiol 69: 4067–4075.

[20] Nanba K, King GM, Dunfield K (2004) Analysis of facultative lithotroph distribution and diversity on volcanic deposits by use of the large subunit of ribulose 1,5-bisphosphate carboxylase/oxygenase. Appl Environ Microbiol 70: 2245–2253.

[21] Dunfield KE, King GM (2004) Molecular analysis of carbon monoxide-oxidizing bacteria associated with recent Hawaiian volcanic deposits. Appl Environ Microbiol 70: 4242–4248.

[22] Dunfield KE, King GM (2005) Analysis of the distribution and diversity in recent Hawaiian volcanic deposits of a putative carbon monoxide dehydrogenase large subunit gene. Environ Microbiol 7: 1405–1412.

[23] Gomez-Alvarez V, King GM, Nusslein K (2007) Comparative bacterial diversity in recent Hawaiian volcanic deposits of different ages. FEMS Microbiol Ecol 60: 60–73.

[24] King GM, Weber CF (2008) Interactions between bacterial carbon monoxide and hydrogen consumption and plant development on recent volcanic deposits. ISME J 2: 195–203.

[25] Weber CF, King GM (2009) Water stress impacts on bacterial carbon monoxide oxidation on recent volcanic deposits. ISME J 3: 1325–1334.

[26] King GM, Weber CF, Nanba K, Sato Y, Ohta H (2008) Atmospheric CO and hydrogen uptake and CO oxidizer phylogeny for Miyake-jima, Japan volcanic deposits. Microbes Environ 23: 299–305.

[27] King CE, King GM (2012) Temperature responses of carbon monoxide and hydrogen uptake by vegetated and unvegetated volcanic cinders. ISME J 6: 1558–1565.

[28] Northup DE, Barns SM, Yu LE, Spilde MN, Schelble RT, et al. (2003) Diverse microbial communities inhabiting ferromanganese deposits in Lechuguilla and Spider Caves. Environ Microbiol 5: 1071–1086.

[29] Northup DE, Melim LA, Spilde MN, Hathaway JJ, Garcia MG, et al. (2011) Lava cave microbial communities within mats and secondary mineral deposits: implications for life detection on other planets. Astrobiology 11: 601–618.

[30] Asencio A, Aboal M (2010) In situ nitrogen fixation by cyanobacteria at the Andragulla Cave, Spain. Journal of Cave and Karst Studies 73: 50–54.

[31] Albertano P (2012) Ecology of Cyanobacteria II, Springer, chapter Cyanobacterial Biofilms in Monuments and Caves. pp. 317–343.

[32] Cushing G, Titus T, Wynne J, Christensen P (2007) THEMIS observes possible cave skylights on Mars. Geophysical Research Letters 34: L17201.

[33] Acuna M, Connerney J, Wasilewski P, Lin R, Anderson K, et al. (1998) Magnetic field and plasma observations at mars: Initial results of the mars global surveyor mission. Science 279: 1676–1680.

[34] Martinez A, Asencio A (2010) Distribution of cyanobacteria at the Gelada Cave(Spain) by physical parameters. Journal of Cave and Karst Studies 72: 11–20.

[35] Jones DS, Albrecht HL, Dawson KS, Schaperdoth I, Freeman KH, et al. (2012) Community genomic analysis of an extremely acidophilic sulfur-oxidizing biofilm. ISME J 6: 158–170.

[36] Ley RE, Harris JK, Wilcox J, Spear JR, Miller SR, et al. (2006) Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat. Appl Environ Microbiol 72: 3685–3695.

[37] Kunin V, Raes J, Harris JK, Spear JR, Walker JJ, et al. (2008) Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. Mol Syst Biol 4: 198.

[38] Bhaya D, Grossman AR, Steunou AS, Khuri N, Cohan FM, et al. (2007) Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. ISME J 1: 703–713.

[39] Klatt CG, Wood JM, Rusch DB, Bateson MM, Hamamura N, et al. (2011) Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential. ISME J 5: 1262–1278.

[40] Liu Z, Klatt CG, Wood JM, Rusch DB, Ludwig M, et al. (2011) Metatranscriptomic analyses of chlorophototrophs of a hot-spring microbial mat. ISME J 5: 1279–1290.

[41] Grotzinger JP, Knoll AH (1999) Stromatolites in Precambrian carbonates: evolutionary mileposts or environmental dipsticks? Annu Rev Earth Planet Sci 27: 313–358.

[42] Rasmussen B, Fletcher IR, Brocks JJ, Kilburn MR (2008) Reassessing the first appearance of eukaryotes and cyanobacteria. Nature 455: 1101–1104.

[43] Couradeau E, Benzerara K, Gerard E, Moreira D, Bernard S, et al. (2012) An early-branching microbialite cyanobacterium forms intracellular carbonates. Science 336: 459–462.

[44] Brown II, Bryant DA, Casamatta D, Thomas-Keprta KL, Sarkisova SA, et al. (2010) Polyphasic characterization of a thermotolerant siderophilic filamentous cyanobacterium that produces intracellular iron deposits. Appl Environ Microbiol 76: 6664–6672.

[45] Riding R (2006) Cyanobacterial calcification, carbon dioxide concentrating mechanisms, and Proterozoic–Cambrian changes in atmospheric composition. Geobiology 4: 299–316.

[46] Stal L (2012) Ecology of Cyanobacteria II, Springer, chapter Cyanobacterial Mats and Stromatolites. pp. 65–125.

[47] Shi T, Falkowski PG (2008) Genome evolution in cyanobacteria: the stable core and the variable shell. Proc Natl Acad Sci U S A 105: 2510–2515.

[48] Gupta RS (2009) Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades. Int J Syst Evol Microbiol 59: 2510–2526.

[49] Schirrmeister BE, Antonelli A, Bagheri HC (2011) The origin of multicellularity in cyanobacteria. BMC Evol Biol 11: 45.

[50] Ashby MK, Houmard J (2006) Cyanobacterial two-component proteins: structure, diversity, distribution, and evolution. Microbiol Mol Biol Rev 70: 472–509.

[51] Couturier J, Jacquot JP, Rouhier N (2009) Evolution and diversity of glutaredoxins in photosynthetic organisms. Cell Mol Life Sci 66: 2539–2557.

[52] Allen AE, Dupont CL, Obornik M, Horak A, Nunes-Nesi A, et al. (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. Nature 473: 203–207.

[53] Dekas AE, Poretsky RS, Orphan VJ (2009) Deep-sea archaea fix and share nitrogen in methane-consuming microbial consortia. Science 326: 422–426.

[54] Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37–43.

[55] Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, et al. (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science 335: 587–590.

[56] Kang Y, Norris MH, Zarzycki-Siek J, Nierman WC, Donachie SP, et al. (2011) Transcript amplification from single bacterium for transcriptome analysis. Genome Res 21: 925–935.

[57] Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, et al. (2010) Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. Nature 464: 90–94.

[58] Woyke T, Xie G, Copeland A, Gonzalez JM, Han C, et al. (2009) Assembling the marine metagenome, one cell at a time. PLoS One 4: e5299.

[59] Hongoh Y, Sharma VK, Prakash T, Noda S, Toh H, et al. (2008) Genome of an endosymbiont coupling N_2 fixation to cellulolysis within protist cells in termite gut. Science 322: 1108–1109.

[60] Hongoh Y, Sharma VK, Prakash T, Noda S, Taylor TD, et al. (2008) Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. Proc Natl Acad Sci U S A 105: 5555–5560.

[61] Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, et al. (2009) Whole genome amplification and *de novo* assembly of single bacterial cells. PLoS One 4: e6864.

[62] Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, et al. (2009) A phylogeny-driven genomic encyclopaedia of *Bacteria* and *Archaea*. Nature 462: 1056–1060.

[63] Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5: e16.

[64] Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest atlantic through eastern tropical pacific. PLoS Biol 5: e77.

[65] Costerton JW, Cheng KJ, Geesey GG, Ladd TI, Nickel JC, et al. (1987) Bacterial biofilms in nature and disease. Annu Rev Microbiol 41: 435–464.

[66] Hall-Stoodley L, Costerton JW, Stoodley P (2004) Bacterial biofilms: from the natural environment to infectious diseases. Nat Rev Microbiol 2: 95–108.

[67] Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75: 7537–7541.

[68] Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. ISME J 3: 1314–1317.

[69] Brady A, Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated markov models. Nat Methods 6: 673–676.

[70] Brady A, Salzberg S (2011) PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat Methods 8: 367.

[71] Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P (2011) iPath2.0: interactive pathway explorer. Nucleic Acids Res 39: W412–W415.

[72] Zerbino DR, Birney E (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res 18: 821–829.

[73] Huang X, Wang J, Aluru S, Yang SP, Hillier L (2003) PCAP: a whole-genome assembly program. Genome Res 13: 2164–2170.

[74] Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33: 5691–5702.

[75] Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.

[76] Vos M, Quince C, Pijl AS, de Hollander M, Kowalchuk GA (2012) A comparison of *rpoB* and 16S rRNA as markers in pyrosequencing studies of bacterial diversity. PLoS One 7: e30600.

[77] Gilbert JA, Meyer F, Knight R, Field D, Kyrpides N, et al. (2010) Meeting report: GSC M5 roundtable at the 13th International Society for Microbial Ecology meeting in Seattle, WA, USA August 22-27, 2010. Stand Genomic Sci 3: 235–239.

[78] Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. Genome Biol 8: R10.

[79] Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N, Chen F, et al. (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. ISME J 4: 642–647.

[80] Gotelli N, Colwell R (2010) Biological diversity: frontiers in measurement and assessment., Oxford University Press, chapter Estimating species richness. pp. 39–54.

[81] Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a web browser. BMC Bioinformatics 12: 385.

[82] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

[83] Wu M, Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. Genome Biol 9: R151.

[84] Turroni F, Ribbera A, Foroni E, van Sinderen D, Ventura M (2008) Human gut microbiota and bifidobacteria: from composition to functionality. Antonie Van Leeuwenhoek 94: 35–50.

[85] Lewin RA (1997) *Saprospira grandis*: A flexibacterium that can catch bacterial prey by "ixotrophy". Microb Ecol 34: 232–236.

[86] Bauer M, Kube M, Teeling H, Richter M, Lombardot T, et al. (2006) Whole genome analysis of the marine Bacteroidetes 'Gramella forsetii' reveals adaptations to degradation of polymeric organic matter. Environ Microbiol 8: 2201–2213.

[87] Cottrell MT, Kirchman DL (2000) Natural assemblages of marine proteobacteria and members of the Cytophaga-Flavobacter cluster consuming low- and high-molecular-weight dissolved organic matter. Appl Environ Microbiol 66: 1692–1697.

[88] Henne A, Bruggemann H, Raasch C, Wiezer A, Hartsch T, et al. (2004) The genome sequence of the extreme thermophile *Thermus thermophilus*. Nat Biotechnol 22: 547–553.

[89] White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, et al. (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 286: 1571–1577.

[90] Makarova KS, Omelchenko MV, Gaidamakova EK, Matrosova VY, Vasilenko A, et al. (2007) *Deinococcus geothermalis*: the pool of extreme radiation resistance genes shrinks. PLoS One 2: e955.

[91] Wagner M, Horn M (2006) The *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr Opin Biotechnol 17: 241–249.

[92] Belland R, Ojcius DM, Byrne GI (2004) Chlamydia. Nat Rev Microbiol 2: 530–531.

[93] Dunfield PF, Yuryev A, Senin P, Smirnova AV, Stott MB, et al. (2007) Methane oxidation by an extremely acidophilic bacterium of the phylum *Verrucomicrobia*. Nature 450: 879–882.

[94] Pol A, Heijmans K, Harhangi HR, Tedesco D, Jetten MS, et al. (2007) Methanotrophy below pH 1 by a new *Verrucomicrobia* species. Nature 450: 874–878.

[95] Islam T, Jensen S, Reigstad LJ, Larsen O, Birkeland NK (2008) Methane oxidation at 55 degrees c and pH 2 by a thermoacidophilic bacterium belonging to the *Verrucomicrobia* phylum. Proc Natl Acad Sci U S A 105: 300–304.

[96] Fukunaga Y, Kurahashi M, Sakiyama Y, Ohuchi M, Yokota A, et al. (2009) *Phycisphaera mikurensis* gen. nov., sp. nov., isolated from a marine alga, and proposal of Phycisphaeraceae fam. nov., Phycisphaerales ord. nov. and Phycisphaerae classis nov. in the phylum *Planctomycetes*. J Gen Appl Microbiol 55: 267–275.

[97] Labutti K, Sikorski J, Schneider S, Nolan M, Lucas S, et al. (2010) Complete genome sequence of *Planctomyces limnophilus* type strain (Mu 290). Stand Genomic Sci 3: 47–56.

[98] Hirsch P, Müller M (1985) *Planctomyces limnophilus* sp. nov., a stalked and budding bacterium from freshwater. Systematic and applied microbiology 6: 276–280.

[99] van Passel MW, Kant R, Palva A, Copeland A, Lucas S, et al. (2011) Genome sequence of the verrucomicrobium *Opitutus terrae* PB90-1, an abundant inhabitant of rice paddy soil ecosystems. J Bacteriol 193: 2367–2368.

[100] Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, et al. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437: 543–546.

[101] Walker CB, de la Torre JR, Klotz MG, Urakawa H, Pinel N, et al. (2010) *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. Proc Natl Acad Sci U S A 107: 8818–8823.

[102] King GM, Weber CF (2007) Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. Nat Rev Microbiol 5: 107–118.

[103] Sun S, Chen J, Li W, Altintas I, Lin A, et al. (2011) Community cyberinfrastructure for advanced microbial ecology research and analysis: the camera resource. Nucleic Acids Res 39: D546–D551.

[104] Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, et al. (2008) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36: D534–D538.

[105] Barabote RD, Xie G, Leu DH, Normand P, Necsulea A, et al. (2009) Complete genome of the cellulolytic thermophile *Acidothermus cellulolyticus* 11b provides insights into its ecophysiological and evolutionary adaptations. Genome Res 19: 1033–1043.

[106] Stott MB, Crowe MA, Mountain BW, Smirnova AV, Hou S, et al. (2008) Isolation of novel bacteria, including a candidate division, from geothermal soils in New Zealand. Environ Microbiol 10: 2030–2041.

[107] Rippka R, Waterbury J, Cohen-Bazire G (1974) A cyanobacterium which lacks thylakoids. Archives of Microbiology 100: 419-436.

[108] Nakamura Y, Kaneko T, Sato S, Mimuro M, Miyashita H, et al. (2003) Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. DNA Res 10: 137–145.

[109] Tyson GW, Lo I, Baker BJ, Allen EE, Hugenholtz P, et al. (2005) Genome-directed isolation of the key nitrogen fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an acidophilic microbial community. Appl Environ Microbiol 71: 6319–6324.

[110] Teske A, Sigalevich P, Cohen Y, Muyzer G (1996) Molecular identification of bacteria from a coculture by denaturing gradient gel electrophoresis of 16S ribosomal DNA fragments as a tool for isolation in pure cultures. Appl Environ Microbiol 62: 4210–4215.

[111] Atlas R (2004) Handbook of microbiological media. CRC.

[112] Wright SW, Jeffrey SW, Mantoura RFC, Llewellyn CA, Bjornland T, et al. (1991) Improved HPLC method for the analysis of chlorophylls and carotenoids from marine phytoplankton. Marine Ecology 77: 183–196.

[113] Bidigare R, Van Heukelem L, Trees C (2005) Analysis of algal pigments by high-performance liquid chromatography. Algal Culturing Techniques Academic Press, New York : 327–345.

[114] Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.

[115] Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17: 540–552.

[116] Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456–463.

[117] Allen MM, Smith AJ (1969) Nitrogen chlorosis in blue-green algae. Arch Mikrobiol 69: 114–120.

[118] Lau RH, MacKenzie MM, Doolittle WF (1977) Phycocyanin synthesis and degradation in the blue-green bacterium *Anacystis nidulans*. J Bacteriol 132: 771–778.

[119] Gugger MF, Hoffmann L (2004) Polyphyly of true branching cyanobacteria (Stigonematales). Int J Syst Evol Microbiol 54: 349–357.

[120] Lamprinou V, Hernandez-Marine M, Canals T, Kormas K, Economou-Amilli A, et al. (2011) Morphology and molecular evaluation of *Iphinoe spelaeobios* gen. nov., sp. nov. and *Loriellopsis cavernicola* gen. nov., sp. nov., two stigonematalean cyanobacteria from Greek and Spanish caves. Int J Syst Evol Microbiol 61: 2907–2915.

[121] Hernandez-Muniz W, Stevens SEJ (1987) Characterization of the motile hormogonia of *Mastigocladus laminosus*. J Bacteriol 169: 218–223.

[122] Thomazeau S, Houdan-Fourmont A, Couté A, Duval C, Couloux A, et al. (2010) The contribution of Sub-Saharan African strains to the phylogeny of *Cyanobacteria*: focusing on the Nostocaceae (Nostocales, Cyanobacteria). Journal of Phycology 46: 564–579.

[123] Casamatta D, Johansen J, Vis M, Broadwater S (2005) molecular and morphological characterization of ten polar and near-polar strains within the *Oscillatoriales* (*Cyanobacteria*) 1. Journal of Phycology 41: 421–438.

[124] Taton A, Grubisic S, Ertz D, Hodgson D, Piccardi R, et al. (2006) Polyphasic study of Antarctic cyanobacterial strains. Journal of phycology 42: 1257–1270.

[125] Roeselers G, Norris TB, Castenholz RW, Rysgaard S, Glud RN, et al. (2007) Diversity of phototrophic bacteria in microbial mats from Arctic hot springs (Greenland). Environ Microbiol 9: 26–38.

[126] Knowles EJ, Castenholz RW (2008) Effect of exogenous extracellular polysaccharides on the desiccation and freezing tolerance of rock-inhabiting phototrophic microorganisms. FEMS Microbiol Ecol 66: 261–270.

[127] Pereira S, Zille A, Micheletti E, Moradas-Ferreira P, De Philippis R, et al. (2009) Complexity of cyanobacterial exopolysaccharides: composition, structures, inducing factors and putative genes involved in their biosynthesis and assembly. FEMS Microbiol Rev 33: 917–941.

[128] Taton A, Lis E, Adin DM, Dong G, Cookson S, et al. (2012) Gene transfer in *Leptolyngbya* sp. strain BL0902, a cyanobacterium suitable for production of biomass and bioproducts. PLoS One 7: e30901.

[129] Holland HD (2006) The oxygenation of the atmosphere and oceans. Philos Trans R Soc Lond B Biol Sci 361: 903–915.

[130] Frei R, Gaucher C, Poulton SW, Canfield DE (2009) Fluctuations in Precambrian atmospheric oxygenation recorded by chromium isotopes. Nature 461: 250–253.

[131] Nelissen B, Van de Peer Y, Wilmotte A, De Wachter R (1995) An early origin of plastids within the cyanobacterial divergence is suggested by evolutionary trees based on complete 16S rRNA sequences. Mol Biol Evol 12: 1166–1173.

[132] Turner S, Pryer KM, Miao VP, Palmer JD (1999) Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. J Eukaryot Microbiol 46: 327–338.

[133] Falcon LI, Magallon S, Castillo A (2010) Dating the cyanobacterial ancestor of the chloroplast. ISME J 4: 777–783.

[134] Mimuro M, Tomo T, Tsuchiya T (2008) Two unique cyanobacteria lead to a traceable approach of the first appearance of oxygenic photosynthesis. Photosynth Res 97: 167–176.

[135] Rexroth S, Mullineaux CW, Ellinger D, Sendtko E, Rogner M, et al. (2011) The plasma membrane of the cyanobacterium *Gloeobacter violaceus* contains segregated bioenergetic domains. Plant Cell 23: 2379–2390.

[136] Gupta RS (2012) Origin and spread of photosynthesis based upon conserved sequence features in key bacteriochlorophyll biosynthesis proteins. Mol Biol Evol *Epub ahead of print*.

[137] Nitschke W, Rutherford AW (1991) Photosynthetic reaction centres: variations on a common structural theme? Trends Biochem Sci 16: 241–245.

[138] Golbeck JH (1993) Shared thematic elements in photochemical reaction centers. Proc Natl Acad Sci U S A 90: 1642–1646.

[139] Blankenship RE (1994) Protein structure, electron transfer and evolution of prokaryotic photosynthetic reaction centers. Antonie Van Leeuwenhoek 65: 311–329.

[140] Olson JM, Blankenship RE (2004) Thinking about the evolution of photosynthesis. Photosynth Res 80: 373–386.

[141] Sadekar S, Raymond J, Blankenship RE (2006) Conservation of distantly related membrane proteins: photosynthetic reaction centers share a common structural core. Mol Biol Evol 23: 2001–2007.

[142] Blankenship RE (2010) Early evolution of photosynthesis. Plant Physiol 154: 434–438.

[143] Margulis L (1991) Symbiosis as a source of evolutionary innovation: speciation and morphogenesis, The MIT Press, chapter Symbiogenesis and Symbionticism. pp. 1–14.

[144] Mulkidjanian AY (2009) On the origin of life in the zinc world: 1. photosynthesizing, porous edifices built of hydrothermally precipitated zinc sulfide as cradles of life on earth. Biol Direct 4: 26.

[145] Mulkidjanian AY, Galperin MY (2009) On the origin of life in the zinc world. 2. validation of the hypothesis on the photosynthesizing zinc sulfide edifices as cradles of life on earth. Biol Direct 4: 27.

[146] De Marais DJ (2000) Evolution. when did photosynthesis emerge on earth? Science 289: 1703–1705.

[147] Xiong J, Fischer WM, Inoue K, Nakahara M, Bauer CE (2000) Molecular evidence for the early evolution of photosynthesis. Science 289: 1724–1730.

[148] Cuzman OA, Ventura S, Sili C, Mascalchi C, Turchetti T, et al. (2010) Biodiversity of phototrophic biofilms dwelling on monumental fountains. Microb Ecol 60: 81–95.

[149] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biol 5: R12.

[150] Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365–386.

[151] Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11: 119.

[152] Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.

[153] Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T (2010) GLOOME: gain loss mapping engine. Bioinformatics 26: 2914–2915.

[154] Auch AF, von Jan M, Klenk HP, Goker M (2010) Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci 2: 117–134.

[155] Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, et al. (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57: 81–91.

[156] Richter M, Rosselló-Móra R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 106: 19126–19131.

[157] Li L, Stoeckert CJJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189.

[158] Chen F, Mackey AJ, Stoeckert CJJ, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34: D363–D368.

[159] Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, et al. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. Science 311: 1768–1770.

[160] Dufresne A, Ostrowski M, Scanlan DJ, Garczarek L, Mazard S, et al. (2008) Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. Genome Biol 9: R90.

[161] Deveau H, Garneau JE, Moineau S (2010) CRISPR/Cas system and its role in phage-bacteria interactions. Annu Rev Microbiol 64: 475–493.

[162] Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, et al. (2011) Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 9: 467–477.

[163] Heidelberg JF, Nelson WC, Schoenfeld T, Bhaya D (2009) Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. PLoS One 4: e4169.

[164] Karp PD, Paley S, Romero P (2002) The Pathway Tools software. Bioinformatics 18 Suppl 1: S225–S232.

[165] Tsuchiya T, Takaichi S, Misawa N, Maoka T, Miyashita H, et al. (2005) The cyanobacterium *Gloeobacter violaceus* PCC 7421 uses bacterial-type phytoene desaturase in carotenoid biosynthesis. FEBS Lett 579: 2125–2129.

[166] Steiger S, Jackisch Y, Sandmann G (2005) Carotenoid biosynthesis in *Gloeobacter violaceus* PCC4721 involves a single *crtI*-type phytoene desaturase instead of typical cyanobacterial enzymes. Arch Microbiol 184: 207–214.

[167] Liang C, Zhao F, Wei W, Wen Z, Qin S (2006) Carotenoid biosynthesis in cyanobacteria: structural and evolutionary scenarios based on comparative genomics. Int J Biol Sci 2: 197–207.

[168] Takaichi S, Mochimaru M (2007) Carotenoids and carotenogenesis in cyanobacteria: unique ketocarotenoids and carotenoid glycosides. Cell Mol Life Sci 64: 2607–2619.

[169] Imasheva ES, Balashov SP, Choi AR, Jung KH, Lanyi JK (2009) Reconstitution of gloeobacter violaceus rhodopsin with a light-harvesting carotenoid antenna. Biochemistry 48: 10948–10955.

[170] Mathies G, van Hemert MC, Gast P, Gupta KB, Frank HA, et al. (2011) Configuration of spheroidene in the photosynthetic reaction center of *Rhodobacter sphaeroides* : a comparison of wild-type and reconstituted R26. J Phys Chem A 115: 9552–9556.

[171] Allen JP, Williams JC (2011) The evolutionary pathway from anoxygenic to oxygenic photosynthesis examined by comparison of the properties of photosystem II and bacterial reaction centers. Photosynth Res 107: 59–69.

[172] Holman TR, Wu Z, Wanner BL, Walsh CT (1994) Identification of the DNA-binding site for the phosphorylated vanr protein required for vancomycin resistance in *Enterococcus faecium*. Biochemistry 33: 4625–4631.

[173] Ulijasz AT, Kay BK, Weisblum B (2000) Peptide analogues of the VanS catalytic center inhibit VanR binding to its cognate promoter. Biochemistry 39: 11417–11424.

[174] Courvalin P (2006) Vancomycin resistance in Gram-positive cocci. Clin Infect Dis 42 Suppl 1: S25–S34.

[175] Rosselló-Móra R (2006) DNA-DNA reassociation methods applied to microbial taxonomy and their critical evaluation. Molecular Identification, Systematics and Population Structure of Prokaryotes : 23–50.

[176] Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 102: 2567–2572.

[177] Borriss R, Chen XH, Rueckert C, Blom J, Becker A, et al. (2011) Relationship of *Bacillus amyloliquefaciens* clades associated with strains DSM 7T and FZB42T: a proposal for *Bacillus amyloliquefaciens* subsp. *amyloliquefaciens* subsp. nov. and *Bacillus amyloliquefaciens* subsp. *plantarum* subsp. nov. based on complete genome sequence comparisons. Int J Syst Evol Microbiol 61: 1786–1801.

[178] Stackebrandt E, Goebel B (1994) Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. International Journal of Systematic Bacteriology 44: 846–849.

[179] Guler S, Seeliger A, Hartel H, Renger G, Benning C (1996) A null mutant of *Synechococcus* sp. PCC7942 deficient in the sulfolipid sulfoquinovosyl diacylglycerol. J Biol Chem 271: 7501–7507.

[180] Guler S, Essigmann B, Benning C (2000) A cyanobacterial gene, *sqdX*, required for biosynthesis of the sulfolipid sulfoquinovosyldiacylglycerol. J Bacteriol 182: 543–545.

[181] Westphal S, Heins L, Soll J, Vothknecht UC (2001) Vipp1 deletion mutant of *Synechocystis*: a connection between bacterial phage shock and thylakoid biogenesis? Proc Natl Acad Sci U S A 98: 4243–4248.

[182] Kroll D, Meierhoff K, Bechtold N, Kinoshita M, Westphal S, et al. (2001) VIPP1, a nuclear gene of *Arabidopsis thaliana* essential for thylakoid membrane formation. Proc Natl Acad Sci U S A 98: 4238–4242.

[183] Mulo P, Sicora C, Aro EM (2009) Cyanobacterial *psbA* gene family: optimization of oxygenic photosynthesis. Cell Mol Life Sci 66: 3697–3710.

[184] Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, et al. (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol 4: e234.

[185] Andam CP, Gogarten JP (2011) Biased gene transfer in microbial evolution. Nat Rev Microbiol 9: 543–555.

[186] Syvanen M (2012) Evolutionary Implications of Horizontal Gene Transfer. Annu Rev Genet 46: *Epub ahead of print*.

[187] Zhaxybayeva O, Doolittle WF, Papke RT, Gogarten JP (2009) Intertwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus marinus*. Genome Biol Evol 1: 325–339.

[188] Swingley W, Blankenship R, Raymond J (2008) The cyanobacteria: molecular biology, genomics, and evolution, Caister Academic Pr, chapter Insights into cyanobacterial evolution from comparative genomics. pp. 21–43.

[189] Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059–3066.

[190] Britton T, Anderson CL, Jacquet D, Lundqvist S, Bremer K (2007) Estimating divergence times in large phylogenetic trees. Syst Biol 56: 741–752.

[191] Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29: 1969–1973.

[192] Lartillot N, Lepage T, Blanquart S (2009) Phylobayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25: 2286–2288.

[193] Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, et al. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61: 539–542.

[194] Hess W (2008) The cyanobacteria: molecular biology, genomics, and evolution, Caister Academic Pr, chapter Comparative genomics of marine cyanobacteria and their phages. pp. 89–116.

[195] Cohen O, Pupko T (2010) Inference and characterization of horizontally transferred gene families using stochastic mapping. Mol Biol Evol 27: 703–713.

[196] Saw JH, Yuryev A, Kanbe M, Hou S, Young AG, et al. (2012) Complete genome sequencing and analysis of *Saprospira grandis* str. Lewin, a predatory marine bacterium. Stand Genomic Sci 6: 84–93.

[197] Honda D, Yokota A, Sugiyama J (1999) Detection of seven major evolutionary lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new sequences of five marine *Synechococcus* strains. J Mol Evol 48: 723–739.

[198] Rippka R, Coursin T, Hess W, Lichtle C, Scanlan DJ, et al. (2000) *Prochlorococcus marinus* Chisholm et al. 1992 subsp. *pastoris* subsp. nov. strain PCC 9511, the first axenic chlorophyll a2/b2-containing cyanobacterium (*Oxyphotobacteria)*. Int J Syst Evol Microbiol 50 Pt 5: 1833–1847.

[199] Paerl HW, Pinckney JL, Steppe TF (2000) Cyanobacterial-bacterial mat consortia: examining the functional unit of microbial survival and growth in extreme environments. Environ Microbiol 2: 11–26.

[200] Boston PJ, Ivanov MV, McKay CP (1992) On the possibility of chemosynthetic ecosystems in subsurface habitats on Mars. Icarus 95: 300–308.

[201] Laloui W, Palinska KA, Rippka R, Partensky F, Tandeau de Marsac N, et al. (2002) Genotyping of axenic and non-axenic isolates of the genus *Prochlorococcus* and the OMF-'*Synechococcus*' clade by size, sequence analysis or RFLP of the Internal Transcribed Spacer of the ribosomal operon. Microbiology 148: 453–465.

[202] Donlan RM, Costerton JW (2002) Biofilms: survival mechanisms of clinically relevant microorganisms. Clin Microbiol Rev 15: 167–193.

[203] Boston PJ, Spilde MN, Northup DE, Melim LA, Soroka DS, et al. (2001) Cave biosignature suites: microbes, minerals, and Mars. Astrobiology 1: 25–55.

[204] Ishida T, Yokota A, Sugiyama J (1997) Phylogenetic relationships of filamentous cyanobacterial taxa inferred from 16S rRNA sequence divergence. J Gen Appl Microbiol 43: 237–241.

[205] de los Rios A, Wierzchos J, Sancho LG, Ascaso C (2003) Acid microenvironments in microbial biofilms of antarctic endolithic microecosystems. Environ Microbiol 5: 231–237.

[206] Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, et al. (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci U S A 100: 10020–10025.

[207] Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, et al. (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424: 1042–1047.

[208] Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554–557.

[209] Jurado V, Groth I, Gonzalez JM, Laiz L, Saiz-Jimenez C (2005) *Agromyces subbeticus* sp. nov., isolated from a cave in southern spain. Int J Syst Evol Microbiol 55: 1897–1901.

[210] Engel AS, Porter ML, Stern LA, Quinlan S, Bennett PC (2004) Bacterial diversity and ecosystem function of filamentous microbial mats from aphotic (cave) sulfidic springs dominated by chemolithoautotrophic "*Epsilonproteobacteria*". FEMS Microbiol Ecol 51: 31–53.

[211] Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A 103: 12115–12120.

[212] Miller SR, Castenholz RW, Pedersen D (2007) Phylogeography of the thermophilic cyanobacterium *Mastigocladus laminosus*. Appl Environ Microbiol 73: 4751–4759.

[213] Knoll AH, Barghoorn ES (1977) Archean microfossils showing cell division from the Swaziland System of South Africa. Science 198: 396–398.

[214] Lacap DC, Barraquio W, Pointing SB (2007) Thermophilic microbial mats in a tropical geothermal location display pronounced seasonal changes but appear resilient to stochastic disturbance. Environ Microbiol 9: 3065–3076.

[215] Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, et al. (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet 3: e231.

[216] Cowie RH, Holland BS (2008) Molecular biogeography and diversification of the endemic terrestrial fauna of the Hawaiian Islands. Philos Trans R Soc Lond B Biol Sci 363: 3363–3376.

[217] Cole JR, Wang Q, Cardenas E, Fish J, Chai B, et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37: D141–D145.

[218] Chen Y, Wu L, Boden R, Hillebrand A, Kumaresan D, et al. (2009) Life without light: microbial diversity and evidence of sulfur- and ammonium-based chemolithotrophy in Movile Cave. ISME J 3: 1093–1104.

[219] Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19: 1639–1645.

[220] Engel AS, Meisinger DB, Porter ML, Payn RA, Schmid M, et al. (2010) Linking phylogenetic and functional diversity to nutrient spiraling in microbial mats from Lower Kane Cave (USA). ISME J 4: 98–110.

[221] Tindall BJ, Rosselló-Móra R, Busse HJ, Ludwig W, Kampfer P (2010) Notes on the characterization of prokaryote strains for taxonomic purposes. Int J Syst Evol Microbiol 60: 249–266.

[222] Pasic L, Kovce B, Sket B, Herzog-Velikonja B (2010) Diversity of microbial communities colonizing the walls of a Karstic cave in Slovenia. FEMS Microbiol Ecol 71: 50–60.

[223] Portillo MC, Gonzalez JM (2010) Differential effects of distinct bacterial biofilms in a cave environment. Curr Microbiol 60: 435–438.

[224] Jorge-Villar SE, Edwards HG (2010) Raman spectroscopy of volcanic lavas and inclusions of relevance to astrobiological exploration. Philos Transact A Math Phys Eng Sci 368: 3127–3135.

[225] West NJ, Lebaron P, Strutton PG, Suzuki MT (2011) A novel clade of *Prochlorococcus* found in high nutrient low chlorophyll waters in the South and Equatorial Pacific Ocean. ISME J 5: 933–944.

[226] dos Reis M, Yang Z (2011) Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol Biol Evol 28: 2161–2172.

[227] Sher D, Thompson JW, Kashtan N, Croal L, Chisholm SW (2011) Response of *Prochlorococcus* ecotypes to co-culture with diverse marine bacteria. ISME J 5: 1125–1132.

[228] Thompson AW, Huang K, Saito MA, Chisholm SW (2011) Transcriptome response of high- and low-light-adapted *Prochlorococcus* strains to changing iron availability. ISME J 5: 1580–1594.

[229] Nguyen TA, Brescic J, Vinyard DJ, Chandrasekar T, Dismukes GC (2012) Identification of an Oxygenic Reaction center *psbADC* Operon in the cyanobacterium *Gloeobacter violaceus* PCC 7421. Mol Biol Evol 29: 35–38.

[230] Sarbu SM, Kane TC, Kinkle BK (1996) A chemoautotrophically based cave ecosystem. Science 272: 1953–1955.

[231] Blankenship RE, Hartman H (1998) The origin and evolution of oxygenic photosynthesis. Trends Biochem Sci 23: 94–97.

[232] Mohagheghi A, Grohmann K, Himmel M, Leighton L, Updegraff D (1986) Isolation and characterization of *Acidothermus cellulolyticus* gen. nov., sp. nov., a new genus of thermophilic, acidophilic, cellulolytic bacteria. Int J Syst Evol Microbiol 36: 435–443.

[233] Ascaso C, Wierzchos J (2002) New approaches to the study of antarctic lithobiontic microorganisms and their inorganic traces, and their application in the detection of life in martian rocks. International Microbiology 5: 215–222.

[234] Bhattacharya D, Medlin L (1995) Phylogeny of plastids: A review based on comparisons of small subunit ribosomal RNA coding regions. J Phycol 31: 487–496.

[235] Lepot K, Benzerara K, Brown G, Philippot P (2008) Microbially influenced formation of 2,724-million-year-old stromatolites. Nature Geoscience 1: 118–121.

[236] Schopf J, Walter M (1982) Origin and early evolution of cyanobacteria: the geological evidence [Algae]. Botanical Monographs 19.

[237] Sheridan P, Freeman K, Brenchley J (2003) Estimated minimal divergence times of the major bacterial and archaeal phyla. Geomicrobiology Journal 20: 1–14.

[238] Steinman A, Parker A (1990) Influence of substrate conditioning on periphytic growth in a heterotrophic woodland stream. Journal of the North American Benthological Society : 170–179.