Title

The Comparative Genomics of Salinispora and the Distribution and Abundance of Secondary Metabolite Genes in Marine Plankton

Author
Penn, Kevin Matthew

Publication Date
2012-03-15

UNIVERSITY OF CALIFORNIA, SAN DIEGO

The Comparative Genomics of *Salinispora* and the Distribution and Abundance of Secondary Metabolite Genes in Marine Plankton

A Dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Marine Biology

by

Kevin Matthew Penn

Committee in charge:

Paul R. Jensen, Chair
Eric Allen
Lin Chao
Bradley Moore
Brian Palenik
Forest Rohwer

2012

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

The Dissertation of Kevin Matthew Penn is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2012

DEDICATION

I dedicate this dissertation to my Mom Gail Penn and my Father Lawrence Penn they deserve more credit then any person could imagine. They have supported me through the good times and the bad times. They have never given up on me and they are always excited to know that I am doing well. They just want the best for me. They have encouraged my education from both a philosophical and financial point of view. I also thank my sister Heather Kalish and brother in-law Michael Kalish for providing me with support during the beginning of my academic career and introducing me to Jonathan Eisen who ended opening the door for me to an endless bounty of intellectual pursuits.

EPIGRAPH

“Nothing in Biology Makes Sense Except in the Light of Evolution”

- Theodosius Dobzhansky, 1973

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Chapter 2, in full, is published in the Journal of the International Society for Microbial Ecology, 2009, Kevin Penn, Caroline Jenkins, Markus Nett, Daniel W. Udwary, Erin A. Gontang, Ryan P. McGlinchey, Brian Foster, Alla Lapidus, Sheila Podell, Eric E. Allen, Bradley S. Moore, and Paul R. Jensen. The dissertation author was the primary investigator and author of the research, which forms the basis for this chapter.

Chapter 3, in full, is in published in the Journal BMC Genomics, 2012, Kevin Penn and Paul R. Jensen. The dissertation author was the primary author and conducted the majority of the research, which forms the basis for this chapter.

Chapter 4, in part, contains methods, data and results that is submitted to PLoS ONE, 2012, Nadine Ziemert, Sheila Podell, Kevin Penn, Eric E. Allen, Paul R. Jensen. The dissertation author helped conceive the methods and aided in parts of the analyses, which forms the basis of this publication. The portions of chapter 4 related to analysis of metagenomes and metatranscriptomes will be submitted to an academic journal. I would like to acknowledge my collaborators at JCVI who collected and supported the collection and sequencing of all the metagenomes and metatransciptomes, specifically I would like to thank Dr. Andrew Allen, Dr. Lisa Zeigler, and James K. McCarthy for providing me with the sequence data for which I based my analyses.

Appendix A, in full, is submitted to Applied and Environmental Microbiology, 2012, Sergio Bucarey, Kevin Penn, Lauren Paul, William Fenical and Paul R. Jensen.

The dissertation author proposed the hypothesis that was tested and aided in the design of experiments.

Appendix B, in full, is submitted to PLoS ONE, 2012, Nadine Ziemert, Sheila Podell, Kevin Penn, Eric E. Allen, Paul R. Jensen. The dissertation author conceived the methods and aided in parts of the analyses, which forms the basis of this publication.

This dissertation is a result of collaboration. I would like to specifically thank my advisor Dr. Paul Jensen. He has always made the time to discuss my research interests. His unwavering determination that I make an outline has helped my scientific writing skills immensely. He has supported me both academically and in my personal life. I would also like to thank my committee members for useful comments and suggestions about my research. Dr. Eric Allen whom set my initial bioinformatics endeavors at Scripps into the right direction. I thank Dr. Jonathan Badger for his guidance, sequencing of *Salinispora* genomes and Ruby scripts that have made much of my analyses possible. I should also thank *Salinispora* for existing, as this Bacterium has been a great challenge to study. I would like to acknowledge Dr. Caroline Jenkins for her help in genome analysis. I thank Sheila Podell for many useful Perl scripts. I thank Dr. Erin Gontang for her help in learning about secondary metabolism. I thank Dr. Daniel Udwary for his help figuring out how to define secondary metabolite gene clusters. I thank Dr. Alejandra Prieto-Davo for providing me with a chance to be a part her family. I thank Dr. Xavier Alvarez Mico for all the good times especially running half marathons. There are numerous people who are

being left out but I would like to acknowledge those that have generated the sequence data that I have used as part of my analyses. I have to welcome all of my lab mates over the past six years Anna Edlund, Kelle Freel, Kelley Gallagher, Nadine Ziemert, Natalie Millan, Eun Ju Choi, Chris Kaufmann, Krystle Chavarria and Anindita Sarkar. Finally, I would like to thank all the members of my committee for useful insight into my research over the years.

VITA

June, 2002 B.S. in Aquatic Biology, University of California, Santa Barbara.

2002-2006 Research Associate at The Institute for Genomic Research (TIGR).

2006-2012 Ph.D. in Marine Biology, Scripps Institution of Oceanography, University of California, San Diego, CA.

PUBLICATIONS

Bucarey S, Penn K, Paul L, Fenical W, Jensen PR (Submitted) Comparative genomics reveals evidence of marine adaptation in Salinispora species. Applied and Environmental Microbiology.

Eustáquio AS, Nam S-J, Penn K, Lechner A, Wilson MC, Fenical W, Jensen PR, Moore BS (2011) The Discovery of Salinosporamide K from the Marine Bacterium "Salinispora pacifica" by Genome Mining Gives Insight into Pathway Evolution. ChemBioChem 12(1): 61-64.

Hubert CRJ, Oldenburg TBP, Fustic M, Gray ND, Larter SR, Penn K, Rowan AK, Seshadri R, Sherry A, Swainsbury R, Voordouw G, Voordouw JK, Head IM Massive dominance of Epsilonproteobacteria in formation waters from a Canadian oil sands reservoir containing severely biodegraded oil. Environmental Microbiology 14(2): 387-404.

Martiny JBH, Eisen JA, Penn K, Allison SD, Horner-Devine MC Drivers of bacterial β-diversity depend on spatial scale. Proceedings of the National Academy of Sciences 108(19): 7850-7854.

Penn K, Jensen PR (2012) Comparative genomics reveals evidence of marine adaptation in Salinispora species. BMC Genomics 13(86).

Penn K, Wu D, Eisen JA, Ward N (2006) Characterization of Bacterial Communities Associated with Deep-Sea Corals on Gulf of Alaska Seamounts. Applied and Environmental Microbiology 72(2): 1680-1683.

Penn K, Jenkins C, Nett M, Udwary DW, Gontang EA, McGlinchey RP, Foster B, Lapidus A, Podell S, Allen EE, Moore BS, Jensen PR (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. ISME J 3(10): 1193-1203.

Udwary DW, Gontang EA, Jones AC, Jones CS, Schultz AW, Winter JM, Yang JY, Beauchemin N, Capson TL, Clark BR, Esquenazi E, Eustáquio AS, Freel K, Gerwick L, Gerwick WH, Gonzalez D, Liu W-T, Malloy KL, Maloney KN, Nett M, Nunnery JK, Penn K, Prieto-Davo A, Simmons TL, Weitz S, Wilson MC, Tisa LS, Dorrestein PC, Moore BS Significant Natural Product Biosynthetic Potential of Actinorhizal Symbionts of the Genus Frankia, as Revealed by Comparative Genomic and Proteomic Analyses. Applied and Environmental Microbiology 77(11): 3617-3625.

Ward N, Steven B, Penn K, Methé B, Detrich W (2009a) Characterization of the intestinal microbiota of two Antarctic notothenioid fish species. Extremophiles 13(4): 679-685.

Ward NL, Challacombe JF, Janssen PH, Henrissat B, Coutinho PM, Wu M, Xie G, Haft DH, Sait M, Badger J, Barabote RD, Bradley B, Brettin TS, Brinkac LM, Bruce D, Creasy T, Daugherty SC, Davidsen TM, DeBoy RT, Detter JC, Dodson RJ, Durkin AS, Ganapathy A, Gwinn-Giglio M, Han CS, Khouri H, Kiss H, Kothari SP, Madupu R, Nelson KE, Nelson WC, Paulsen I, Penn K, Ren Q, Rosovitz MJ, Selengut JD, Shrivastava S, Sullivan SA, Tapia R, Thompson LS, Watkins KL, Yang Q, Yu C, Zafar N, Zhou L, Kuske CR (2009b) Three Genomes from the Phylum Acidobacteria Provide Insight into the Lifestyles of These Microorganisms in Soils. Applied and Environmental Microbiology 75(7): 2046-2056.

Ziemert N, Podell S, Penn K, Allen EE, Jensen PR (In Press) The Natural Product Domain Seeker NaPDoS: a bioinformatics tool to classify secondary metabolite gene diversity. PLoS ONE.

FIELD OF STUDY

Marine microbial genomics

Evolution and Ecology of marine microbes

Using phylogenetics to identify gene that can make known and novel secondary metabolites

ABSTRACT OF THE DISSERTATION

The Comparative Genomics of *Salinispora* and the Distribution and Abundance of Secondary Metabolite Genes in Marine Plankton

by

Kevin Matthew Penn

Doctor of Philosophy in Marine Biology

University of California, San Diego, 2012

Paul Jensen, Chair

This dissertation is based on a bioinformatics approach to study microbiology, ecology, evolution, marine biology and secondary metabolites. Comparative genomics was applied to identify the similarities and differences between two marine Actinobacteria *Salinispora tropica* and *S. arenicola*. The first step in this analysis was to identify orthologous genes between the two species and create a gene-by-gene alignment of the genomes in order to identify synteny of orthologs. The second step was to identify all secondary metabolite gene clusters and mobile genetic elements followed by a thorough analysis of the evidence for horizontal gene transfer. The first two steps reveal that the main differences between these species lie on genomic islands that harbor secondary metabolites and mobile genetic elements. The *Salinispora* genomes were used as the basis for comparison against other Actinobacteria to identify possible marine adaptation genes. Several marine adaptation genes were identified based on two fundamental approaches, a comparative genomic approach and a study of gene annotation previously linked to marine adaptation. These two approaches, coupled with phylogenetic analyses, identified genes that show a close relationship to marine bacteria and appear to be involved in marine adaptation. During this study, a gene that encodes a mechanosensitive channel was identified as having been lost in *Salinispora* relative to almost all other terrestrial Actinobacteria. This gene is likely a contributing factor to the inability of *Salinispora* to grow when seawater based media is replaced with DI based growth media. In this dissertation, I also describe a method to identify sequence tags related to polyketide synthase and non-ribosomal peptide synthetases. I applied this method to study a metagenome of surface water collected in the California current and metatranscriptomes of a dinoflagellate bloom in surface water of the coast of California and water beneath sea ice in Antarctica. This study revealed an abundance of protist-associated secondary metabolite genes and evidence that extensive sequencing efforts will be required to detect rare functional genes such as those involved in secondary metabolism.

Chapter 1: Introduction

Bacteria are significant numerical and ecological players in marine ecosystems (Azam et al. 1983). Their diversity and function is relevant to a complete understanding of marine habitats. The analysis of genome sequence data has recently emerged as an effective way to identify similarities and differences between bacterial species and determine the implications of the similarities and differences as they relate to evolution and ecology (Tettelin et al. 2005; Coleman et al. 2006). Prior to the invention of high-throughput sequencing methods, marine microbiology suffered from a lack of tools to study the roles of bacterial populations in their natural habitats (Ducklow 1983). Now machines from Roche and Illumina can generate massive sets of DNA sequence data and the analysis of genomic and metagenomic sequences has become an effective way to study marine microbial communities. As of 2012 over 1,500 publications have cited the use of 454 Roche flx sequencing technology, which is capable of generating approximately 450 megabases of data in one sequencing experiment, and over 2000 publications have cited the use of illumina's sequencing by synthesis method, which produces 600 gigabases of sequence data in one sequencing experiment.

Cultured based comparisons of closely related bacterial species have shown the genomic differences that confer the ability to occupy different ecological niches (Fleischmann et al. 1995; Coleman et al. 2006). Metagenomics based comparisons of bacteria have provided clues to how an entire microbial community functions and how species and genes vary across different environmental gradients (Tyson et al. 2004;

Rusch et al. 2007; Dinsdale et al. 2008). As part of this thesis, the genome analyses of cultured isolates of the marine actinomycete *Salinispora* were used to make inferences about its ecology and evolution. From these studies of *Salinispora,* a phylogenetic guide was established that could be used to predict known and novel types of natural products. The guide was then used to study the distribution of genes that can produce natural products in metagenomic data from the ocean.

Marine Biology

Marine Biology has been at the forefront of science from the earliest of times. When Aristotle spent two years studying on the island of Lesbos, he sketched the anatomy of octopus, cuttlefish and other marine invertebrates along with distinguishing whales and dolphins from fish, which earned him the title "father of zoology" (Barnes 1995). In the time of Columbus (circa 1492), the sea represented an unknown where people fell off the edge of the world. Even Charles Darwin was intrigued by the ocean and correctly hypothesized about the formation of coral reef atolls (Darwin 1896). Despite human interest in the ocean, many beliefs and scientific assumptions have proven to be inaccurate. In the early 19th century, it was thought that life did not exist below a depth of more then 300 fathoms, which was termed the azoic zone (Kunzig 2003). This belief was destroyed in the 1850's when encrusting animals were observed on telegraph cables brought to the surface for repair from 1200 fathoms (Murray and Great Britain. Challenger 1895). Scientists originally did not understand why phytoplankton are so diverse because they believed most of the ocean was isotropic and unstructured and only limited types of resources exist. This conflict between belief and observation was termed the paradox of the plankton (Hutchinson 1961). The paradox existed mainly because scientists at the time did not understand the significant structure that could exist at the microscopic level in the seemingly mundane and vast sea. Research on microbes in the 1970's and 1980's led to the hypothesis that microbes play a significant and fundamental role in planktonic food webs (Azam et al. 1983; Ducklow 1983; Fenchel 2008). The term microbial loop was coined to describe the major role microbes have in the transformation of matter and energy in the plankton. The findings from this and other research have shown that the ocean is not isotropic and unstructured but is highly heterogeneous with specific habitats associated with different particulates. The extensive diversity of microbial life has provided scientists with enormous opportunities to learn about microbial ecology and evolution. Microbes are gaining more and more attention as studies continue to reveal details about their extensive and unexpected diversity and role in our oceans.

Microbial Loop

It was the realization that aerobic heterotrophic bacteria make up a very large and dynamic component of the biomass in the illuminated surface layers of the coastal and open oceans that has driven humans to embark on highly detailed studies in microbial oceanography. The concept of a microbial loop put microbes in perspective and made people think about looking more specifically at the types of microbes involved in different ecological processes. Without knowing the types and functions of the bacteria in different areas of the ocean, it is impossible to completely understand bacterial food webs.

The goal of learning more about the ecology and evolution of marine microbes gathered momentum due to advances in whole genome sequencing technologies and, most recently, high-throughput gene sequencing efforts on both whole communities and cultured bacterial populations. First, the dominant bacteria from the ocean were sequenced and then some of the rare taxonomic groups were analyzed. The results of sequencing *Vibrio*, *Prochlorococus* and *Roseobacter* genomes from the ocean revealed extensive genomic diversity even among closely related species (Acinas et al. 2004; Coleman et al. 2006; Moran et al. 2007). Now the goal of scientists is to learn about dominant functional types of genes in different environments and understand the diversity among closely related groups in the hopes of establishing and understanding what a bacterial species is and what level of divergence is ecologically relevant. One basic question researched in chapter 3 of this dissertation is what are the differences between marine and non-marine species.

Marine Sediment

The first life observed below the "azoic zone" was actually from marine sediment (Thomson et al. 1873). C. Wyville Thompson said, "The land of promise for the naturalist is the bottom of the sea" (Thomson et al. 1873). There are 10^6 microbial cells per ml of seawater throughout the world oceans but on average 10^9 microbial cells per ml of marine sediments (Schallenberg and Kalff 1993). Many studies have focused on marine microbial planktonic communities but few studies have attempted to characterize microbial diversity or function in marine sediments. In particular, it is unknown what the dominant bacteria are and how diversity and function vary across environmental gradients. Recently, culture dependent and culture independent studies of marine sediments show there is a great amount of diversity and some support that redox gradients play a role in structuring bacterial communities (Edlund et al. 2008). For example, the variation in Baltic Sea sediment microbial communities correlates to dramatic changes in redox potential. In addition, organic carbon and total nitrogen were significant variables associated with bacterial community structure over horizontal scales of up to 1 km (Edlund et al. 2008). A metagenomic fosmid library from a China Sea sediment revealed that Proteobacteria and planctomycetes were the dominant members and that sulfate reducing, anaerobic ammonium oxidizing bacteria dominate (Hu et al. 2010). This study also provided evidence that metabolism of one-carbon compounds, methanogenesis and the biodegradation of xenobiotics are common in marine sediments (Hu et al. 2010).

It has been suggested that many marine sediment bacteria are derived from terrestrial runoff (Munn 2004; Bull et al. 2005). It is logical that marine sediments would harbor similar types of bacteria as those inhabiting terrestrial soil but until recently there was little data to support the concept. A recent study of Gram-positive bacteria in marine sediment along the coast of California showed that some marine sediment bacteria are terrestrial in origin but found that there are also specific populations of marine bacteria (Prieto-Davó et al. 2008). In this study, it did not appear that the numbers of bacteria with a requirement of seawater for growth increased as distance from shore increased. One third of the operational taxonomic units were marine-specific, suggesting that sediment communities include considerable diversity that does not occur on land. However, the seawater requiring actinomycetes isolated from marine sediments did not form any deeply rooted clades in the actinomycete phylogenetic tree suggesting that marine actinomycetes have secondarily been introduced to the ocean (Prieto-Davó et al. 2008). This study is in agreement with previous discoveries of specific actinomycete taxa residing in tropical marine sediments (Mincer et al. 2002).

A study of tropical marine sediments in Palau showed that there is a wealth of new bacteria to be discovered (Gontang et al. 2007). In this study (Gontang et al. 2007), phylogenetic diversity of Gram-positive bacteria cultured from marine sediments suggests that Gram-positive bacteria comprise a relatively large proportion of marine sediment communities. Within 22 Gram-positive families in marine sediments a total of 78 Gram-positive OTUs were cultured of which 21 were considered to be new phylotypes based on the sharing of <98% 16S rRNA gene sequence identity with any previously cultured isolates. Using relatively easy cultivation techniques, the study showed that much new Gram-positive diversity could be found thus emphasizing that no one has thoroughly tried to culture bacteria in marine sediments.

Studies of marine sediment bacteria are usually application driven with a focus on looking for enzymes with industrial application such as lipases or enzymes that reduce heavy metals such as nickel and iron and natural products for pharmaceutical applications (Hu et al. 2010). Marine sediments along with sponges are one of the few places in the ocean where bacteria that produce natural products have consistently been recovered (Fenical and Jensen 2006). A strong interest in marine sediments comes from the wealth of Gram-positive bacterial inhabitants.

Natural Product Discovery

Natural product research seeks to identify molecules that can be developed into pharmaceuticals (Fenical and Jensen 2006). The fact that Actinobacteria are the largest producers of natural products makes them of great importance to society but the mystery surrounding the actual ecological function and evolutionary history of natural products makes them attractive to study by microbiologists. Microbes in the sea were not just ignored for their role in marine food webs, they have also been ignored for the potential to produce cures for disease. Scientists have known that microbes in the soil produce antibiotics since the 1940's (Kresge et al. 2004), however there was little effort to look in the ocean. The stage was set for the incorporation of marine microbes into natural product research, when in 1977 Dr. William Fenical joined Scripps Institution of Oceanography to search for "medicine in the sea" (Balzar 2006). From the period of 1977 to 2001, mostly larger organisms from the ocean were studied for their ability to produce natural products using a method referred to in terms of jargon as "grind and find". A process where larger organism are collected and ground up to have their chemicals extracted and examined for bioactivities against various diseases. Perhaps with the realization that many of the bioactive molecules are produced by bacteria associated with the larger molecules Dr. William Fenical in collaboration with Dr. Paul Jensen began to look specifically at microbes for possible novel drugs. This collaboration led to the discovery of the first marine obligate actinomycete genus, formally known as *Salinispora* (Maldonado et al. 2005).

Actinobacteria

The Actinobacteria are Gram-positive bacteria, as are the low GC bacteria in the phylum Firmicutes. Marine sediments have indeed been the source of soil related bacteria, in particular species from the Phylum Actinobacteria. The Actinobacteria are well known soil bacteria and produce the majority of antibiotics (Berdy 2005). The Actinobacteria found in marine sediments also produce natural products but can also be specifically adapted to grow and live in the marine environment (Fenical and Jensen 2006). The name Gram-positive is taken from the fact that these bacteria score "positive" or perhaps more appropriately purple in the gram test (Kaplan and Kaplan 1933). A positive test result is indicative of a thick outer peptidoglycan layer outside of the cell membrane relative to that observed in Gram-negative bacteria. Actinobacteria are differentiated from the Firmicutes both phylogenetically and because the genomes generally have a GC content well above 50% while Firmicutes typically have a GC content below 50%.

The class Actinomycetales is the group of organisms from which the majority of natural products have been found (Berdy 2005). Natural products are also called secondary metabolites because of their non-ubiquity and non-essential role in survival

(Challis and Hopwood 2003). The two main classes of secondary metabolites are polyketides and non-ribosomally derived peptides. However, there are other types of natural products that are produced by microorganisms. These include small ribosomally produced peptides and terpene molecules that have been found to have medicinal properties (Schmidt 2010).

Salinispora

The genus *Salinispora* is composed of three species and is part of the phylum Actinobacteria. Two complete genomes and four draft *Salinispora* genomes are available. All known species of *Salinispora* fail to grow on typical growth media when seawater is replaced with deionized water (Mincer et al. 2002). Typically, terrestrial type Actinomycetes have no seawater requirement for growth and grow on DI based growth media. On Petri dishes, they form substrate mycelia on which spores can form. *Salinispora* has an interesting species distribution (Freel et al. 2011), they require seawater for growth, and have only been found in the ocean. A study of their genomes promised to provide insights for the field of microbial ecology and evolution along with natural products researchers hoping to link biosynthetic pathways to molecules.

Salinispora makes many highly bioactive secondary metabolites. The only major phenotypic difference observed to date among *Salinispora* species is the set of secondary metabolites that they produce (Jensen et al. 2007). Specific secondary metabolites appear to correlate with each species (Jensen et al. 2007). As part of this dissertation, studies of the genome sequences of two *Salinispora* species revealed an abundance of polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) genes (Udwary et al. 2007; Penn et al. 2009). Thus far, *Salinispora* studies have yielded several varieties of secondary metabolites including polyketides and non-ribosomally produced peptides, some of which have been developed for the treatment of human disease (Feling et al. 2003). There are several well-known molecules produced by the PKS and NRPS genes in *Salinispora* (Fenical and Jensen 2006). One well-known compound produced by *S. tropica* is called salinosporamide A and is in clinical trials for the treatment of cancer (Feling et al. 2003).

Genome sequencing is directly influencing the methods of natural product research. The known PKS and NRPS genes in the *Salinispora* genomes have been used to learn about the evolution of these genes and design methods to use phylogenetics to predict the presence of known and novel secondary metabolites. Natural products researchers look for new molecules, but they also look for drugs with novel mechanisms of action. One way to find new natural products is to look for phylogenetically distinct biosynthetic pathways. As part of this dissertation, I worked with others to produce a tool that can use phylogenetics to identify C and KS domains from genetic data. This information was then used to predict novel biosynthetic pathways and natural product structures.

History of Evolutionary Biology

One could simply believe that the story of evolutionary biology has already been written. Charles Darwin wrote the book on natural selection and showed how natural selection is the mechanism of evolution (Darwin 1871). Gregor Mendel figured out modes of inheritance (Henig 2001) and then James Watson and Francis Crick determined the structure of DNA (Watson and Crick 1953) providing a mechanism by which information is passed from generation to generation. Julian Huxley then put it all together in the modern synthesis (Huxley 1942). The next monumental steps in understanding evolution occurred in 1977 when Carl Woese used the 16S rRNA gene to show that life can be split into three domains thus defining the Bacteria, Archae and Eukarea (Woese and Fox 1977; Woese et al. 1978). The discovery of fossils inside the rocks of the burgess shale and other places around the globe have shown that early animal life looked drastically different than today (Gould 1990). Stephen Gould proposed punctuated equilibrium, a process where evolution is accelerated and thus change occurs too quickly to be observed in the fossil record thus explaining why there are often large gaps (Gould and Eldredge 1993). More recently, shotgun DNA sequencing employed first by Craig Venter and Hamilton Smith has been exploited to sequence complete genomes and has brought biology into a new phase of research (Fleischmann et al. 1995). Now several generations of sequencing technology have passed and generating sequence data is no longer a limiting factor. Indeed, as science requires, the predictions set out by evolutionary biologists are constantly being tested and methods upgraded to deal with new information such as that derived from massive genome sequencing efforts. The foundation for studies in evolution is set (Gould 2002) but now scientists are beginning to go through and comprehensively study the details about the relationships of organisms. High-powered computers and new algorithms allow scientists to build gene trees containing thousand of sequences and compare these to species trees in a field called phylogenomics (Eisen 1998). These species tree comparisons have allowed us to learn more about how to reconstruct evolutionary pathways. New approaches for understanding and defining relationships are always emerging and, in fact, the story of evolution is far from complete and new evidence is continuously being discovered and ideas about how evolution occurs are constantly being refined.

Microbial Ecology and Evolution

Evolutionary microbiology can be defined as the study of the patterns (relationships between genes and organisms) and processes (mechanisms generating diversity and the selection operating on it) of evolution in microbes (Case and Boucher 2011). Without considering microbes, much of the biology observed appears to fall into the two categories of plant or animal. Initial confusion related to classifying organisms was due in part to the fact that many microbes can fit into both plant and animal categories. To deal with the part animal, part plant paradox, biologists divided life into prokaryote and eukaryote and used a five-kingdom system. Then based on work using small ribosomal RNA genes by Carl Woese and George Fox (Woese and Fox 1977) it became clear that life could be divided into three domains currently called Archaea, Bacteria and Eukaraea. Norman Pace then used the high level of conservation of ribosomal genes in all living things to show that much of Archaeal and Bacterial diversity has not been cultured (Stahl et al. 1984). Morphology (*e.g.*

phenotype) had previously been the major criterion used by microbiologist to understand the phylogenetic relationships of bacteria. It was the studies of bacteria using the 16S rRNA genes that had a profound impact on the way phenotypes are related to phylogeny in bacteria. Phylogenetic studies objectively relate organisms (Hugenholtz et al. 1998). Most importantly, it has been the realization that bacterial phenotypes are particularly unreliable as an indicator of phylogeny as is thought to be the case for many animal characteristics. Furthermore rampant horizontal transfer makes interpretation of gene based phylogeny diffictult (Philippe et al. 2011). As the gene-based analysis of microbes progressed, it became clear that much of the phenotypic characteristics are not phylogenetically informative and horizontal gene transfer blurs species boundaries among organism that were once considered clonal, which gave birth to the controversy of whether actual species analogous to eukaryotic species exist among prokaryotes (Cohan 2002; Gogarten et al. 2002). This is mainly where the field lies today. Most of the current research is focused on understanding the fundamental units of diversity that microbes can be divided into.

Conclusion

The remarkable part of science is that no matter what field is studied the same rules apply. Evidence in the form of testable hypotheses and repeatable results from experiments must be provided to explain observations no matter how intuitive the explanation of the observation may seem. Charles Darwin was the first to detail the evidence that all life evolved and shares a common ancestry. Without Darwin's contribution to the structure of the theory of 'the origin of species', comparative studies of life would most certainly struggle to draw conclusions. His evidence provided the first real support of life's common ancestry, which now seems so intuitive. If the genomic era had dawned before Darwin's time would his theory on the origin of species seemed so revolutionary? Presently, comparative studies of different species are aided by genome sequence data. Specifically, studies of microbes have benefited from the ability to sequence whole genomes and allow scientists to understand the similarities and differences. A speech teacher of mine in college once explained that man should not be called homo sapien, which is latin for wise man but instead homo narrare (narrating man). I agree because we are indeed the only organisms that can take something and turn it into a great story. Here I will give a narrative of three distinct stories from evolutionary biology, marine biology, genomics and natural products research. I use the stories to introduce the results of previous scientific endeavors and provide a guide for the different fields I have drawn from to produce this dissertation.

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF (2004) Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430(6999): 551-554.

Azam F, Fenchel T, Field JG, Gray JS, Meyer LA, Thingstad F (1983) The ecological role of water column microbes in the sea. Marine Ecology Progress Series 10: 257-263.

Balzar J (2006) Neptune's Medicine Chest. Los Angeles Times. Los Angeles.

Barnes J (1995) The Cambridge companion to Aristotle: Cambridge University Press.

Berdy J (2005) Bioactive Microbial Metabolites. J Antibiot 58(1): 1-26.

Bull A, Stach J, Ward A, Goodfellow M (2005) Marine actinobacteria: perspectives, challenges, future directions. Antonie van Leeuwenhoek 87(1): 65-79.

Case RJ, Boucher Y (2011) Molecular musings in microbial ecology and evolution. Biology direct 6(1): 58.

Challis GL, Hopwood DA (2003) Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by Streptomyces species. Proceedings of the National Academy of Sciences of the United States of America 100(Suppl 2): 14555-14561.

Cohan FM (2002) What are bacterial species? Annual Review of Microbiology 56: 457-487.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, DeLong EF, Chisholm SW (2006) Genomic Islands and the Ecology and Evolution of *Prochlorococcus*. Science 311(5768): 1768-1770.

Darwin C (1871) On the origin of species. New York: D. Appleton and Company.

Darwin C (1896) The structure and distribution of coral reefs. New York: D. Appleton and Company.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F (2008) Functional metagenomic profiling of nine biomes. Nature 452(7187): 629-632.

Ducklow HW (1983) Production and Fate of Bacteria in the Oceans. BioScience 33(8): 494-501.

Edlund A, Hårdeman F, Jansson JK, Sjöling S (2008) Active bacterial community structure along vertical redox gradients in Baltic Sea sediment. Environmental Microbiology 10(8): 2051-2063.

Eisen JA (1998) Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary, Analysis. Genome Research 8(3): 163-167.

Feling RH, Buchanan GO, Mincer TJ, Kauffman CA, Jensen PR, Fenical W (2003) Salinosporamide A: A Highly Cytotoxic Proteasome Inhibitor from a Novel Microbial Source, a Marine Bacterium of the New Genus *Salinospora*. Angewandte Chemie 115(3): 369-371.

Fenchel T (2008) The microbial loop - 25 years later. Journal of Experimental Marine Biology and Ecology 366(1-2): 99-103.

Fenical W, Jensen PR (2006) Developing a new resource for drug discovery: marine actinomycete bacteria. Nature Chemical Biology 2(12): 666-673.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM (1995a) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269(5223): 496-512.

Freel KC, Nam S-J, Fenical W, Jensen PR (2011) Evolution of Secondary Metabolite Genes in Three Closely Related Marine Actinomycete Species. Applied and Environmental Microbiology 77(20): 7261-7270.

Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic Evolution in Light of Gene Transfer. Molecular Biology and Evolution 19(12): 2226-2238.

Gontang EA, Fenical W, Jensen PR (2007) Phylogenetic Diversity of Gram-Positive Bacteria Cultured from Marine Sediments. Applied and Environmental Microbiology 73(10): 3272-3282.

Gould SJ (1990) Wonderful life: the Burgess Shale and the nature of history: Norton.

Gould SJ (2002) The structure of evolutionary theory: Belknap Press of Harvard University Press.

Gould Sj, Eldredge N (1993) Punctuated equilibrium comes of age. Nature 366(6452): 223-227.

Henig RM (2001) The Monk in the Garden: The Lost and Found Genius of Gregor Mendel, the Father of Genetics: Houghton Mifflin.

Hu Y, Fu C, Yin Y, Cheng G, Lei F, Yang X, Li J, Ashforth E, Zhang L, Zhu B (2010) Construction and Preliminary Analysis of a Deep-Sea Sediment Metagenomic Fosmid Library from Qiongdongnan Basin, South China Sea. Marine Biotechnology 12(6): 719-727.

Hugenholtz P, Goebel BM, Pace NR (1998) Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. The Journal of Bacteriology 180(18): 4765-4774.

Hutchinson GE (1961) The Paradox of the Plankton. The American Naturalist 95(882): 137-145.

Huxley J (1942) Evolution: the modern synthesis: George Allen & Unwin.

Jensen PR, Williams PG, Oh D-C, Zeigler L, Fenical W (2007) Species-Specific Secondary Metabolite Production in Marine Actinomycetes of the Genus *Salinispora*. Applied and Environmental Microbiology 73(4): 1146-1152.

Kaplan ML, Kaplan L (1933) The Gram Stain and Differential Staining. Journal of Bacteriology 25(3): 309-321.

Kresge N, Simoni RD, Hill RL (2004) Selman Waksman: the Father of Antibiotics. Journal of Biological Chemistry 279(48): e7.

Kunzig R (2003) Deep-Sea Biology: Living with the Endless Frontier. Science 302(5647): 991.

Maldonado LA, Fenical W, Jensen PR, Kauffman CA, Mincer TJ, Ward AC, Bull AT, Goodfellow M (2005) *Salinispora arenicola* gen. nov., sp. nov. and *Salinispora tropica* sp. nov., obligate marine actinomycetes belonging to the family Micromonosporaceae. International Journal of Systematic and Evolutionary Microbiology 55(5): 1759-1766.

Mincer TJ, Jensen PR, Kauffman CA, Fenical W (2002) Widespread and Persistent Populations of a Major New Marine Actinomycete Taxon in Ocean Sediments. Applied and Environmental Microbiology 68(10): 5005-5011.

Moran MA, Belas R, Schell MA, Gonzalez JM, Sun F, Sun S, Binder BJ, Edmonds J, Ye W, Orcutt B, Howard EC, Meile C, Palefsky W, Goesmann A, Ren Q, Paulsen I, Ulrich LE, Thompson LS, Saunders E, Buchan A (2007) Ecological Genomics of Marine Roseobacters. Applied and Environmental Microbiology 73(14): 4559-4569.

Munn CB (2004) Marine microbiology: ecology and applications: Garland Science/BIOS Scientific Publishers.

Murray J, Great Britain. Challenger O (1895) Selections from Report on the scientific results of the voyage of H.M.S. Challenger during the years 1872-76: Arno Press.

Penn K, Jenkins C, Nett M, Udwary DW, Gontang EA, McGlinchey RP, Foster B, Lapidus A, Podell S, Allen EE, Moore BS, Jensen PR (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. ISME J 3(10): 1193-1203.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D (2011) Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biology 9(3): e1000602.

Prieto-Davó A, Fenical W, Jensen PR (2008) Comparative actinomycete diversity in marine sediments. Aquat Microb Ecol 52(1): 11.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers Y-H, Falcón LI, Souza V, Bonilla-Rosso Gn, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC (2007) The *Sorcerer* II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biology 5(3): e77.

Schallenberg M, Kalff J (1993) The Ecology of Sediment Bacteria in Lakes and Comparisons with Other Aquatic Ecosystems. Ecology 74(3): 919-934.

Schmidt E (2010) The hidden diversity of ribosomal peptide natural products. BMC Biology 8(1): 83.

Stahl DA, Lane DJ, Olsen GJ, Pace NR (1984) Analysis of Hydrothermal Vent-Associated Symbionts by Ribosomal RNA Sequences. Science 224(4647): 409-411.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences 102(39): 13950-13955.

Thomson CW, Carpenter WB, Jeffreys JG (1873) The depths of the sea. An account of the general results of the dredging cruises of H. M. SS. 'Porcupine' and 'Lighting' during the summers of 1868, 1869, and 1870, under the scientific direction of Dr. Carpenter, F. R. S., J. Gwyn Jeffreys, F. R. S., and Dr. Wyville Thomson, F. R. S.: New York and London, Macmillan and co.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428(6978): 37-43.

Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W, Jensen PR, Moore BS (2007) Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. Proceedings of the National Academy of Sciences 104(25): 10376-10381.

Watson JD, Crick FHC (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature 171(4356): 737-738.

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences 74(11): 5088-5090.

Woese CR, Magrum LJ, Fox GE (1978) Archaebacteria. Journal of Molecular Evolution 11(3): 245-252.

Chapter 2: Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria

Abstract

Genomic islands have been shown to harbor functional traits that differentiate ecologically distinct populations of environmental bacteria. A comparative analysis of the complete genome sequences of the marine Actinobacteria *Salinispora tropica* and *S. arenicola* reveals that 75% of the species-specific genes are located in 21 genomic islands. These islands are enriched in genes associated with secondary metabolite biosynthesis providing evidence that secondary metabolism is linked to functional adaptation. Secondary metabolism accounts for 8.8% and 10.9% of the genes in the *S. tropica* and *S. arenicola* genomes, respectively, and represents the major functional category of annotated genes that differentiates the two species. Genomic islands harbor all 25 of the species-specific biosynthetic pathways, the majority of which occur in *S. arenicola* and may contribute to the cosmopolitan distribution of this species. Genome evolution is dominated by gene duplication and acquisition, which in the case of secondary metabolism provide immediate opportunities for the production of new bioactive products. Evidence that secondary metabolic pathways are exchanged horizontally, coupled with prior evidence for fixation among globally distributed populations, supports a functional role and suggests that the acquisition of natural product biosynthetic gene clusters represents a previously unrecognized force

driving bacterial diversification. Species-specific differences observed in CRISPR (clustered regularly interspaced short palindromic repeat) sequences suggest that *S. arenicola* may possess a higher level of phage immunity, while a highly duplicated family of polymorphic membrane proteins provides evidence of a new mechanism of marine adaptation in Gram-positive bacteria.

Introduction

Linking functional traits to bacterial phylogeny remains a fundamental but elusive goal of microbial ecology (Hunt et al. 2008). Without this information, it becomes difficult to resolve meaningful units of diversity and the mechanisms by which bacteria interact with each other and adapt to environmental change. Most bacterial diversity is delineated among clusters of sequences that share >99% 16S rRNA gene sequence identity (Acinas et al. 2004). These sequence clusters are believed to represent fundamental units of diversity, while intra-cluster microdiversity is thought to persist due to weak selective pressures (Acinas et al. 2004) suggesting little ecological or taxonomic relevance. Recently, progress has been made in terms of delineating units of diversity that possess the fundamental properties of species by linking genetic diversity with ecology and evolutionary theory (Achtman and Wagner 2008; Fraser et al. 2009). Despite these advances, there remains no widely accepted species concept for prokaryotes (Gevers et al. 2005), and sequence-based analyses reveal widely varied levels of diversity within assigned species boundaries.

The comparative analysis of bacterial genome sequences has revealed considerable differences among closely related strains (Joyce et al. 2002; Welch et al. 2002; Thompson et al. 2005) and provides a new perspective on genome evolution and prokaryotic species concepts. Genomic differences among closely related strains are concentrated in islands, strain-specific regions of the chromosome that are generally acquired by horizontal gene transfer (HGT) and harbor functionally adaptive traits (Dobrindt et al. 2004) that can be linked to niche adaptation. The pelagic cyanobacterium *Prochlorococcus* is an important model for the study of island genes, which in this case are differentially expressed under low nutrient and high light stress in ecologically distinct populations (Coleman et al. 2006). Despite convincing evidence for the adaptive significance of island genes among environmental bacteria, the precise functions of their products have seldom been characterized and their potential role in the evolution of independent bacterial lineages remains poorly understood.

The marine sediment inhabiting genus *Salinispora* belongs to the Order Actinomycetales, a group of Actinobacteria commonly referred to as actinomycetes. Actinomycetes are a rich source of structurally diverse secondary metabolites and account for the majority of antibiotics discovered as of 2002 (Berdy 2005). *Salinispora* spp. have likewise proven to be a rich source of secondary metabolites (Fenical and Jensen 2006) including salinosporamide A, which is currently in clinical trials for the treatment of cancer (Fenical et al. 2009). At present, the genus is comprised of three species that collectively constitute a microdiverse sequence cluster (*sensu* (Acinas et al. 2004), i.e., they share ≥99% 16S rRNA gene sequence identity (Jensen and Mafnas 2006). Although the microdiversity within this cluster has been formally delineated into species-level taxa (Maldonado et al. 2005), it remains to be determined if these taxa represent ecologically or functionally distinct lineages.

Here we report the comparative analysis of the complete genome sequences of *S. tropica* (strain CNB-440, the type strain for the species and thus a contribution to the Genomic Encyclopedia of Bacteria and Archaea project), hereafter referred to as ST, and *S. arenicola* (strain CNS-205), hereafter referred to as SA, the first obligately marine Actinobacteria to be obtained in culture (Mincer et al. 2002). The aims of this study were to describe, compare, and contrast the gene content and organization of the two genomes in the context of prevailing species concepts, identify the functional attributes that differentiate the two species, assess the processes that have driven genome evolution, and search for evidence of marine adaptation in this unusual group of Gram-positive marine bacteria.

Methods

Sequencing and ortholog identification

The sequencing and annotation of the SA genome was as previously reported for ST (Udwary et al. 2007). Both genomes were sequenced as part of the Department of Energy, Joint Genome Institute, Community Sequencing Program. Genome sequences have been deposited in GenBank under accession numbers CP000667 (*S.*

tropica) and CP000850 (*S. arenicola*). Orthologs within the two genomes were predicted using the Reciprocal Smallest Distance (RSD) method (Wall et al. 2003), which includes a maximum likelihood estimate of amino acid substitutions. A linear alignment of positional orthologs was created and the positions of rearranged orthologs and species-specific genes identified. Genomic islands were defined as regions >20 kb that are flanked by regions of conservation and within which <40% of the island genes possess a positional ortholog in the reciprocal genome. Paralogs within each genome were identified using the blastclust algorithm (Dondoshansky and Wolf 2000) with a cut-off of 30% identity over 40% of the sequence length. The automated phylogenetic inference system (APIS) was used to identify recent gene duplications (Badger et al. 2005).

Horizontal Gene Transfer

All genes were assessed for evidence of HGT based on abnormal DNA composition, phylogenetic, taxonomic, and sequence-based relationships, and comparisons to known Mobile Genetic Elements (MGEs). Genes identified by ≥2 different methodologies were counted as positive for HGT. To reflect confidence in the assignments, genes displaying positive evidence of HGT were color coded from yellow to red corresponding to total scores from 2 to 6. The results were mapped onto the genome to reveal HGT clustering patterns and adjacent clusters were merged (Figure 2.1a). Four DNA compositional analyses included G+C content (obtained from the JGI annotation), codon adaptive index, calculated with the CAI calculator (Wu et al. 2005) using a suite of housekeeping genes as reference, dinucleotide frequency differences (δ*), calculated using IslandPath (Hsiao et al. 2003), and DNA composition, calculated using Alien_Hunter (Vernikos and Parkhill 2006). G+C content or codon usage values >1.5 standard deviations from the genomic mean and dinucleotide frequency differences >1 standard deviation from the mean were scored positive for HGT. Taxonomic relationships in the form of lineage probability index (LPI) values for all protein coding genes were assigned using the Darkhorse algorithm (Podell and Gaasterland 2007). Genes with an LPI of <0.5, indicating the orthologs are not in closely related genomes, were scored positive for HGT. A reciprocal Darkhorse analysis (Podell et al. 2008) was then performed on the orthologs of all positives, and if these genes had an LPI score >0.5, indicating the match sequence is phylogenetically typical within its own lineage, they were assigned an additional positive score.

A phylogenetic approach using the APIS program (Badger et al. 2005) was also employed to assess HGT. Using this program, bootstrapped neighbor-joining trees of all predicted protein coding genes within each genome were created. All genes cladding with non-Actinobacterial homologs were binned into their respective taxonomic groups and given a positive HGT score. Evidence of HGT was also inferred from RSD analyses of each genome against a compiled set of 27 finished Actinobacterial genomes that included at least two representatives of each genus for which sequences were available. Genes present in SA and/or ST and not observed among the 27 Actinobacterial genomes were assigned a positive HGT score. Bacteriophage were identified using Prophage (Bose and Barber 2006) and Phage

Finder (Fouts 2006). Other insertion elements were identified as prophage or transposon in origin through blastX homology searches. Gene annotation based on searches for identity across PFAM, SPTR, KEGG and COG databases was also used to help identify mobile genetic elements (MGEs). Each gene associated with an MGE was assigned a positive HGT score. Test scores were amalgamated and those genes showing evidence of HGT in two or more tests (maximum score 6) were classified as horizontally acquired. The results were mapped onto the genome and genes identified by only one test but associated with clusters of genes that scored in two or more tests were added to the total HGT pool. Adjacent clusters were merged.

CRISPRs were identified using CRISPR finder (http://crispr.u-psud.fr/Server/CRISPRfinder.php) while repeats larger than 35 bases were identified using Reputer (Kurtz et al. 2001). Secondary metabolite gene clusters were manually annotated as in (Udwary et al. 2007). Cluster boundaries were predicted using previously reported gene clusters when available as in the case of rifamycin. For unknown clusters, loss of gene conservation across the Actinobacteria was used to aid boundary predictions. In the future, programs such as “ClustScan” may prove useful for pathway annotation and product prediction (Starcevic et al. 2008). However, many biosynthetic genes are large (5-10 kb) and highly repetitive creating challenges associated with gene calling and assembly, eg., (Udwary et al. 2007) and the interpretation of operon structure. The ratio of non-synonymous to synonymous mutations (dN/dS) for all orthologs was calculated using the perl progam SNAP (http://www.hiv.lanl.gov) with the alignments for all values >1 checked manually.

Results and Discussion

The ST and SA genomes share 3606 orthologs, representing 79.4% and 73.2% of the respective genomes (Table 2.1). The average nucleotide identity among these orthologs is 87.2%, well below the 94% cut-off that has been suggested to delineate bacterial species (Konstantinidis and Tiedje 2005). Despite differing by only seven nucleotides (99.7% identity) in the 16S rRNA gene, the genome of SA is 603 kb (11.6%) larger and possesses 1505 species-specific genes compared to 987 in ST. Seventy-five percent of these species-specific genes are located in 21 genomic islands (Tables 2.1, 2.2), none of which are comprised of genes originating entirely from one genome (Figure 2.1). The presence of genomic islands in the same location on the chromosomes of closely related bacteria is well recognized (Coleman et al. 2006) and facilitated by the presence of tRNAs (Tuanyok et al. 2008). Twelve islands in the *Salinispora* alignment share at least one tRNA between both genomes and of those, four share two or more tRNAs within a single island indicating multiple insertion sites. In addition to tRNAs, direct repeats detected in the same location in both genomes could also act as insertion sites to help create islands. These islands are enriched with large clusters of genes devoted to the biosynthesis of secondary metabolites (Figure 2.1). They house all 25 of the species-specific secondary metabolic pathways, while eight of the 12 shared pathways occur in the genus-specific core (Tables 2.3, 2.4). We have isolated and identified the products of eight of these pathways, which include the highly selective proteasome inhibitor salinosporamide A

(Feling et al. 2003) as well as sporolide A (Buchanan et al. 2005), which is derived from an enediyne polyketide precursor (Udwary et al. 2007), one of the most potent classes of biologically active agents discovered to date. A previous analysis of 46 *Salinispora* strains revealed that secondary metabolite production is the major phenotypic difference among the three species (Jensen et al. 2007), an observation supported by the analysis of the *S. tropica* genome (Udwary et al. 2007).

Of the eight secondary metabolites that have been isolated from the two strains, all but salinosporamide A, sporolide A, and salinilactam have been reported from unrelated taxa (Figure 2.1), providing strong evidence of HGT. Further evidence for HGT comes from a phylogenetic analysis of the polyketide synthase (PKS) genes associated with the rifamycin biosynthetic gene cluster (*rif*) in SA and *Amycolatopsis mediterranei*, the original source of this antibiotic (Yu et al. 1999). This analysis confirms prior observations of HGT in this pathway (Kim et al. 2006) and reveals that all 10 of the ketosynthase domains are perfectly interleaved, as would be predicted if the entire PKS gene cluster had been exchanged between the two strains (Figure 2.2). Evidence of HGT coupled with prior evidence for the fixation of specific pathways such as *rif* among globally distributed SA populations (Jensen et al. 2007) supports vertical inheritance following pathway acquisition (Ochman et al. 2005). This evolutionary history is what might be expected if pathway acquisition fostered ecotype diversification or a selective sweep (Cohan 2002) resulting from strong selection for the acquired pathway, either of which provide compelling evidence that secondary metabolites represent functional traits with important ecological roles. The concept that gene acquisition provides a mechanism for ecological diversification that may ultimately drive the formation of independent bacterial lineages has been previously proposed (Ochman et al. 2000). The inclusion of secondary metabolism among the functional categories of acquired genes that may have this effect sheds new light on the functional importance and evolutionary significance of this class of genes. Although the ecological functions of secondary metabolites remain largely unknown, and thus it is not clear how these molecules might facilitate ecological diversification, there is mounting evidence that they play important roles in chemical defense (Haeder et al. 2009) or as signaling molecules involved in population or community communication (Yim et al. 2007).

Differences between the two species also occur in CRISPR sequences, which are non-continuous direct repeats separated by variable (spacer) sequences that have been shown to confer immunity to phage (Barrangou et al. 2007). The ST genome carries three intact prophage and three CRISPRs (35 spacers), while only one prophage has been identified in the genome of SA, which possesses eight different CRISPRs (140 spacers). The SA prophage is unprecedented among bacterial genomes in that it occurs in two adjacent copies that share 100% sequence identity. These copies are flanked by tRNA *att* sites and separated by an identical 45 bp *att* site, suggesting double integration as opposed to duplication (te Poele et al. 2008). Remarkably, four of the SA CRISPRs possess a spacer that shares 100% identity with portions of three different genes found in ST prophage 1 (Figure 2.3). These spacer sequences have no similar matches to genes in the SA prophage or in any prophage sequences deposited in the NCBI, CAMERA, or the SDSU Center for Universal Microbial Sequencing databases. The detection of these spacer sequences provides evidence that SA has been exposed to a phage related to one that currently infects ST and that SA now maintains acquired immunity to this phage genotype as has been previously reported in other bacteria (Barrangou et al. 2007). This is a rare example in which evidence has been obtained for CRISPR-mediated acquired immunity to a prophage that resides in the genome of a closely related environmental bacterium. Given that SA strain CNS-205 was isolated from Palau while ST strain CNB-440 was recovered 15 years earlier from the Bahamas, it appears that actinophage have broad temporal-spatial distributions or that resistance is maintained on temporal scales sufficient for the global distribution of a bacterial species.

Enhanced phage immunity, as evidenced by 140 relative to 35 CRISPR spacer sequences, coupled with a larger genome size and a greater number of species-specific secondary metabolic pathways may account for the cosmopolitan distribution of SA relative to ST, which to date has only been recovered from the Caribbean (Jensen and Mafnas 2006). Also included among the SA-specific gene pool is a complete phospho-transferase system (PTS, Sare4844-4850). PTSs are centrally involved in carbon source uptake and regulation (Parche et al. 2000) and may provide growth advantages that also factor into the relatively broad distribution of SA. However, additional strains will need to be studied before any of these differences can be firmly linked to species distributions.

The 21 genomic islands are not contiguous regions of species-specific DNA but were instead created by a complex process of gene acquisition, loss, duplication, and inactivation (Figure 2.4). The overall composition, evolutionary history, and function of the island genes are similar in both strains, with duplication and HGT accounting for the majority of genes and secondary metabolism representing the largest functionally annotated category. Remarkably, 42% of the rearranged island orthologs fall within other islands indicating that inter-island movement or "island hopping" is common, thus providing support for the hypothesis that islands undergo continual rearrangement (Coleman et al. 2006). There is dramatic, operon-scale evidence of this process in the shared yersiniabactin pathways (ST *sid2* and SA *sid1*), which occur in islands 15 and 10, respectively, and in the unknown dipeptide pathways (ST *nrps1* and SA *nrps3*), which occur in islands 4 and 15, respectively. In both cases, these pathways remain intact yet are located in different islands in the two strains (Figure 2.1, Table 2.3, 2.4). There is also evidence of cluster fragmentation in the 10-membered enediyne gene set SA *pks3*, which contains the core set of genes associated with calicheamicin biosynthesis (Figure 2.5) (Ahlert et al. 2002), yet is split by the introduction of 145 kb of DNA from three different biosynthetic loci (island 10, Figure 2.1). The conserved fragments appear to encode the biosynthesis of a calicheamicin anolog, while flanking genes display a high level of gene duplication and rearrangement indicative of active pathway evolution. Cluster fragmentation is also observed in the 9-membered enediyne PKS cluster SA *pks1* (A-C), which is scattered across the genome in islands 4, 10, and 21 (Figure 2.1, Table 2.4).

The genomic islands are also enriched in mobile genetic elements including prophage, integrases, and actinobacterial integrative and conjugative elements (AICEs) (Burrus et al. 2002) (Tables 2.5, 2.6), the later of which are known to play a role in gene acquisition and rearrangement. The *Salinispora* AICEs possess *traB* homologs, which promote conjugal plasmid transfer in mycelial streptomycetes (Reuther et al. 2006), suggesting that hyphal tip fusion is a prominent mechanism driving gene exchange in these bacteria. AICEs have been linked to the acquisition of secondary metabolite gene clusters (te Poele et al. 2007) and their occurrence in island 7 (SA AICE1), which includes the entire 90 kb *rif* cluster, and island 10 (SA AICE3), which contains biosynthetic gene clusters for enediyne, siderophore, and amino acid-derived secondary metabolites, provides a mechanism for the acquisition of these pathways (Figure 2.1). Six additional secondary metabolite gene clusters (ST *nrps1*, ST *spo*, SA *nrps3*, SA *pks5*, SA *cym*, and SA *pks2*) are flanked by direct repeats, providing further support for HGT. In the case of *cym* (Schultz et al. 2008), which is clearly inserted into a tRNA, the pseudogenes preceding and following it are all related to transposases or integrases providing a mechanism for chromosomal integration.

Despite exhaustive analyses of HGT, only 22% of the 127 genes in the five biosynthetic pathways (*rif, sta, des, lym, cym*) whose products have also been observed in other bacteria (Figure 2.1, Table 2.4) scored positive for HGT. This observation suggests that the pathways either originated in *Salinispora* or that the exchange of these biosynthetic genes has occurred largely among closely related bacteria and therefore gone undetected with the HGT methods applied in this study. The latter scenario is supported by the observation that all five of the shared biosynthetic pathways were previously reported in other actinomycetes. The acquisition of genes from closely related bacteria likely accounts for many of the species-specific island genes for which no evidence of evolutionary history could be determined (Figure 2.4b). These genes were poorly conserved among 27 Actinobacterial genomes (Figure 2.4d) providing additional support that they were acquired, most likely from environmental Actinobacteria that are not well represented among sequenced genomes. Although gene loss was not quantified, this process is also a likely contributor to island formation. In support of an adaptive role for island genes, 7.6% (44/573) of the orthologs show evidence of positive selection (dN/dS >1) compared to 1.6% (49/3027) of the non-island pairs. Given that the majority of island genes display evidence of HGT, the increased dN/dS ratio is in agreement with the observation that acquired genes experience relaxed functional constraints (Hao and Golding, 2006).

Functional differences between related organisms can be obscured when orthologs are taken out of the context of the gene clusters in which they reside. For example, the PKS genes Sare1250 and Stro2768 are orthologous and likely perform similar functions, yet they reside in the *rif* and *slm* pathways, respectively, and thus contribute to the biosynthesis of dramatically different secondary metabolites. Likewise, intra-cluster PKS gene duplication (Sare3151 and Sare3152, Figure 2.1) has an immediate effect on the product of the pathway by the introduction of an additional acyl group into the carbon skeleton of the macrolide, as opposed to the more traditional concept of parology facilitating mutation-driven functional divergence (Prince and Pickett 2002). Sub-genic, modular duplications are also observed (Sare3156 modules 4 and 5, Figure 2.1), which likewise have an immediate effect on the structure of the secondary metabolite produced by the pathway. While HGT is considered a rapid method for ecological adaptation in bacteria (Ochman et al. 2000), PKS gene duplication provides a complementary evolutionary strategy (Fischbach et al. 2008) that could lead to the rapid production of new secondary metabolites that subsequently drive the creation of new adaptive radiations.

Salinispora species are the first marine Actinobacteria reported to require seawater for growth (Maldonado et al. 2005). Unlike Gram-negative marine bacteria, in which seawater requirements are linked to a specific sodium ion requirement (Kogure 1998), *Salinispora* strains are capable of growth in osmotically adjusted, sodium-free media (Tsueng and Lam 2008). An analysis of the *Salinispora* core for evidence of genes associated with this unusual osmotic requirement reveals a highly duplicated family of 29 polymorphic membrane proteins (PMPs) that include homologs associated with polymorphic outer membrane proteins (POMPs). POMPs remain functionally uncharacterized however there is strong evidence that they are type V secretory systems (Henderson and Lam 2001), making this the first report of type V autotransporters outside of the Proteobacteria (Henderson et al. 2004). Phylogenetic analyses provide evidence that the *Salinispora* PMPs were acquired from aquatic, Gram-negative bacteria and that they have continued to undergo considerable duplication subsequent to divergence of the two species (Figure 2.6). The occurrence of this large family of PMP autotransporters in marine Actinobacteria may represent a low nutrient adaptation that renders cells susceptible to lysis in low osmotic environments.

Conclusions

In conclusion, the comparative analysis of two closely related marine Actinobacterial genomes provides new insight into the functional traits associated with genomic islands. It has been possible to assign precise, physiological functions to island genes and link differences in secondary metabolism to fine-scale phylogenetic architecture in two distinct bacterial lineages, which by all available metrics maintain the fundamental characteristics of species-level units of diversity. It is clear that gene clusters devoted to secondary metabolite biosynthesis are dynamic entities that are readily acquired, rearranged, and fragmented in the context of genomic islands, and that the results of these processes create natural product diversity that can have an immediate effect on fitness or niche utilization. The high level of species specificity associated with secondary metabolism suggests that this functional trait may represent a previously unrecognized force driving ecological diversification among closely related, sediment inhabiting bacteria.

Acknowldegements

Chapter 2, in full, is published in the Journal of the International Society for Microbial Ecology, 2009, Kevin Penn, Caroline Jenkins, Markus Nett, Daniel W. Udwary, Erin A. Gontang, Ryan P. McGlinchey, Brian Foster, Alla Lapidus, Sheila Podell, Eric E. Allen, Bradley S. Moore, and Paul R. Jensen. The dissertation author was the primary investigator and author of the research, which forms the basis for this chapter.

References

Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. Nat Rev Micro 6(6): 431-440.

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF (2004) Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430(6999): 551-554.

Ahlert J, Shepard E, Lomovskaya N, Zazopoulos E, Staffa A, Bachmann BO, Huang K, Fonstein L, Czisny A, Whitwam RE, Farnet CM, Thorson JS (2002) The Calicheamicin Gene Cluster and Its Iterative Type I Enediyne PKS. Science 297(5584): 1173-1176.

Badger JH, Eisen JA, Ward NL (2005) Genomic analysis of *Hyphomonas neptunium* contradicts 16S rRNA gene-based phylogenetic analysis: implications for the taxonomy of the orders '*Rhodobacterales*' and *Caulobacterales*. International Journal of Systematic and Evolutionary Microbiology 55(3): 1021-1026.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P (2007) CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science 315(5819): 1709-1712.

Berdy J (2005) Bioactive Microbial Metabolites. J Antibiot 58(1): 1-26.

Bose M, Barber RD (2006) Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. In Silico Biology 6: 223-227.

Buchanan GO, Williams PG, Feling RH, Kauffman CA, Jensen PR, Fenical W (2005) Sporolides A and B: Structurally Unprecedented Halogenated Macrolides from the Marine Actinomycete *Salinispora tropica*. Organic Letters 7(13): 2731-2734.

Burrus V, Pavlovic G, Decaris B, GuÈdon G (2002) Conjugative transposons: the tip of the iceberg. Molecular Microbiology 46(3): 601-610.

Cohan FM (2002) What are bacterial species? Annual Review of Microbiology 56: 457-487.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, DeLong EF, Chisholm SW (2006) Genomic Islands and the Ecology and Evolution of *Prochlorococcus*. Science 311(5768): 1768-1770.

Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic and environmental microorganisms. Nat Rev Micro 2(5): 414-424.

Dondoshansky I, Wolf Y (2000) BLASTCLUST. 2.2 ed: National Institutes of Health, Bethesda, MD.

Feling RH, Buchanan GO, Mincer TJ, Kauffman CA, Jensen PR, Fenical W (2003) Salinosporamide A: A Highly Cytotoxic Proteasome Inhibitor from a Novel Microbial Source, a Marine Bacterium of the New Genus *Salinospora*. Angewandte Chemie 115(3): 369-371.

Fenical W, Jensen PR (2006) Developing a new resource for drug discovery: marine actinomycete bacteria. Nature Chemical Biology 2(12): 666-673.

Fenical W, Jensen PR, Palladino MA, Lam KS, Lloyd GK, Potts BC (2009) Discovery and development of the anticancer agent salinosporamide A (NPI-0052). Bioorganic &amp; Medicinal Chemistry 17(6): 2175-2180.

Fischbach MA, Walsh CT, Clardy J (2008) The evolution of gene collectives: How natural selection drives chemical innovation. Proceedings of the National Academy of Sciences 105(12): 4601-4608.

Fouts DE (2006) Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Research: gkl732.

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP (2009) The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity. Science 323(5915): 741-746.

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL, Swings J (2005) Opinion: re-evaluating prokaryotic species. Nat Rev Microbiol 3: 733 - 739.

Haeder S, Wirth R, Herz H, Spiteller D (2009) Candicidin-producing Streptomyces support leaf-cutting ants to protect their fungus garden against the pathogenic fungus Escovopsis. Proceedings of the National Academy of Sciences 106(12): 4742-4746.

Henderson IR, Lam AC (2001) Polymorphic proteins of *Chlamydia* spp. - autotransporters beyond the Proteobacteria. Trends in Microbiology 9(12): 573-578.

Henderson IR, Navarro-Garcia F, Desvaux M, Fernandez RC, Ala'Aldeen D (2004) Type V protein secretion pathways: the autotransporter story. Microbiology and Molecular Biology Reviews 68(4): 692-744.

Hsiao W, Wan I, Jones SJ, Brinkman FSL (2003) IslandPath: aiding detection of genomic islands in prokaryotes. Bioinformatics 19(3): 418-420.

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF (2008) Resource Partitioning and Sympatric Differentiation Among Closely Related Bacterioplankton. Science 320(5879): 1081-1085.

Jensen PR, Mafnas C (2006) Biogeography of the marine actinomycete *Salinispora*. Environmental Microbiology 8(11): 1881-1888.

Jensen PR, Williams PG, Oh D-C, Zeigler L, Fenical W (2007) Species-Specific Secondary Metabolite Production in Marine Actinomycetes of the Genus *Salinispora*. Applied and Environmental Microbiology 73(4): 1146-1152.

Joyce EA, Chan K, Salama NR, Falkow S (2002) Redefining bacterial populations: a post-genomic reformation. Nat Rev Genet 3(6): 462-473.

Kim TK, Hewavitharana AK, Shaw PN, Fuerst JA (2006) Discovery of a New Source of Rifamycin Antibiotics in Marine Sponge Actinobacteria by Phylogenetic Prediction. Applied and Environmental Microbiology 72(3): 2118-2125.

Kogure K (1998) Bioenergetics of marine bacteria. Current Opinion in Biotechnology 9: 278-282.

Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences 102(7): 2567-2572.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Research 29(22): 4633-4642.

Maldonado LA, Fenical W, Jensen PR, Kauffman CA, Mincer TJ, Ward AC, Bull AT, Goodfellow M (2005) *Salinispora arenicola* gen. nov., sp. nov. and *Salinispora tropica* sp. nov., obligate marine actinomycetes belonging to the family Micromonosporaceae. International Journal of Systematic and Evolutionary Microbiology 55(5): 1759-1766.

Mincer TJ, Jensen PR, Kauffman CA, Fenical W (2002) Widespread and Persistent Populations of a Major New Marine Actinomycete Taxon in Ocean Sediments. Applied and Environmental Microbiology 68(10): 5005-5011.

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299-304.

Ochman H, Lerat E, Daublin V (2005) Examining bacterial species under the specter of gene transfer and exchange. Proceedings of the National Academy of Sciences 102: 6595-6599.

Parche S, Nothaft H, Kamionka A, Titgemeyer F (2000) Sugar uptake and utilisation in Streptomyces coelicolor: a PTS view to the genome. Antonie van Leeuwenhoek 78(3): 243-251.

Podell S, Gaasterland T (2007) DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome Biology 8(2): R16.

Podell S, Gaasterland T, Allen E (2008) A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. BMC Bioinformatics 9(1): 419.

Prince VE, Pickett FB (2002) Splitting pairs: the diverging fates of duplicated genes. Nat Rev Gen 3: 827-837.

Reuther J, Gekeler C, Tiffert Y, Wohlleben W, Muth G (2006) Unique conjugative mechanism in mycelial streptomycetes: a DNA-binding ATPase translocates unprocessed plasmid DNA at the hyphal tip. Molecular Microbiology 61(2): 436-446.

Schultz AW, Oh D-C, Carney JR, Williamson RT, Udwary DW, Jensen PR, Gould SJ, Fenical W, Moore BS (2008) Biosynthesis and structures of cyclomarins and cyclomarazines, prenylated cyclic peptides of marine actinobacterial origin. Journal of the American Chemical Society 130(13): 4507-4516.

Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Research 36(21): 6882-6892.

te Poele E, Bolhuis H, Dijkhuizen L (2008) Actinomycete integrative and conjugative elements. Antonie van Leeuwenhoek 94(1): 127-143.

te Poele EM, Habets MN, Tan GYA, Ward AC, Goodfellow M, Bolhuis H, Dijkhuizen L (2007) Prevalence and distribution of nucleotide sequences typical for pMEA-like accessory genetic elements in the genus *Amycolatopsis*. FEMS Microbiology Ecology 61(2): 285-294.

Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, Sarma-Rupavtarm R, Distel DL, Polz MF (2005) Genotypic Diversity Within a Natural Coastal Bacterioplankton Population. Science 307(5713): 1311-1313.

Tsueng G, Lam K (2008) A low-sodium-salt formulation for the fermentation of salinosporamides by *Salinispora tropica* strain NPS21184. Applied Microbiology and Biotechnology 78(5): 821-826.

Tuanyok A, Leadem B, Auerbach R, Beckstrom-Sternberg S, Beckstrom-Sternberg J, Mayo M, Wuthiekanun V, Brettin T, Nierman W, Peacock S, Currie B, Wagner D, Keim P (2008) Genomic islands from five strains of *Burkholderia pseudomallei*. BMC Genomics 9(1): 566.

Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W, Jensen PR, Moore BS (2007) Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. Proceedings of the National Academy of Sciences 104(25): 10376-10381.

Vernikos GS, Parkhill J (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics 22(18): 2196-2203.

Wall DP, Fraser HB, Hirsh AE (2003) Detecting putative orthologs. Bioinformatics 19(13): 1710-1711.

Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HLT, Donnenberg MS, Blattner FR (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proceedings of the National Academy of Sciences 99(26): 17020-17024.

Wu G, Culley DE, Zhang W (2005) Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. Microbiology 151(7): 2175-2187.

Yim G, Huimi Wang H, Davies Frs J (2007) Antibiotics as signalling molecules. Philosophical Transactions of the Royal Society B: Biological Sciences 362(1483): 1195-1200.

Yu T-W, Shen Y, Doi-Katayama Y, Tang L, Park C, Moore BS, Hutchinson CR, Floss HG (1999) Direct evidence that rifamycin polyketide synthase assembles polyketide chains processively. Proceedings of the National Academy of Sciences 96(16): 9051-9056.

Figures

Figure 2.1: Linear alignment of the *S. tropica* and *S. arenicola* genomes starting with the origins of replication. (**a**) Positional orthologs (core) flanked by islands (E, F), heat-mapped HGT genes (D, G), rearranged orthologs (C, H), species-specific genes (B, I), secondary metabolite genes (green), MGEs (pink) with prophage (P) and AICES (E) indicated (A, J). For genomic islands, predicted (lower case) and isolated (uppercase with structures) secondary metabolites are given (not shown are six non-island secondary metabolic gene clusters of unknown function). Shared positional (blue) and rearranged (red) secondary metabolite clusters are indicated. *Previously isolated from other bacteria. (**b**) Expanded view of SA *pks5* revealing gene and modular architecture. (**c**) Neighbor-joining phylogenetic tree of KS domains from SA *pks5* revealing gene and modular duplication events (erythromycin root, % bootstrap values from 1000 re-samplings).

Figure 1

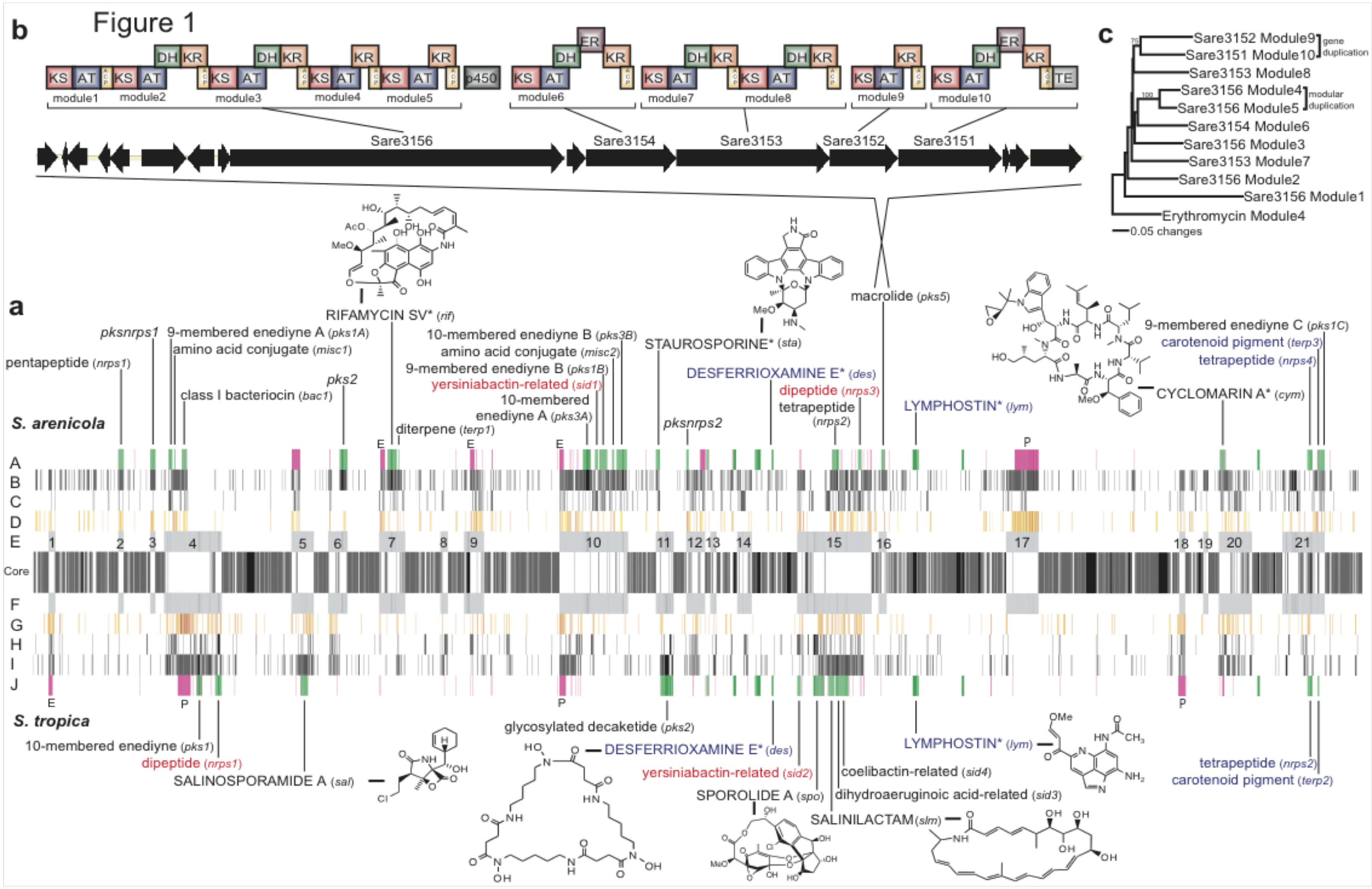
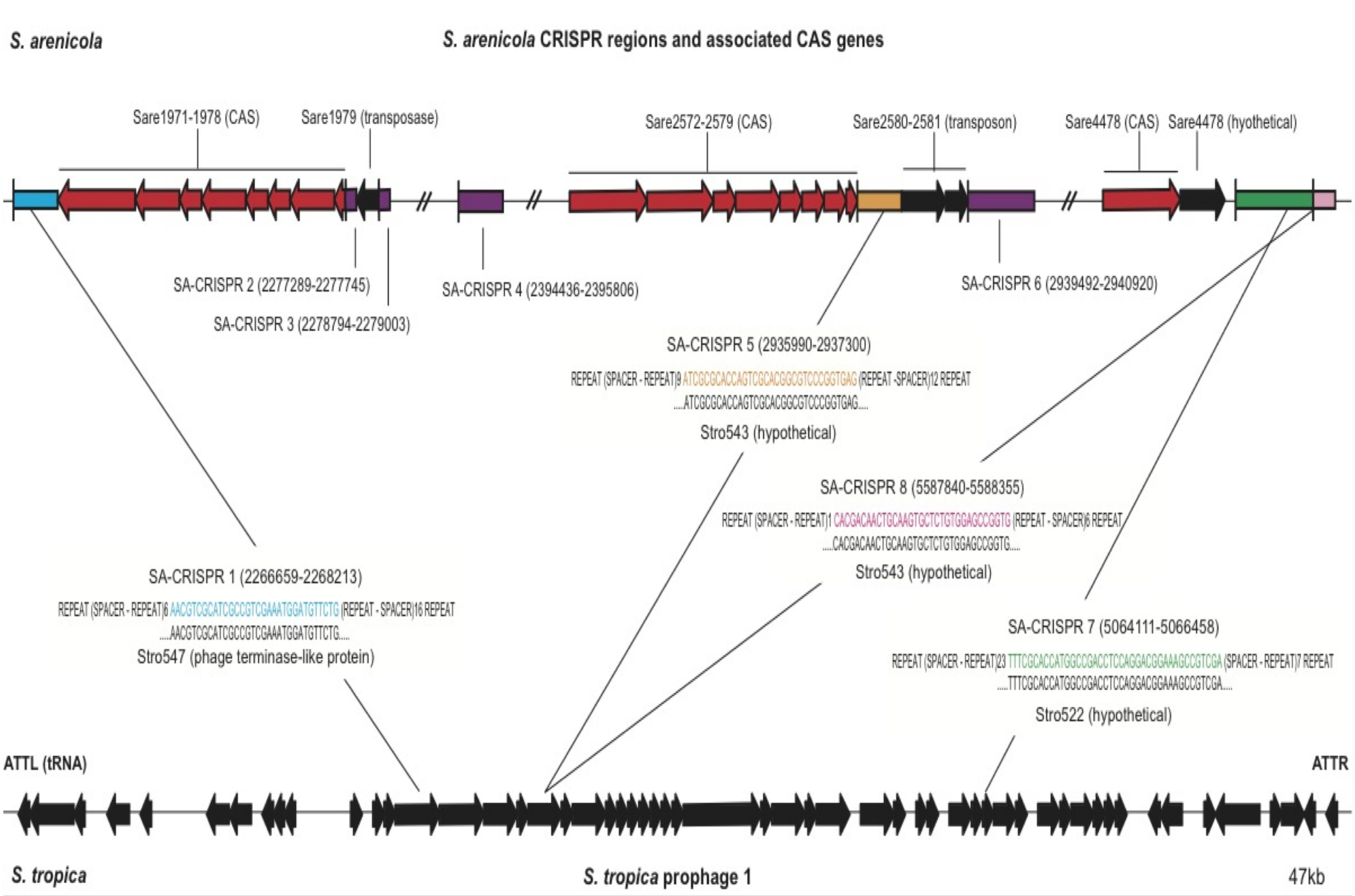

b
KS
AT
ACP
DH
KR
ER
TE
p450
module1
module2
module3
module4
module5
module6
module7
module8
module9
module10
Sare3156
Sare3154
Sare3153
Sare3152
Sare3151
c
Sare3152 Module9
Sare3151 Module10
gene duplication
Sare3153 Module8
Sare3156 Module4
Sare3156 Module5
modular duplication
Sare3154 Module6
Sare3156 Module3
Sare3153 Module7
Sare3156 Module2
Sare3156 Module1
Erythromycin Module4
0.05 changes
a
S. arenicola
pentapeptide (nrps1)
pksnrps1
9-membered enediyne A (pks1A)
amino acid conjugate (misc1)
class I bacteriocin (bac1)
pks2
RIFAMYCIN SV* (rif)
diterpene (terp1)
10-membered enediyne B (pks3B)
amino acid conjugate (misc2)
9-membered enediyne B (pks1B)
yersiniabactin-related (sid1)
10-membered enediyne A (pks3A)
STAUROSPORINE* (sta)
pksnrps2
DESFERRIOXAMINE E* (des)
dipeptide (nrps3)
tetrapeptide (nrps2)
macrolide (pks5)
LYMPHOSTIN* (lym)
9-membered enediyne C (pks1C)
carotenoid pigment (terp3)
tetrapeptide (nrps4)
CYCLOMARIN A* (cym)
A
B
C
D
E
Core
F
G
H
I
J
E
P
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
S. tropica
10-membered enediyne (pks1)
dipeptide (nrps1)
SALINOSPORAMIDE A (sal)
glycosylated decaketide (pks2)
DESFERRIOXAMINE E* (des)
yersiniabactin-related (sid2)
SPOROLIDE A (spo)
coelibactin-related (sid4)
dihydroaeruginoic acid-related (sid3)
SALINILACTAM (slm)
LYMPHOSTIN* (lym)
tetrapeptide (nrps2)
carotenoid pigment (terp2)

Figure 2.2: *S. tropica* prophage and *S. arenicola* CRISPRs. Four of 8 SA CRISPRs (1, 5, 7, 8) have spacers (color coded) that share 100% sequence identity with genes (Stro numbers and annotation given) in ST prophage 1 (Table S2, inverted for visual purposes). Other CRISPRs are colored purple. SA CRISPRs 2-3 and 5-6 share the same direct repeats and may have at one time been a single allele. CRISPR associated (CAS) genes (red) and genes interrupting CRISPRs (black) are indicated. None of the spacer sequences possessed 100% identity to prophage in the NCBI non-redundant sequence database, the SDSU Center for Universal Microbial Sequencing database, or the CAMERA metagenomic database.


S. arenicola
S. arenicola CRISPR regions and associated CAS genes
Sare1971-1978 (CAS)
Sare1979 (transposase)
Sare2572-2579 (CAS)
Sare2580-2581 (transposon)
Sare4478 (CAS)
Sare4478 (hyothetical)
SA-CRISPR 2 (2277289-2277745)
SA-CRISPR 3 (2278794-2279003)
SA-CRISPR 4 (2394436-2395806)
SA-CRISPR 6 (2939492-2940920)
SA-CRISPR 5 (2935990-2937300)
REPEAT (SPACER - REPEAT)9 ATCGCGCACCAGTCGCACGGCGTCCCGGTGAG (REPEAT -SPACER)12 REPEAT
.....ATCGCGCACCAGTCGCACGGCGTCCCGGTGAG.....
Stro543 (hypothetical)
SA-CRISPR 8 (5587840-5588355)
REPEAT (SPACER - REPEAT)1 CACGACAACTGCAAGTGCTCTGTGGAGCCGGTG (REPEAT - SPACER)6 REPEAT
.....CACGACAACTGCAAGTGCTCTGTGGAGCCGGTG.....
Stro543 (hypothetical)
SA-CRISPR 1 (2266659-2268213)
REPEAT (SPACER - REPEAT)6 AACGTCGCATCGCCGTCGAAATGGATGTTCTG (REPEAT - SPACER)16 REPEAT
.....AACGTCGCATCGCCGTCGAAATGGATGTTCTG.....
Stro547 (phage terminase-like protein)
SA-CRISPR 7 (5064111-5066458)
REPEAT (SPACER - REPEAT)23 TTTCGCACCATGGCCGACCTCCAGGACGGAAAGCCGTCGA (SPACER - REPEAT)7 REPEAT
.....TTTCGCACCATGGCCGACCTCCAGGACGGAAAGCCGTCGA.....
Stro522 (hypothetical)
ATTL (tRNA)
ATTR
S. tropica
S. tropica prophage 1
47kb

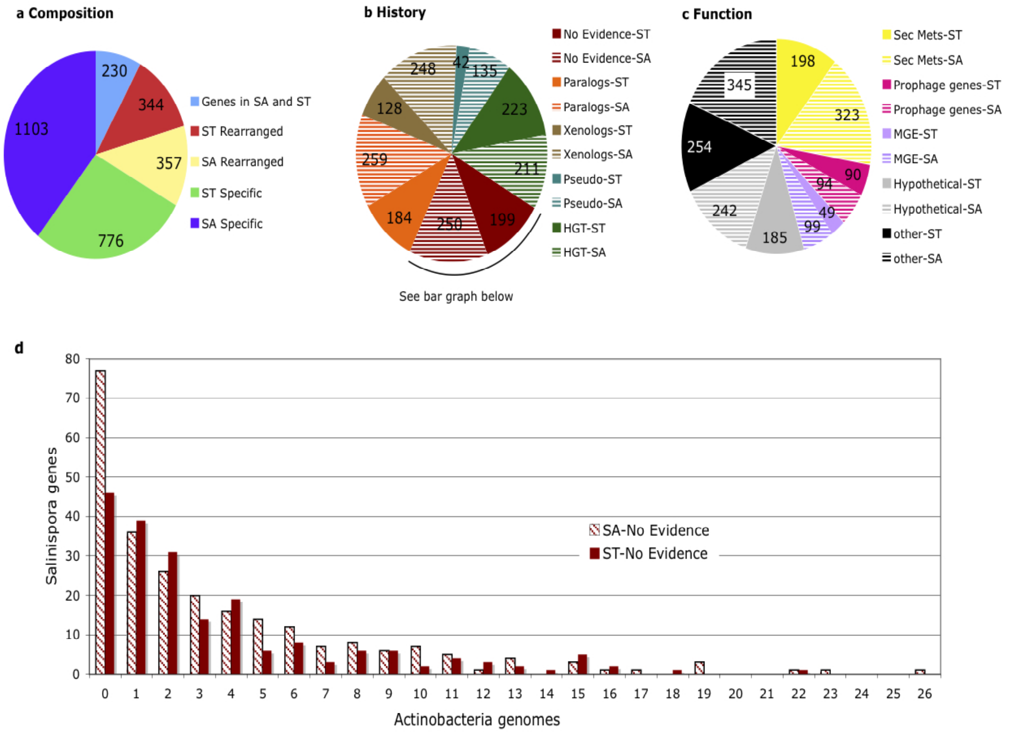


Figure 2.3: Composition, evolutionary history, and function of island genes in *S. tropica* (ST) and *S. arenicola* (SA). (**a**) 3040 genes comprising 21 genomic islands were analyzed for positional orthology (ie., the gene is part of the shared "core" genome), re-arranged orthology (ie., the gene is present in the other genome but not in the same position or island), and species-specificity (gene totals presented in wedges). (**b**) The ST and SA species-specific island genes were analyzed for evidence of parology, xenology, and HGT. Pseudogenes and the number of genes with no evidence for any of these processes were also identified. (**c)** Functional annotation of the species-specific island genes. (**d**) Distribution of species-specific island genes that have no evidence for HGT or parology among 27 Actinobacterial genomes.

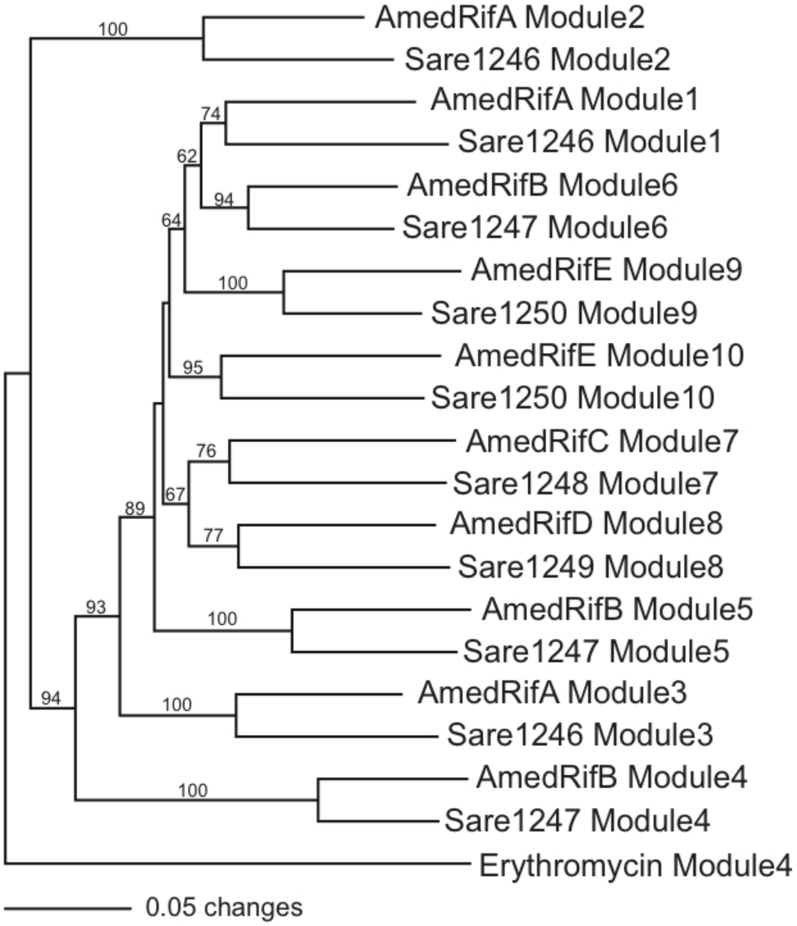


Figure 2.4: Polyketide synthase phylogeny. Neighbor-joining distance tree constructed using the aligned amino acid sequences of the *rif* KS domains from *A. mediterranei* and *S. arenicola*. Bootstrap values (in percent) calculated from 1000 re-samplings are shown at their respective nodes for values greater than or equal to 60%. The KS domain from module 4 of the erythromycin biosynthetic pathway (*Saccharopolyspora erythraea*) was used to position the root.

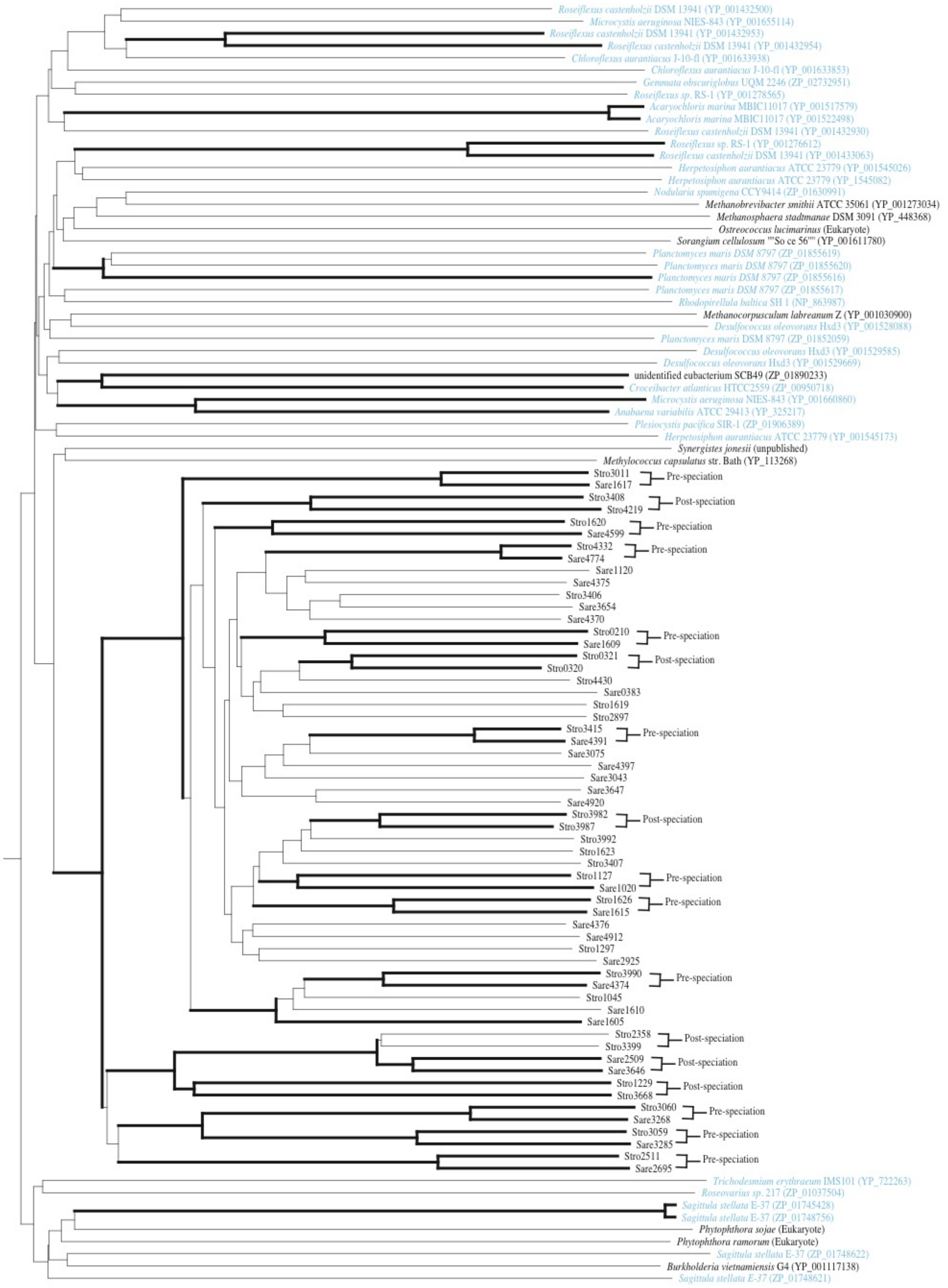

Roseiflexus castenholzii DSM 13941 (YP_001432500)
Microcystis aeruginosa NIES-843 (YP_001655114)
Roseiflexus castenholzii DSM 13941 (YP_001432953)
Roseiflexus castenholzii DSM 13941 (YP_001432954)
Chloroflexus aurantiacus J-10-fl (YP_001633938)
Chloroflexus aurantiacus J-10-fl (YP_001633853)
Gemmata obscuriglobus UQM 2246 (ZP_02732951)
Roseiflexus sp. RS-1 (YP_001278565)
Acaryochloris marina MBIC11017 (YP_001517579)
Acaryochloris marina MBIC11017 (YP_001522498)
Roseiflexus castenholzii DSM 13941 (YP_001432930)
Roseiflexus sp. RS-1 (YP_001276612)
Roseiflexus castenholzii DSM 13941 (YP_001433063)
Herpetosiphon aurantiacus ATCC 23779 (YP_001545026)
Herpetosiphon aurantiacus ATCC 23779 (YP_1545082)
Nodularia spumigena CCY9414 (ZP_01630991)
Methanobrevibacter smithii ATCC 35061 (YP_001273034)
Methanosphaera stadtmanae DSM 3091 (YP_448368)
Ostreococcus lucimarinus (Eukaryote)
Sorangium cellulosum ""So ce 56"" (YP_001611780)
Planctomyces maris DSM 8797 (ZP_01855619)
Planctomyces maris DSM 8797 (ZP_01855620)
Planctomyces maris DSM 8797 (ZP_01855616)
Planctomyces maris DSM 8797 (ZP_01855617)
Rhodopirellula baltica SH 1 (NP_863987)
Methanocorpusculum labreanum Z (YP_001030900)
Desulfococcus oleovorans Hxd3 (YP_001528088)
Planctomyces maris DSM 8797 (ZP_01852059)
Desulfococcus oleovorans Hxd3 (YP_001529585)
Desulfococcus oleovorans Hxd3 (YP_001529669)
unidentified eubacterium SCB49 (ZP_01890233)
Croceibacter atlanticus HTCC2559 (ZP_00950718)
Microcystis aeruginosa NIES-843 (YP_001660860)
Anabaena variabilis ATCC 29413 (YP_325217)
Plesiocystis pacifica SIR-1 (ZP_01906389)
Herpetosiphon aurantiacus ATCC 23779 (YP_001545173)
Synergistes jonesii (unpublished)
Methylococcus capsulatus str. Bath (YP_113268)
Stro3011
Sare1617
Pre-speciation
Stro3408
Stro4219
Post-speciation
Stro1620
Sare4599
Pre-speciation
Stro4332
Sare4774
Pre-speciation
Sare1120
Sare4375
Stro3406
Sare3654
Sare4370
Stro0210
Sare1609
Pre-speciation
Stro0321
Stro0320
Post-speciation
Stro4430
Sare0383
Stro1619
Stro2897
Stro3415
Sare4391
Pre-speciation
Sare3075
Sare4397
Sare3043
Sare3647
Sare4920
Stro3982
Stro3987
Post-speciation
Stro3992
Stro1623
Stro3407
Stro1127
Sare1020
Pre-speciation
Stro1626
Sare1615
Pre-speciation
Sare4376
Sare4912
Stro1297
Sare2925
Stro3990
Sare4374
Pre-speciation
Stro1045
Sare1610
Sare1605
Stro2358
Stro3399
Post-speciation
Sare2509
Sare3646
Post-speciation
Stro1229
Stro3668
Post-speciation
Stro3060
Sare3268
Pre-speciation
Stro3059
Sare3285
Pre-speciation
Stro2511
Sare2695
Pre-speciation
Trichodesmium erythraeum IMS101 (YP_722263)
Roseovarius sp. 217 (ZP_01037504)
Sagittula stellata E-37 (ZP_01745428)
Sagittula stellata E-37 (ZP_01748756)
Phytophthora sojae (Eukaryote)
Phytophthora ramorum (Eukaryote)
Sagittula stellata E-37 (ZP_01748622)
Burkholderia vietnamiensis G4 (YP_001117138)
Sagittula stellata E-37 (ZP_01748621)

Figure 2.5: Polymorphic Membrane Protein (PMP) phylogeny. Neighbor-joining distance tree constructed in APIS (J. Badger, unpublished) using the aligned amino acid sequences of SA and ST PMPs as well as those observed in other genomes. Bold lines indicate boot-strap values >50% and blue indicates strains other than SA and ST that were derived from aquatic environments. Accession numbers in parentheses.

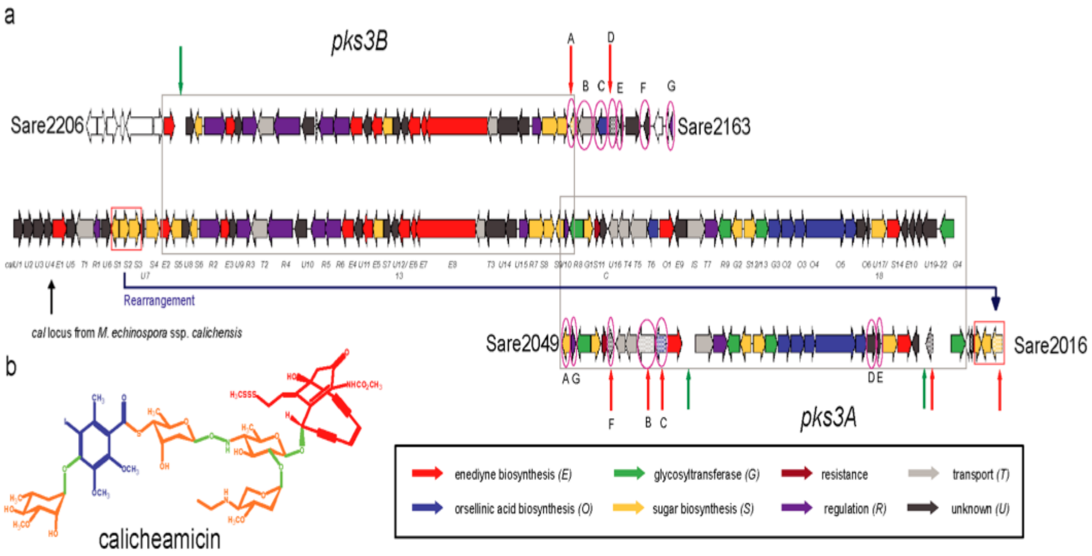

a
pks3B
A
D
B
C
E
F
G
Sare2206
Sare2163
calU1 U2 U3 U4 E1 U5 T1 R1 U6 S1 S2 S3 S4 E2 S5 U8 S6 R2 E3 U9 R3 T2 R4 U10 R5 R6 E4 U11 E5 S7 U12/13 E6 E7 E8 T3 U14 U15 R7 S8 S9/10 R8 G1 S11 U16 T4 T5 T6 O1 E9 IS T7 R9 G2 S12/13 G3 O2 O3 O4 O5 O6 U17/18 S14 E10 U19-22 G4
U7
C
Rearrangement
cal locus from M. echinospora ssp. calichensis
Sare2049
Sare2016
A G
D E
F
B C
pks3A
b
CH_3
O
H_3C
OCH_3
OH
HO
H_3CSSS
$NHCO_2CH_3$
calicheamicin
enediyne biosynthesis (E)
glycosyltransferase (G)
resistance
transport (T)
orsellinic acid biosynthesis (O)
sugar biosynthesis (S)
regulation (R)
unknown (U)

Figure 2.6: Cluster SA *pks3A* and *pks3B* in comparison with the *cal* locus from *M. echinospora*. **a** Grey boxes indicate regions of gene conservation. Duplicated genes are circled in red with parologs identified by letter. Red arrows indicate pseudogenes (which are also checkered). Genes missing (green arrows) and unique (colored white) relative to the *cal* locus are indicated. **b** structure of calicheamicin.

Tables

Table 2.1: General genome features.

Feature	*S. tropica* (ST)	ST%	*S. arenicola* (SA)	SA %
No. base pairs	5183331	NA	5786361	NA
% G+C	69.4	NA	69.5	NA
Total genes	4536	NA	4919	NA
Pseudogenes	57	1.26%	192	3.90%
Hypotheticals (% genome)	1140	25.10%	1418	28.80%
No. rRNA operons (% identity)	3	100%	3	100%
Orthologs (% genome)	3606	79.40%	3606	73.20%
Positional orthologs (% genome)	3178	70.10%	3178	64.60%
Rearranged orthologs (% genome)	428	9.40%	428	8.70%
Species-specific genes (% genome)	987	21.80%	1505	30.60%
Island genes (% genome)	1350	29.80%	1690	34.30%
Total genes with evidence of HGT (% genome)	652	14.30%	750	14.70%
Species-specific genes with evidence of HGT (% species-specific)	405	41.00%	573	38.10%
Total island genes with evidence of HGT (% HGT)	473	72.50%	555	74.00%
Paralogs[a] (% genome)	1819	39.60%	2179	42.60%
Species-specific paralogs (% species-specific genes)	391	39.70%	647	43.00%
Secondary metabolism (% genome)	405	8.80%	556	10.90%

[a]Totals include parental gene.

NA: not applicable.

Table 2.2: Genomic islands.

Island no. strain[a]	Start position	Stop position	Size (bp)	ST+SA total (bp)	Start gene	Stop gene	No. genes	Total genes	Island no. strain[a]	Start position	Stop position	Size (bp)	ST+SA total (bp)	Start gene	Stop gene	No. genes	Total genes
1 ST	67688	92154	24466		58	83	26		12 ST	2575986	2626461	50475		2282	2333	52	
1 SA	73610	95007	21397	45863	63	80	18	44	12 SA	2756098	2832944	76846	127321	2400	2476	77	129
2 ST	340915	355342	14427		300	304	5		13 ST	2650556	2667541	16985		2355	2373	19	
2 SA	381999	427379	45380	59807	345	367	23	28	13 SA	2856961	2871093	14132	31117	2500	2512	13	32
3 ST	471193	472396	1203		410	411	2		14 ST	2750480	2781381	30901		2445	2473	29	
3 SA	547253	570209	22956	24159	478	499	22	24	14 SA	2967508	3029499	61991	92892	2601	2656	56	85
4 ST	512154	781349	269195		449	694	246		15 ST	2968640	3325240	356600		2645	2909	265	
4 SA	608623	723155	114532	383727	537	641	105	351	15 SA	3227645	3533832	306187	662787	2842	3109	268	533
5 ST	1107318	1215323	108005		988	1068	81		16 ST	3357796	3368101	10305		2937	2946	10	
5 SA	1040020	1082195	42175	150180	924	958	35	116	16 SA	3568463	3657421	88958	99263	3143	3170	33	43
6 ST	1271151	1324135	52984		1127	1180	54		17 ST	3910860	3921251	10391		3407	3417	11	
6 SA	1139966	1202883	62917	115901	1018	1073	56	110	17 SA	4217838	4322435	104597	114988	3655	3794	140	151
7 ST	1477636	1495949	18313		1315	1357	43		18 ST	4543960	4565093	21133		3991	4016	26	
7 SA	1354284	1521916	167632	185945	1204	1314	111	154	18 SA	4942782	4969601	26819	47952	4375	4397	23	49
8 ST	1702965	1734332	31367		1492	1524	33		19 ST	4634866	4636953	2087		4077	4077	1	
8 SA	1685105	1694512	9407	40774	1457	1466	10	43	19 SA	5057490	5084903	27413	29500	4476	4497	22	23
9 ST	1803340	1855755	52415		1585	1631	47		20 ST	4688038	4738420	50382		4121	4239	119	
9 SA	1771879	1850282	78403	130818	1536	1617	82	129	20 SA	5136948	5290253	153305	203687	4543	4669	127	246
10 ST	2206426	2298319	91893		1931	2067	137		21 ST	4936430	4954143	17713		4357	4441	85	
10 SA	2218415	2546802	328387	420280	1922	2210	289	426	21 SA	5432928	5629894	196966	214679	4799	4956	158	243
11 ST	2460444	2522674	62230		2172	2230	59										
11 SA	2672473	2701243	28770	91000	2326	2347	22	81									

[a]ST: *S. tropica*, SA: *S. arenicola*

Table 2.3: Secondary metabolite gene clusters in *S. tropica* (ST).

No.	Cluster name	Equivalent cluster	Biosynthetic class	Product	Biological activity/target	Island	Gene start	Gene stop	No. genes
1	ST *pks1*	none	polyketide	10-membered enediyne	cytotoxin/DNA	4	586	610	25
2	ST *nrps1*	SA *nrps3*[a]	non-ribosomal peptide	dipeptide	N/D	4/15	667	694	28
3	ST *sal*	none	polyketide/non-ribosomal peptide	**salinosporamide**	**cytotoxin/proteasome**	5	1012	1043	32
4	ST *pks2*	none	polyketide	glycosylated decaketide	N/D	11	2174	2227	54
5	ST *amc*	SA *amc*	carbohydrate	aminocyclitol	N/D	NI/NI	2340	2346	7
6	ST *bac1*	SA *bac2*	ribosomal peptide	class I bacteriocin (non-lantibiotic)	antimicrobial	NI/NI	2428	2440	13
7	ST *pks3*	SA *pks4*	polyketide	aromatic polyketide	N/D	NI/NI	2486	2510	25
8	ST *des*[b]	SA *des*	hydroxamate	**desferrioxamine**[c]	**siderophore/iron chelation**	NI/NI	2541	2555	15
9	ST *sid2*	SA *sid1*[a]	non-ribosomal peptide	yersiniabactin-related	siderophore/iron chelation	15/10	2645	2659	15
10	ST *spo*	none	polyketide	**sporolide**	N/D	15	2691	2737	47
11	ST *slm*	none	polyketide	**salinilactam**	N/D	15	2757	2781	25
12	ST *sid3*	none	non-ribosomal peptide	dihydroaeruginoic acid-related siderophore	siderophore/iron chelation	15	2786	2813	28
13	ST *sid4*	none	non-ribosomal peptide	coelibactin-related siderophore	siderophore/iron chelation	15	2814	2842	29
14	ST *bac2*	SA *bac3*	ribosomal peptide	class I bacteriocin (non-lantibiotic)	antimicrobial	NI/NI	3042	3054	13
15	ST *lym*	SA *lym*	polyketide/non-ribosomal peptide	**lymphostin**[c]	**immunosuppressant**	NI/NI	3055	3066	12
16	ST *terp1*	SA *terp2*	terpenoid	carotenoid pigment	antioxidant	NI/NI	3244	3253	10
17	ST *pks4*	SA *pks6*	polyketide	phenolic lipids	cell wall lipid	NI/NI	4264	4267	4
18	ST *nrps2*	SA *nrps4*	non-ribosomal peptide	tetrapeptide	N/D	21/21	4410	4429	20
19	ST *terp2*	SA *terp3*	terpenoid	carotenoid pigment	antioxidant	21/21	4437	4441	5
								Total	407

NI: non-island. Italics: predicted product or activity. Bold: observed product or activity. N/D: not determined.

[a]Partial cluster. [b]Previously designated ST Sid1 (32). [c]Product observed in other bacteria.

Table 2.4: Secondary metabolite clusters in *S. arenicola* (SA).

No.	Cluster name	Equivalent cluster	Biosynthetic class	Product	Biological activity/target	Island	Gene start	Gene stop	No. genes
1	SA *nrps1*	none	non-ribosomal peptide	pentapeptide	N/D	2	345	367	23
2	SA *pksnrps1*	none	polyketide/non-ribosomal peptide	N/D	N/D	3	478	499	22
3	SA *pks1A*	none	polyketide	9-membered enediyne unit/kedarcidin-related, fragment A	cytotoxin/DNA	4	545	560	16
4	SA *misc1*	none	aminoacyl tRNA synthetase-derivec	amino acid conjugate	N/D	4	570	573	4
5	SA *bac1*	none	ribosomal peptide	class I bacteriocin (lantibiotic)	antimicrobial	4	602	623	22
6	SA *pks2*	none	polyketide	N/D	N/D	6	1041	1073	33
7	SA *rif*	none	polyketide	**rifamycin**[b]	**antibiotic/RNA polymerase**	7	1240	1278	39
8	SA *terp1*	none	terpenoid	diterpene	N/D	7	1286	1288	3
9	SA *pks3A*	none	polyketide	10-membered enediyne unit/calicheamicin-related, fragment A	cytotoxin/DNA	10	2017	2049	33
10	SA *sid1*[b]	ST *sid2*	non-ribosomal peptide	yersiniabactin-related	siderophore/iron chelation	10/15	2070	2081	12
11	SA *pks1B*	none	polyketide-associated	nodified tyrosine and deoxysugar units/kedarcidin-related, fragment l	cytotoxin/DNA	10	2088	2121	34
12	SA *misc2*	none	aminoacyl tRNA synthetase-derivec	amino acid conjugate	N/D	10	2144	2151	8
13	SA *pks3B*	none	polyketide-related	aryltetrasaccharide unit/calicheamicin-related, fragment B	cytotoxin/DNA	10	2163	2206	44
14	SA *sta*	none	indolocarbazole	**staurosporine**[b]	**cytotoxin/protein kinase**	11	2326	2342	17
15	SA *pksnrps2*	none	polyketide/non-ribosomal peptide	N/D	N/D	12	2400	2409	10
16	SA *amc*	ST *amc*	carbohydrate	aminocyclitol	N/D	NI/NI	2483	2491	9
17	SA *bac2*	ST *bac1*	ribosomal peptide	class I bacteriocin (non-lantibiotic)	antimicrobial	NI/NI	2583	2595	13
18	SA *pks4*	ST *pks3*	polyketide	aromatic polyketide	N/D	NI/NI	2669	2694	26
19	SA *des*	ST *des*	hydroxamate	**desferrioxamine**[b]	**siderophore/iron chelation**	NI/NI	2728	2744	17
20	SA *nrps2*	none	non-ribosomal peptide	tetrapeptide	N/D	15	2939	2968	30
21	SA *nrps3*[a]	ST *nrps1*	non-ribosomal peptide	dipeptide	N/D	15/4	3051	3063	13
22	SA *pks5*	none	polyketide	macrolide	N/D	16	3148	3163	16
23	SA *bac3*	ST *bac2*	ribosomal peptide	class I bacteriocin (non-lantibiotic)	antimicrobial	NI/NI	3268	3280	13
24	SA *lym*	ST l*ym*	polyketide	**lymphostin**[b]	**immunosuppressant**	NI/NI	3281	3293	13
25	SA *terp2*	ST *terp1*	terpenoid	carotenoid pigment	antioxidant	NI/NI	3471	3480	10
26	SA *cym*	none	non-ribosomal peptide	**cyclomarin**[b]	**anti-inflammatory, antiviral**	20	4547	4569	23
27	SA *pks6*	ST *pks4*	polyketide	phenolic lipids	cell wall lipid	NI/NI	4694	4697	4
28	SA *nrps4*	ST *nrps2*	non-ribosomal peptide	tetrapeptide	N/D	21/21	4885	4904	20
29	SA *terp3*	ST *terp2*	terpenoid	carotenoid pigment	antioxidant	21/21	4927	4931	5
30	SA *pks1C*	none	polyketide	naphthoic acid unit/kedarcidin-related, fragment C	cytotoxin/DNA	21	4932	4956	25
								Total	540

NI: non-island. Italics: predicted product or activity. Bold: observed product or activity. N/D: not determined.

[a] Partial cluster. [b] Product observed in other bacteria.

Table 2.5: *S. tropica* mobile genetic elements (MGEs).

	Gene	Gene	No.	
MGE	start	stop	genes	Island
AICE1	58	74	17	1
Phage integrase	505	505	1	4
Prophage 1	507	559	53	4
IS1380	570	570	1	4
IS256	586	587	2	4
ISNCY	608	608	1	4
ISNCY	609	609	1	4
IS3	648	648	1	4
Unknown MGE	988	994	7	5
IS1380	1014	1014	1	5
IS3	1164	1165	2	6
IS5	1315	1315	1	7
phage gene	1317	1317	1	7
IS701	1506	1518	13	8
Unknown MGE	1602	1609	8	9
IS5	1614	1614	1	9
Prophage 2	1931	1957	27	10
Phage gene	1980	1980	1	10
Phage gene	1983	1983	1	10
Phage gene	2002	2002	1	10
Phage gene	2013	2013	1	10
IS630	2021	2022	2	10
Tn3	2304	2304	1	12
IS110	2305	2305	1	12
Tn3	2369	2369	1	13
IS110	2466	2466	1	14
IS5	2661	2661	1	15
IS1380	2716	2717	2	15
IS630	2729	2730	2	15
IS701	2752	2753	2	15
IS630	2845	2846	2	15
IS110	2861	2861	1	15
IS630	2891	2891	1	15
unk IS	2899	2899	1	15
Unknown MGE	2908	2909	2	15
IS5	2941	2941	1	16
Phage gene	3417	3417	1	17
Prophage 3	3986	4017	32	18
IS630	4122	4123	2	20
Tn3	4134	4134	1	20
Tn3	4137	4137	1	20
ISL3	4138	4138	1	20
Rev transcriptase	4139	4139	1	20
IS5	4140	4140	1	20
IS3	4141	4141	1	20
transposase	4142	4142	1	20
IS5	4179	4179	1	20
IS30	368	368	1	NI
Unknown MGE	749	756	8	NI
IS66	1556	1556	1	NI
IS110	1662	1662	1	NI
Phage gene	2334	2334	1	NI
Phage gene	2347	2347	1	NI
IS3	3350	3351	2	NI
Phage gene	3352	3352	1	NI
IS5	3501	3506	6	NI
IS630	3656	3662	7	NI
Total			153	

NI: non-island.

Table 2.6: *S. arenicola* mobile genetic elements (MGEs).

MGE	Gene start	Gene stop	No. genes	Island	MGE	Gene start	Gene stop	No. genes	Island
Tn3	346	346	1	2	Phage gene	3074	3074	1	15
Recombinase	612	612	1	4	Recombinase	3094	3094	1	15
Plasmid	925	958	34	5	IS4	3105	3105	1	15
IS21	1024	1025	2	6	IS4	3106	3106	1	15
ICE1	1208	1227	20	7	IS630	3107	3107	1	15
ICE2	1562	1580	19	9	IS21	3160	3161	2	16
IS21	1590	1591	2	9	Prophage 1A	3692	3743	52	17
Phage gene	1612	1613	2	9	Prophage 1B	3744	3794	51	17
IS701	1650	1650	1	10	IS5	4558	4558	1	20
IS256	1651	1651	1	10	IS630	4571	4571	1	20
ICE3	1922	1939	18	10	IS21	4925	4926	2	21
IS21	1968	1969	2	10	Recombinase	413	413	1	NI
IS5	1979	1979	1	10	Old Plasmid	1501	1502	2	NI
IS3	1991	1991	1	10	IS630	1649	1649	1	NI
IS5	1998	1998	1	10	Recombinase	1915	1915	1	NI
Recombinase	2051	2051	1	10	IS630	2285	2285	1	NI
Unknown MGE	2456	2477	22	12	Recombinase	2492	2492	1	NI
IS21	2854	2855	2	15	IS21	2580	2581	2	NI
Phage gene	2857	2857	1	15	IS630	3178	3178	1	NI
Plasmid gene	2979	2979	1	15	IS630	3576	3576	1	NI
IS110	2982	2982	1	15	IS	4038	4038	1	NI
IS5	3023	3023	1	15	ISL3	4192	4192	1	NI
IS4	3041	3041	1	15	Phage gene	4977	4977	1	NI
					Total			128	

NI: non-island.

Chapter 3: Comparative genomics reveals evidence of marine adaptation in *Salinispora* species

Abstract

Gram-positive bacteria represent a consistent component of most marine bacterial communities yet little is known about the mechanisms by which they adapt to life in the marine environment. Here we employed a phylogenomic approach to identify marine adaptation genes in marine Actinobacteria. The focus was on the obligate marine actinomycete genus *Salinispora* and the identification of marine adaptation genes that have been acquired from other marine bacteria. Functional annotation, comparative genomics, and evidence of a shared evolutionary history with bacteria from hyperosmotic environments were used to identify a pool of more than 50 marine adaptation genes. An Actinobacterial species tree was used to infer the likelihood of gene gain or loss in accounting for the distribution of each gene. Acquired marine adaptation genes were associated with electron transport, sodium and ABC transporters, and channels and pores. In addition, the loss of a mechanosensitive channel gene appears to have played a major role in the inability of *Salinispora* strains to grow following transfer to low osmotic strength media. The marine Actinobacteria for which genome sequences are available are broadly distributed throughout the Actinobacterial phylogenetic tree and closely related to non-marine forms suggesting they have been independently introduced relatively recently into the marine

environment. It appears that the acquisition of transporters in *Salinispora* spp. represents a major marine adaptation while gene loss is proposed to play a role in the inability of this genus to survive outside of the marine environment. This study reveals fundamental differences between marine adaptations in Gram-positive and Gram-negative bacteria and no common genetic basis for marine adaptation among the Actinobacteria analyzed.

Introduction

Microbiologists have long sought to define the physiological characteristics of marine bacteria (Macleod 1965). These studies have largely focused on seawater-inhabiting Gram-negative bacteria. None-the-less, Gram-positive bacteria are consistently reported from marine samples (Pommier et al. 2007). Among these, representatives of the phylum Actinobacteria are particularly well represented (Rappe et al. 1999; Prieto-Davó et al. 2008). To date, the genetic basis for marine adaptation in the Actinobacteria remains uncharacterized.

Early attempts to define marine bacteria centered on the observation that some marine-derived strains failed to grow when seawater was replaced with deionized (DI) water in the growth medium (Macleod 1965). Subsequently, this physiological response was linked to a specific sodium ion requirement, which led to the realization that seawater was not simply required for osmotic balance (Drapeau et al. 1966). Based on this, marine bacteria were further defined by a demonstrable requirement of sodium for growth (Macleod 1965). This requirement was subsequently linked to electron transport (Drapeau et al. 1966) and the possession of the sodium pumping respiratory NADH dehydrogenase Nqr (sodium quinone reductase) (Unemoto and Hayashi 1993). In addition to electron transport, it has also been reported that sodium is required for amino acid transporters and for the oxidation of compounds such as alanine and galactose in some marine bacteria (Drapeau et al. 1966). The ionic requirements of marine bacteria can also include calcium and magnesium (Macleod 1965), but the genetic basis for these requirements is unknown. At present, it remains unclear if similar marine adaptations occur in Gram-positive taxa.

The discovery of the sodium-pumping NADH dehydrogenase Nqr (Unemoto and Hayashi 1993) and the associated genes *nqrA-F* (Mulkidjanian et al. 2008) represented the first genetic link to sodium dependence in Gram-negative marine bacteria. Nqr is one of three types of respiratory NADH dehydrogenases and is known to occur in many Gram-negative marine bacteria and some clinical pathogens (Unemoto and Hayashi 1993; Hase et al. 2001). When present, Nqr does not preclude the occurrence of other NADH dehydrogenases in a genome (Hase et al. 2001). The more common prokaryotic NADH dehydrogenase is the proton-pumping NDH-1, which is also known as complex I (Bogachev and Verkhovsky 2005). NDH-1 is composed of 14 genes (*nuoA-N*) and displays no homology with Nqr yet both are energy-coupling enzyme complexes that create an ionic motive force used to generate ATP and drive other cellular processes (Schäfer et al. 2008). Interestingly, the membrane-bound, ion pumping *nuo* genes display significant sequence similarity to the six genes that make up the multi-subunit Na^+/H^+ antiporter Mrp (*mrpA-G*) (Swartz et al. 2005). The third type of NADH deydrogenase is NDH-2, which is typically composed of one to a few proteins (Schäfer et al. 2008) and is not an energy-coupling complex or been linked to marine adaptation.

The ability of bacteria to adapt to external changes in the osmotic environment is fundamental to survival (Sleator and Hill 2002). Osmoadaptation in bacteria typically involves the intracellular accumulation of compatible solutes such as glycine and betaine. These compounds are acquired either by de novo biosynthesis or directly from the environment. Bacteria also have mechanisms to survive osmotic down-shock that usually involve a combination of specific (secondary transport) and non-specific (stretch-activated channel) mechanisms of solute efflux together with aquaporin-mediated water efflux (Sleator and Hill 2002). One important mechanism of solute efflux is mediated by the mechanosensitive channel of large conductance (MscL). This membrane bound, stretch-activated channel is common in bacteria and believed to act as an emergency value to release turgor pressure following sudden osmotic downshock (Sukharev et al. 1997). In the marine halophile *Vibrio alginolyticus*, the introduction of *mscL* alleviated cell lysis following osmotic downshock (Nakamaru et al. 1999) and thus the product of this gene may represent an important mechanism to survive the transition from marine to freshwater environments.

In addition to specific ionic requirements and mechanisms to survive osmotic stress, comparative genomics has been used to identify marine adaptation genes in bacteria. For example, ABC branched chain amino acid (BCAA) transporters are enriched in *Bacillus* spp. adapted to alkaline and marine environments (Takami et al. 2000). Once taken into the cell, BCAAs are converted into L-glutamate, which would help acidify an otherwise basic cytoplasm (Takami et al. 2002). More recently, an abundance of BCAA transporters was observed in several marine *Roseobacter* strains (Moran et al. 2007). BCAA transporters also represent a significant portion of the genes observed in marine metagenomes (Morris et al. 2010) and thus appear to represent an important marine adaptation. Marine adaptation genes were also identified in the marine cyanobacterium *Synechoccocus*, which has a greater capacity to transport Na^+ than freshwater species (Palenik et al. 2003).

Actinomycetes belonging to the genus *Salinispora* occur broadly in tropical and sub-tropical marine sediments (Mincer et al. 2002). To date, two species (*S. tropica* and *S. arenicola*) have been formally described while a third ("*S. pacifica*") has been proposed (Fenical and Jensen 2006). This taxon was described as the first obligate marine actinomycete genus based on a failure to grow when seawater was replaced with DI water in a complex growth medium (Maldonado et al. 2005). It was recently demonstrated that *Salinispora* spp. are capable of growth with as little as 5 mM Na^+ if the appropriate osmotic environment is provided (Tsueng and Lam 2008a). However, it was also demonstrated that cells lyse in low osmotic strength media (Tsueng and Lam 2008b) suggesting a high level of marine adaptation.

The genome sequences of *S. tropica* strain CNB-440 and *S. arenicola* strain CNS-205 along with four unrelated marine Actinobacteria (*Aeromicrobium marinum*, *Janibacter* sp., 'marine actinobacterium PHSC20C1', and *Rhodococcus erythropolis*

PR4) and a large number of non-marine strains provided an opportunity to use comparative genomics to identify genes associated with marine adaptation. An earlier comparison of the two *Salinispora* genomes revealed a large paralogous family of genes encoding polymorphic membrane proteins (Pmps) (Penn et al. 2009). Although functionally uncharacterized, Pmps appear to be type V autotransporters. The large number of copies observed in the two genomes led to the proposal that they represent an adaptation to life in low nutrient environments and that they form pores that render *Salinispora* spp. susceptible to lysis in low osmotic conditions (Penn et al. 2009). The present study expands on that initial observation by employing a phylogenomic approach targeting gene gain and loss events to identify additional marine adaptation genes (MAGs). These analyses reveal that the mechanisms of marine adaptation in *Salinispora* spp. are fundamentally different from those reported for Gram-negative bacteria and that there is no common genetic basis for marine adaptation among the Actinobacteria for which genome sequences are currently available. In addition, the results provide strong evidence that gene loss plays a critical role in the inability of *Salinispora* spp. to survive when seawater replaces DI water in complex growth media.

Methods

Genome strains and analyses

The genomes of *S. tropica* strain CNB-440 (accession # CP000667) and *S. arenicola* strain CNS-205 (accession # CP000850) were downloaded from the U.S. Department of Energy Joint Genome Institute website (genome.ornl.gov/microbial/stro/03jan07 and genome.ornl.gov/microbial/sare/18jul07). Strains CNB-440 and CNS-205 were cultured from sediments collected at a depth of 20 m from the Bahamas and Palau, respectively. Artemis was used to visualize gene arrangement and annotation in each genome (Rutherford et al. 2000). A Fasta file of predicted protein sequences from the two genomes served as a database for BLAST searches (Altschul et al. 1990). Candidate marine adaptation genes (MAGs) were identified based on 1) gene function (annotation-derived) and 2) comparative genomics. The resulting pool of candidate MAGs was then analyzed using phylogenetic approaches and those with evidence of a shared ancestry with bacteria associated with hyper-osmotic environments were kept in the final MAG pool. Thus, this study is largely focused on the identification of marine adaptation genes that were acquired from other marine bacteria.

Function-based MAG identification

Keyword and BLAST searches were performed on the two *Salinispora* genomes using proteins previously linked to marine adaptation in studies of marine bacteria. The key words searched were associated with electron transport (complex I), sodium transporters, ABC transporters, and pores (Table 3.1). To improve the annotation of transporters prior to the key word searches, the two *Salinispora* genomes were submitted to transportDB (Ren et al. 2007), which annotates transporters according to the transport classification system (Saier et al. 2009). The BLAST searches were performed using complex I genes or *mscL* (Table 3.1). All sequences identified using these methods were subject to phylogenetic analysis as described below.

Comparative genomics-based MAG identification

Pair-wise comparisons were performed between *S. tropica* CNB-440 and 37 Actinobacterial genomes (including *S. arenicola* CNS-205) to identify orthologs that are present in both *Salinispora* genomes but absent in other Actinobacteria. The genomes selected for comparison include a broad phylogenetic range of Actinobacteria, three *Micromonospora* spp., and all marine Actinobacteria for which genomes sequences were available as of March 31, 2011. Orthologs were identified using the program Reciprocal Smallest Distance (Wall et al. 2003) based on e-values <1e-5, no more than 50% sequence divergence over the entire alignment of the sequence, and the remainder of the parameters set at default. Orthologs were eliminated if they were <350 amino acids in length or part of a mobile genetic element or secondary metabolite gene cluster as previously defined (Penn et al. 2009). Orthologs that passed these criteria were then evaluated phylogenetically to determine if they had a shared evolutionary history with bacteria derived from hyper-osmotic environments.

The RSD test was also used to identify genes that were lost in the two *Salinispora* genomes relative to other Actinobacteria. In this case, the

Micromonospora sp. L5 genome served as the reference for the pair-wise prediction of orthologs in 27 representative Actinobacterial genomes, including both *Salinispora* genomes. Sequences present in >24 Actinobacterial genomes based on the above RSD criteria for orthology, but not in the two *Salinispora* genomes, were considered as candidates for gene loss. Functional annotation was then used to determine if gene loss could represent a marine adaptation.

MAG phylogeny

All *Salinispora* protein sequences identified as candidate MAGs based on functional class and comparative genomics were subject to phylogenetic analysis to test for a shared evolutionary history with bacteria derived from hyper-osmotic environments. If a candidate MAG was part of an operon, the entire operon was tested. Maximum likelihood phylogenies were constructed for each candidate MAG using the online program MABL (Dereeper et al. 2008) with default settings (phylogeny.fr/version2_cgi/simple_phylogeny.cgi). The top 100 BLASTP hits were downloaded from the NCBI protein database and those with an e-value <1e-5 and length greater than 50% of the alignment were included in the tree. Genes that claded with orthologs from hyper-osmotic environments and ≤25 Actinobacterial species were kept in the final MAG pool. In cases where the nearest clade was not entirely comprised of strains from hyper-osmotic environments but a majority of strains in all other major clades were, the gene was included in the final MAG pool. Exceptions included trees that contained two or more *Micromonospora* sequences, as this was viewed as evidence of vertical inheritance. The files used to create the trees shown in Figures 3.2 and 3.4 are available at http://purl.org/phylo/treebase/phylows/study/TB2:S12306.

Species tree

All finished and several draft Actinobacterial genomes were downloaded from the NCBI FTP site on March 31, 2011. For Actinobacterial species with several draft genomes, at least two strains were included. In addition, any unnamed Actinobacterial species that contained a MAG were also included. The program AMPHORA (Wu and Eisen 2008) was then used to retrieve, align, and trim phylogenetic markers from each genome. Any marker that was not found in all species was excluded. If more than one version of a marker was found in a genome, the longest version that most closely fit the expected species phylogeny was selected. If the two versions were the same size and fit the expected phylogeny, one was selected randomly. Draft genomes were removed from the dataset if any marker gene was ≤25% of the size of all other sequences. However if the draft genome contained a MAG then none of the sequence data was removed. Finally, all aligned genes were concatenated and trimmed with Gblocks. The resulting alignment was input to PhyML for the construction of an Actinobacterial species tree.

Quantification of gene gain and loss

The species tree was used to calculate whether horizontal gene transfer or vertical inheritance is the most parsimonious explanation for the observed evolutionary history of each MAG. This was done by first documenting the distribution of each MAG in the species tree. The minimum number of gene loss events was then calculated by identifying the deepest branches in the tree within which all strains lacked the MAG. These branches or points were then summed. The calculation started at the last common ancestor of all strains that possessed the gene. The maximum number of gene gain events was calculated assuming each MAG was acquired independently and summing the terminal branch tips representing each lineage in which the gene was observed. The ratio of the minimum number of gene loss events to the maximum number of gene gain events was then calculated and values above one considered to support the hypothesis that the gene was acquired (ie, a higher number of gene loss events would be required to account for the observed distribution and thus gene gain is the more likely explanation) while values below one were used to support gene loss.

Results

Marine adaptation genes

Two fundamental approaches were used to identify genes associated with marine adaptation in the marine actinomycete genus *Salinispora*. The function-based approach relied on BLAST analyses using key words derived from previously reported marine adaptation genes (MAGs). The comparative genomics approach was annotation independent and detected genes that were present in *Salinispora* species but absent or rare in other Actinobacteria. Thus, the first approach tested for common mechanisms of marine adaptation among marine bacteria while the later had the potential to detect new or unknown gene functions that may be relevant to marine adaptation. All genes detected using these two approaches were then tested for evidence of a recent common ancestry with bacteria associated with hyperosmotic environments.

The function-based approach yielded the largest number of candidate marine adaptation genes (MAGs), however the vast majority identified using both approaches did not pass the phylogenetic test and therefore did not advance to the final MAG pool (Table 3.2). Ultimately, 60 and 58 MAGs were identified in the *S. tropica* and *S. arenicola* genomes, respectively. Of the MAGs identified in each species based on gene function, 13 are involved in electron transport, 12 encode transporters, and 28-30 (depending upon species) encode channels or pores. Based on comparative genomics, more genes related to marine adaptation appear to have been gained than lost from the two *Salinispora* spp. (Table 3.2).

Species tree

An Actinobacterial species tree was constructed using 19 of 31 AMPHORA marker genes (Wu and Eisen 2008) (Table 3.3) derived from 186 Actinobacterial genome sequences (Figure 3.1). This phylogeny is largely congruent to that previously published (Stackebrandt et al. 1997) with the notable exception of the close relationship of *Stackebrandtia nassauensis* DSM 4478 (family Glycomycetaceae) to the Micromonosporaceae. This relationship is supported by all of the individual gene trees and has also been reported by others (Wu et al. 2009). The tree clearly shows that the marine Actinobacteria for which genome sequences are available are polyphyletic and not deeply rooted. It is also notable that the order Actinomycetales is paraphyletic with respect to the Bifidobacteriales and that the previously reported polyphyly of the families Frankineae and Streptosporangineae is maintained in this tree (Garrity 2005).

Function-based identification of MAGs

Genes associated with the sodium-dependent NADH dehydrogenase (Nqr), which have been reported in Gram-negative marine bacteria, were not detected in either *Salinispora* genome or in any available Gram-positive marine bacterial genomes. Thus, when it comes to respiratory electron transport, there appear to be fundamentally different mechanisms by which Gram-negative and Gram-positive bacteria have adapted to the marine environment. None-the-less, 35 candidate MAGs with annotation linked to NDH-1 were initially detected in both *Salinispora* genomes (Table 3.2). These genes comprise three partial and one complete NDH-1 operon (Table 3.4). The 14 genes in the complete NDH-1 operon (*nuoA-N*) as well as those in the first partial NDH-1 operon were not considered further because their phylogenies are in general agreement with the Actinobacterial species tree, and thus there was no evidence they had been acquired from marine bacteria.

In contrast, phylogenetic analyses of all 13 genes in the second and third partial NDH-1 operons revealed close relationships with marine bacteria and thus these genes remained in the final MAG pool (Table 3.2). The annotation of the seven genes in the second partial NDH-1 operon predict that they encode the membranous portion of the enzyme complex, which pumps sodium ions or protons to generate an ionic motive force (Tokuda and Unemoto 1982). Among these seven genes, Stro769 and Sare711 are annotated as hypothetical proteins but likely encode NuoJ because top BLAST hits are annotated as such. The phylogenies of the corresponding seven Nuo protein sequences are similar and place them in a clade with nine other Actinobacteria (Figure 3.2A). The next three most closely related clades are comprised of nine Proteobacteria of which six are of marine origin. The Actinobacteria that possess these *nuo* genes are scattered throughout the species tree (Figure 3.3A), which could be interpreted as evidence for common ancestry within the Actinobacteria. To more formally infer the likelihood of gene loss (vertical inheritance) vs. gene gain (horizontal acquisition) in accounting for the distribution of these genes, the minimum number of loss events and maximum number of gain events was calculated. The resulting loss to gain ratio of 2.8 indicates that nearly three times as many loss events would be required to explain the observed distribution and thus provides support for the horizontal acquisition of this partial NDH-1 complex in *Salinispora* spp. (Figure 3.3A).

The six genes in the third partial NDH-1 complex have annotation related to *nuo* genes however upon closer analysis these genes appear to encode the sodium proton antiporter Mrp. The ambiguous annotation is not surprising as *mrp* genes are known to have sequence similarity to *nuo* genes (Swartz et al. 2005). Both *Salinispora* strains have *mrpA-G*, which are required for a functional antiporter (Ito et al. 2000), however *mrpA* and *B* are fused indicating that this is a group two Mrp operon (Swartz et al. 2005). *MrpG* was incorrectly predicted by auto-annotation but subsequently resolved based on homology with *B. halodurans*. MrpCEF and G are each too short to produce a robust phylogeny, however, the blast matches for these genes, and the fused *mrpAB* gene, were similar to the longer MrpD sequence and therefore it is inferred they share the same evolutionary history. The phylogeny of MrpD (Figure 3.2B) places the two *Salinispora* spp. in a primary clade that includes five *Corynebacteria* spp. and the Gram-negative marine bacterium *A. marinum*. This clade then shares a common ancestor with a large and diverse group of bacteria that contains at least four phyla including many marine and alkaliphilic species. The ratio of gene loss to gain events for each gene in the *mrp* operon is 2.3 (Figure 3.3B), thus supporting gene gain as the most parsimonious explanation for the occurrence of this gene in the two *Salinispora* spp.

S. tropica and S. arenicola contain 18 and 19 candidate sodium transporter genes respectively (Table 3.2), three of which were confirmed as MAGs following phylogenetic analysis (Table 3.4). One of these MAGs constitutes a Na$^+$/bile acid symporter (Stro2582 and Sare2779). The orthologs in the two *Salinispora* genomes group phylogenetically with 15 Actinobacteria including two other marine Actinobacteria (Figure 3.4A). The next clade contains *Acinetobacter* spp. followed by a single Actinobacterium and a large clade of *Pseudomonas* spp., many of which are human pathogens, and one *Myxococcus* sp. Subsequent clades include five Gram-negative marine bacteria. The apparent acquisition of this symporter may provide a mechanism to exploit a natural sodium gradient to import bile salts, which can be converted to compatible solutes such as glycine or taurine (Ridlon et al. 2006). Interestingly, genes for the biosynthesis of the compatible solute glycine betaine were not found in either *Salinispora* genome while genes for the uptake of this compound displayed no evidence of acquisition from marine bacteria and thus were not identified as MAGs (data not shown). The second sodium transport gene is a Na^{+}/Ca^{+2} exchanger (Stro449 and Sare538) with phylogenetic links to three different Actinobacteria and then *Nitrococcus mobilis*, a member of the Gamma-proteobacteria derived from surface waters of the equatorial pacific (Figure 3.4B) followed by a group comprised entirely of marine proteobacteria (see treeBASE link provided in the Methods). The third gene is a Na^{+}/Ca^{+2} antiporter (Stro4216 and Sare4649) that is largely related to genes observed in marine Alpha-proteobacteria (data not shown). The gene loss to gain ratios of 1.7 and 3.8 for the Na^{+}/bile acid symporter and Na^{+}/Ca^{+2} exchanger, respectively, supports the hypothesis that these genes were acquired. A gene loss to gain ratio was not calculated for the Na^{+}/Ca^{+2} antiporter because it was only observed in distantly related Actinobacteria and thus was assumed acquired. These calcium transporters may be related to the calcium requirement reported for *Salinispora* (Tsueng and Lam 2010).

TransportDB was used to identify 225 ABC transporters in each *Salinispora* genome (Table 3.4). After phylogenetic analysis of each protein, it was shown that the phosphate transporter Pst and branched chain amino acid transporter Liv have phylogenetic links to both marine and human associated bacteria (Figure 3.4C and D) and therefore advanced to the final MAG pool (Table 3.2). The four genes encoding the Pst transporter (Stro286-Stro289) display the same phylogenetic relationships and are closely related to homologs in marine cyanobacteria (Figure 3.4C). This transporter may be more efficient at scavenging phosphate from seawater than the form observed in soil Actinobacteria. The gene loss to gain ratio of 3.9 for each *pst* gene (Figure 3.3F) provides additional support for the acquisition of these genes in *Salinispora* spp. The five Liv proteins (Stro1801-1805) maintain the same phylogeny and reveal a close relationship to homologs from the marine Actinobacterium *Janibacter* sp. and then four bacteria from the Phylum Deinococcus-Thermus (Figure 3.4D). The next clade contains marine and pathogenic Proteobacteria (data not shown). The gene loss to gain ratio of 3.3 for each gene in this operon supports gene acquisition (Figure 3.3E).

Of the 35 and 33 channel and pore genes identified as candidate MAGs based on functional annotation in *S. tropica* and *S. arenicola,* respectively, 30 and 28 passed the phylogenetic test (Table 3.2). All of these were previously identified polymorphic membrane proteins (Pmps) that showed a strong phylogenetic relationship with homologs in marine bacteria (Penn et al. 2009). These genes are in high copy number (≥28) in both *Salinispora* genomes relative to the closely related genus

Micromonospora, in which only two copies are observed. A structural alignment of the predicted Pmp proteins indicates that each forms a beta-barrel structure, which likely forms a pore in the membrane, and contains a signal sequence common to all Pmps supporting that these proteins target the cell membrane.

Comparative genomics based identification of MAGs

A representative dataset comprised of 36 Actinobacterial genomes was used to identify 105 genes that are unique to both *Salinispora* spp. based on the RSD test of orthology (Table 3.2). Phylogenetic analyses revealed that seven of these genes shared a close relationship with homologs in marine bacteria and therefore advanced to the final MAG pool. However all seven of these genes were included among the MAGs previously identified based on gene function and thus comparative genomics revealed no new MAGS based on gene gain.

To assess gene loss based on comparative genomics, the *Micromonospora* sp. L5 genome was used as the reference sequence for the pair-wise RSD test of orthology in 27 representative Actinobacterial genomes, including both *Salinispora* spp. Four of 430 genes with predicted orthologs in at least 24 of the 27 genomes are absent in both *Salinispora* sequences (Table 3.4). These four genes are 1) a large conductance mechanosensitive channel (*mscL*) 2) an ABC transporter phosphate-binding protein (*pstS*), 3) a HAD-superfamily hydrolase, and 4) a peptidoglycan synthetase (*ftsI*). Homologs of *mscL* play a role in osmotic adaptation in halotolerant bacteria (Le Dain et al. 1998) and provide a mechanism to survive osmotic down shock (Sleator and Hill

2002; Roberts 2005). Thus, the loss of this gene may play a key role in the inability of *Salinispora* strains to survive when transferred to low osmotic strength media. The gene loss to gain ratio for *mscL* is 0.04 and thus highly supports gene loss in both *Salinispora* spp. (Figure 3.3G). Based on the RSD analysis, *pstS* was also identified as being lost in both *Salinispora* spp. However, all four genes in the *pst* operon are present in both *Salinispora* genomes and were already identified as MAGs based on functional annotation and evidence they were acquired from marine cyanobacteria (Figure 3.4C). Thus, it appears that the *pst* genes observed in both *Salinispora* spp. were too divergent to be detected as orthologs based on a comparison with the *Micromonospora* L5 genome. In support of this, a synteny plot in the region of the *pst* operon suggests that a homologous recombination event has resulted in the replacement of the entire *Salinispora* operon with a cyanobacterial version (Figure 3.5). The HAD-superfamily hydrolase and *ftsI* were not considered further as MAGs based on functional annotation.

Discussion

The marine Actinobacteria for which genome sequences are available are broadly distributed throughout the Actinobacterial phylogenetic tree and closely related to non-marine forms suggesting they have been independently introduced relatively recently into the marine environment. There is no evidence for a common set of genes linked to marine adaptation in these bacteria suggesting they have responded in different ways to the environmental pressures associated with survival in the marine environment. None of these bacteria, including the obligate marine genus *Salinispora*, possess Nqr, the sodium dependent respiratory NADH dehydrogenase that has frequently been linked to marine adaptation in Gram-negative marine bacteria (Oh et al. 1991). Thus, there appear to be fundamental differences in the ways Gram-negative bacteria and the Gram-positive bacteria studied here have adapted to the marine environment.

Given that gene acquisition represents a major force driving bacterial evolution (Ochman et al. 2000), it can be inferred that bacteria secondarily introduced into the marine environment will, over time, acquire adaptive traits from other marine bacteria. Using annotation as a guide, it was possible to identify a pool of genes in the two *Salinispora* genomes that are both relevant to marine adaptation and share a common evolutionary history with homologs from bacteria that inhabit hyper-osmotic environments. Despite the absence of Nqr, this pool includes 13 genes related to electron transport. These genes comprise two partial copies of NDH-1. One copy appears to encode the membranous portion of complex I, which pumps sodium ions or protons to generate an ionic motive force. The second copy contains *mrp* genes that likely encode a sodium antiporter that may help maintain a low cytoplasmic concentration of sodium. While Mrp is commonly found in bacteria and known to play a role in intracellular pH regulation (Swartz et al. 2005), homologs in the two *Salinispora* spp. are distantly related to any previously described and may represent a new type of Mrp antiporter. Taken together, the two partial NDH-1 complexes likely give *Salinispora* spp. the ability to keep excess sodium out of the cytoplasm while helping to meet the challenges of maintaining a proton gradient in seawater, which typically has a pH of 8.3. None of the MAGs were related to the biosynthesis or acquisition of compatible solutes such as glycine betaine, and there was no evidence that any proteins have excessive amounts of acidic amino acids or hydrophobic residues (data not shown), suggesting they do not accumulate intracellular salts as a mechanism of osmoregulation.

Genome sequences for six Actinobacteria isolated from the marine environment were available at the time of this study. While the MAG pool identified in the two *Salinispora* genomes is not shared by any of these strains, the Na^+/bile acid symporter is present in both *Janibacter* sp. and *A. marinum*. In addition, *A. marinum* also shares the MAGs *mrpD* and *pstS* with both *Salinispora* spp. while *livK* is also observed in *Janibacter* sp. The strain labeled 'marine actinobacterium' has none of the marine adaptation genes identified in the two *Salinispora* genome sequences. While all of the MAGs identified by gene gain were also identified by functional annotation, the *mscL* gene was uniquely identified as a MAG based on gene loss in *Salinispora* relative to other Actinobacteria. The loss of *mscL* is also observed in eight *Mobiluncus* species, *Streptomyces viridochromogenes*, *Streptomyces clavuligerus*, *Nocardiopsis dassonvillei*, *Rubrobacter xylanophilus*, and two *Collinsella* species and thus is not unique to *Salinispora* spp. These bacteria come either from sludge or a human source, two potentially consistent, hyper-osmotic environments where the loss of this gene may not prove disadvantageous. No other marine Actinobacteria have lost *mscL* and no Actinobacteria missing *mscL* have any of the *Salinispora* MAGs. These observations led to a series of genetic experiments that demonstrate the importance of MscL in allowing *Salinispora* strains to survive osmotic downshock (Appendix A).

The phylogenies of all but one *Salinispora* MAG (Na^+/Ca^{+2} antiporter) contain non-marine Actinobacteria, which suggests these genes may also prove adaptive in other environments. For example, the human pathogen *Nocardiopsis dassonvillei* has three of the MAGs while *Brevibacterium linens*, *Streptomyces roseosporus*, *Streptosporangium roseum*, *Corynebacterium kroppensteddti*, *and Geodermatophilus obscurus* each possess two. In total MAG homologs were found in 32 non-marine Actinobacteria. As with the non-marine Actinobacteria that have lost *mscL*, many of these strains are human pathogens or were derived from activated sludge.

The key word searches and comparative genomics approaches used here yielded a pool of candidate MAGs that were subsequently tested for phylogenetic links to bacteria associated with hyperosmotic environments. The final list of MAGs is almost certainly incomplete, as the key word searches were limited and it is possible that adaptations to survival in marine sediments may be very different from those previously reported for seawater inhabiting bacteria. It is also possible that some genes involved in marine adaptation are widely distributed among Actinobacteria and thus would remain undetected using the comparative genomics approach. Likewise, gene mutation or duplication may lead to environmentally relevant adaptive traits that were not detected with the methods employed. While this study was not a comprehensive assessment of marine adaptation, it nonetheless identified a pool of acquired genes that appear to be highly relevant to the survival of *Salinispora* spp. in the marine environment.

Conclusions

Functional annotation and comparative genomics were used to identify candidate marine adaptation genes in two *Salinispora* genome sequences. Using a phylogenomic approach, evidence of acquisition from bacteria associated with hyperosmotic environments was obtained for 57 and 59 genes in *S. arenicola* and *S. tropica*, respectively. An analysis of these genes reveals that the mechanisms of marine adaptation in *Salinispora* spp. are fundamentally different from those reported for Gram-negative bacteria and other marine Actinobacteria. While not comprehensive, the MAGs identified are largely associated with electron transport, sodium transporters, and ABC transporters and are predicted to represent marine adaptations based on evidence of acquisition from marine bacteria. The results also indicate that the loss of the *mscL* gene may play a key role in the inability of *Salinispora* strains to survive osmotic down shock (Appendix A). Given that *Salinispora* spp. are a useful source of secondary metabolites with applications in human medicine (Feling et al. 2003), identifying the genetic basis for the osmotic requirements reported for this genus may prove useful for future industrial development.

Acknowledgments

Chapter 3, in full, is in published in the Journal BMC Genomics, 2012, Kevin Penn and Paul R. Jensen. The dissertation author was the primary author and conducted the majority of the research, which forms the basis for this chapter.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215(3): 403-410.

Bogachev AV, Verkhovsky MI (2005) Na+-Translocating NADH: Quinone Oxidoreductase: Progress Acheived and Prospects of Investigations. Biochemistry (Moscow) 70(2): 143-149.

Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Research 36(suppl 2): W465-W469.

Drapeau GR, Matula TI, MacLeod RA (1966) Nutrition and Metabolism of Marine Bacteria XV. Relation of Na+-Activated Transport to the Na+ Requirement of a Marine Pseudomonad for Growth. The Journal of Bacteriology 92(1): 63-71.

Feling RH, Buchanan GO, Mincer TJ, Kauffman CA, Jensen PR, Fenical W (2003) Salinosporamide A: A Highly Cytotoxic Proteasome Inhibitor from a Novel Microbial Source, a Marine Bacterium of the New Genus *Salinospora*. Angewandte Chemie 115(3): 369-371.

Fenical W, Jensen PR (2006) Developing a new resource for drug discovery: marine actinomycete bacteria. Nature Chemical Biology 2(12): 666-673.

Garrity GM (2005) Bergey's Manual of Systematic Bacteriology: Springer-Verlag.

Hase CC, Fedorova ND, Galperin MY, Dibrov PA (2001) Sodium Ion Cycle in Bacterial Pathogens: Evidence from Cross-Genome Comparisons. Microbiology and Molecular Biology Reviews 65(3): 353-370.

Ito M, Guffanti AA, Wang W, Krulwich TA (2000) Effects of Nonpolar Mutations in Each of the Seven *Bacillus subtilis* mrp Genes Suggest Complex Interactions among the Gene Products in Support of Na+ and Alkali but Not Cholate Resistance. The Journal of Bacteriology 182(20): 5663-5670.

Le Dain AC, Saint N, Kloda A, Ghazi A, Martinac B (1998) Mechanosensitive Ion Channels of the Archaeon *Haloferax volcanii*. Journal of Biological Chemistry 273(20): 12116-12119.

Macleod RA (1965) The Question of the Existence of Specific Marine Bacteria. Microbiology and Molecular Biology Reviews 29(1): 9-23.

Maldonado LA, Fenical W, Jensen PR, Kauffman CA, Mincer TJ, Ward AC, Bull AT, Goodfellow M (2005) *Salinispora arenicola* gen. nov., sp. nov. and *Salinispora tropica* sp. nov., obligate marine actinomycetes belonging to the family Micromonosporaceae. International Journal of Systematic and Evolutionary Microbiology 55(5): 1759-1766.

Mincer TJ, Jensen PR, Kauffman CA, Fenical W (2002) Widespread and Persistent Populations of a Major New Marine Actinomycete Taxon in Ocean Sediments. Applied and Environmental Microbiology 68(10): 5005-5011.

Moran MA, Belas R, Schell MA, Gonzalez JM, Sun F, Sun S, Binder BJ, Edmonds J, Ye W, Orcutt B, Howard EC, Meile C, Palefsky W, Goesmann A, Ren Q, Paulsen I, Ulrich LE, Thompson LS, Saunders E, Buchan A (2007) Ecological

Genomics of Marine Roseobacters. Applied and Environmental Microbiology 73(14): 4559-4569.

Morris RM, Nunn BL, Frazar C, Goodlett DR, Ting YS, Rocap G (2010) Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. ISME J 4(5): 673-685.

Mulkidjanian AY, Dibrov P, Galperin MY (2008) The past and present of sodium energetics: May the sodium-motive force be with you. Biochimica et Biophysica Acta 1777(7-8): 985-992.

Nakamaru Y, Takahashi Y, Unemoto T, Nakamura T (1999) Mechanosensitive channel functions to alleviate the cell lysis of marine bacterium, *Vibrio alginolyticus*, by osmotic downshock. FEBS Letters 444(2-3): 170-172.

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299-304.

Oh S, Kogure K, Ohwada K, Simidu U (1991) Correlation between Possession of a Respiration-Dependent Na+ Pump and Na+ Requirement for Growth of Marine Bacteria. Applied and Environmental Microbiology 57(6): 1844-1846.

Palenik B, Brahamsha B, Larimer FW, Land M, Hauser L, Chain P, Lamerdin J, Regala W, Allen EE, McCarren J, Paulsen I, Dufresne A, Partensky F, Webb EA, Waterbury J (2003) The genome of a motile marine *Synechococcus*. Nature 424(6952): 1037-1042.

Penn K, Jenkins C, Nett M, Udwary DW, Gontang EA, McGlinchey RP, Foster B, Lapidus A, Podell S, Allen EE, Moore BS, Jensen PR (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. ISME J 3(10): 1193-1203.

Pommier T, CanbÄCk B, Riemann L, BostrÖM KH, Simu K, Lundberg P, Tunlid A, HagstrÖM Å (2007) Global patterns of diversity and community structure in marine bacterioplankton. Molecular Ecology 16(4): 867-880.

Prieto-Davó A, Fenical W, Jensen PR (2008) Comparative actinomycete diversity in marine sediments. Aquat Microb Ecol 52(1): 11.

Rappe MS, Gordon DA, Vergin KL, Giovannoni SJ (1999) Phylogeny of actinobacteria small subunit (SSU) rRNA gene clones recovered from marine bacterioplankton. Anglais 22(1): 106-112.

Ren Q, Chen K, Paulsen IT (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Research 35(suppl 1): D274-D279.

Ridlon JM, Kang D-J, Hylemon PB (2006) Bile salt biotransformations by human intestinal bacteria. Journal of Lipid Research 47(2): 241-259.

Roberts M (2005) Organic compatible solutes of halotolerant and halophilic microorganisms. Saline Systems 1(1): 5.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-Al, Barrell B (2000) Artemis: sequence visualization and annotation. Bioinformatics 16(10): 944-945.

Saier MH, Yen MR, Noto K, Tamang DG, Elkan C (2009) The Transporter Classification Database: recent advances. Nucleic Acids Research 37(suppl 1): D274-D278.

Schäfer G, Penefsky H, Kerscher S, Dröse S, Zickermann V, Brandt U (2008) The Three Families of Respiratory NADH Dehydrogenases. Bioenergetics: Springer Berlin / Heidelberg. pp. 185-222.

Sleator RD, Hill C (2002) Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence. FEMS Microbiology Reviews 26(1): 49-71.

Stackebrandt E, Rainey FA, Ward-Rainey NL (1997) Proposal for a new hierarchic classification system, Actinobacteria class s nov. International Journal of Systematic Bacteriology 47(2): 479-491.

Sukharev SI, Blount P, Martinac B, Kung aC (1997) Mechanosensitive Channels of *Escherichia coli*: The MscL Gene, Protein, and Activities. Annual Review of Physiology 59(1): 633-657.

Swartz T, Ikewada S, Ishikawa O, Ito M, Krulwich T (2005) The Mrp system: a giant among monovalent cation/proton antiporters? Extremophiles 9(5): 345-354.

Takami H, Takaki Y, Uchiyama I (2002) Genome sequence of *Oceanobacillus iheyensis* isolated from the Iheya Ridge and its unexpected adaptive capabilities to extreme environments. Nucleic Acids Research 30(18): 3927-3935.

Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, Masui N, Fuji F, Hirama C, Nakamura Y, Ogasawara N, Kuhara S, Horikoshi K (2000) Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. Nucleic Acids Research 28(21): 4317-4331.

Tokuda H, Unemoto T (1982) Characterization of the respiration-dependent Na+ pump in the marine bacterium *Vibrio alginolyticus*. Journal of Biological Chemistry 257(17): 10007-10014.

Tsueng G, Lam K (2008a) A low-sodium-salt formulation for the fermentation of salinosporamides by *Salinispora tropica* strain NPS21184. Applied Microbiology and Biotechnology 78(5): 821-826.

Tsueng G, Lam KS (2008b) Growth of *Salinispora tropica* strains CNB440, CNB476, and NPS21184 in nonsaline, low-sodium media. Applied Microbiology and Biotechnology 80(5): 873-880.

Tsueng G, Lam K (2010) A preliminary investigation on the growth requirement for monovalent cations, divalent cations and medium ionic strength of marine actinomycete *Salinispora*. Applied Microbiology and Biotechnology.

Unemoto T, Hayashi M (1993) Na+-translocating NADH-quinone reductase of marine and halophilic bacteria. Journal of Bioenergetics and Biomembranes 25(4): 385-391.

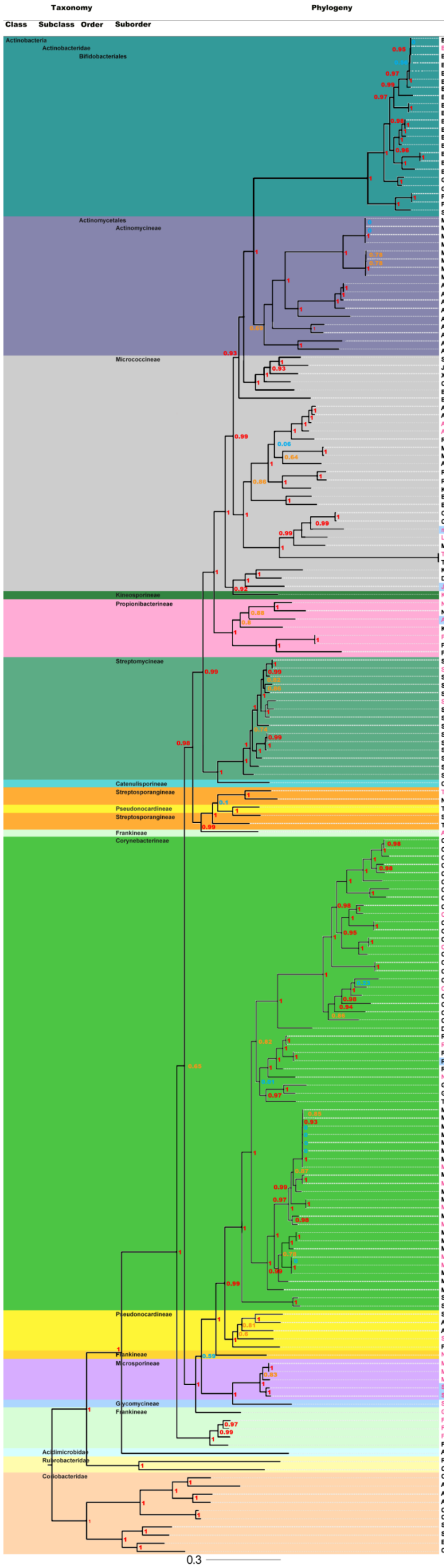
Wall DP, Fraser HB, Hirsh AE (2003) Detecting putative orthologs. Bioinformatics 19(13): 1710-1711.

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng J-F, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk H-P, Eisen JA (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462(7276): 1056-1060.

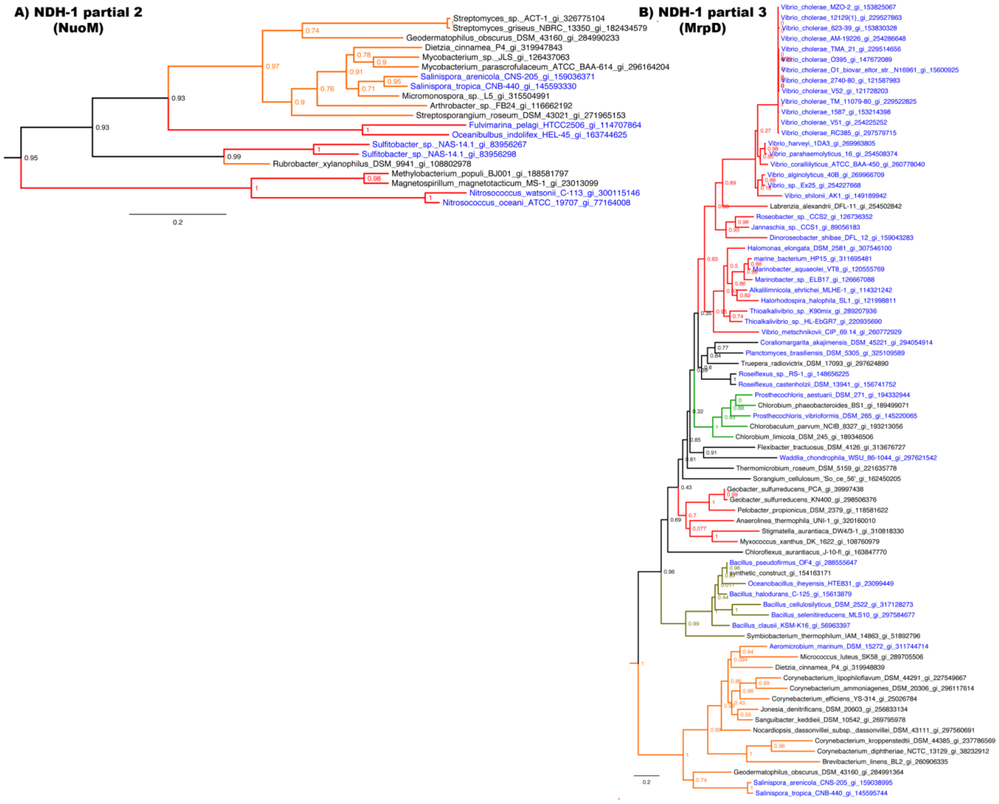
Wu M, Eisen J (2008) A simple, fast, and accurate method of phylogenomic inference. Genome Biology 9(10): R151.

Figures

Figure 3.1: Actinobacterial species tree showing the distribution of marine adaptation genes (MAGs). The Actinobacteria are color coded according to major taxonomic affiliations. Species names are listed vertically and MAGs listed horizontally across the top of the table. Colored boxes indicate the distribution of each MAG. The names of the 38 strains used in the comparative genomic analyses are colored in pink while the last two columns indicate the genome sequences used for the gene gain and loss analyses. Strains highlighted in blue are of marine origin. Branch support is listed on each node; red values indicate a likelihood of 90 or higher, orange indicates values between 60 and 89 while blue indicates support values below 60. The Coriobacteridae were chosen as the root. Pst, Liv, Partial 2, and Mrp represent all genes in the respective operons. See table 3.4 for detailed tree parameters.



Species	Presence or absence of marine adaptation genes								Genome status	Genome used in RSD test	
	Pst	Liv	Na+/Ca+2 exchanger	Na+/Ca+2 antiporter	Na+/Bile symporter	Partial 2 Nuo	Mrp	MscL		Gene gain	Gene loss
Bifidobacterium longum subsp. longum BBMN68								■	finished		
Bifidobacterium longum DJO10A								■	finished	YES	YES
Bifidobacterium longum subsp. longum JCM 1217								■	finished		
Bifidobacterium longum subsp. infantis 157F								■	finished		
Bifidobacterium longum subsp. longum JDM301								■	finished		
Bifidobacterium longum subsp. infantis ATCC 15697								■	finished		
Bifidobacterium breve DSM 20213								■	draft		
Bifidobacterium angulatum DSM 20098								■	draft		
Bifidobacterium bifidum S17								■	finished		
Bifidobacterium bifidum PRL2010								■	finished		
Bifidobacterium pseudocatenulatum DSM 20438								■	draft		
Bifidobacterium catenulatum DSM 16992								■	draft		
Bifidobacterium adolescentis ATCC 15703								■	finished		
Bifidobacterium dentium Bd1								■	finished		
Bifidobacterium animalis subsp. lactis DSM 10140								■	finished		
Bifidobacterium animalis subsp. lactis Bl-04								■	finished		
Bifidobacterium gallicum DSM 20093								■	draft		
Gardnerella vaginalis ATCC 14019								■	finished		
Gardnerella vaginalis 409-05								■	finished		
Parascardovia denticolens F0305								■	draft		
Parascardovia denticolens DSM 10105								■	draft		
Scardovia inopinata F0304								■	draft		
Mobiluncus mulieris ATCC 35239									draft		
Mobiluncus mulieris 28-1									draft		
Mobiluncus mulieris FB024-16									draft		
Mobiluncus mulieris ATCC 35243									draft		
Mobiluncus curtisii subsp. curtisii ATCC 35241									draft		
Mobiluncus curtisii ATCC 43063									finished		
Mobiluncus curtisii subsp. holmesii ATCC 35242									draft		
Mobiluncus curtisii ATCC 51333									draft		
Actinomyces odontolyticus F0309								■	draft		
Actinomyces odontolyticus ATCC 17982								■	draft		
Actinomyces sp. oral taxon 180 str. F0310								■	draft		
Actinomyces sp. oral taxon 178 str. F0338								■	draft		
Actinomyces coleocanis DSM 15436								■	draft		
Actinomyces viscosus C505								■	draft		
Actinomyces urogenitalis DSM 15434								■	draft		
Arcanobacterium haemolyticum DSM 20595								■	finished		
Actinomyces sp. oral taxon 848 str. F0332								■	draft		
Sanguibacter keddieii DSM 10542							■	■	finished		
Jonesia denitrificans DSM 20603							■	■	finished		
Xylanimonas cellulosilytica DSM 15894								■	finished		
Cellulomonas flavigena DSM 20109								■	finished		
Beutenbergia cavernae DSM 12333								■	finished		
Brachybacterium faecium DSM 4810								■	finished		
Arthrobacter phenanthrenivorans Sphe3								■	finished		
Arthrobacter chlorophenolicus A6								■	finished		
Arthrobacter sp. FB24						■		■	finished	YES	YES
Arthrobacter aurescens TC1								■	finished	YES	YES
Renibacterium salmoninarum ATCC 33209								■	finished		
Micrococcus luteus NCTC 2665							■	■	finished		
Micrococcus luteus SK58								■	draft		
Arthrobacter arilaitensis Re117								■	finished		
Rothia mucilaginosa DY-18								■	finished		
Rothia dentocariosa ATCC 17931								■	finished		
Kocuria rhizophila DC2201								■	finished		
Brevibacterium mcbrellneri ATCC 49030								■	draft		
Brevibacterium linens BL2					■		■	■	draft		
Clavibacter michiganensis subsp. sepedonicus								■	finished		
Clavibacter michiganensis subsp. michiganensis NCPPB 382								■	finished		
marine actinobacterium PHSC20C1								■	draft	YES	
Leifsonia xyli subsp. xyli str. CTCB07								■	finished	YES	YES
Microbacterium testaceum StLB037								■	finished		
Tropheryma whipplei str. Twist								■	finished	YES	YES
Tropheryma whipplei TW08/27								■	finished		
Kytococcus sedentarius DSM 20547								■	finished		
Dermacoccus sp. Ellin185								■	draft		
Janibacter sp. HTCC2649		■			■			■	draft	YES	
Kineococcus radiotolerans SRS30216								■	finished	YES	YES
Nocardioides sp. JS614	■							■	finished	YES	
Nocardioidaceae bacterium Broad-1		■			■			■	draft		
Aeromicrobium marinum DSM 15272	■				■		■	■	draft	YES	
Kribbella flavida DSM 17836					■			■	finished		
Propionibacterium acnes SK137								■	finished	YES	YES
Propionibacterium acnes KPA171202								■	finished		
Propionibacterium freudenreichii subsp. shermanii CIRM-BIA1								■	finished		
Streptomyces lividans TK24								■	draft		
Streptomyces coelicolor A32								■	finished	YES	YES
Streptomyces griseoflavus Tu4000								■	draft		
Streptomyces viridochromogenes DSM 40736									draft		
Streptomyces ghanaensis ATCC 14672								■	draft		
Streptomyces avermitilis MA-4680								■	finished	YES	YES
Streptomyces scabiei 87.22					■			■	finished		
Streptomyces sviceus ATCC 29083								■	draft		
Streptomyces albus J1074			■					■	draft		
Streptomyces griseus subsp. griseus NBRC 13350						■		■	finished		
Streptomyces cf. griseus XylebKG-1 (aka sp. ACT-1)						■		■	finished		
Streptomyces roseosporus NRRL 15998	■				■			■	draft		
Streptomyces pristinaespiralis ATCC 25486								■	draft		
Streptomyces clavuligerus ATCC 27064									draft		
Streptomyces violaceusniger Tu 4113								■	draft		
Catenulispora acidiphila DSM 44928								■	finished		
Thermobifida fusca YX									finished	YES	YES
Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111			■		■		■	■	finished		
Thermobispora bispora DSM 43833								■	finished		
Streptosporangium roseum DSM 43021					■	■		■	finished		
Thermomonospora curvata DSM 43183								■	finished		
Acidothermus cellulolyticus 11B								■	finished	YES	
Corynebacterium tuberculostearicum SK141								■	draft		
Corynebacterium pseudogenitalium ATCC 33035								■	draft		
Corynebacterium accolens ATCC 49725								■	draft		
Corynebacterium striatum ATCC 6940								■	draft		
Corynebacterium aurimucosum ATCC 700975								■	finished		
Corynebacterium ammoniagenes DSM 20306							■	■	draft		
Corynebacterium lipophiloflavum DSM 44291							■	■	draft		
Corynebacterium genitalium ATCC 33030								■	draft		
Corynebacterium pseudotuberculosis FRC41								■	finished		
Corynebacterium diphtheriae NCTC 13129							■	■	finished	YES	
Corynebacterium matruchotii ATCC 33806								■	draft		
Corynebacterium matruchotii ATCC 14266								■	draft		
Corynebacterium glutamicum R								■	finished		
Corynebacterium glutamicum ATCC 13032								■	finished	YES	YES
Corynebacterium efficiens YS-314							■	■	finished		
Corynebacterium glucuronolyticum ATCC 51867								■	draft		
Corynebacterium glucuronolyticum ATCC 51866								■	draft		
Corynebacterium urealyticum DSM 7109								■	finished		
Corynebacterium jeikeium K411								■	finished	YES	YES
Corynebacterium resistens DSM 45100								■	draft		
Corynebacterium variabile DSM 44702								■	draft		
Corynebacterium kroppenstedtii DSM 44385							■	■	finished		
Corynebacterium amycolatum SK46								■	draft		
Dietzia cinnamea P4						■	■	■	draft		
Rhodococcus opacus B4	■							■	finished		
Rhodococcus jostii RHA1								■	finished	YES	YES
Rhodococcus erythropolis SK121 (aka SK121)					■			■	draft		
Rhodococcus erythropolis PR4					■			■	finished		
Rhodococcus equi 103S					■			■	finished		
Nocardia farcinica IFM 10152								■	finished	YES	YES
Gordonia neofelifaecis NRRL B-59395								■	draft		
Gordonia bronchialis DSM 43247								■	finished		
Tsukamurella paurometabola DSM 20162								■	finished		
Mycobacterium bovis BCG str. Tokyo 172								■	finished		
Mycobacterium bovis BCG str. Pasteur 1173P2								■	finished		
Mycobacterium bovis AF2122/97								■	finished		
Mycobacterium tuberculosis F11								■	finished		
Mycobacterium tuberculosis KZN 1435								■	finished		
Mycobacterium tuberculosis H37Rv								■	finished		
Mycobacterium tuberculosis H37Ra								■	finished		
Mycobacterium tuberculosis CDC1551								■	finished	YES	YES
Mycobacterium ulcerans Agy99								■	finished		
Mycobacterium marinum M								■	finished	YES	
Mycobacterium kansasii ATCC 12478								■	draft		
Mycobacterium leprae TN								■	finished		
Mycobacterium leprae Br4923								■	finished	YES	YES
Mycobacterium parascrofulaceum ATCC BAA-614						■		■	draft		
Mycobacterium avium subsp. paratuberculosis K-10								■	finished	YES	YES
Mycobacterium sp. Spyr1								■	finished		
Mycobacterium gilvum PYR-GCK								■	finished		
Mycobacterium vanbaalenii PYR-1								■	finished		
Mycobacterium sp. KMS								■	finished	YES	YES
Mycobacterium sp. JLS						■		■	finished	YES	YES
Mycobacterium sp. MCS								■	finished		
Mycobacterium smegmatis str. MC2 155								■	finished		
Mycobacterium abscessus ATCC 19977								■	finished		
Segniliparus rugosus ATCC BAA-974								■	draft		
Segniliparus rotundus DSM 44985								■	finished		
Saccharomonospora viridis DSM 43017			■					■	finished		
Amycolatopsis mediterranei U32								■	finished		
Actinosynnema mirum DSM 43827								■	finished		
Saccharopolyspora erythraea NRRL 2338								■	finished	YES	
Pseudonocardia sp. P1								■	draft		
Nakamurella multipartita DSM 44233					■			■	finished	YES	
Micromonospora aurantiaca ATCC 27029						■		■	finished	YES	YES
Micromonospora sp. ATCC 39149					■			■	draft	YES	YES
Micromonospora sp. L5								■	finished	YES	YES
Salinispora tropica CNB-440	■	■	■	■	■	■	■		finished	YES	YES
Salinispora arenicola CNS-205	■	■	■	■	■	■	■		finished	YES	YES
Stackebrandtia nassauensis DSM 44728					■			■	finished	YES	
Geodermatophilus obscurus DSM 43160						■	■	■	finished	YES	
Frankia alni ACN14a								■	finished	YES	YES
Frankia sp. Ccl3								■	finished	YES	YES
Frankia sp. EAN1pec								■	finished	YES	YES
Frankia sp. EuI1c								■	finished		
Acidimicrobium ferrooxidans DSM 10331								■	finished		
Rubrobacter xylanophilus DSM 9941									finished		
Conexibacter woesei DSM 14684	■							■	finished		
Olsenella uli DSM 7084								■	finished		
Atopobium vaginae PB189-T1-4								■	draft		
Atopobium rimae ATCC 49626								■	draft		
Atopobium parvulum DSM 20469								■	finished		
Collinsella stercoris DSM 13279									finished		
Collinsella intestinalis DSM 13280									draft		
Slackia heliotrinireducens DSM 20476								■	finished		
Slackia exigua ATCC 700122								■	draft		
Eggerthella lenta DSM 2243								■	finished		
Cryptobacterium curtum DSM 15641								■	finished		



A) NDH-1 partial 2 (NuoM)

B) NDH-1 partial 3 (MrpD)

Streptomyces_sp._ACT-1_gi_326775104
Streptomyces_griseus_NBRC_13350_gi_182434579
Geodermatophilus_obscurus_DSM_43160_gi_284990233
Dietzia_cinnamea_P4_gi_319947843
Mycobacterium_sp._JLS_gi_126437063
Mycobacterium_parascrofulaceum_ATCC_BAA-614_gi_296164204
Salinispora_arenicola_CNS-205_gi_159036371
Salinispora_tropica_CNB-440_gi_145593330
Micromonospora_sp._L5_gi_315504991
Arthrobacter_sp._FB24_gi_116662192
Streptosporangium_roseum_DSM_43021_gi_271965153
Fulvimarina_pelagi_HTCC2506_gi_114707864
Oceanibulbus_indolifex_HEL-45_gi_163744625
Sulfitobacter_sp._NAS-14.1_gi_83956267
Sulfitobacter_sp._NAS-14.1_gi_83956298
Rubrobacter_xylanophilus_DSM_9941_gi_108802978
Methylobacterium_populi_BJ001_gi_188581797
Magnetospirillum_magnetotacticum_MS-1_gi_23013099
Nitrosococcus_watsonii_C-113_gi_300115146
Nitrosococcus_oceani_ATCC_19707_gi_77164008
0.74
0.97
0.78
0.9
0.91
0.95
0.71
0.76
0.9
0.93
0.93
0.95
0.99
0.98
1
0.2
Vibrio_cholerae_MZO-2_gi_153825067
Vibrio_cholerae_12129(1)_gi_229527863
Vibrio_cholerae_623-39_gi_153830328
Vibrio_cholerae_AM-19226_gi_254286648
Vibrio_cholerae_TMA_21_gi_229514656
Vibrio_cholerae_O395_gi_147672089
Vibrio_cholerae_O1_biovar_eltor_str._N16961_gi_15600925
Vibrio_cholerae_2740-80_gi_121587983
Vibrio_cholerae_V52_gi_121728203
Vibrio_cholerae_TM_11079-80_gi_229522825
Vibrio_cholerae_1587_gi_153214398
Vibrio_cholerae_V51_gi_254225252
Vibrio_cholerae_RC385_gi_297579715
Vibrio_harveyi_1DA3_gi_269963805
Vibrio_parahaemolyticus_16_gi_254508374
Vibrio_coralliilyticus_ATCC_BAA-450_gi_260778040
Vibrio_alginolyticus_40B_gi_269966709
Vibrio_sp._Ex25_gi_254227668
Vibrio_shilonii_AK1_gi_149189942
Labrenzia_alexandrii_DFL-11_gi_254502842
Roseobacter_sp._CCS2_gi_126736352
Jannaschia_sp._CCS1_gi_89056183
Dinoroseobacter_shibae_DFL_12_gi_159043283
Halomonas_elongata_DSM_2581_gi_307546100
marine_bacterium_HP15_gi_311695481
Marinobacter_aquaeolei_VT8_gi_120555769
Marinobacter_sp._ELB17_gi_126667088
Alkalilimnicola_ehrlichei_MLHE-1_gi_114321242
Halorhodospira_halophila_SL1_gi_121998811
Thioalkalivibrio_sp._K90mix_gi_289207936
Thioalkalivibrio_sp._HL-EbGR7_gi_220935690
Vibrio_metschnikovii_CIP_69.14_gi_260772929
Coraliomargarita_akajimensis_DSM_45221_gi_294054914
Planctomyces_brasiliensis_DSM_5305_gi_325109589
Truepera_radiovictrix_DSM_17093_gi_297624890
Roseiflexus_sp._RS-1_gi_148656225
Roseiflexus_castenholzii_DSM_13941_gi_156741752
Prosthecochloris_aestuarii_DSM_271_gi_194332944
Chlorobium_phaeobacteroides_BS1_gi_189499071
Prosthecochloris_vibrioformis_DSM_265_gi_145220065
Chlorobaculum_parvum_NCIB_8327_gi_193213056
Chlorobium_limicola_DSM_245_gi_189346506
Flexibacter_tractuosus_DSM_4126_gi_313676727
Waddlia_chondrophila_WSU_86-1044_gi_297621542
Thermomicrobium_roseum_DSM_5159_gi_221635778
Sorangium_cellulosum_'So_ce_56'_gi_162450205
Geobacter_sulfurreducens_PCA_gi_39997438
Geobacter_sulfurreducens_KN400_gi_298506376
Pelobacter_propionicus_DSM_2379_gi_118581622
Anaerolinea_thermophila_UNI-1_gi_320160010
Stigmatella_aurantiaca_DW4/3-1_gi_310818330
Myxococcus_xanthus_DK_1622_gi_108760979
Chloroflexus_aurantiacus_J-10-fl_gi_163847770
Bacillus_pseudofirmus_OF4_gi_288555647
synthetic_construct_gi_154163171
Oceanobacillus_iheyensis_HTE831_gi_23099449
Bacillus_halodurans_C-125_gi_15613879
Bacillus_cellulosilyticus_DSM_2522_gi_317128273
Bacillus_selenitireducens_MLS10_gi_297584677
Bacillus_clausii_KSM-K16_gi_56963397
Symbiobacterium_thermophilum_IAM_14863_gi_51892796
Aeromicrobium_marinum_DSM_15272_gi_311744714
Micrococcus_luteus_SK58_gi_289705506
Dietzia_cinnamea_P4_gi_319948839
Corynebacterium_lipophiloflavum_DSM_44291_gi_227549667
Corynebacterium_ammoniagenes_DSM_20306_gi_296117614
Corynebacterium_efficiens_YS-314_gi_25026784
Jonesia_denitrificans_DSM_20603_gi_256833134
Sanguibacter_keddieii_DSM_10542_gi_269795978
Nocardiopsis_dassonvillei_subsp._dassonvillei_DSM_43111_gi_297560691
Corynebacterium_kroppenstedtii_DSM_44385_gi_237786569
Corynebacterium_diphtheriae_NCTC_13129_gi_38232912
Brevibacterium_linens_BL2_gi_260906335
Geodermatophilus_obscurus_DSM_43160_gi_284991364
Salinispora_arenicola_CNS-205_gi_159038995
Salinispora_tropica_CNB-440_gi_145595744
0.27
0.89
0.83
0.35
0.77
0.84
0.6
0.32
0.85
0.91
0.81
0.43
0.69
0.98
0.99
0.7
0.077
0.94
0.094
0.99
0.96
0.43
0.55
0.74
0.2

Figure 3.2: NADH dehydrogenase-related gene phylogenies. Representative phylogenies for (A) the NDH-1 partial 2 operon (NuoM) and (B) the NDH-1 partial 3 operon (MrpD). Branch colors: orange = Actinobacteria, red = Proteobacteria, brown = Firmicutes, green = Chlorbi, Pink = Cyanobacteria, gray = Deinococus-Thermus, and black = other bacterial phyla. Names of marine bacteria are colored blue and non-marine black. Midpoint rooting was used and likelihood values shown for each node. Scale bar represents changes per site. See table 3.4 for detailed tree parameters. The NuoL homolog from *S. tropica* was used as an outgroup in the MrpD tree but is not shown.

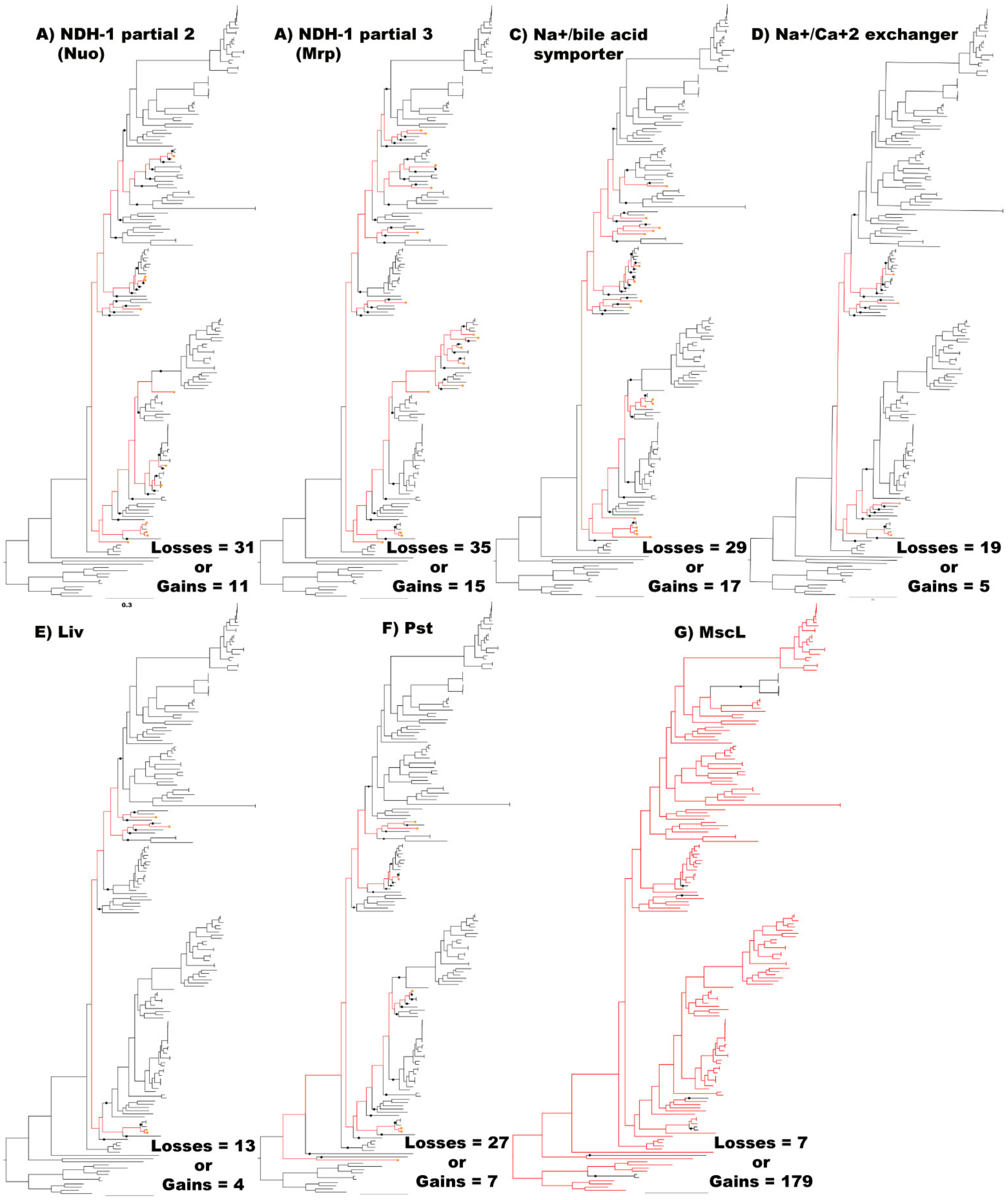


Figure 3.3: Phylogenetic distributions of marine adaptation genes (MAGs) among the Actinobacteria. Red branches in the species tree trace the occurrence of each MAG starting from the ancestor that accounts for all strains that maintain the MAG (+). Black circles indicate the point in a lineage within which all strains lack the MAG. The minimum number of gene loss events was calculated by summing the black circles. The maximum number of gene gain events was calculated by summing the red circles. Pst, Liv, Partial 2, and Mrp represent all genes in the respective operons. S = *Salinispora*.

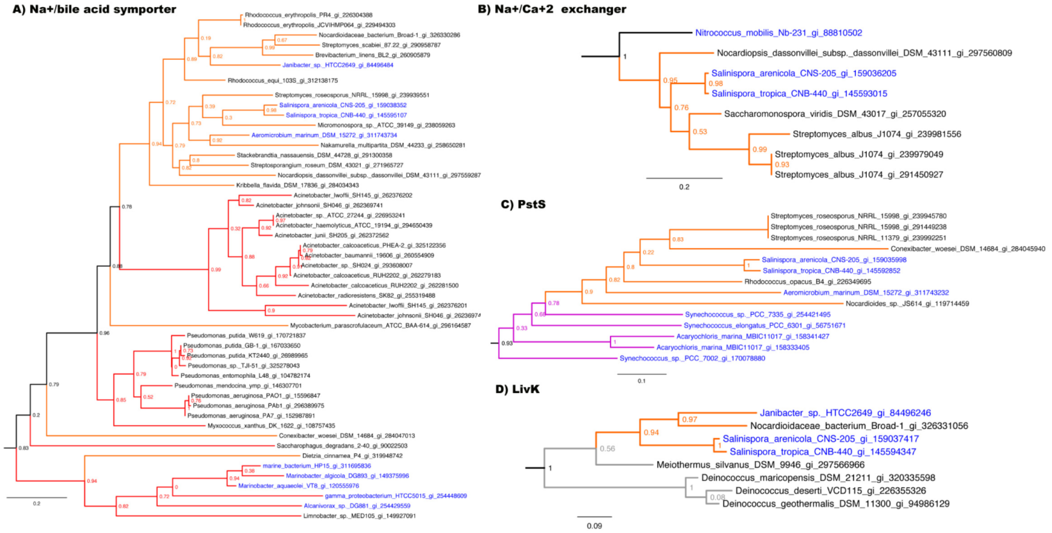

A) Na+/bile acid symporter
Rhodococcus_erythropolis_PR4_gi_226304388
Rhodococcus_erythropolis_JCVIHMP064_gi_229494303
Nocardioidaceae_bacterium_Broad-1_gi_326330286
Streptomyces_scabiei_87.22_gi_290958787
Brevibacterium_linens_BL2_gi_260905879
Janibacter_sp._HTCC2649_gi_84496484
Rhodococcus_equi_103S_gi_312138175
Streptomyces_roseosporus_NRRL_15998_gi_239939551
Salinispora_arenicola_CNS-205_gi_159038352
Salinispora_tropica_CNB-440_gi_145595107
Micromonospora_sp._ATCC_39149_gi_238059263
Aeromicrobium_marinum_DSM_15272_gi_311743734
Nakamurella_multipartita_DSM_44233_gi_258650281
Stackebrandtia_nassauensis_DSM_44728_gi_291300358
Streptosporangium_roseum_DSM_43021_gi_271965727
Nocardiopsis_dassonvillei_subsp._dassonvillei_DSM_43111_gi_297559287
Kribbella_flavida_DSM_17836_gi_284034343
Acinetobacter_lwoffii_SH145_gi_262376202
Acinetobacter_johnsonii_SH046_gi_262369741
Acinetobacter_sp._ATCC_27244_gi_226953241
Acinetobacter_haemolyticus_ATCC_19194_gi_294650439
Acinetobacter_junii_SH205_gi_262372562
Acinetobacter_calcoaceticus_PHEA-2_gi_325122356
Acinetobacter_baumannii_19606_gi_260554909
Acinetobacter_sp._SH024_gi_293608007
Acinetobacter_calcoaceticus_RUH2202_gi_262279183
Acinetobacter_calcoaceticus_RUH2202_gi_262281500
Acinetobacter_radioresistens_SK82_gi_255319488
Acinetobacter_lwoffii_SH145_gi_262376201
Acinetobacter_johnsonii_SH046_gi_26236974
Mycobacterium_parascrofulaceum_ATCC_BAA-614_gi_296164587
Pseudomonas_putida_W619_gi_170721837
Pseudomonas_putida_GB-1_gi_167033650
Pseudomonas_putida_KT2440_gi_26989965
Pseudomonas_sp._TJI-51_gi_325278043
Pseudomonas_entomophila_L48_gi_104782174
Pseudomonas_mendocina_ymp_gi_146307701
Pseudomonas_aeruginosa_PAO1_gi_15596847
Pseudomonas_aeruginosa_PAb1_gi_296389975
Pseudomonas_aeruginosa_PA7_gi_152987891
Myxococcus_xanthus_DK_1622_gi_108757435
Conexibacter_woesei_DSM_14684_gi_284047013
Saccharophagus_degradans_2-40_gi_90022503
Dietzia_cinnamea_P4_gi_319948742
marine_bacterium_HP15_gi_311695836
Marinobacter_algicola_DG893_gi_149375996
Marinobacter_aquaeolei_VT8_gi_120555976
gamma_proteobacterium_HTCC5015_gi_254448609
Alcanivorax_sp._DG881_gi_254429559
Limnobacter_sp._MED105_gi_149927091
0.2
B) Na+/Ca+2 exchanger
Nitrococcus_mobilis_Nb-231_gi_88810502
Nocardiopsis_dassonvillei_subsp._dassonvillei_DSM_43111_gi_297560809
Salinispora_arenicola_CNS-205_gi_159036205
Salinispora_tropica_CNB-440_gi_145593015
Saccharomonospora_viridis_DSM_43017_gi_257055320
Streptomyces_albus_J1074_gi_239981556
Streptomyces_albus_J1074_gi_239979049
Streptomyces_albus_J1074_gi_291450927
0.2
C) PstS
Streptomyces_roseosporus_NRRL_15998_gi_239945780
Streptomyces_roseosporus_NRRL_15998_gi_291449238
Streptomyces_roseosporus_NRRL_11379_gi_239992251
Conexibacter_woesei_DSM_14684_gi_284045940
Salinispora_arenicola_CNS-205_gi_159035998
Salinispora_tropica_CNB-440_gi_145592852
Rhodococcus_opacus_B4_gi_226349695
Aeromicrobium_marinum_DSM_15272_gi_311743232
Nocardioides_sp._JS614_gi_119714459
Synechococcus_sp._PCC_7335_gi_254421495
Synechococcus_elongatus_PCC_6301_gi_56751671
Acaryochloris_marina_MBIC11017_gi_158341427
Acaryochloris_marina_MBIC11017_gi_158333405
Synechococcus_sp._PCC_7002_gi_170078880
0.1
D) LivK
Janibacter_sp._HTCC2649_gi_84496246
Nocardioidaceae_bacterium_Broad-1_gi_326331056
Salinispora_arenicola_CNS-205_gi_159037417
Salinispora_tropica_CNB-440_gi_145594347
Meiothermus_silvanus_DSM_9946_gi_297566966
Deinococcus_maricopensis_DSM_21211_gi_320335598
Deinococcus_deserti_VCD115_gi_226355326
Deinococcus_geothermalis_DSM_11300_gi_94986129
0.09

Figure 3.4: Partial phylogenetic trees of four marine adaptation genes. (A) Na^+/bile acid symporter, (B) Na^+/Ca^{+2} exchanger, (C) PstS of the high affinity phosphate transporter, and (D) LivK from the branched chain amino acid transporter. Note: deep branches within the Actinobacteria are incongruent with the species phylogeny. Branches and species are colored as in Figure 3.2. See table 3.5 for detailed tree parameters.

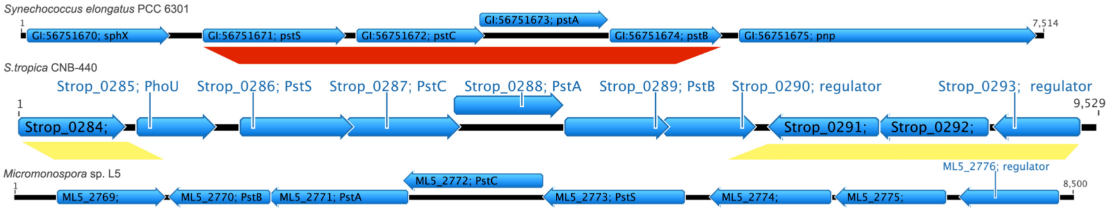


Figure 3.5: Phosphate transport (*pst*) operon and surrounding region in *S. tropica* CNB-440. Red box indicates synteny of *pst* between *S. tropica* CNB-440 and *Synechococcus elongatus* PCC6301. Yellow box indicates synteny between *S. tropica* CNB-440 and *Micromonospora* sp. L5. The GI numbers are listed for *S. elongatus* PCC6301, the locus tags are given for the *S. tropica* and *Micromonospora* genomes.

Tables

Table 3.1: Keyword searches and BLAST query sequences.

Functional categories and keywords

ETS (Complex I)	Sodium transporters	ABC transporters	Pores
NADH dehydrogenase	Na+	ABC	Pores
Nuo	Sodium		Channels
Nqr			Msc
Na+-quinone reductase			Porins
			Mechanosensitive

BLAST query sequences

Species	Gene	Accesion #	Functional category
Vibrio alginolyticus 40B	*nqrA*	ZP_06180303.1	Complex I
Vibrio alginolyticus 40B	*nqrB*	ZP_06180304.1	Complex I
Vibrio alginolyticus 40B	*nqrC*	ZP_06180305.1	Complex I
Vibrio alginolyticus 40B	*nqrD*	ZP_06180306.1	Complex I
Vibrio alginolyticus 40B	*nqrE*	ZP_06180307.1	Complex I
Vibrio alginolyticus 40B	*nqrF*	ZP_06180308.1	Complex I
Bacillus haloduran C-125	*MrpG*	NP_242179.1	Complex I
Bacillus haloduran C-125	*MrpF*	NP_242180.1	Complex I
Bacillus haloduran C-125	*MrpE*	NP_242181.1	Complex I
Bacillus haloduran C-125	*MrpD*	NP_242182.1	Complex I
Bacillus haloduran C-125	*MrpC*	NP_242183.1	Complex I
Bacillus haloduran C-125	*MrpB*	NP_242184.1	Complex I
Bacillus haloduran C-125	*MrpA*	NP_242185.1	Complex I
Escherichia coli K-12	*Ndh2*	NP_415627	Complex I
Micromonospora sp. L5	*MscL*	YP_004079991.1	Pores

Table 3.2: Marine adaptation genes. MAGs identified based on functional class and comparative genomics.

Species	MAG status	Functional class				Subtotal	Comparative genomics		Total
		ETS (complex 1)	Na+ transport	ABC tranport	Channels and pores		Gene gain	Gene loss	
S. tropica	Candidate	35	18	225	35	313	105	4	422
S. tropica	Final	13 (2 operons)	3	9 (2 operons)	30	55	7	1*	60**
S. arenicola	Candidate	35	19	225	33	312	105	4	421
S. arenicola	Final	13 (2 operons)	3	9 (2 operons)	28	53	7	1*	58**

*Based on annotation.

**Total is not additive because 3 of the 7 Genes in the gene gain category were also found in functional analysis.

ETS: electron transport system.

Table 3.3: Genes used for species tree construction.

Phylogenetic markers
rpsC
rplE
rplK
tsf
rplF
rplM
rplA
rpsK
rpmA
rplP
rplC
rpoB
rplD
rplL
rpsM
frr
rpsE
rplB
smpB

Table 3.4: Complete list of candidate MAGs. Those considered in the final MAG pool based on phylogentic links to marine bacteria are highlighted in gray.

		MAGs based on annotation and BLAST searches		
S. tropica gene	S. arenicola ortholog	Annotation*	Description*	Functional class
Stro0119	Sare0118	sodium/hydrogen exchanger		Sodium transporter
Stro0120	Sare0119	TrkA-C domain protein		Sodium transporter
Stro0242	Sare0283	TrkA-N domain protein		Sodium transporter
Stro0372	Sare0443	TrkA-N domain protein		Sodium transporter
Stro0373	Sare0444	H (+)-transporting two-sector ATPase		Sodium transporter
Stro0449	Sare0538	sodium/calcium exchanger membrane region		Sodium transporter
Stro0697	Sare0644	Na+ transporter		Sodium transporter
Stro1152	Sare2512	Na+/H+ antiporter NhaA		Sodium transporter
Stro1358	Sare1315	sodium:dicarboxylate symporter		Sodium transporter
Stro1485	Sare1450	TrkA-N domain protein		Sodium transporter
Stro1486	Sare1451	TrkA-N domain protein		Sodium transporter
Stro1666	Sare1658	Na+/H+ antiporter NhaA		Sodium transporter
Stro1844	Sare1837	sodium/hydrogen exchanger		Sodium transporter
Stro2118	Sare2262	Na+/H+ antiporter NhaA		Sodium transporter
Stro2582	Sare2779	Bile acid:sodium symporter		Sodium transporter
Stro4216	Sare4649	Na+/Ca+ antiporter, CaCA family		Sodium transporter
Stro4497	Sare5011	Na+ solute symporter		Sodium transporter
Stro4508	Sare5018	Na+ solute transporter		Sodium transporter
	Sare0724	Na+/solute symporter		Sodium transporter
Stro0390	Sare0461	nuoB	NDH-1 partial1	Electron transport
Stro0391	Sare0462			
Stro0392	Sare0463	nuoC	NDH-1 partial1	Electron transport
Stro0393	Sare0464	nuoH	NDH-1 partial1	Electron transport
Stro0394	Sare0466			
Stro0395	Sare1603			
Stro0396	Sare1602			
Stro0397	Sare1601			
Stro0398	Sare1600			
Stro0399	Sare0467			
Stro0400	Sare0468	nuoI	NDH-1 partial1	Electron transport
Stro0401	Sare0469	nuoJ	NDH-1 partial1	Electron transport
Stro0402	Sare0470	nuoK	NDH-1 partial1	Electron transport
Stro0403	Sare0471	nuoL	NDH-1 partial1	Electron transport
Stro0404	Sare0472	nuoM	NDH-1 partial1	Electron transport
Stro0405	Sare0473	nuoN	NDH-1 partial1	Electron transport
Stro0766	Sare0708	nuoA	NDH-1 partial2	Electron transport
Stro0767	Sare0709	Prophage tail		
Stro0768	Sare0710	nuoH	NDH-1 partial2	Electron transport
Stro0769	Sare0711	nuoJ	NDH-1 partial2	Electron transport
Stro0770	Sare0712	nuoK	NDH-1 partial2	Electron transport
Stro0771	Sare0713	nuoL	NDH-1 partial2	Electron transport
Stro0772	Sare0714	nuoM	NDH-1 partial2	Electron transport
Stro0773	Sare0715	nuoN	NDH-1 partial2	Electron transport
Stro3226	Sare3452	mrpG	NDH-1 partial3	Electron transport
Stro3227	Sare3453	mrpF	NDH-1 partial3	Electron transport
Stro3228	Sare3454	mrpE	NDH-1 partial3	Electron transport
Stro3229	Sare3455	mrpD	NDH-1 partial3	Electron transport
Stro3230	Sare3456	mrpC	NDH-1 partial3	Electron transport
Stro3231	Sare3457	mrpAB	NDH-1 partial3	Electron transport
Stro4052	Sare4450	nouN	NDH-1 complete	Electron transport
Stro4053	Sare4451	nuoM	NDH-1 complete	Electron transport
Stro4054	Sare4452	nuoL	NDH-1 complete	Electron transport
Stro4055	Sare4453	nuoK	NDH-1 complete	Electron transport
Stro4056	Sare4454	nuoJ	NDH-1 complete	Electron transport
Stro4057	Sare4455	nuoI	NDH-1 complete	Electron transport
Stro4058	Sare4456	nuoH	NDH-1 complete	Electron transport
Stro4059	Sare4457	nuoG	NDH-1 complete	Electron transport
Stro4060	Sare4458	nuoF	NDH-1 complete	Electron transport
Stro4061	Sare4459	nuoE	NDH-1 complete	Electron transport
Stro4062	Sare4460	nuoD	NDH-1 complete	Electron transport
Stro4063	Sare4461	nuoC	NDH-1 complete	Electron transport
Stro4064	Sare4462	nuoB	NDH-1 complete	Electron transport
Stro4065	Sare4463	nuoA	NDH-1 complete	Electron transport
Stro0165	Sare0174	binding protein	branched-chain amino acid	ABC transporter
Stro0216	Sare0255	binding protein	dipeptide/oligopeptide	ABC transporter
Stro0217	Sare0256	membrane	dipeptide/oligopeptide	ABC transporter
Stro0218	Sare0257	membrane	dipeptide/oligopeptide	ABC transporter
Stro0219	Sare0258	ABC	dipeptide/oligopeptide	ABC transporter
Stro0220	Sare0259	ABC	oligopeptide	ABC transporter
Stro0249		ABC	multidrug	ABC transporter
Stro0256	Sare0296	ABC	multidrug	ABC transporter
Stro0257	Sare0297	membrane	multidrug	ABC transporter
Stro0286	Sare0330	binding protein	phosphate	ABC transporter

Table 3.4 (continued)

MAGs based on annotation and BLAST searches				
S. tropica gene	S. arenicola ortholog	Annotation*	Description*	Functional class
Stro0287	Sare0331	membrane	phosphate	ABC transporter
Stro0288	Sare0332	membrane	phosphate	ABC transporter
Stro0289	Sare0333	ABC	phosphate	ABC transporter
Stro0397	Sare1601	ABC	multidrug	ABC transporter
Stro0431	Sare0519	membrane	polysaccharide export	ABC transporter
Stro0432	Sare0520	ABC	polysaccharide export	ABC transporter
Stro0435		ABC	sugar (maltose?)	ABC transporter
Stro0454		ABC	multidrug	ABC transporter
Stro0503		ABC	multidrug	ABC transporter
Stro0617	Sare4550	ABC	multidrug	ABC transporter
Stro0618		membrane	polysaccharide export	ABC transporter
Stro0750		ABC	? (Uup homolog/duplicated ATPase)	ABC transporter
Stro0751		ABC	? (Uup homolog/duplicated ATPase)	ABC transporter
Stro0759	Sare0701	binding protein	branched-chain amino acid	ABC transporter
Stro0761	Sare0703	membrane	branched-chain amino acid	ABC transporter
Stro0762	Sare0704	membrane	branched-chain amino acid	ABC transporter
Stro0763	Sare0705	ABC	sugar (ribose?)	ABC transporter
Stro0784	Sare0728	ABC	? (Uup homolog/duplicated ATPase)	ABC transporter
Stro0803	Sare0747	binding protein	sugar (glycerol-3-phosphate?)	ABC transporter
Stro0804	Sare0748	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro0805	Sare0749	membrane	sugar (maltose?)	ABC transporter
Stro0807	Sare0751	binding protein	sugar	ABC transporter
Stro0808	Sare0752	ABC	sugar	ABC transporter
Stro0809	Sare0753	membrane	sugar	ABC transporter
Stro0810	Sare0754	membrane	sugar	ABC transporter
Stro0822	Sare0766	membrane	CydC/CydD homolog	ABC transporter
Stro0823	Sare0767	membrane	CydC/CydD homolog	ABC transporter
Stro0842	Sare0785	membrane	multidrug	ABC transporter
Stro0843	Sare0786	membrane	multidrug	ABC transporter
Stro0982	Sare0918	ABC	cell division	ABC transporter
Stro0983	Sare0919	membrane	cell division	ABC transporter
Stro1118	Sare1008	membrane	cobalt ion	ABC transporter
Stro1119	Sare1009	ABC	cobalt ion	ABC transporter
Stro1187	Sare1080	binding protein	sugar (glycerol-3-phosphate?)	ABC transporter
Stro1188	Sare1081	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro1189	Sare1082	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro1201	Sare1093	ABC	molybdate	ABC transporter
Stro1342		ABC	efflux (antimicrobial peptide?)	ABC transporter
Stro1357		membrane	multidrug	ABC transporter
Stro1428	Sare1392	membrane	amino aicd (glutamine/glutamate/aspartate?)	ABC transporter
Stro1429	Sare1393	membrane	amino aicd (glutamine/glutamate/aspartate?)	ABC transporter
Stro1430	Sare1394	binding protein	amino acid (glutamine/glutamate/aspartate?)	ABC transporter
Stro1431	Sare1395	ABC	amino acid (glutamine/glutamate/aspartate?)	ABC transporter
Stro1524		membrane	toxin secretion	ABC transporter
Stro1557	Sare1506	binding protein	sugar (glycerol-3-phosphate?)	ABC transporter
Stro1558	Sare1507	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro1559	Sare1508	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro1628		ABC	cobalamin/Fe3+-siderophores	ABC transporter
Stro1629		membrane	ferric enterobactin	ABC transporter
Stro1630		membrane	cobalamin/Fe3+-siderophores	ABC transporter
Stro1631		binding protein	?	ABC transporter
Stro1633	Sare1620	binding protein	glycine betaine/L-proline/carnitine/choline	ABC transporter
Stro1634	Sare1621	membrane	glycine betaine/L-proline/carnitine/choline	ABC transporter
Stro1635	Sare1622	ABC	glycine betaine/L-proline/carnitine/choline	ABC transporter
Stro1636	Sare1623	membrane	glycine betaine/L-proline/carnitine/choline	ABC transporter
Stro1641	Sare1626	ABC	glycine betaine/L-proline	ABC transporter
Stro1642	Sare1627	membrane	glycine betaine/L-proline	ABC transporter
Stro1643	Sare1628	binding protein	glycine betaine/L-proline	ABC transporter
Stro1669	Sare1661	binding protein	dipeptide/oligopeptide	ABC transporter
Stro1670	Sare1662	membrane	dipeptide/oligopeptide	ABC transporter
Stro1671	Sare1663	membrane	dipeptide/oligopeptide	ABC transporter
Stro1672	Sare1664	ABC	dipeptide/oligopeptide	ABC transporter
Stro1673	Sare1665	ABC	oligopeptide	ABC transporter
Stro1685	Sare1680	ABC	polysaccharide export	ABC transporter
Stro1686	Sare1681	membrane	polysaccharide export	ABC transporter
Stro1794	Sare1780	membrane	sodium ion efflux	ABC transporter
Stro1795	Sare1781	ABC	sodium ion efflux	ABC transporter
Stro1801	Sare1791	ABC	branched-chain amino acid	ABC transporter
Stro1802	Sare1792	ABC	branched-chain amino acid	ABC transporter
Stro1803	Sare1793	binding protein	branched-chain amino acid	ABC transporter
Stro1804	Sare1794	membrane	branched-chain amino acid	ABC transporter
Stro1805	Sare1795	membrane	branched-chain amino acid	ABC transporter

Table 3.4 (continued)

MAGs based on annotation and BLAST searches				
S. tropica gene	S. arenicola ortholog	Annotation*	Description*	Functional class
Stro1913	Sare1904	binding protein	?	ABC transporter
Stro1967		binding protein	cobalamin/Fe3+-siderophores	ABC transporter
Stro1968		ABC	cobalamin/Fe3+-siderophores	ABC transporter
Stro1969		membrane	cobalamin/Fe3+-siderophores	ABC transporter
Stro2003	Sare2628	membrane	spermidine/putrescine	ABC transporter
Stro2004	Sare2629	membrane	sulfate	ABC transporter
Stro2005	Sare2630	ABC	spermidine/putrescine	ABC transporter
Stro2006	Sare2631	binding protein	iron(III)	ABC transporter
Stro2015	Sare2987	ABC	? (Uup homolog/duplicated ATPase)	ABC transporter
Stro2033	Sare2620	binding protein	manganese/zinc ion	ABC transporter
Stro2035	Sare2618	ABC	manganese/zinc ion	ABC transporter
Stro2036	Sare2617	membrane	manganese/zinc ion	ABC transporter
Stro2103	Sare2246	binding protein	cobalamin/Fe3+-siderophores	ABC transporter
Stro2175		ABC	multidrug	ABC transporter
Stro2208		membrane	multidrug	ABC transporter
Stro2233	Sare2350	membrane	?	ABC transporter
Stro2234	Sare2351	ABC	efflux (antimicrobial peptide?)	ABC transporter
Stro2401	Sare2550	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro2402	Sare2551	ABC	sugar (maltose?)	ABC transporter
Stro2403	Sare2552	binding protein	sugar (glycerol-3-phosphate?)	ABC transporter
Stro2404	Sare2553	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro2429	Sare2584	membrane	multidrug	ABC transporter
Stro2430	Sare2585	ABC	multidrug	ABC transporter
Stro2461		binding protein	oligopeptide	ABC transporter
Stro2462		binding protein	oligopeptide	ABC transporter
Stro2539	Sare2718	membrane	multidrug	ABC transporter
Stro2540	Sare2719	ABC	multidrug	ABC transporter
Stro2545	Sare2732	binding protein	molybdate	ABC transporter
Stro2546	Sare2733	membrane	molybdate	ABC transporter
Stro2547	Sare2734	ABC	spermidine/putrescine	ABC transporter
Stro2553	Sare2742	membrane	cobalamin/Fe3+-siderophores	ABC transporter
Stro2554	Sare2743	membrane	ferric enterobactin	ABC transporter
Stro2555	Sare2744	ABC	cobalamin/Fe3+-siderophores	ABC transporter
Stro2584	Sare2781	membrane	multidrug	ABC transporter
Stro2585	Sare2782	ABC	multidrug	ABC transporter
Stro2592	Sare2790	ABC	cobalt ion	ABC transporter
Stro2593	Sare2791	membrane	cobalt ion	ABC transporter
Stro2594	Sare2792	binding protein	cobalt ion	ABC transporter
Stro2650	Sare2076	ABC	dipeptide/oligopeptide	ABC transporter
Stro2651	Sare2075	membrane	dipeptide/oligopeptide	ABC transporter
Stro2652	Sare2074	membrane	dipeptide/oligopeptide	ABC transporter
Stro2653	Sare2073	binding protein	dipeptide/oligopeptide	ABC transporter
Stro2722		binding protein	sugar (glycerol-3-phosphate?)	ABC transporter
Stro2759		membrane	?	ABC transporter
Stro2760		ABC	efflux (antimicrobial peptide?)	ABC transporter
Stro2841		membrane	multidrug	ABC transporter
Stro2842		membrane	multidrug	ABC transporter
Stro2851		membrane	multidrug?	ABC transporter
Stro2945		ABC	? (Uup homolog/duplicated ATPase)	ABC transporter
Stro2981	Sare3205	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro2982	Sare3206	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro2983	Sare3207	binding protein	sugar (glycerol-3-phosphate?)	ABC transporter
Stro3071	Sare3298	membrane	multidrug	ABC transporter
Stro3074	Sare3301	ABC	? (Uup homolog/duplicated ATPase)	ABC transporter
Stro3080	Sare3307	ABC	? (Fe-S assembly/SufBCD system)	ABC transporter
Stro3082	Sare3309	membrane	? (Fe-S assembly/SufBCD system)	ABC transporter
Stro3083	Sare3310	membrane	? (Fe-S assembly/SufBCD system)	ABC transporter
Stro3146	Sare3373	ABC	branched-chain amino acid	ABC transporter
Stro3147	Sare3374	ABC	branched-chain amino acid	ABC transporter
Stro3148	Sare3375	membrane	branched-chain amino acid	ABC transporter
Stro3149	Sare3376	membrane	branched-chain amino acid	ABC transporter
Stro3150	Sare3377	binding protein	branched-chain amino acid	ABC transporter
Stro3162	Sare3387	ABC	multidrug	ABC transporter
Stro3163	Sare3388	membrane	multidrug	ABC transporter
Stro3177	Sare3402	ABC	multidrug	ABC transporter
Stro3178	Sare3403	membrane	multidrug	ABC transporter
Stro3180	Sare3406	membrane	multidrug	ABC transporter
Stro3181	Sare3407	membrane	multidrug	ABC transporter
Stro3262	Sare3492	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro3263	Sare3493	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro3264	Sare3494	binding protein	sugar (glycerol-3-phosphate?)	ABC transporter
Stro3421	Sare3798	membrane	polysaccharide export	ABC transporter
Stro3422	Sare3799	membrane	multidrug	ABC transporter
Stro3423	Sare3800	ABC	multidrug	ABC transporter

Table 3.4 (continued)

MAGs based on annotation and BLAST searches				
S. tropica gene	S. arenicola ortholog	Annotation*	Description*	Functional class
Stro3437	Sare3814	binding protein	manganese/zinc ion	ABC transporter
Stro3438	Sare3815	ABC	manganese/zinc ion	ABC transporter
Stro3439	Sare3816	membrane	manganese/zinc ion	ABC transporter
Stro3443	Sare3822	ABC	dipeptide/oligopeptide	ABC transporter
Stro3444	Sare3823	membrane	cobalt ion	ABC transporter
Stro3542	Sare3917	ABC	? (Uup homolog/duplicated ATPase)	ABC transporter
Stro3587	Sare3967	binding protein	sugar (glycerol-3-phosphate?)	ABC transporter
Stro3588	Sare3968	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro3589	Sare3969	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro3622	Sare4004	ABC	multidrug	ABC transporter
Stro3623	Sare4005	membrane	multidrug	ABC transporter
Stro3688	Sare4068	ABC	multidrug	ABC transporter
Stro3785	Sare4165	membrane	multidrug	ABC transporter
Stro3786	Sare4166	membrane	multidrug	ABC transporter
Stro3787	Sare4167	membrane	multidrug	ABC transporter
Stro3788	Sare4168	ABC	multidrug	ABC transporter
Stro3790	Sare4170	ABC	multidrug	ABC transporter
Stro3796	Sare4176	binding protein	oligopeptide	ABC transporter
Stro3797	Sare4177	membrane	dipeptide/oligopeptide	ABC transporter
Stro3798	Sare4178	membrane	dipeptide/oligopeptide	ABC transporter
Stro3799	Sare4179	ABC	dipeptide/oligopeptide	ABC transporter
Stro3800	Sare4180	ABC	oligopeptide	ABC transporter
Stro3819	Sare4209	ABC	oligopeptide	ABC transporter
Stro3820	Sare4210	ABC	dipeptide/oligopeptide	ABC transporter
Stro3821	Sare4211	membrane	dipeptide/oligopeptide	ABC transporter
Stro3822	Sare4212	membrane	dipeptide/oligopeptide	ABC transporter
Stro3823	Sare4213	binding protein	oligopeptide	ABC transporter
Stro3846	Sare4236	binding protein	sugar (xylose?)	ABC transporter
Stro3847	Sare4237	ABC	sugar (ribose?)	ABC transporter
Stro3848	Sare4238	membrane	sugar (xylose?)	ABC transporter
Stro3876	Sare4267	binding protein	branched-chain amino acid	ABC transporter
Stro3877	Sare4268	membrane	nitrate/sulfonate/taurine	ABC transporter
Stro3878	Sare4269	binding protein	nitrate/sulfonate/taurine	ABC transporter
Stro3879	Sare4270	ABC	nitrate/sulfonate/taurine	ABC transporter
Stro3891	Sare4282	ABC	? (Uup homolog/duplicated ATPase)	ABC transporter
Stro4079	Sare4499	ABC	cobalamin/Fe3+-siderophores	ABC transporter
Stro4080	Sare4500	membrane	cobalamin/Fe3+-siderophores	ABC transporter
Stro4081	Sare4501	binding protein	cobalamin/Fe3+-siderophores	ABC transporter
Stro4095	Sare4515	membrane	heme export	ABC transporter
Stro4130		ABC	multidrug	ABC transporter
Stro4171	Sare4597	membrane	cobalt ion	ABC transporter
Stro4172	Sare4598	ABC	sugar (ribose?)	ABC transporter
Stro4185		ABC	efflux (antimicrobial peptide?)	ABC transporter
Stro4186		membrane	lipoprotein releasing	ABC transporter
Stro4220	Sare4657	ABC	sugar (maltose?)	ABC transporter
Stro4231		binding protein	?	ABC transporter
Stro4232		membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro4233		membrane	spermidine/putrescine	ABC transporter
Stro4234		ABC	spermidine/putrescine	ABC transporter
Stro4288	Sare4723	binding protein	glycine betaine/L-proline/carnitine/choline	ABC transporter
Stro4289	Sare4724	membrane	glycine betaine/L-proline/carnitine/choline	ABC transporter
Stro4290	Sare4725	membrane	glycine betaine/L-proline/carnitine/choline	ABC transporter
Stro4291	Sare4726	ABC	glycine betaine/L-proline/carnitine/choline	ABC transporter
Stro4293	Sare4728	membrane	sugar (ribose?)	ABC transporter
Stro4336	Sare4778	binding protein	dipeptide/oligopeptide	ABC transporter
Stro4337	Sare4779	membrane	dipeptide/oligopeptide	ABC transporter
Stro4338	Sare4780	membrane	dipeptide/oligopeptide	ABC transporter
Stro4339	Sare4781	ABC	dipeptide/oligopeptide	ABC transporter
Stro4383	Sare4874	membrane	polysaccharide export	ABC transporter
Stro4384	Sare4875	ABC	multidrug	ABC transporter
Stro4399		membrane	cobalt ion	ABC transporter
Stro4400		ABC	dipeptide/oligopeptide	ABC transporter
Stro4410	Sare4885	membrane	toxin secretion	ABC transporter
Stro4423	Sare4898	ABC	multidrug	ABC transporter
Stro4444		membrane	multidrug	ABC transporter
Stro4445		membrane	multidrug	ABC transporter
Stro4500	Sare5012	ABC	multidrug	ABC transporter
Stro4501	Sare5013	membrane	?	ABC transporter
Stro4528	Sare5038	binding protein	branched-chain amino acid	ABC transporter
	Sare0178	binding protein	branched-chain amino acid	ABC transporter
	Sare0222	ABC	efflux (antimicrobial peptide?)	ABC transporter
	Sare0223	membrane	efflux (antimicrobial peptide)?	ABC transporter
	Sare0391	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
	Sare0392	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
	Sare0393	binding protein	sugar (glycerol-3-phosphate?)	ABC transporter

Table 3.4 (continued)

MAGs based on annotation and BLAST searches				
S. tropica gene	S. arenicola ortholog	Annotation*	Description*	Functional class
Stro3437	Sare3814	binding protein	manganese/zinc ion	ABC transporter
Stro3438	Sare3815	ABC	manganese/zinc ion	ABC transporter
Stro3439	Sare3816	membrane	manganese/zinc ion	ABC transporter
Stro3443	Sare3822	ABC	dipeptide/oligopeptide	ABC transporter
Stro3444	Sare3823	membrane	cobalt ion	ABC transporter
Stro3542	Sare3917	ABC	? (Uup homolog/duplicated ATPase)	ABC transporter
Stro3587	Sare3967	binding protein	sugar (glycerol-3-phosphate?)	ABC transporter
Stro3588	Sare3968	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro3589	Sare3969	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro3622	Sare4004	ABC	multidrug	ABC transporter
Stro3623	Sare4005	membrane	multidrug	ABC transporter
Stro3688	Sare4068	ABC	multidrug	ABC transporter
Stro3785	Sare4165	membrane	multidrug	ABC transporter
Stro3786	Sare4166	membrane	multidrug	ABC transporter
Stro3787	Sare4167	membrane	multidrug	ABC transporter
Stro3788	Sare4168	ABC	multidrug	ABC transporter
Stro3790	Sare4170	ABC	multidrug	ABC transporter
Stro3796	Sare4176	binding protein	oligopeptide	ABC transporter
Stro3797	Sare4177	membrane	dipeptide/oligopeptide	ABC transporter
Stro3798	Sare4178	membrane	dipeptide/oligopeptide	ABC transporter
Stro3799	Sare4179	ABC	dipeptide/oligopeptide	ABC transporter
Stro3800	Sare4180	ABC	oligopeptide	ABC transporter
Stro3819	Sare4209	ABC	oligopeptide	ABC transporter
Stro3820	Sare4210	ABC	dipeptide/oligopeptide	ABC transporter
Stro3821	Sare4211	membrane	dipeptide/oligopeptide	ABC transporter
Stro3822	Sare4212	membrane	dipeptide/oligopeptide	ABC transporter
Stro3823	Sare4213	binding protein	oligopeptide	ABC transporter
Stro3846	Sare4236	binding protein	sugar (xylose?)	ABC transporter
Stro3847	Sare4237	ABC	sugar (ribose?)	ABC transporter
Stro3848	Sare4238	membrane	sugar (xylose?)	ABC transporter
Stro3876	Sare4267	binding protein	branched-chain amino acid	ABC transporter
Stro3877	Sare4268	membrane	nitrate/sulfonate/taurine	ABC transporter
Stro3878	Sare4269	binding protein	nitrate/sulfonate/taurine	ABC transporter
Stro3879	Sare4270	ABC	nitrate/sulfonate/taurine	ABC transporter
Stro3891	Sare4282	ABC	? (Uup homolog/duplicated ATPase)	ABC transporter
Stro4079	Sare4499	ABC	cobalamin/Fe3+-siderophores	ABC transporter
Stro4080	Sare4500	membrane	cobalamin/Fe3+-siderophores	ABC transporter
Stro4081	Sare4501	binding protein	cobalamin/Fe3+-siderophores	ABC transporter
Stro4095	Sare4515	membrane	heme export	ABC transporter
Stro4130		ABC	multidrug	ABC transporter
Stro4171	Sare4597	membrane	cobalt ion	ABC transporter
Stro4172	Sare4598	ABC	sugar (ribose?)	ABC transporter
Stro4185		ABC	efflux (antimicrobial peptide?)	ABC transporter
Stro4186		membrane	lipoprotein releasing	ABC transporter
Stro4220	Sare4657	ABC	sugar (maltose?)	ABC transporter
Stro4231		binding protein	?	ABC transporter
Stro4232		membrane	sugar (glycerol-3-phosphate?)	ABC transporter
Stro4233		membrane	spermidine/putrescine	ABC transporter
Stro4234		ABC	spermidine/putrescine	ABC transporter
Stro4288	Sare4723	binding protein	glycine betaine/L-proline/carnitine/choline	ABC transporter
Stro4289	Sare4724	membrane	glycine betaine/L-proline/carnitine/choline	ABC transporter
Stro4290	Sare4725	membrane	glycine betaine/L-proline/carnitine/choline	ABC transporter
Stro4291	Sare4726	ABC	glycine betaine/L-proline/carnitine/choline	ABC transporter
Stro4293	Sare4728	membrane	sugar (ribose?)	ABC transporter
Stro4336	Sare4778	binding protein	dipeptide/oligopeptide	ABC transporter
Stro4337	Sare4779	membrane	dipeptide/oligopeptide	ABC transporter
Stro4338	Sare4780	membrane	dipeptide/oligopeptide	ABC transporter
Stro4339	Sare4781	ABC	dipeptide/oligopeptide	ABC transporter
Stro4383	Sare4874	membrane	polysaccharide export	ABC transporter
Stro4384	Sare4875	ABC	multidrug	ABC transporter
Stro4399		membrane	cobalt ion	ABC transporter
Stro4400		ABC	dipeptide/oligopeptide	ABC transporter
Stro4410	Sare4885	membrane	toxin secretion	ABC transporter
Stro4423	Sare4898	ABC	multidrug	ABC transporter
Stro4444		membrane	multidrug	ABC transporter
Stro4445		membrane	multidrug	ABC transporter
Stro4500	Sare5012	ABC	multidrug	ABC transporter
Stro4501	Sare5013	membrane	?	ABC transporter
Stro4528	Sare5038	binding protein	branched-chain amino acid	ABC transporter
	Sare0178	binding protein	branched-chain amino acid	ABC transporter
	Sare0222	ABC	efflux (antimicrobial peptide?)	ABC transporter
	Sare0223	membrane	efflux (antimicrobial peptide)?	ABC transporter
	Sare0391	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
	Sare0392	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
	Sare0393	binding protein	sugar (glycerol-3-phosphate?)	ABC transporter

Table 3.4 (continued)

MAGs based on annotation and BLAST searches				
S. tropica gene	S. arenicola ortholog	Annotation*	Description*	Functional class
	Sare0411	ABC	efflux (antimicrobial peptide?)	ABC transporter
	Sare0412	membrane	efflux (antimicrobial peptide)?	ABC transporter
	Sare0621	ABC	multidrug	ABC transporter
	Sare1429	ABC	spermidine/putrescine	ABC transporter
	Sare1430	membrane	iron(III)	ABC transporter
	Sare1431	binding protein	iron(III)	ABC transporter
	Sare2038	binding protein	dipeptide/oligopeptide	ABC transporter
	Sare2042	ABC	multidrug	ABC transporter
	Sare2043	membrane	multidrug	ABC transporter
	Sare2145	binding protein	dipeptide/oligopeptide	ABC transporter
	Sare2171	binding protein	oligopeptide	ABC transporter
	Sare2401	ABC	iron compound	ABC transporter
	Sare2402	membrane	iron compound	ABC transporter
	Sare2403	membrane	iron compound	ABC transporter
	Sare2404	binding protein	iron compound	ABC transporter
	Sare2447	membrane	multidrug	ABC transporter
	Sare2939	membrane	polysaccharide export	ABC transporter
	Sare2940	ABC	multidrug	ABC transporter
	Sare2954	binding protein	branched-chain amino acid	ABC transporter
	Sare2999	binding protein	sugar (glycerol-3-phosphate?)	ABC transporter
	Sare3000	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
	Sare3001	membrane	sugar (glycerol-3-phosphate?)	ABC transporter
	Sare3169	ABC	? (Uup homolog/duplicated ATPase)	ABC transporter
	Sare3224	binding protein	glycine betaine/L-proline	ABC transporter
	Sare3486	ABC	glycine betaine/L-proline	ABC transporter
	Sare3487	membrane	glycine betaine/L-proline	ABC transporter
	Sare3488	binding protein	glycine betaine/L-proline	ABC transporter
	Sare3644	binding protein	glycine betaine/L-proline	ABC transporter
	Sare4186	membrane	toxin secretion	ABC transporter
	Sare4381	binding protein	oligopeptide	ABC transporter
	Sare4551	membrane	multidrug	ABC transporter
	Sare4936	membrane	polysaccharide export	ABC transporter
	Sare4937	ABC	multidrug	ABC transporter
Stro0096	Sare0093	channel protein, hemolysin III family		Channels and pores
Stro0210	Sare1609	hypothetical protein		Channels and pores
Stro0320		Polymorphic membrane protein Chlamydia		Channels and pores
Stro0321		hypothetical protein		Channels and pores
Stro0495	Sare4609	MIP family channel protein		Channels and pores
Stro1045	Sare1605	polymorphic outer membrane protein		Channels and pores
Stro1127	Sare1020	hypothetical protein		Channels and pores
Stro1229		hypothetical protein		Channels and pores
Stro1297	Sare2925	hypothetical protein		Channels and pores
Stro1619		hypothetical protein		Channels and pores
Stro1620	Sare4599	Polymorphic membrane protein Chlamydia		Channels and pores
Stro1623		hypothetical protein		Channels and pores
Stro1626	Sare1615	hypothetical protein		Channels and pores
Stro1771	Sare1758	MscS Mechanosensitive ion channel		Channels and pores
Stro1861	Sare1854	guanylate kinase/L-type calcium channel region		Channels and pores
Stro2358	Sare2509	hypothetical protein		Channels and pores
Stro2511	Sare2695	Polymorphic membrane protein Chlamydia		Channels and pores
Stro2897		polymorphic outer membrane protein		Channels and pores
Stro3011	Sare1617	Polymorphic membrane protein Chlamydia		Channels and pores
Stro3059	Sare3285	hypothetical protein		Channels and pores
Stro3060	Sare3286	hypothetical protein		Channels and pores
Stro3399	Sare3646	polymorphic outer membrane protein		Channels and pores
Stro3406	Sare3654	Parallel beta-helix repeat		Channels and pores
Stro3407		hypothetical protein		Channels and pores
Stro3408		polymorphic outer membrane protein		Channels and pores
Stro3415	Sare4391	hypothetical protein		Channels and pores
Stro3536	Sare3911	MscS Mechanosensitive ion channel		Channels and pores
Stro3668		hypothetical protein		Channels and pores
Stro3982	Sare4370	polymorphic outer membrane protein		Channels and pores
Stro3987		hypothetical protein		Channels and pores
Stro3990	Sare4374	hypothetical protein		Channels and pores
Stro3992		hypothetical protein		Channels and pores
Stro4219		hypothetical protein		Channels and pores
Stro4332	Sare4774	hypothetical protein		Channels and pores
Stro4430		polymorphic outer membrane protein		Channels and pores
	Sare0383	conserved hypothetical protein		Channels and pores
	Sare1120	polymorphic outer membrane protein		Channels and pores
	Sare1610	polymorphic outer membrane protein		Channels and pores
	Sare3043	conserved hypothetical protein		Channels and pores
	Sare3075	polymorphic outer membrane protein		Channels and pores
	Sare3647	conserved hypothetical protein		Channels and pores
	Sare4375	parallel beta-helix repeat		Channels and pores
	Sare4376	parallel beta-helix repeat		Channels and pores
	Sare4397	conserved hypothetical protein		Channels and pores
	Sare4912	polymorphic outer membrane protein		Channels and pores
	Sare4920	conserved hypothetical protein		Channels and pores

*Annotation and descriptions for ABC transporters was generated by TransporterDB.

Table 3.4 (continued)

MAGs based on comparative genomics/gene gain		
S. tropica gene	S. arenicola ortholog	Annotation
Stro0170	Sare0183	Abortive infection protein
Stro2721	Sare2097	condensation domain protein
Stro0562	Sare1942	conserved hypothetical protein
Stro1168	Sare1038	conserved hypothetical protein
Stro2025	Sare2424	conserved hypothetical protein
Stro2948	Sare3172	conserved hypothetical protein
Stro3044	Sare3270	conserved hypothetical protein
Stro3047	Sare3273	conserved hypothetical protein
Stro4209	Sare4640	conserved hypothetical protein
Stro1659	Sare1644	cyclic nucleotide-binding
Stro0741	Sare2985	Endonuclease/exonuclease/phosphatase
Stro2359	Sare2111	GCN5-related N-acetyltransferase
Stro2057	Sare2175	helix-turn-helix- domain containing protein AraC type
Stro2693	Sare0560	helix-turn-helix- domain containing protein AraC type
Stro2026	Sare2425	Helix-turn-helix type 11 domain protein
Stro0143	Sare0149	hypothetical protein
Stro0210	Sare1609	hypothetical protein (polymorphic membrane protein)
Stro0488	Sare0616	hypothetical protein
Stro0506	Sare0613	hypothetical protein
Stro0655	Sare0631	hypothetical protein
Stro0686	Sare3053	hypothetical protein
Stro0995	Sare0944	hypothetical protein
Stro1055	Sare1652	hypothetical protein
Stro1127	Sare1020	hypothetical protein (polymorphic membrane protein)
Stro1163	Sare1031	hypothetical protein
Stro1297	Sare2925	hypothetical protein (polymorphic membrane protein)
Stro1356	Sare1314	hypothetical protein
Stro1419	Sare1385	hypothetical protein
Stro1510	Sare1461	hypothetical protein
Stro1514	Sare2972	hypothetical protein
Stro1515	Sare0304	hypothetical protein
Stro1626	Sare1615	hypothetical protein (polymorphic membrane protein)
Stro1652	Sare1637	hypothetical protein
Stro1899	Sare1375	hypothetical protein
Stro1935	Sare1589	hypothetical protein
Stro1970	Sare4619	hypothetical protein
Stro1976	Sare2994	hypothetical protein
Stro2071	Sare2214	hypothetical protein
Stro2219	Sare2346	hypothetical protein
Stro2315	Sare1284	hypothetical protein
Stro2417	Sare2569	hypothetical protein
Stro2423	Sare3332	hypothetical protein
Stro2574	Sare2771	hypothetical protein
Stro2575	Sare2772	hypothetical protein
Stro2627	Sare2824	hypothetical protein
Stro2664	Sare1587	hypothetical protein
Stro2666	Sare4809	hypothetical protein
Stro2705	Sare0547	hypothetical protein
Stro2893	Sare2473	hypothetical protein
Stro2966	Sare3187	hypothetical protein
Stro3013	Sare2722	hypothetical protein
Stro3043	Sare3269	hypothetical protein
Stro3050	Sare3276	hypothetical protein
Stro3383	Sare3625	hypothetical protein
Stro3401	Sare3649	hypothetical protein
Stro3402	Sare3650	hypothetical protein
Stro3415	Sare4391	hypothetical protein (polymorphic membrane protein)
Stro3416	Sare3683	hypothetical protein
Stro3508	Sare3884	hypothetical protein
Stro3658	Sare1433	hypothetical protein
Stro3954	Sare4345	hypothetical protein
Stro3986	Sare4371	hypothetical protein
Stro3990	Sare4374	hypothetical protein (polymorphic membrane protein)
Stro4000	Sare3775	hypothetical protein
Stro4001	Sare3772	hypothetical protein
Stro4003	Sare3677	hypothetical protein
Stro4004	Sare3777	hypothetical protein
Stro4010	Sare3785	hypothetical protein
Stro4012	Sare3789	hypothetical protein
Stro4126	Sare1582	hypothetical protein
Stro4174	Sare0438	hypothetical protein
Stro4176	Sare4942	hypothetical protein
Stro4377	Sare4865	hypothetical protein
Stro4379	Sare4870	hypothetical protein
Stro4386	Sare1026	hypothetical protein
Stro4436	Sare4910	hypothetical protein
Stro4419	Sare4893	Kynurenine 3-monooxygenase
Stro3046	Sare3272	Lantibiotic dehydratase domain protein
Stro3054	Sare3280	Lantibiotic dehydratase domain protein
Stro1179	Sare1072	LPXTG-motif cell wall anchor domain
Stro2231	Sare2348	major facilitator superfamily MFS_1
Stro2926	Sare3129	major facilitator superfamily MFS_1
Stro2962	Sare3195	major facilitator superfamily MFS_1
Stro4112	Sare4533	major facilitator superfamily MFS_1
Stro1136	Sare3663	Methionine adenosyltransferase
Stro2648	Sare2078	Methyltransferase type 12
Stro2649	Sare2077	Methyltransferase type 12
Stro2920	Sare3120	MMPL domain protein
Stro4148	Sare3014	nucleoside diphosphate kinase
Stro3406	Sare3654	Parallel beta-helix repeat (polymorphic membrane protein)
Stro1466	Sare1426	peptidase S8 and S53 subtilisin kexin sedolisin
Stro3049	Sare3275	peptidase U62 modulator of DNA gyrase
Stro1139	Sare2468	protein of unknown function DUF129
Stro2701	Sare0549	protein of unknown function DUF1702
Stro1269	Sare1159	protein of unknown function DUF397
Stro1517	Sare2969	protein of unknown function DUF397
Stro1588	Sare1548	protein of unknown function DUF397
Stro4463	Sare1786	protein of unknown function DUF397
Stro3985	Sare4369	protein of unknown function DUF81
Stro2327	Sare2449	protein phosphatase 2C domain protein
Stro1418	Sare1384	response regulator receiver
Stro4181	Sare4603	steroid delta-isomerase domain protein
Stro3056	Sare3282	thioester reductase domain
Stro2328	Sare2450	UbiC transcription regulator-associated domain protein
Stro2706	Sare0559	Wyosine base formation domain protein

Table 3.4 (continued)

MAGS based on comparative genomics/gene loss**	
Micromonospora L5 gene ID	Annotation
2501944239	peptidoglycan synthetase FtsI
2501945835	phosphate ABC transporter substrate-binding protein, PhoT family
2501943698	HAD-superfamily hydrolase, subfamily IIB
2501942438	large conductance mechanosensitive channel protein

**Gene are from the *Micromonospora* L5 genome

Table 3.5: Tree parameters and statistics generated by MABL.

Protein	Gene used for BLAST	Length of gene used for BLAST	Positions in alignment	Positions after Gblocks alignment	Patterns in alignment	Proportion of invariant sites	Gamma shape parameter	Tree log likelihood
MrpAB	Stro3231	916	1391	407	358	0.141	1.219	-20602.74
MrpC	Stro3230	107	360	57	57	0.091	1.159	-3483.29
MrpD	Stro3229	488	699	323	310	0.068	1.236	-24038.83
MrpE	Stro3228	129	258	30	30	0	0.982	-2450.32
MrpF	Stro3227	85	152	16	16	0.127	0.872	-974.48
MrpG.SA	Sare3452	118	251	39	39	0	0.958	-2389.19
NuoA	Stro766	114	170	73	71	0.055	1.201	-3861.47
NuoH	Stro768	309	465	214	209	0.042	1.372	-14834.48
NuoJ	Stro769	187	292	140	126	0.131	2.158	-2800.21
NuoK	Stro770	100	168	89	89	0.01	1.848	-4004.51
NuoL	Stro771	621	1123	118	118	0.114	1.022	-6835.58
NuoM	Stro772	496	700	273	252	0.09	0.918	-15371.32
NuoN	Stro773	462	738	214	207	0.074	1.662	-13803.99
LivF	Stro1802	230	293	223	190	0.113	0.694	-10801.13
LivG	Stro1801	253	328	236	200	0.174	0.796	-10396.82
LivH	Stro1804	293	711	256	218	0.133	0.834	-9589.13
LivK	Stro1803	417	464	353	339	0.039	1.164	-16775.30
LivM	Stro1805	337	446	232	202	0.102	0.796	-10218.97
PstA	Stro288	306	845	169	159	0.069	0.977	-10307.90
PstB	Stro289	288	315	227	185	0.207	0.742	-12227.82
PstC	Stro287	315	396	206	193	0.086	0.997	-13075.50
PstS	Stro286	315	1171	199	191	0.075	1.172	-17448.51
Na$^+$/Ca^{+2} exchanger	Stro449	347	706	127	124	0.028	1.122	-10830.54
Na$^+$/bile symporter	Stro2582	296	575	138	136	0.006	0.831	-12479.00
Na$^+$/Ca^{+2} antiporter	Stro4216	352	446	205	187	0.084	0.774	-10742.77
Species tree proteins	N/A	N/A	5927	3367	2654	0.176	0.837	-210637.61

Chapter 4: The abundance and expression of secondary metabolism genes in marine plankton reveals new phylogenetic diversity of protistan like ketosynthase domains

Abstract

Metagenomics research has provided evidence for genetic diversity that was not observed in culture-based studies. As modern sequencing methods provide deeper sequencing of environmental DNA, tools are needed to exploit as much data as possible. A recently designed online tool called Natural Product Domain Seeker is well suited to identify ketosynthase (KS) domains from polyketide synthases and condensation (C) domains from non-ribosomal peptide synthetases from short sequence reads. This tool was applied to metagenomes from distinct water masses collected off the coast of California and metatranscriptomes from California surface waters and Antarctic plankton collected beneath sea ice. Analyses of these metagenomic data provide evidence for extensive new KS diversity associated with protists and offers evidence that different bodies of water contain different amounts of natural product-related genes. By comparing different size fractions, it was revealed that larger size particles contain more bacterial KS and C domains related to secondary metabolism. The domains from the metagenomic sequences are closest phylogenetically to KS domains from genomes of cultured marine bacteria thus suggesting that the KS and C domains detected are specific to marine environments.

The results of this study emphasize that metagenomic approaches can provide new insight into environments that may be rich in organisms that produce new natural products.

Introduction

Metagenomic analyses have identified thousands of genes that were previously unknown from the genomes of cultured organisms and provided insight into novel biological processes at work in the environment (Béjà et al. 2000; Rusch et al. 2007). In addition to finding novel genes, metagenomic approaches have allowed scientists to characterize different marine habitats by focusing on the most abundant genes in the environment such as genes for the metabolism of carbohydrates and amino acids (Dinsdale et al. 2008). Metagenomic approaches have also been used to measure the enormous taxonomic diversity in the environment (Kembel et al. 2011) that was originally discovered through PCR amplification of the rRNA genes from environmental DNA for example (Penn et al. 2006; Sogin et al. 2006). Metatranscriptomics is an extension of the metagenomic concept in which cDNA is reverse transcribed from environmental RNA. This approach has expanded our understanding of the dynamics of gene expression under natural conditions (Frias-Lopez et al. 2008). Metagenomics has been well suited to study marine habitats and helped overcome the difficulty of making in situ observations of marine microbes.

Despite the growing understanding of natural assemblages of organisms in the marine environment, no study has systematically studied the abundance and distribution of secondary metabolite genes, which likely play an important ecological role (Penn et al. 2009). These genes are relatively rare in metagenomic datasets and thus have not been a focus of past studies. In addition, bioinformatics tools by which informed interpretations of secondary metabolism could be made from highly fragmented datasets were generally not available. The recently released web tool called the Natural Product Domain Seeker (NaPDoS) provides a new opportunity to analyze the secondary metabolite genes associated with metagenomic and metatranscriptomic datasets (npdomainseeker.sdsc.edu). This tool can extract and classify sequence tags from two types of natural product biosynthetic pathways, polyketide synthases and non-ribosomal peptide synthetases. It can be used to determine the abundance, diversity, and expression of potential secondary metabolite genes in DNA or amino acid sequence data.

Polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs) are large enzyme families that account for many clinically important pharmaceutical agents. These enzymes sequentially construct a diverse array of natural products from relatively simple carboxylic acid and amino acid building blocks using an assembly line process (Finking and Marahiel 2004; Hertweck 2009). The molecular architectures of PKS and NRPS genes have been reviewed in detail and minimally consist of activation (AT or A), thiolation (ACP or PCP), and condensation (KS or C) domains (Shen 2003; Lautru and Challis 2004; Weissman 2004; Sieber and Marahiel

2005). These genes are among the largest found in microbial genomes and can include highly repetitive modules that create considerable challenges to accurate assembly and subsequent bioinformatics analysis (Udwary et al. 2007). Many tools have been developed to analyze complete PKS and NRPS genes and their associated gene clusters (Bachmann et al. 2009; Yadav et al. 2009; Medema et al.). The web tool NaPDoS extracts and rapidly classifies KS and C domains from a wide range of sequence data (Appendix B). NaPDoS is well suited to study secondary metabolism from metagenomes obtained using next generation sequencing technologies because the sequences that it targets are small relative the entire proteins.

Although PKS and NRPS products are well known to treat human diseases, the ecological functions of these compounds remain largely unknown. In a few cases, it is known that secondary metabolites function in defense and communication (Oh et al. 2009) (Böhnert et al. 2004). These functions were determined for simple systems including two or three organisms. The abundance and expression of secondary metabolites in complex ecological settings also remains unclear. The planktonic environment represents a highly competitive environment where resources and structures are ephemeral (Azam and Malfatti 2007). It is thus logical that microbes in some instances to help compete for space and nutrients would use chemical warfare (Long and Azam 2001).

This study uses NaPDoS to identify and classify KS and C domains in sequence data derived from planktonic marine communities. Metagenomic datasets (DNA) were generated for three different filtrate size classes collected at seven locations off the coast of California (Allen et al. 2012). Two metatranscriptomic datasets (cDNA) were generated from filtered plankton samples. One metatranscriptome was created from four water samples collected from a dinoflagellate bloom of *Lingulodinium polyedrum* off the coast of California and four samples collected from water beneath ice in Antarctica. The results of the phylogenetic classifications show the complexities of the evolutionary history of the KS and C domains. They also reveal the prevalence of fatty acid biosynthesis in both prokaryotes and eukaryotes relative to secondary metabolism.

Most of the KS domains identified were classified as protist like modular ketosynthases domains. As expected, most eukaryotic KS domains were found on the larger filter sizes. However, most of the bacterial KS and C domains related to secondary metabolite production were also found on the larger filter sizes suggesting that the bacteria associated with particulate matter contain more secondary metabolites conceivably because chemical defenses are needed to occupy such a niche. The numbers of both KS and C domains varied among sample sites and include novel groups within known functional classes. The C domains identified were mostly related to siderophore biosynthesis. The KS and C domains associated with secondary metabolism were the most abundant in a sample labeled “aged-upwelled” and was previously determined to have increased levels of Actinobacteria. Although notably missing from the metagenomes are KS and C domains from the Actinobacteria modular domain class, which are the typical natural product producers (Berdy 2005). C domains with exact matches to the siderophore pyoverdine were found in the expression data from Antarctica providing the first expression data of NRPS genes in a natural setting.

Methods

The metagenomic and metatranscriptomic datasets were derived from samples collected by researchers at the J. Craig Venter Institute (JCVI). The samples were also processed and the sequences generated at this facility. I was given access to these sequences for the studies described in this chapter. Some details about how the samples were collected and processed are provided below for clarification.

Metagenome sampling. Metagenome samples were collected from seven sites during a CalCOFI cruise in July 2007 (Allen et al. 2012). Three distinct size classes were created for each sample by filtering seawater through a 200 μm nytex-net followed by 3.0 μm, 0.8μm and 0.1 μm Supor 293mm disc filters (Pall Life Sciences, Ann Arbor, MI, USA). The DNA was extracted from each Supor filter and sequenced with a combination of Sanger and 454 GS FLX Titanium sequencer (Allen et al. 2012). The sequences were not assembled and open reading frames were predicted by metagene (Noguchi et al. 2006).

Metatranscriptome sampling. Different stages of a dinoflagellate bloom were sampled during CalCOFI cruise transects in April, May and June 2010. A 20 μm plankton net was towed for approximately one km four times through a red tide composed of *Lingulodinum polyedrum* (Lisa Allen pers. comm.). The Antarctica samples were collected in January and November 2009 from underneath sea ice and serially filtered in a similar manner as the metagenomic dataset. All RNA samples were flash frozen in liquid Nitrogen for processing later. RNA was amplified in a linear fashion and converted to cDNA for sequencing (Frias-Lopez et al. 2008). The cDNA was sequenced with a 454 GS FLX Titanium sequencer.

Analysis of KS and C domains. The complete scheme for the analysis of both DNA (metagenomes) and cDNA (metatranscriptomes) is presented in figure 4.1. The Basic Local Alignment Search Tool (BLAST) algorithm (Altschul et al. 1990), with an e-value cutoff set at $<1e^{-5}$, using the KS or C domain reference sequences as a query (see methods below) was used to identify candidate KS or C domains from the CalCOFI dataset. The online tool NaPDoS confirmed the sequences as KS or C domains and assigned an initial domain classification (Appendix B). After clustering by CD-hit (Huang et al.), one reference sequence from each cluster and all singletons were further classified based on their phylogenetic relationships with the reference KS and C domain sequences in the NaPDoS database and their top BLAST hits (see methods below). All phylogenetic trees were constructed with FastTree (Price et al.) then visualized and manipulated with archaeopteryx (Han and Zmasek 2009).

An initial blastx of the metatranscriptome cDNA, with an evalue $<1e^{-5}$ and a low complexity sequence filter (Wootton and Federhen 1993), was performed against the NaPDoS reference dataset to identify a pool of candidate KS or C domains. The low complexity filter helped eliminate matches to repetitive sequence that can be present in cDNA libraries. The matches were confirmed by NaPDoS, which was also used to generate amino acid sequences for the domains detected. Often times more than one reading frame would have a match to a domain but in most cases all reading frame matches had the same NaPDoS classification and thus this did not affect the results. If there were discrepancies between reading frame classifications, the longest blast match would be considered the proper classification.

Generation of reference datasets. A carefully aligned reference dataset of KS and C domain sequences that are linked to the production of specific compounds (natural products) was compiled as part of a separate study (Appendix B). This dataset was used in initial blast searches of the meta-DNA and cDNA data. To generate a more comprehensive set of KS and C domains, the NaPDoS reference datasets were used in a blastp search against the NCBI non-redundant (nr) protein database. All protein sequences with an evalue of $<1e^{-5}$ were collected into a fasta file. A search using Hidden Markov Models (HMMs), for KS and C domains was then performed to provide the coordinates of the domains within the proteins (Eddy 2009). The HMM match cutoff was set at an e-value $<1e^{-5}$. All nr KS and C domains were used as a database for further comparison against the metagenomic data. This approach was used because it identifies mostly complete domains and prevents false positives that may go undetected without manual curation.

Results

*Metagenome and Metatranscriptome sequencin*g. Meta data and filter sizes associated with the metagenome and metatranscriptome samples are presented in table 4.1. The number of sequences from each dataset and sample are shown in table 4.2. Data for the metagenome sequences includes predicted open reading frames from the JCVI prokaryotic annotation pipeline. Samples for the metatranscriptomes were collected from Antarctica and off the coast of California (Table 4. 1). The two metatranscriptomes were not initially translated and all analysis was done on the 454 outputs. The average length of sequence for the metagenomes is 304 base pairs (bp) and the average for the metatranscriptomes is 281 bps for all Antarctica sequences and 339 bps for all dinoflagellate bloom sequences.

CalCOFI KS domains. A blastp analysis of the metagenomic data against the NaPDoS reference KS sequences (Table 4. 3) was used to identify a pool of 2774 KS domains (Table 4.4). After NaPDoS analysis, there remained 2750 KS domains. These were assigned an initial domain classification based on top blast match in the NaPDoS database (Table 4.5). The size of the sequence pool decreased to 1080 following clustering with the program CD-hit and elimination of the sequences that were <124 amino acids long (Table 4.5). The initial NaPDoS classification indicated that 598 (55%) of the CalCOFI metagenome sequences were fatty acid synthases (FASs). These sequences were analyzed separately to confirm their classification. After removal of FASs, the remaining 482 KS domains were used for a blastp against the nr KS database and the top two blast hits were added to the set of sequences for phylogenetic analysis. The phylogenetic tree contained 513 unique nr KS hits, 197

NaPDoS KS reference sequences and 482 CalCOFI KS domains (Figure 4.2). The NaPDoS reference sequences from each domain class fall into distinct clades on the phylogenetic tree. Therefore, a domain classification can be assigned when non-reference sequences fall within a clade containing reference sequences. After phylogenetic analysis, 97 CalCOFI sequences were classified as KS domains related to those involved with known bacterial secondary metabolism. Notably, none of the KS domains had a high level of sequence identity to KS sequences from nr and none showed convincing evidence that they are from the well-studied actinobacterial KS modular clade.

The NaPDoS classification system recognizes 10 bacterial KS domain classes (Appendix B). The FAS and PUFAs classes are not considered associated with secondary metabolism. Of the remaining eight domain classes, representatives of six are found in the CalCOFI metagenomes (Table 4.5). Most of these sequences (337) were classified as modular. After phylogenetic analysis, 303 of the 337 sequences formed a distinct sister lineage to the bacterial modular KS clade. Based on the annotation of related sequences derived from NCBI, this lineage can be defined as a modular protist KS clade (Figure 4.2)(John et al. 2008; Monroe and Van Dolah 2008; Sasso et al. 2011). Part of the protist modular clade contains 50 CalCOFI sequences that group among KS domains derived from genome sequences of the protist Chlorophytes *Micromonas*, *Volvox*, *Ostreococcus* and *Aureococcus* (Figure 4.2). An analysis of the PKS genes from which these domains were obtained confirms they are modular. Eleven CalCOFI sequences are identical to KS domains from the

Ostreococcus lucimarinus CCE9901 genome (Palenik et al. 2007). However, 250 CalCOFI sequences form a diverse and distinct branch within the modular protist clade (Figure 4.3A), likely reflecting the vast diversity of uncultured protistan plankton species that have no genome data in nr (Worden 2006). Ten sequences group among the NaPDoS modular (all bacteria) reference sequences. Only one sequence groups within the Actinobacteria modular clade although the branch is anomalously long relative to others in the group (Figure 4.2).

Between the eukaryotic and bacterial modular clade are 24 CalCOFI sequences comprising five distinct lineages that do not contain any of the NaPDoS reference sequences. These are colored as unclassified in figure 4.2. Eukaryotic and bacterial KS sequences obtained from nr also fell into these sequences and thus were labeled "mixed modular" in figure 4.2. Four metagenome sequences in the mixed modular group form a cluster with domains from a modular PKS in *Vibrio nigripulchritudo* (Figure 4.3B). Seven CalCOFI sequences in this group are most closely related to KS domains from genome sequences in the eukaryotes *Aureococcus anophagefferens, Ectocarpus silculosus, Karenia brevis and Micromonas*. Thirteen CalCOFI sequences are nearest to KS domains from the bacterial genomes of *Terdinibacter turnerae*, *Legionella pneumophila* and *Burkholderia ambifaria*. Other KS domains in the mixed modular group were derived from eukaryotes including the sponge *Discodermia dissoluta*, the apicomplexans *Toxoplasma gondii,* and *Cryptosporidium gondii*, and the fungus *Neospora caninum.* This clade is similar to the one previously described in which multiple eukaryotic KS clades were observed (John et al. 2008; Monroe and Van Dolah 2008).

There are 37 KS domains found within the iterative, trans-AT, or hybrid classes of KS domains (Figure 4.4 and 5). One of the iterative CalCOFI sequences groups with a KS domain derived from the genome of *Synechococcus* sp. CC9311, which has never had a PKS type secondary metabolite described (Figure 4.4). Another iterative KS domain groups with a sequence from the genome of *Teredinibacter turnerae*, a marine bacterium that thrives on decomposing wood and is known to contain several secondary metabolite gene clusters although this KS domain has not been linked to a specific molecule (Yang et al. 2009). Eight CalCOFI sequences fall within the hybrid KS clade. Six of these show a close affiliation with the KS domains that produce yersiniabactin (Figure 4.5A). One hybrid KS sequence is closely related to a sequence from the genome of *Lyngbya* and another to a KS domain found in the *Rhodobacter* genome sequence (Figure 4.5A). The trans-AT clade contains 19 CalCOFI sequences that are distinct from any reference or nr sequences however, they clearly fall within the trans-AT clade (Figure 4.5B).

The NaPDoS reference sequences delineate three distinct groups of type II secondary metabolite related KS domains, called alpha, beta and JamG-CurC. The CalCOFI metagenome has 11 KS beta type domains only one of which groups with the reference KS beta domains (Figure 4.6). The remaining ten beta KS domains form a distinct group that is sister to the reference KS beta domains and contains KS domains from cultured strains of the marine bacterium *Pirellula*. There are two

CalCOFI sequences that group with JamG-CurC sequences (data not shown). These are classified as modular in NaPDoS due to the structures of the genes in which they reside however they are distantly related to the type II clade and are believed to be involved with decarboxylation as opposed to condensation reactions (Appendix B).

The majority of identified KS sequences were initially classified as FASs. To confirm the FAS classification, the sequences were placed in a phylogenetic tree separate from the rest of the KS domains. The tree confirmed these sequences as FASs and showed the vast taxonomic and phylogenetic diversity of FAS sequences from California plankton communities (data not shown). Many of the sequences had high percent identity to known FAS sequences. For example many sequences group with *Pelagibacter* (Allen et al. 2012), as would be expected because this genus is known to be a dominant member of the plankton along with *Prochlorococcus* (Allen et al. 2012), which also has many FASs closely related to it.

The CalCOFI sequences were clustered prior to phylogenetic classification with CD-hit at a 90% threshold therefore sequences may actually be in the metagenome more times than the phylogenetic tree shows. Consequently, all of the secondary metabolite KS domains were checked to determine how many other sequences were in their cluster. Based on results of clustering, three sequences that are classified as KS beta are part of clusters. One cluster contains four sequences, one has three sequences and one has two sequences. Three sequences classified as trans KSs occurred more than once, one sequence was part of a three-sequence cluster and two sequences were part of different two-sequence sized clusters. The hybrids, JamG-

CurC and unclassified KS clade each have a representative sequence that is part of a two-sequence cluster.

Each CalCOFI sample site was in a different nutrient state and contained different groups of bacteria (Allen et al. 2012) therefore the number of KS domains from each site was counted to determine if differences were observable in secondary metabolite distribution. GS258, a site composed of "aged up-welled" water and dominated by Actinobacteria (Allen et al. 2012), contained twice the number of KS sequences related to secondary metabolism than any other site and when normalized to total bases per sample still showed the highest percentage (Figure 4.7A). The different size filters were also analyzed to determine the numbers and types of KS sequences detected. As expected, the modular protist domains were observed from the two largest filter sizes. The smallest size fraction contained the most FAS sequences (Figure 4.7B), although the number difference is not as dramatic as the modular protist class likely because FASs are found in both prokaryotic and eukaryotic genomes and bacteria may remain attached to larger particles during filtration. The smallest size fraction contained the least number of KSs associated with secondary metabolism while the middle size contained the most. When combined, the two largest size fractions contain triple the amount of bacterial type secondary metabolite KSs suggesting that secondary metabolites are more abundant in particle-associated bacteria.

CalCOFI C domains. A blastp search of the NaPDoS reference C domains against the CalCOFI metagenome found 301 candidate C domains (Table 4.4). These sequences were analyzed in the NaPDoS pipeline and 194 sequences (Table 4.4) were confirmed as C domains. These sequences were clustered with CD-hit at a 90% threshold, resulting in 109 clusters or singletons. One representative of each cluster and all singletons were then subjected to phylogenetic analysis to assign a final classification (Table 4.6). The phylogenetic tree revealed 59 LCL, 25 DCL, 5 cyclization, 6 dual, 11 epimerases, and 3 starter domains (Figure 4.8).

A BLAST analysis of the CalCOFI C domains against nr revealed that one of the cyclization C domains has ≥90% sequence identity (data not shown) to a cyclization C domain in *Vibrio anguillarum* 775. This domain is in the *angR* gene, which is a biochemically verified cyclization C domain that participates in the biosynthesis of the siderophore anguibactin (Di Lorenzo et al. 2004). The remaining blast hits in the nr database have no more than 79% sequence identity to the CalCOFI C domain sequences (data not shown).

Phylogenetic analysis of the CalCOFI condensation domains reveals six sequences in the LCL clade that are closely related to sequences from *Pseudoalteromonas tunicata* D2 (Figure 4.9A). All of these sequences are from the 0.8 μm and 3.0 μm filters, which make sense because *P. tunicata* is thought to reside on living surfaces (Thomas et al. 2008). A set of 11 sequences in the LCL clade branches with C domains from a predicted cyclic peptide in *Salinispora arenicola,* a marine obligate bacterium typically from tropical sediments (Penn et al. 2009). Two sequences group with C domains from the cyclomarin biosynthetic pathway albeit distantly and with relatively low branch support (Figure 4.9B). Eleven CalCOFI sequences fall in the epimerase C domain clade. These are all distantly related to any of the reference sequences (Figure 4.10).

The number of C domains from each CalCOFI site and filter size were computed for each domain class and the fraction of domains relative to the total number of analyzed sequences were graphed (Figure 4.11A). Site GS258, the same site that has the most KS domains, has the largest number of C domains and happens to be the site with increased Actinobacteria relative to other sample sites (Allen et al. 2012). Site GS257 has the least amount of C domains related to bacterial secondary metabolism. The diversity of C domains appears to remain high at sites with fewer sequences except at site GS260 where diversity was reported to be extremely low with planctomycete bacteria dominating (Allen et al. 2012). As reported for the KS domains, the largest size fractions contain the most C domains.

Metatranscriptomes KS domains. Based on a blastx of all metatranscriptomes versus the reference KS dataset, 97 KS domains were found (Table 4.4). These sequences were placed in NaPDoS for verification and classification leaving 96 confirmed KS sequences. In the Antarctica metatranscriptome, 66 KS domains were classified as FAS and two as modular. The dinoflagellate bloom had more diversity of KS domain types with seven FAS, six modular, one KS1, one trans, two hybrids and one iterative. All of these sequences except four were smaller then 124 amino acids after translation and were thus characterized based on blast hits to the database of KS domain sequences compiled from nr (Table 4.7).

There were 66 Antarctic KS domains with a eukaryotic top blast hit and 12 with a top hit to bacteria (Table 4.7). The Antarctica data has 11 sequences greater then 90% sequence identity to nr KS domains. One close match was observed in the genome sequence of *Maribacter* sp. HTCC2170, a Flavobacterium from Oregon coastal water and another was observed in *Robiginitalea biformata,* a bacterium isolated from the Sargasso Sea. Both of these sequences are FASs. The remaining high percent matches are to KS domains derived from genomes of the micro-eukaryotes *Thalassiosira pseudonana* CCMP1335, *Phaeodactylum tricornutum* CCAP 1055/1, and *Aureococcus anophagefferens.* The KS domains from the microeukaryotes are classified as FAS by NaPDoS.

The dinoflagellate bloom metatranscriptome contains 11 top hits to bacteria and 17 top hits to eukaryotes. Five top hits are to *Streptomyces*. One *Streptomyces* like sequence is classified as FAS and the other four are classified as modular. The percent identity for three are ~30% but the fourth has 64% identity to a sequence observed in *Streptomyces cyaneogriseus subsp. noncyanogenus* over an 81 amino acid alignment, thus pointing to the expression of at least one modular KS domain typically associated with natural product biosynthesis in the Actinobacteria. Of the eukaryote related sequences, one of the dinoflagellate bloom sequences had 91% sequence identity to a *Thalassiosira pseudonana* CCMP1335 sequence; once again, this is classified as FAS by NaPDoS. All other eukaryotic sequences from the dinoflagellate bloom dataset have <68% sequence identity to genome sequences derived from

Karenia brevis, *Cryptosporidium muris*, *Aureococcus anophagefferens*, *Salpingoeca* sp. ATCC 50818, and *Pseudopfiesteria shumwayae* (Table 4.7).

Metatranscriptomes C domains. The initial blastx detected 92 C domain in the metatranscriptomes. These sequences were placed in NaPDoS for verification and classification leaving 21 confirmed C domains that were further analyzed. NaPDoS classified four LCL, three DCL, three epimerases, two starter and one dual C domain while one domain could not be classified. In the dinoflagellate bloom, the diversity was low compared to the Antarctic data with six DCL and one dual C domain (Table 4.7). There are nine sequences in the Antarctica data set that have >90% sequence identity to C domains in *Pseudomonas fluorescens* Pf-5 (Table 4.7). Eight of these sequences have better then 92% identity (Table 4.7) to domains of the siderophore pyoverdine biosynthetic cluster (Paulsen et al. 2005).

Discussion

The online tool NaPDoS was used to identify a wide diversity of KS and C domains from a metagenomic dataset collected from the surface waters off the coast of California. Phylogenetic analysis using reference sequences and a database of domains from nr was used to classify the metagenomic KS and C domains. None of the bacterial KS domains are similar to sequences from an experimentally characterized pathway and thus no predictions can be made about the potential small molecules they may produce. The low abundance of secondary metabolism genes relative to fatty acid biosynthesis is probably indicative of the limited distribution of polyketide synthases across bacterial phyla (Jenke-Kodama et al. 2005). The larger number of condensation domains is similarly indicative of their broader distribution of NRPSs in bacterial phyla (Rausch et al. 2007). Despite the small percentage of secondary metabolism genes relative to total sequences, some patterns related to location, specificity and particle size emerged from the data. Expression data contained only four KS domains with weak links to secondary metabolism and there was no clear evidence that C domains are abundant in metatranscriptomic data. Six transcripts were found that have between 92-100% sequence identity (Table 4.7) to different C domains from the pyoverdine biosynthetic pathway (Meyer 2000) suggesting that bacteria in this sample are responding to the iron-limiting conditions typical in the southern ocean (Church et al. 2000) by producing siderophores (Hopkinson and Barbeau).

KS domains

A clade of modular KS domains related to those observed in marine protists contained new diversity that may represent biosynthetic pathways of never before detected natural products. Inspection of the alignment revealed the active site cysteine is present in these KS domains (data not shown). The sequences in this group have uniform long branches and, because variation of function is related to evolutionary distance, these KS domains may represent many novel KS biosynthetic functions. Alternatively, this diversity may reflect the taxonomic diversity of protists (Worden 2006). Two pieces of evidence support that these sequences are from protists first, the most closely related sequences are from multi-modular PKS proteins in protists and second the sequences are most abundant in larger size fractions. The closely related proteins are from the genomes of the Chlorophytes: *Ostreococcus*, *Volvox*, *Chlamydomonas*, *Chlorella*, and *Micromonas*. The CalCOFI sequences are closely related to 13 modular polyketide synthases that have between 9 and 12 KS domains. Transcripts for the protist group of KS domains have been observed from *Karenia brevis* (Monroe and Van Dolah 2008) and *Chrysochromulina polylepis* (John et al. 2010) but specific compounds have not been linked to these genes. Furthermore, BLAST has been used to identify protist KSs in metagenomes before, but the phylogeny of these was not constructed (John et al. 2008). Despite their large abundance in the metagenomes, no protist KSs were detected in the metatranscriptomes. Surprisingly no PKS or NRPS expression was observed in the dinoflagellate bloom, as these blooms are known to contain toxins that are likely produced by PKS or NRPS biosynthetic pathways (John et al. 2010).

Although 10 bacterial modular sequences were detected, all were phylogenetically nearest to single KS domain proteins from genome sequences of Cyanobacteria (data not shown). Are there really no multi-modular actinobacterial KS domains (the ones responsible for so many natural products) in marine plankton communities? Analysis of environmentally derived 16S rRNA sequences showed that Actinobacteria are present in marine plankton (Jensen and Lauro 2008) but these Actinobacteria have not been cultured and may not produce polyketides. However Actinobacteria have been cultured from the sea but mostly sediments (Prieto-Davó et al. 2008). The most important clue to the lack of Actinobacterial modular KS domains is that sequences with > 60% GC content in all CalCOFI data are present at <0.01%, all Actinomycetes have between 60% and 70% GC content (Allen et al. 2012). Thus as shown in (Figure 4.2) the Actinobacteria modular clade has very short branch lengths and is a highly derived group of sequences. This is not the first study to search for KS domains from the environment. Other culture-independent studies have looked for but not found sequences closely related to the Actinobacterial modular KS clade. One metagenomics study of the sponge *Cymbastela concentrica* found only three genes identified by COG as related to secondary metabolism and none of these were non-ribosomal peptide synthetases or polyketide synthases (Thomas et al.). Using a direct PCR approach did not do much better as only five KS domains were retrieved from the marine sponge *Pseduoceratina clavata* (Kim and Fuerst 2006). However, a metagenomic analysis of the sponge *Discodermia dissolute* detected actinobacterial modular KS domains through a targeted approach where only fosmids with KS domains were sequenced (0.7% of the clones in the fosmid library contained PKS genes) (Schirmer et al. 2005).

Perhaps there is a methodological problem related to the missing Actinobacterial clade. This may be related to an extensive secondary structure in the sequences that prevent proper primer annealing during linker addition for 454 sequencing or possibly, it is related to the fact that high GC sequences do not sequence well (Dabney and Meyer 2012). Furthermore, not detecting domains from Actinobacteria may be related to DNA isolation methods, which are biased against spores and thick peptidoglycan layers present in Gram-positive bacteria (Mincer et al. 2005).

Even if modular Actinobacterial KS domains are not widespread in marine plankton the methods still detected other types of KS domains. Interestingly, four CalCOFI sequences in the mixed modular clade contained moderate branch support for recent common ancestry with four KS domains from the genome of *Vibrio nigripulchritudo* (Figure 4.4). *Vibrio* is a marine genus of bacteria thus it is not surprising to find metagenomic sequences from the sea that group with *Vibrio* KS domains (Figure 4.3B). Recent work has shown that *Vibrio* bacteria make a number of secondary metabolites and live attached to different particulates (Mansson et al.; Preheim et al.). Although the PKS from *V. nigripulchritudo* has not been linked to a molecule, one can speculate that this is related to an antibiotic that helps *Vibrio* spp. compete for space. Furthermore, it could be inferred that the CalCOFI sequences, which were found on large filter sizes, are coding for a similar antibiotic possibly in another *Vibrio* species.

While all the CalCOFI sequences have low similarity to bacterial secondary metabolite KS domains, some of the results suggest that the marine environment contains specific KS types relative to other habitats. For example, two iterative, six hybrid and all of the type II beta KS sequences group with bacteria derived from the marine environment. Besides chemical warfare, iron limitation in the ocean may be one ecological pressure that causes bacteria to have specific KS domains. For example, the CalCOFI KS sequences in the hydrid class group with KS domains involved with the biosynthesis of yersiniabactin, a type of siderophore. In addition, the abundance of halides and a basic environment may provide a selective pressure for bacteria to modify KS domains relative to non-marine habitats.

Although the trans-AT KS domains do not group with any cultured bacteria in the nr database, a completely novel clade with moderate support for the branch is present (Figure 4.5B). The trans-AT clade has recently been defined (for a review (Piel)) and therefore the diversity is less explored compared to for example the modular Actinobacteria clade.

CalCOFI sites GS257 and GS264 contain most of the FASs but contain the least amount of KS diversity (Figure 4.7). The number of FASs at these sites may indicate that bacteria with very few secondary metabolite genes are abundant. Although GS258 has the most bacterial KS domains, all other sites have very similar numbers of secondary metabolite KS domains. Perhaps related to the increased KS domains in GS258 is the increased numbers, albeit still a small fraction, of Actinobacteria in this sample (Allen et al. 2012). Eukaryotic modular sequences are the most abundant in GS263 this may be related to the increased amount of Chla, NO 3 and silicate (SiO3) in the region where this sample was collected (Allen et al. 2012).

This project did show that the initial blast search done by NaPDoS effectively classifies FASs. Thus, NaPDoS is useful to identify and separate FAS sequences from KS domains related to secondary metabolism. It also showed that unique KS and C domains still await discovery either through culturing and alternate or improved metagenomic methods. Many of the FASs that were analyzed grouped with genome sequences of *Pelagibacter* but there was an enormous diversity of sequences from other sequenced genomes from the marine environment. Only a few secondary metabolite genes were part of a CD-hit cluster and most clusters contained FASs. Interestingly, the modular protist sequences were not found in any of the clusters, conceivably because genomes of protists are large and thus sampling the same sequence twice is less likely.

The total FAS can be used to estimate the number of bacterial KS domains per genome based on a few assumptions and observations. Assume that every bacterial genome has on average two FASs (*fabF* and *fabB*) which explains the larger number of FAS relative to other bacterial KS domains. Then half the measured FAS domains can be used as an estimate for total genomes sampled. The number of secondary metabolism KS domains per genome is unknown. This calculation is valid because the sequences are both the same length. Therefore, the total measured secondary metabolism KS domains divided by the estimated number of genomes; based on FASs gives the total number of KS domains per genome. The result is an estimate of 18, 45 and 36 KS domains per 100 genomes for the 0.1, 0.8 and 3.0 μm filters, respectively. The estimate of KS per genome in each filter size indicates that the bacteria associated with particles contain more KS domains. Unfortunately the estimated number of KSs per genome does not distinguish whether KS domains are more numerous in some genomes or present in many different genomes.

C domains

Condensation domains present a slightly less confusing process of identifying specific domain classes of secondary metabolites compared to KS domains. NaPDoS does not detect C domains that are not related to secondary metabolism and thus there are no "false positives" as in the case of fatty acid biosynthesis. Also NRPS genes seem to be limited to Bacteria and Fungi. In addition, there are fewer C domain classes. Fewer known domain classes may mean that there is a higher chance to discover novel functional classes. In fact, a novel C domain clade was detected but low branch support makes it so that the uniqueness cannot be unequivocally determined. Longer sequences of these domains are needed to obtain better resolution. In addition, C domains have larger evolutionary distances compared to the KS phylogeny. This may reflect more flexibility in the protein's ability to tolerate changes while retaining function relative to KS domains.

More genes identified by blast were lost after NaPDoS analysis for C domains then for KS domains. Likely because NaPDoS has only C domains specific to secondary metabolite type C domains but a blast search may be finding sequences that code for a similar type of amino-acyl condensation reaction.

The LCL group is the largest group of C domains (Figure 4.8). The two groups that are shown in figure 4.9 are distinct; figure 4.9A shows a group of closely related sequences. The gi numbers for the *Pseudoalteromonas tunicata* proteins indicate that C domains in this group are all in the same chromosomal region and thus could be part of the same NRPS biosynthetic pathway. Although the branch support is quite low for the nodes, the relatively short branches give some indication that the CalCOFI C domains are coming from a larger biosynthetic pathway similar to the one in *P. tunicata*. The CalCOFI sequences come from three different sites but are present in the two larger size fractions once again pointing to particulate associated bacteria containing secondary metabolite pathways. Additionally both LCL groups (Figure 4.9) contain CalCOFI sequences closely related to marine bacteria from nr indicating that there is not a lot of movement of these genes from marine to non-marine environments. The epimerases represent a large diversity of distantly related sequences and again top hits are to marine sequences.

The distribution of C domains at each CalCOFI site is similar to the KS secondary metabolite distribution. Site GS258 has the most C domains and GS257 has the least amount. The other sites have similar amount of C domains, as was the case for KS except GS259 has a higher fraction of C domains than the remaining four sites. By using the FAS based estimate of genomes sampled per filter size an estimate of 19, 41 and 55 C domains per 100 genomes for the 0.1, 0.8 and 3.0 μm filters, respectively, was determined. The larger number of C domains in larger filter sizes signifies that the bacteria associated with particulate matter contain the most number of C domains as was determined for bacterial KS domains.

Conclusions

The known diversity of KS and C domains related to secondary metabolism are poorly represented in metagenomes and practically non-existent in metatranscriptomes. This can be interpreted to mean that secondary metabolism is not a major factor in marine plankton communities or the datasets analyzed did not contain enough sequence data to capture these genes. Regardless of the problems and reasons that I did not detect a large number of secondary metabolite genes, the genes that were found reveal that there is still considerable diversity that has yet to be linked to specific secondary metabolites. This study also shows that in order to get a complete picture of secondary metabolism in plankton all size fractions should be studied.

This study is not quantitative in the sense that I did not normalize the number of domains relative to other proteins. However, the number of genome equivalents was previously calculated for the different filter sizes and in the 1μm filters there are approximately 100 genome equivalents per filter and the larger filter sizes has ~40 genome equivalents. The fewer genomes equivalents in larger filter sizes scales equally, although different numerically from the FAS based calculation and supports the trend that few bacterial genomes were sampled in larger filter sizes and corroborates that more KS domains are in bacteria associated with particulate. To say KS and C domains are rare would be perhaps an overstatement. However out of 1.6 billion bases sequenced, 14kb are part of genes likely dedicated to secondary metabolism. Future studies of KS and C domains from uncultured bacteria therefore should focus on bacteria associated with particulates and target samples dominated by Actinobacteria.

Metagenomics and metatranscriptomics have been touted as a way to access the massive uncultured diversity of microbes in our world. And polyketides and non-ribosomal peptides have provided modern medicine with amazing cures for what were once fatal diseases. Undoubtedly if metagenomic approaches and natural product discovery can be successfully combined, a completely new revolution in natural product chemistry can begin. However based on this and other studies, the prospects to access natural products through metagenomics remain unfulfilled

Acknowledgements

Chapter 4, in part, contains methods, data and results that is submitted to PLoS ONE, 2012, Nadine Ziemert, Sheila Podell, Kevin Penn, Eric E. Allen, Paul R. Jensen. The dissertation author conceived the methods and aided in parts of the analyses, which forms the basis of this publication. The portions of chapter 4 related to analysis of metagenomes and metatranscriptomes will be submitted to an academic journal. I would like to acknowledge my collaborators at JCVI who collected and supported the collection and sequencing of all the metagenomes and metatransciptomes, specifically I would like to thank Dr. Andrew Allen, Dr. Lisa Zeigler, and James K. McCarthy for providing me with the sequence data for which I based my analyses.

References

Allen LZ, Allen EE, Badger JH, McCrow JP, Paulsen IT, Elbourne LDH, Thiagarajan M, Rusch DB, Nealson KH, Williamson SJ, Venter JC, Allen AE (2012) Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. ISME J.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215(3): 403-410.

Azam F, Malfatti F (2007) Microbial structuring of marine ecosystems. Nat Rev Micro 5(10): 782-791.

Bachmann BO, Ravel J, David AH (2009) Chapter 8 Methods for In Silico Prediction of Microbial Polyketide and Nonribosomal Peptide Biosynthetic Pathways from DNA Sequence Data. Methods in Enzymology: Academic Press. pp. 181-217.

Béjà O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, Spudich EN, DeLong EF (2000) Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea. Science 289(5486): 1902-1906.

Berdy J (2005) Bioactive Microbial Metabolites. J Antibiot 58(1): 1-26.

Böhnert HU, Fudal I, Dioh W, Tharreau D, Notteghem J-L, Lebrun M-H (2004) A Putative Polyketide Synthase/Peptide Synthetase from Magnaporthe grisea Signals Pathogen Attack to Resistant Rice. The Plant Cell Online 16(9): 2499-2513.

Church MJ, Hutchins DA, Ducklow HW (2000) Limitation of Bacterial Growth by Dissolved Organic Matter and Iron in the Southern Ocean. Applied and Environmental Microbiology 66(2): 455-466.

Dabney J, Meyer M (2012) Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. Biotechniques 52(2): 87-94.

Di Lorenzo M, Poppelaars S, Stork M, Nagasawa M, Tolmasky ME, Crosa JH (2004) A Nonribosomal Peptide Synthetase with a Novel Domain Organization Is Essential for Siderophore Biosynthesis in *Vibrio anguillarum*. Journal of Bacteriology 186(21): 7327-7336.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F (2008) Functional metagenomic profiling of nine biomes. Nature 452(7187): 629-632.

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform 23(1): 205-211.

Finking R, Marahiel MA (2004) Biosynthesis of nonribosomal peptides. Annual Review of Microbiology 58: 453 - 488.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, DeLong EF (2008) Microbial community gene expression in ocean surface waters. Proceedings of the National Academy of Sciences 105(10): 3805-3810.

Han M, Zmasek C (2009) phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinformatics 10(1): 356.

Hertweck C (2009) The Biosynthetic Logic of Polyketide Diversity. Angewandte Chemie International Edition 48(26): 4688-4716.

Hopkinson BM, Barbeau KA (2012) Iron transporters in marine prokaryotic genomes and metagenomes. Environmental Microbiology 14(1): 114-128.

Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26(5): 680-682.

Jenke-Kodama H, Sandmann A, Müller R, Dittmann E (2005) Evolutionary Implications of Bacterial Polyketide Synthases. Molecular Biology and Evolution 22(10): 2027-2039.

Jensen P, Lauro F (2008) An assessment of actinobacterial diversity in the marine environment. Antonie van Leeuwenhoek 94(1): 51-62.

John U, Beszteri S, Glockner G, Singh R, Medlin L, Cembella AD (2010) Genomic characterisation of the ichthyotoxic prymnesiophyte Chrysochromulina polylepis, and the expression of polyketide synthase genes in synchronized cultures. European Journal of Phycology 45(3): 215-229.

John U, Beszteri B, Derelle E, Van de Peer Y, Read B, Moreau H, Cembella A (2008) Novel Insights into Evolution of Protistan Polyketide Synthases through Phylogenomic Analysis. Protist 159(1): 21-30.

Kembel SW, Eisen JA, Pollard KS, Green JL (2011) The Phylogenetic Diversity of Metagenomes. PLoS ONE 6(8): e23214.

Kim TK, Fuerst JA (2006) Diversity of polyketide synthase genes from bacteria associated with the marine sponge *Pseudoceratina clavata*: culture-dependent and culture-independent approaches. Environmental Microbiology 8(8): 1460-1470.

Lautru S, Challis GL (2004) Substrate recognition by nonribosomal peptide synthetase multi-enzymes. Microbiology 150(Pt 6): 1629 - 1636.

Long RA, Azam F (2001) Antagonistic Interactions among Marine Pelagic Bacteria. Applied and Environmental Microbiology 67(11): 4975-4983.

Mansson M, Gram L, Larsen TO (2011) Production of Bioactive Secondary Metabolites by Marine Vibrionaceae. Marine Drugs 9(9): 1440-1468.

Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Research 39(suppl 2): W339-W346.

Meyer J-M (2000) Pyoverdines: pigments, siderophores and potential taxonomic markers of fluorescent *Pseudomonas* species. Archives of Microbiology 174(3): 135-142.

Mincer TJ, Fenical W, Jensen PR (2005) Culture-Dependent and Culture-Independent Diversity within the Obligate Marine Actinomycete Genus Salinispora. Applied and Environmental Microbiology 71(11): 7019-7028.

Monroe EA, Van Dolah FM (2008) The Toxic Dinoflagellate *Karenia brevis* Encodes Novel Type I-like Polyketide Synthases Containing Discrete Catalytic Domains. Protist 159(3): 471-482.

Noguchi H, Park J, Takagi T (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Research 34(19): 5623-5630.

Oh D-C, Poulsen M, Currie CR, Clardy J (2009) Dentigerumycin: a bacterial mediator of an ant-fungus symbiosis. Nat Chem Biol 5(6): 391-393.

Palenik B, Grimwood J, Aerts A, Rouze P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S, Zhou K, Otillar R, Merchant SS, Podell S, Gaasterland T, Napoli C, Gendler K, Manuell A, Tai V, Vallon O, Piganeau G, Jancek Sv, Heijde M, Jabbari K, Bowler C, Lohr M, Robbens S, Werner G, Dubchak I, Pazour GJ, Ren Q, Paulsen I, Delwiche C, Schmutz J, Rokhsar D, Van de Peer Y, Moreau H, Grigoriev IV (2007) The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. Proceedings of the National Academy of Sciences 104(18): 7705-7710.

Paulsen IT, Press CM, Ravel J, Kobayashi DY, Myers GSA, Mavrodi DV, DeBoy RT, Seshadri R, Ren Q, Madupu R, Dodson RJ, Durkin AS, Brinkac LM, Daugherty SC, Sullivan SA, Rosovitz MJ, Gwinn ML, Zhou L, Schneider DJ, Cartinhour SW, Nelson WC, Weidman J, Watkins K, Tran K, Khouri H, Pierson EA, Pierson LS, Thomashow LS, Loper JE (2005) Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. Nat Biotech 23(7): 873-878.

Penn K, Wu D, Eisen JA, Ward N (2006) Characterization of Bacterial Communities Associated with Deep-Sea Corals on Gulf of Alaska Seamounts. Applied and Environmental Microbiology 72(2): 1680-1683.

Penn K, Jenkins C, Nett M, Udwary DW, Gontang EA, McGlinchey RP, Foster B, Lapidus A, Podell S, Allen EE, Moore BS, Jensen PR (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. ISME J 3(10): 1193-1203.

Piel J (2010) Biosynthesis of polyketides by trans-AT polyketide synthases. Natural Product Reports 27(7).

Preheim SP, Timberlake S, Polz MF (2011) Merging Taxonomy with Ecological Population Prediction in a Case Study of Vibrionaceae. Applied and Environmental Microbiology 77(20): 7195-7206.

Price MN, Dehal PS, Arkin AP (2010) FastTree 2 -Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 5(3): e9490.

Prieto-Davó A, Fenical W, Jensen PR (2008) Comparative actinomycete diversity in marine sediments. Aquat Microb Ecol 52(1): 11.

Rausch C, Hoof I, Weber T, Wohlleben W, Huson D (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. BMC Evolutionary Biology 7(1): 78.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers Y-H, Falcón LI, Souza V, Bonilla-Rosso Gn, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC (2007) The *Sorcerer* II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biology 5(3): e77.

Sasso S, Pohnert G, Lohr M, Mittag M, Hertweck C (2011) Microalgae in the postgenomic era: a blooming reservoir for new natural products. FEMS Microbiology Reviews: 1574-6976.

Schirmer A, Gadkari R, Reeves CD, Ibrahim F, DeLong EF, Hutchinson CR (2005) Metagenomic Analysis Reveals Diverse Polyketide Synthase Gene Clusters in Microorganisms Associated with the Marine Sponge *Discodermia dissoluta*. Applied and Environmental Microbiology 71(8): 4840-4849.

Shen B (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. Current Opinion in Chemical Biology 7(2): 285-295.

Sieber SA, Marahiel MA (2005) Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics. Chem Rev 105(2): 715 - 738.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored , "rare biosphere". Proceedings of the National Academy of Sciences 103(32): 12115-12120.

Thomas T, Rusch D, DeMaere MZ, Yung PY, Lewis M, Halpern A, Heidelberg KB, Egan S, Steinberg PD, Kjelleberg S (2010) Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. ISME J 4(12): 1557-1567.

Thomas T, Evans FF, Schleheck D, Mai-Prochnow A, Burke C, Penesyan A, Dalisay DS, Stelzer-Braid S, Saunders N, Johnson J, Ferriera S, Kjelleberg S, Egan S (2008) Analysis of the *Pseudoalteromonas tunicata* Genome Reveals Properties of a Surface-Associated Life Style in the Marine Environment. PLoS ONE 3(9): e3252.

Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W, Jensen PR, Moore BS (2007) Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. Proceedings of the National Academy of Sciences 104(25): 10376-10381.

Weissman KJ (2004) Polyketide biosynthesis: understanding and exploiting modularity. Philosophical Transactions of the Royal Society of London Series A: Mathematical, Physical and Engineering Sciences 362(1825): 2671-2690.

Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. Computers & Chemistry 17(2): 149-163.

Worden AZ (2006) Picoeukaryote diversity in coastal waters of the Pacific Ocean. Aquatic Microbial Ecology 43(2): 165-175.

Yadav G, Gokhale RS, Mohanty D (2009) Towards Prediction of Metabolic Products of Polyketide Synthases: An *In Silico* Analysis. PLoS Comput Biol 5(4): e1000351.

Yang JC, Madupu R, Durkin AS, Ekborg NA, Pedamallu CS, Hostetler JB, Radune D, Toms BS, Henrissat B, Coutinho PM, Schwarz S, Field L, Trindade-Silva AE, Soares CAG, Elshahawi S, Hanora A, Schmidt EW, Haygood MG, Posfai J, Benner J, Madinger C, Nove J, Anton B, Chaudhary K, Foster J, Holman A, Kumar S, Lessard PA, Luyten YA, Slatko B, Wood N, Wu B, Teplitski M, Mougous JD, Ward N, Eisen JA, Badger JH, Distel DL (2009) The Complete Genome of *Teredinibacter turnerae* T7901: An Intracellular Endosymbiont of Marine Wood-Boring Bivalves (Shipworms). PLoS ONE 4(7): e6085.

Figures

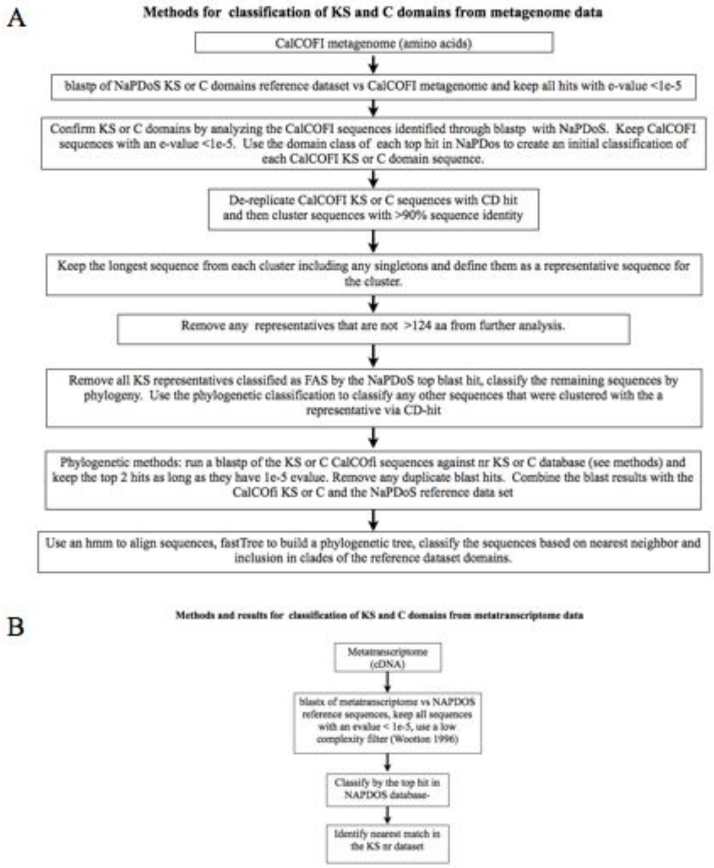

A
Methods for classification of KS and C domains from metagenome data
CalCOFI metagenome (amino acids)
blastp of NaPDoS KS or C domains reference dataset vs CalCOFI metagenome and keep all hits with e-value <1e-5
Confirm KS or C domains by analyzing the CalCOFI sequences identified through blastp with NaPDoS. Keep CalCOFI sequences with an e-value <1e-5. Use the domain class of each top hit in NaPDos to create an initial classification of each CalCOFI KS or C domain sequence.
De-replicate CalCOFI KS or C sequences with CD hit and then cluster sequences with >90% sequence identity
Keep the longest sequence from each cluster including any singletons and define them as a representative sequence for the cluster.
Remove any representatives that are not >124 aa from further analysis.
Remove all KS representatives classified as FAS by the NaPDoS top blast hit, classify the remaining sequences by phylogeny. Use the phylogenetic classification to classify any other sequences that were clustered with the a representative via CD-hit
Phylogenetic methods: run a blastp of the KS or C CalCOfi sequences against nr KS or C database (see methods) and keep the top 2 hits as long as they have 1e-5 evalue. Remove any duplicate blast hits. Combine the blast results with the CalCOfi KS or C and the NaPDoS reference data set
Use an hmm to align sequences, fastTree to build a phylogenetic tree, classify the sequences based on nearest neighbor and inclusion in clades of the reference dataset domains.
B
Methods and results for classification of KS and C domains from metatranscriptome data
Metatranscriptome (cDNA)
blastx of metatranscriptome vs NAPDOS reference sequences, keep all sequences with an evalue < 1e-5, use a low complexity filter (Wootton 1996)
Classify by the top hit in NAPDOS database-
Identify nearest match in the KS nr dataset

Figure 4.1: Methods scheme to detect secondary metabolite domains in A) metagenomes and B) metatranscriptomes.

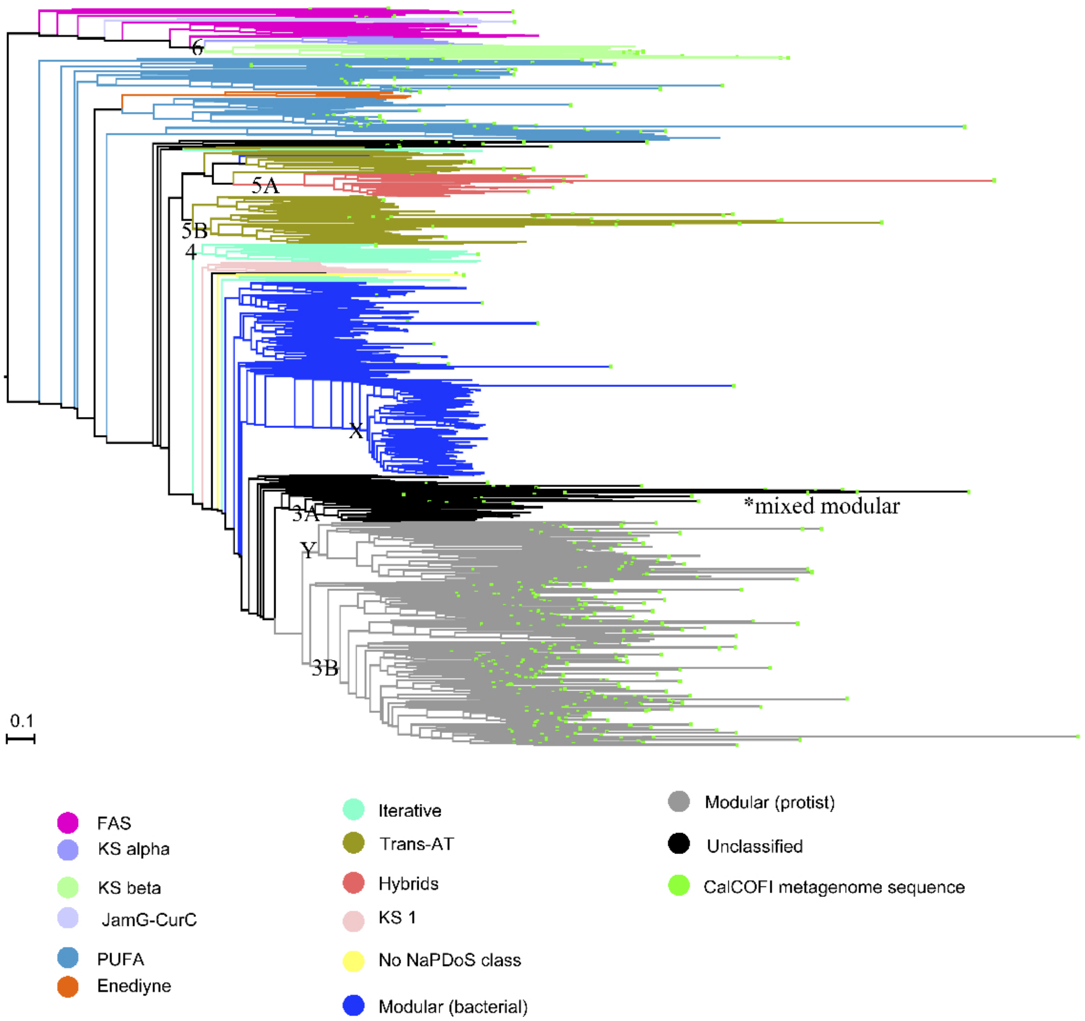


Figure 4.2: Phylogenetic tree with CalCOFI sequences, NaPDoS reference KS domains and nr KS domains. The nr KS domains are the top two hits from blastP of CalCOFI KS domains against the nr KS database. Domain classes are color-coded and are based on the presence of NaPDoS reference sequences except the protist clade and the mixed modular clade that are defined by nr KS sequences. The JamG-CurC clade is not defined in NaPDoS although the reference sequences are from the NaPDoS reference sequences. The mixed modular group as defined in this study is demarcated with a * , it contains both eukaryotic and prokaryotic derived sequences. The number on each node indicates the figure number with a more detailed view. The X indicates the Actinobacterial modular clade. The Y indicates the branch of the protist modular clade that contains sequences from genomes, the other portion, which is shown in figure 4.3A, contains the majority of new diversity. Scale bar represents changes per site.

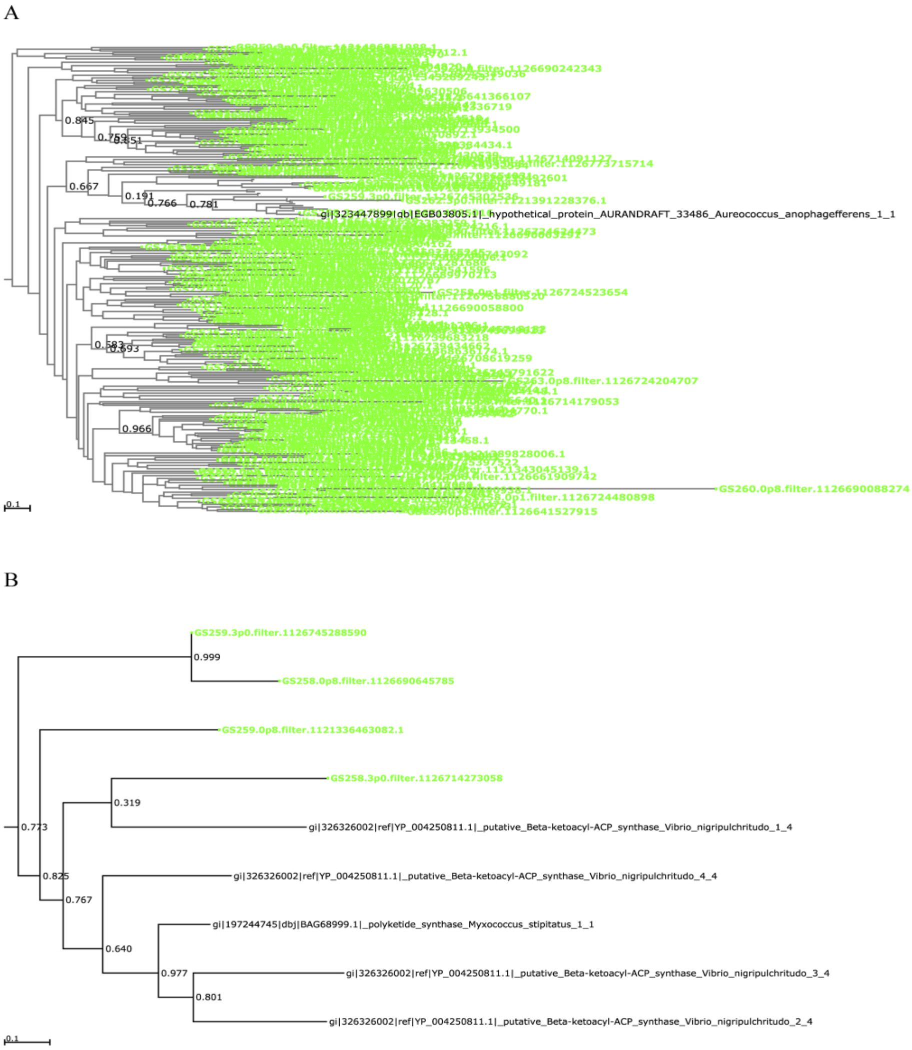

A
0.845
0.759
0.851
0.667
0.191
0.766
0.781
gi|323447899|gb|EGB03805.1|_hypothetical_protein_AURANDRAFT_33486_Aureococcus_anophagefferens_1_1
0.583
0.593
0.966
GS263.0p8.filter.1126724204707
GS260.0p8.filter.1126690088274
0.1
B
GS259.3p0.filter.1126745288590
0.999
GS258.0p8.filter.1126690645785
GS259.0p8.filter.1121336463082.1
GS258.3p0.filter.1126714273058
0.319
0.773
gi|326326002|ref|YP_004250811.1|_putative_Beta-ketoacyl-ACP_synthase_Vibrio_nigripulchritudo_1_4
0.825
gi|326326002|ref|YP_004250811.1|_putative_Beta-ketoacyl-ACP_synthase_Vibrio_nigripulchritudo_4_4
0.767
gi|197244745|dbj|BAG68999.1|_polyketide_synthase_Myxococcus_stipitatus_1_1
0.640
0.977
gi|326326002|ref|YP_004250811.1|_putative_Beta-ketoacyl-ACP_synthase_Vibrio_nigripulchritudo_3_4
0.801
gi|326326002|ref|YP_004250811.1|_putative_Beta-ketoacyl-ACP_synthase_Vibrio_nigripulchritudo_2_4
0.1

Figure 4.3: Select portion of the modular group from figure 4.2. The CalCOFI sequences (green text) have the site name listed first, followed by the size of filter given in the form of 0p1 to indicate for example 0.1µm filters. The nr sequences are genbank names but the last two numbers for each name are the domain number and the total number of KS domains in the protein. The NaPDoS reference sequences are also in black further information for each sequence can be found in appendix B. Branches colored as in figure 4.2. Node numbers are pseudo-likelihood values generated by FastTree. Scale bar represents changes per site. A) Portion of the protist clade showing the extensive phylogenetic diversity with no closely related sequences from the nr KS database. B) Portion of the mixed modular clade showing CalCOFI sequences that appear to match the multiple domains from one biosynthetic gene cluster in *Vibrio nigripulchritudo*. The sequences from *Vibrio* were sequenced as part of a whole genome-sequencing project. Scale bar represents changes per site.

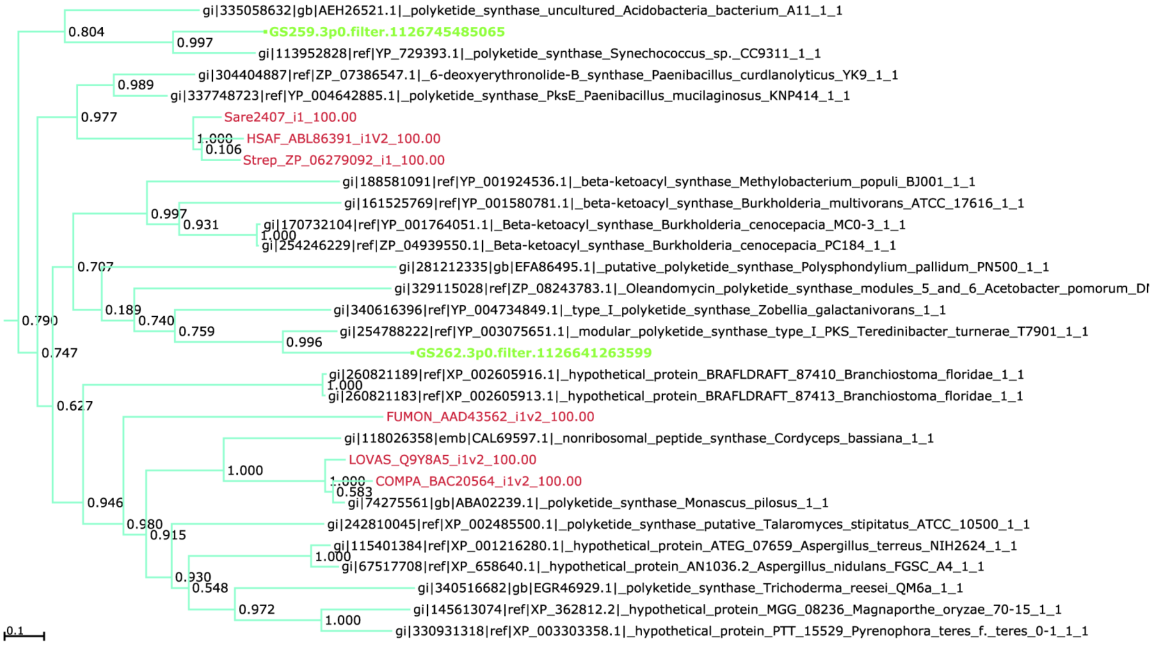

gi|335058632|gb|AEH26521.1|_polyketide_synthase_uncultured_Acidobacteria_bacterium_A11_1_1
0.804
GS259.3p0.filter.1126745485065
0.997
gi|113952828|ref|YP_729393.1|_polyketide_synthase_Synechococcus_sp._CC9311_1_1
gi|304404887|ref|ZP_07386547.1|_6-deoxyerythronolide-B_synthase_Paenibacillus_curdlanolyticus_YK9_1_1
0.989
gi|337748723|ref|YP_004642885.1|_polyketide_synthase_PksE_Paenibacillus_mucilaginosus_KNP414_1_1
0.977
Sare2407_i1_100.00
1.000
HSAF_ABL86391_i1V2_100.00
0.106
Strep_ZP_06279092_i1_100.00
gi|188581091|ref|YP_001924536.1|_beta-ketoacyl_synthase_Methylobacterium_populi_BJ001_1_1
gi|161525769|ref|YP_001580781.1|_beta-ketoacyl_synthase_Burkholderia_multivorans_ATCC_17616_1_1
0.997
0.931
gi|170732104|ref|YP_001764051.1|_Beta-ketoacyl_synthase_Burkholderia_cenocepacia_MC0-3_1_1
1.000
gi|254246229|ref|ZP_04939550.1|_Beta-ketoacyl_synthase_Burkholderia_cenocepacia_PC184_1_1
0.707
gi|281212335|gb|EFA86495.1|_putative_polyketide_synthase_Polysphondylium_pallidum_PN500_1_1
gi|329115028|ref|ZP_08243783.1|_Oleandomycin_polyketide_synthase_modules_5_and_6_Acetobacter_pomorum_DI
0.189
0.790
0.740
gi|340616396|ref|YP_004734849.1|_type_I_polyketide_synthase_Zobellia_galactanivorans_1_1
0.759
gi|254788222|ref|YP_003075651.1|_modular_polyketide_synthase_type_I_PKS_Teredinibacter_turnerae_T7901_1_1
0.996
0.747
GS262.3p0.filter.1126641263599
gi|260821189|ref|XP_002605916.1|_hypothetical_protein_BRAFLDRAFT_87410_Branchiostoma_floridae_1_1
1.000
gi|260821183|ref|XP_002605913.1|_hypothetical_protein_BRAFLDRAFT_87413_Branchiostoma_floridae_1_1
0.627
FUMON_AAD43562_i1v2_100.00
gi|118026358|emb|CAL69597.1|_nonribosomal_peptide_synthase_Cordyceps_bassiana_1_1
LOVAS_Q9Y8A5_i1v2_100.00
1.000
1.000
COMPA_BAC20564_i1v2_100.00
0.583
0.946
gi|74275561|gb|ABA02239.1|_polyketide_synthase_Monascus_pilosus_1_1
0.980
gi|242810045|ref|XP_002485500.1|_polyketide_synthase_putative_Talaromyces_stipitatus_ATCC_10500_1_1
0.915
gi|115401384|ref|XP_001216280.1|_hypothetical_protein_ATEG_07659_Aspergillus_terreus_NIH2624_1_1
1.000
gi|67517708|ref|XP_658640.1|_hypothetical_protein_AN1036.2_Aspergillus_nidulans_FGSC_A4_1_1
0.930
0.548
gi|340516682|gb|EGR46929.1|_polyketide_synthase_Trichoderma_reesei_QM6a_1_1
0.972
gi|145613074|ref|XP_362812.2|_hypothetical_protein_MGG_08236_Magnaporthe_oryzae_70-15_1_1
1.000
gi|330931318|ref|XP_003303358.1|_hypothetical_protein_PTT_15529_Pyrenophora_teres_f._teres_0-1_1_1
0.1

Figure 4.4: Selected portion of the iterative group from the KS CalCOFI phylogenetic tree. Sequences in green are from the CalCOFI dataset. The site name is listed first followed by the size of filter given in the form of 0p1 to indicate for example 0.1 µm filters. The red sequences are the NaPDoS reference sequences (Appendix B). The sequence names in black are derived from the NCBI nr database but the last two numbers for each name are the domain number and the total number of KS domains in the protein. All sequences shown are derived from genome sequencing projects of cultured organisms except the "uncultured Acidobacteria bacterium A11" sequence is derived from a metagenome library. Scale bar represents changes per site.

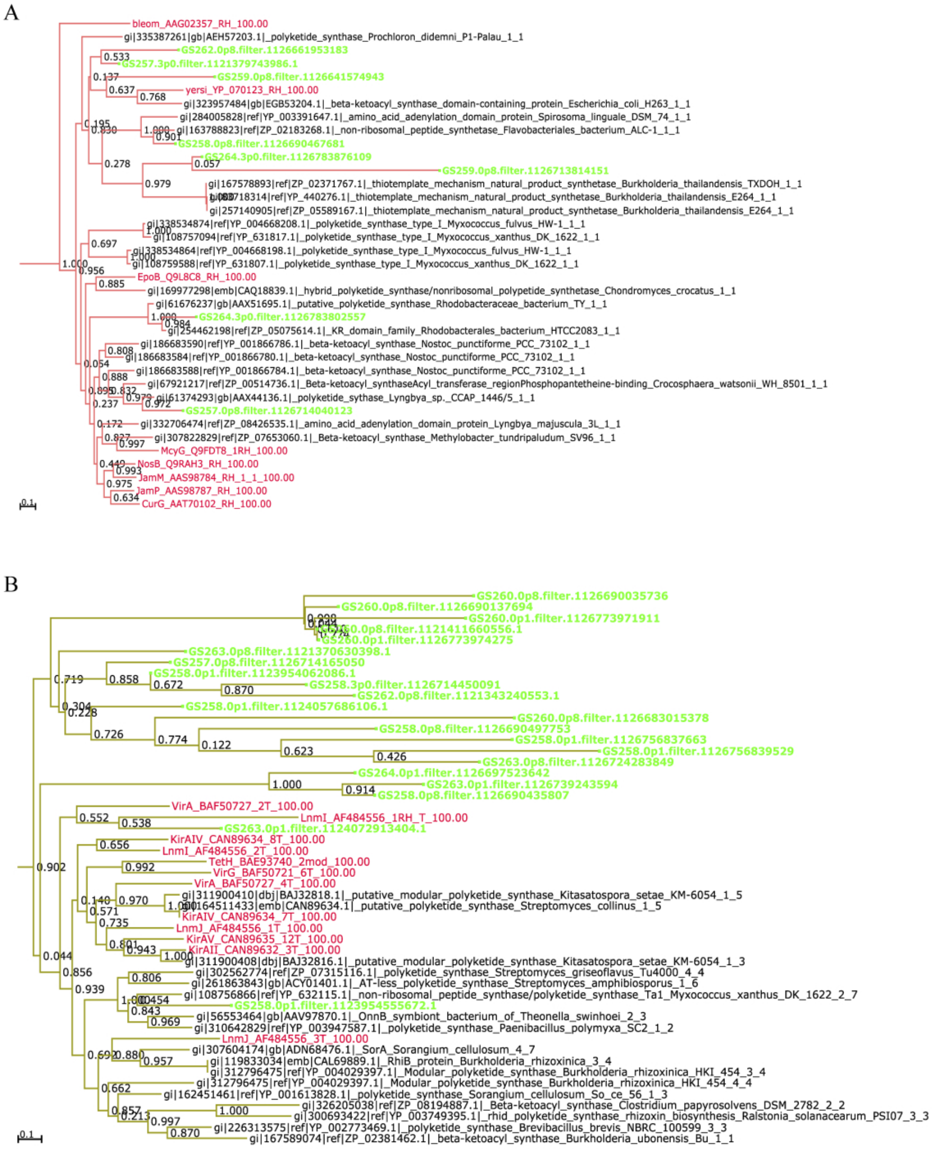

A
bleom_AAG02357_RH_100.00
gi|335387261|gb|AEH57203.1|_polyketide_synthase_Prochloron_didemni_P1-Palau_1_1
GS262.0p8.filter.1126661953183
0.533
GS257.3p0.filter.1121379743986.1
0.137
GS259.0p8.filter.1126641574943
0.637
yersi_YP_070123_RH_100.00
0.768
gi|323957484|gb|EGB53204.1|_beta-ketoacyl_synthase_domain-containing_protein_Escherichia_coli_H263_1_1
gi|284005828|ref|YP_003391647.1|_amino_acid_adenylation_domain_protein_Spirosoma_linguale_DSM_74_1_1
0.195
0.830
1.000
gi|163788823|ref|ZP_02183268.1|_non-ribosomal_peptide_synthetase_Flavobacteriales_bacterium_ALC-1_1_1
0.901
GS258.0p8.filter.1126690467681
GS264.3p0.filter.1126783876109
0.278
0.057
GS259.0p8.filter.1126713814151
0.979
gi|167578893|ref|ZP_02371767.1|_thiotemplate_mechanism_natural_product_synthetase_Burkholderia_thailandensis_TXDOH_1_1
1.000
gi|83718314|ref|YP_440276.1|_thiotemplate_mechanism_natural_product_synthetase_Burkholderia_thailandensis_E264_1_1
gi|257140905|ref|ZP_05589167.1|_thiotemplate_mechanism_natural_product_synthetase_Burkholderia_thailandensis_E264_1_1
gi|338534874|ref|YP_004668208.1|_polyketide_synthase_type_I_Myxococcus_fulvus_HW-1_1_1
1.000
gi|108757094|ref|YP_631817.1|_polyketide_synthase_type_I_Myxococcus_xanthus_DK_1622_1_1
0.697
gi|338534864|ref|YP_004668198.1|_polyketide_synthase_type_I_Myxococcus_fulvus_HW-1_1_1
1.000
1.000
gi|108759588|ref|YP_631807.1|_polyketide_synthase_type_I_Myxococcus_xanthus_DK_1622_1_1
0.956
EpoB_Q9L8C8_RH_100.00
0.885
gi|169977298|emb|CAQ18839.1|_hybrid_polyketide_synthase/nonribosomal_polypetide_synthetase_Chondromyces_crocatus_1_1
gi|61676237|gb|AAX51695.1|_putative_polyketide_synthase_Rhodobacteraceae_bacterium_TY_1_1
1.000
GS264.3p0.filter.1126783802557
0.984
gi|254462198|ref|ZP_05075614.1|_KR_domain_family_Rhodobacterales_bacterium_HTCC2083_1_1
gi|186683590|ref|YP_001866786.1|_beta-ketoacyl_synthase_Nostoc_punctiforme_PCC_73102_1_1
0.808
gi|186683584|ref|YP_001866780.1|_beta-ketoacyl_synthase_Nostoc_punctiforme_PCC_73102_1_1
0.054
gi|186683588|ref|YP_001866784.1|_beta-ketoacyl_synthase_Nostoc_punctiforme_PCC_73102_1_1
0.888
gi|67921217|ref|ZP_00514736.1|_Beta-ketoacyl_synthaseAcyl_transferase_regionPhosphopantetheine-binding_Crocosphaera_watsonii_WH_8501_1_1
0.895
0.832
0.979
gi|61374293|gb|AAX44136.1|_polyketide_sythase_Lyngbya_sp._CCAP_1446/5_1_1
0.237
0.972
GS257.0p8.filter.1126714040123
0.172
gi|332706474|ref|ZP_08426535.1|_amino_acid_adenylation_domain_protein_Lyngbya_majuscula_3L_1_1
0.827
gi|307822829|ref|ZP_07653060.1|_Beta-ketoacyl_synthase_Methylobacter_tundripaludum_SV96_1_1
0.997
McyG_Q9FDT8_1RH_100.00
0.449
NosB_Q9RAH3_RH_100.00
0.993
JamM_AAS98784_RH_1_1_100.00
0.975
JamP_AAS98787_RH_100.00
0.634
CurG_AAT70102_RH_100.00
0.1
B
GS260.0p8.filter.1126690035736
GS260.0p8.filter.1126690137694
0.998
GS260.0p1.filter.1126773971911
0.044
GS260.0p8.filter.1121411660556.1
0.774
GS260.0p1.filter.1126773974275
GS263.0p8.filter.1121370630398.1
GS257.0p8.filter.1126714165050
0.719
0.858
GS258.0p1.filter.1123954062086.1
0.672
GS258.3p0.filter.1126714450091
0.870
GS262.0p8.filter.1121343240553.1
0.304
0.228
GS258.0p1.filter.1124057686106.1
GS260.0p8.filter.1126683015378
0.726
GS258.0p8.filter.1126690497753
0.774
GS258.0p1.filter.1126756837663
0.122
0.623
GS258.0p1.filter.1126756839529
0.426
GS263.0p8.filter.1126724283849
GS264.0p1.filter.1126697523642
1.000
GS263.0p1.filter.1126739243594
0.914
GS258.0p8.filter.1126690435807
VirA_BAF50727_2T_100.00
0.552
LnmI_AF484556_1RH_T_100.00
0.538
GS263.0p1.filter.1124072913404.1
KirAIV_CAN89634_8T_100.00
0.656
LnmI_AF484556_2T_100.00
TetH_BAE93740_2mod_100.00
0.902
0.992
VirG_BAF50721_6T_100.00
VirA_BAF50727_4T_100.00
gi|311900410|dbj|BAJ32818.1|_putative_modular_polyketide_synthase_Kitasatospora_setae_KM-6054_1_5
0.140
0.970
1.000
gi|164511433|emb|CAN89634.1|_putative_polyketide_synthase_Streptomyces_collinus_1_5
0.571
KirAIV_CAN89634_7T_100.00
0.735
LnmJ_AF484556_1T_100.00
KirAV_CAN89635_12T_100.00
0.801
0.943
KirAII_CAN89632_3T_100.00
0.044
1.000
gi|311900408|dbj|BAJ32816.1|_putative_modular_polyketide_synthase_Kitasatospora_setae_KM-6054_1_3
0.856
gi|302562774|ref|ZP_07315116.1|_polyketide_synthase_Streptomyces_griseoflavus_Tu4000_4_4
0.806
gi|261863843|gb|ACY01401.1|_AT-less_polyketide_synthase_Streptomyces_amphibiosporus_1_6
0.939
1.000
0.454
gi|108756866|ref|YP_632115.1|_non-ribosomal_peptide_synthase/polyketide_synthase_Ta1_Myxococcus_xanthus_DK_1622_2_7
GS258.0p1.filter.1123954555672.1
0.843
gi|56553464|gb|AAV97870.1|_OnnB_symbiont_bacterium_of_Theonella_swinhoei_2_3
0.969
gi|310642829|ref|YP_003947587.1|_polyketide_synthase_Paenibacillus_polymyxa_SC2_1_2
LnmJ_AF484556_3T_100.00
0.692
0.880
gi|307604174|gb|ADN68476.1|_SorA_Sorangium_cellulosum_4_7
0.957
gi|119833034|emb|CAL69889.1|_RhiB_protein_Burkholderia_rhizoxinica_3_4
gi|312796475|ref|YP_004029397.1|_Modular_polyketide_synthase_Burkholderia_rhizoxinica_HKI_454_3_4
gi|312796475|ref|YP_004029397.1|_Modular_polyketide_synthase_Burkholderia_rhizoxinica_HKI_454_4_4
0.662
gi|162451461|ref|YP_001613828.1|_polyketide_synthase_Sorangium_cellulosum_So_ce_56_1_3
gi|326205038|ref|ZP_08194887.1|_Beta-ketoacyl_synthase_Clostridium_papyrosolvens_DSM_2782_2_2
0.857
1.000
0.213
gi|300693422|ref|YP_003749395.1|_rhid_polyketide_synthase_rhizoxin_biosynthesis_Ralstonia_solanacearum_PSI07_3_3
0.997
gi|226313575|ref|YP_002773469.1|_polyketide_synthase_Brevibacillus_brevis_NBRC_100599_3_3
0.870
gi|167589074|ref|ZP_02381462.1|_beta-ketoacyl_synthase_Burkholderia_ubonensis_Bu_1_1
0.1

Figure 4.5: Selected portion of the hybrid (A) and trans-AT (B) groups from the KS CalCOFI phylogenetic tree. Sequences in green are from the CalCOFI dataset. The site name is listed first followed by the size of filter given in the form of 0p1 to indicate for example 0.1 µm filters. The red sequences are the NaPDoS reference sequences (Appendix B). The sequence names in black are derived from the NCBI nr database but the last two numbers for each name are the domain number and the total number of KS domains in the protein. All NCBI nr derived sequences shown are from genome sequencing projects of cultured organisms. Scale bar represents changes per site.

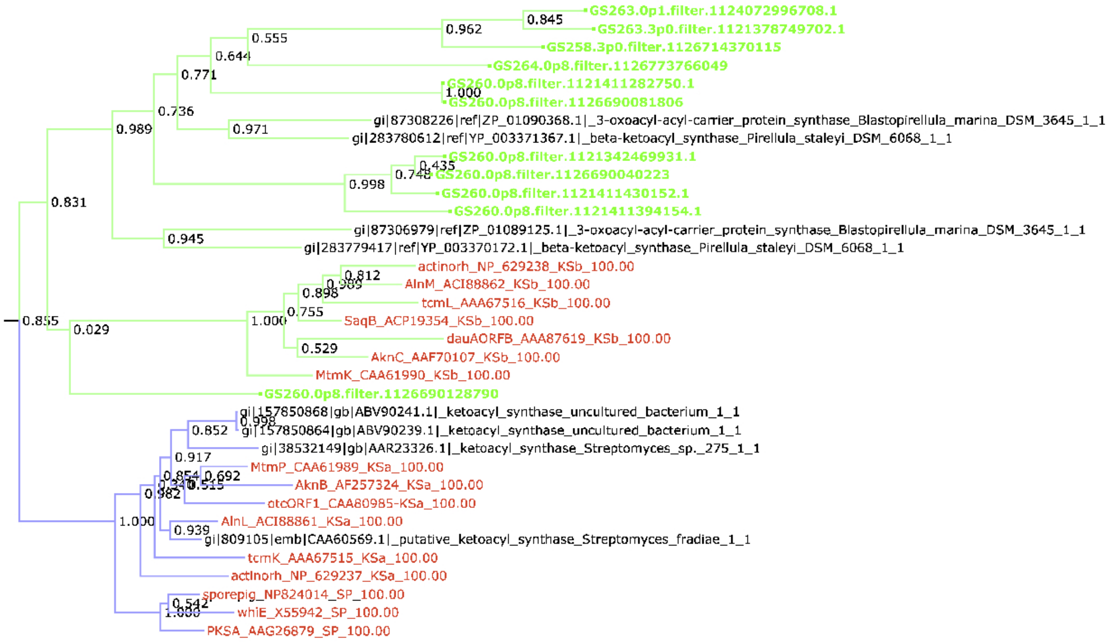

0.845
GS263.0p1.filter.1124072996708.1
0.962
GS263.3p0.filter.1121378749702.1
0.555
GS258.3p0.filter.1126714370115
0.644
GS264.0p8.filter.1126773766049
0.771
GS260.0p8.filter.1121411282750.1
1.000
GS260.0p8.filter.1126690081806
0.736
gi|87308226|ref|ZP_01090368.1|_3-oxoacyl-acyl-carrier_protein_synthase_Blastopirellula_marina_DSM_3645_1_1
0.989
0.971
gi|283780612|ref|YP_003371367.1|_beta-ketoacyl_synthase_Pirellula_staleyi_DSM_6068_1_1
GS260.0p8.filter.1121342469931.1
0.435
0.748
GS260.0p8.filter.1126690040223
0.998
GS260.0p8.filter.1121411430152.1
0.831
GS260.0p8.filter.1121411394154.1
gi|87306979|ref|ZP_01089125.1|_3-oxoacyl-acyl-carrier_protein_synthase_Blastopirellula_marina_DSM_3645_1_1
0.945
gi|283779417|ref|YP_003370172.1|_beta-ketoacyl_synthase_Pirellula_staleyi_DSM_6068_1_1
actinorh_NP_629238_KSb_100.00
0.812
0.989
AlnM_ACI88862_KSb_100.00
0.898
tcmL_AAA67516_KSb_100.00
0.755
0.855
1.000
SaqB_ACP19354_KSb_100.00
0.029
dauAORFB_AAA87619_KSb_100.00
0.529
AknC_AAF70107_KSb_100.00
MtmK_CAA61990_KSb_100.00
GS260.0p8.filter.1126690128790
gi|157850868|gb|ABV90241.1|_ketoacyl_synthase_uncultured_bacterium_1_1
0.998
0.852
gi|157850864|gb|ABV90239.1|_ketoacyl_synthase_uncultured_bacterium_1_1
gi|38532149|gb|AAR23326.1|_ketoacyl_synthase_Streptomyces_sp._275_1_1
0.917
MtmP_CAA61989_KSa_100.00
0.854
0.692
0.340
0.515
AknB_AF257324_KSa_100.00
0.982
otcORF1_CAA80985-KSa_100.00
1.000
AlnL_ACI88861_KSa_100.00
0.939
gi|809105|emb|CAA60569.1|_putative_ketoacyl_synthase_Streptomyces_fradiae_1_1
tcmK_AAA67515_KSa_100.00
actinorh_NP_629237_KSa_100.00
sporepig_NP824014_SP_100.00
0.542
1.000
whiE_X55942_SP_100.00
PKSA_AAG26879_SP_100.00

Figure 4.6: Selected portion of the type II group from the KS CalCOFI phylogenetic tree. Sequences in green are from the CalCOFI dataset. The site name is listed first followed by the size of filter given in the form of 0p1 to indicate for example 0.1 µm filters. The red sequences are the NaPDoS reference sequences (Appendix B). The sequence names in black are derived from the NCBI nr database but the last two numbers for each name are the domain number and the total number of KS domains in the protein. All NCBI nr derived sequences shown are from genome sequencing projects of cultured organisms except two sequences named uncultured are from PCR amplification of environmental DNA. Scale bar represents changes per site.

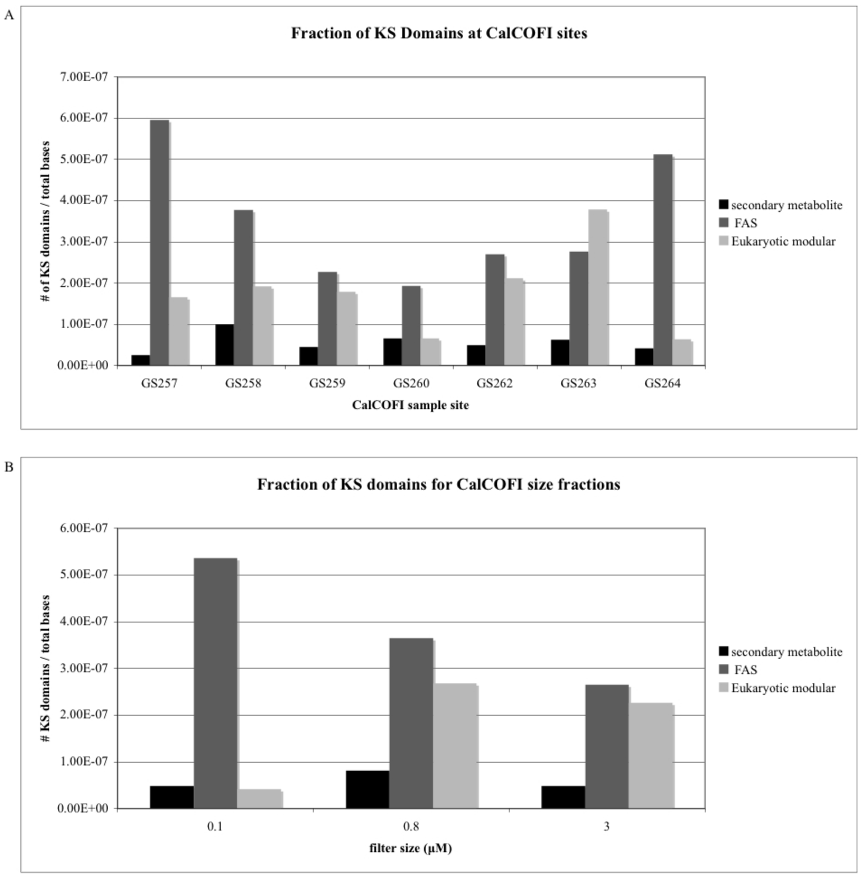


Figure 4.7: Chart of KS domains for different CalCOFI sizes and sample. Bar graphs representing the fraction of KS domains A) found at each sample site and B) in each size fraction. The numbers used to derive the values can be found in table 4.1 and table 4.2.

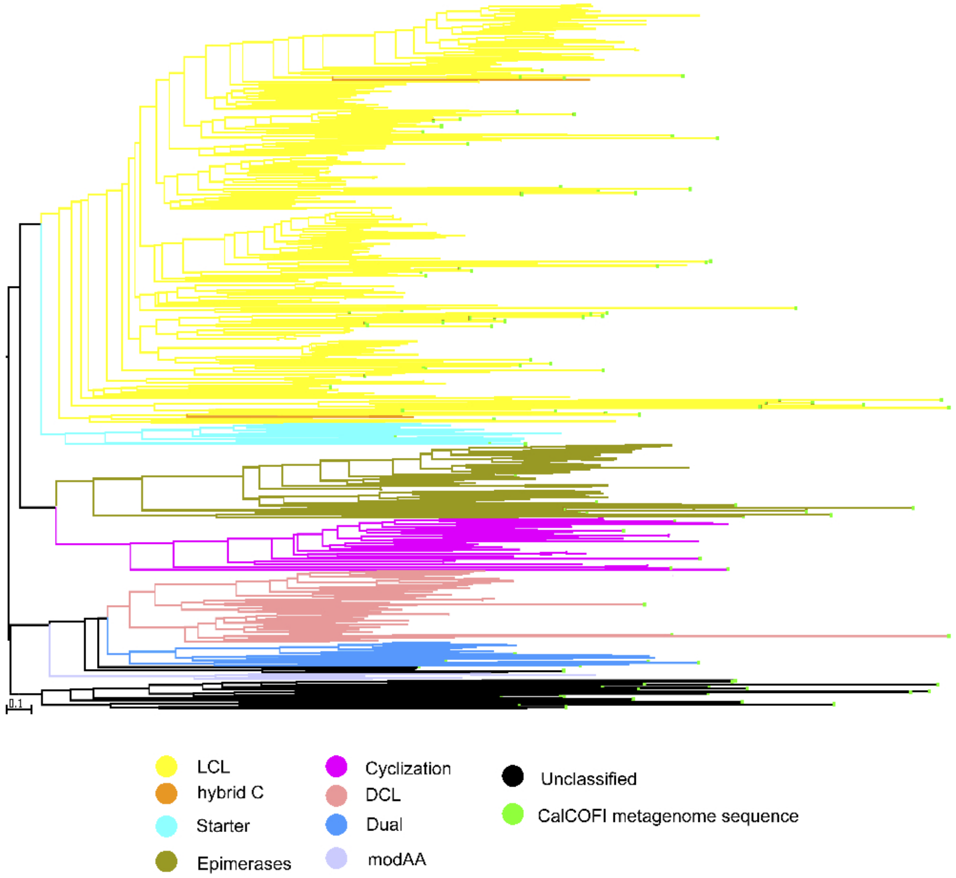


Figure 4.8: Phylogenetic tree with CalCOFI sequences, NaPDoS reference C domains and nr condensation domains. The nr C domains are the top two hits from blastP of CalCOFI C domains against the nr C domain database. Domain classes are color-coded and are based on the presence of NaPDoS reference sequences. The number on each node indicates the figure number with a more detailed view. Scale bar represents changes per site.

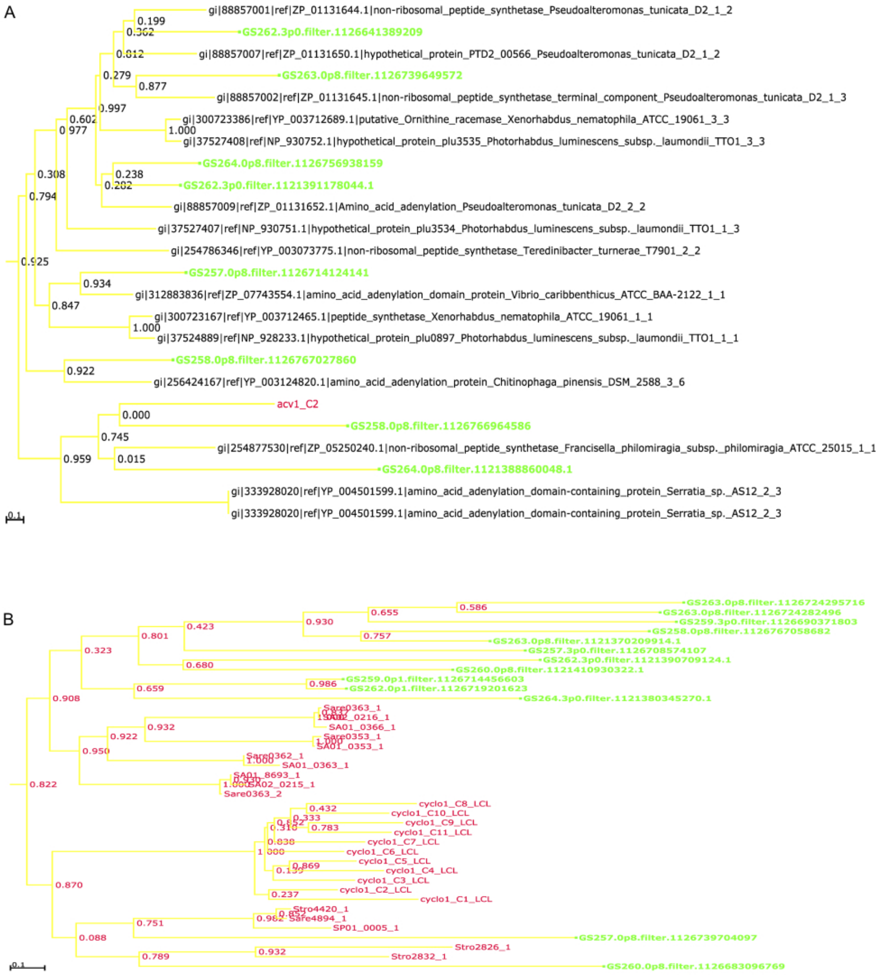

A
gi|88857001|ref|ZP_01131644.1|non-ribosomal_peptide_synthetase_Pseudoalteromonas_tunicata_D2_1_2
GS262.3p0.filter.1126641389209
gi|88857007|ref|ZP_01131650.1|hypothetical_protein_PTD2_00566_Pseudoalteromonas_tunicata_D2_1_2
GS263.0p8.filter.1126739649572
gi|88857002|ref|ZP_01131645.1|non-ribosomal_peptide_synthetase_terminal_component_Pseudoalteromonas_tunicata_D2_1_3
gi|300723386|ref|YP_003712689.1|putative_Ornithine_racemase_Xenorhabdus_nematophila_ATCC_19061_3_3
gi|37527408|ref|NP_930752.1|hypothetical_protein_plu3535_Photorhabdus_luminescens_subsp._laumondii_TTO1_3_3
GS264.0p8.filter.1126756938159
GS262.3p0.filter.1121391178044.1
gi|88857009|ref|ZP_01131652.1|Amino_acid_adenylation_Pseudoalteromonas_tunicata_D2_2_2
gi|37527407|ref|NP_930751.1|hypothetical_protein_plu3534_Photorhabdus_luminescens_subsp._laumondii_TTO1_1_3
gi|254786346|ref|YP_003073775.1|non-ribosomal_peptide_synthetase_Teredinibacter_turnerae_T7901_2_2
GS257.0p8.filter.1126714124141
gi|312883836|ref|ZP_07743554.1|amino_acid_adenylation_domain_protein_Vibrio_caribbenthicus_ATCC_BAA-2122_1_1
gi|300723167|ref|YP_003712465.1|peptide_synthetase_Xenorhabdus_nematophila_ATCC_19061_1_1
gi|37524889|ref|NP_928233.1|hypothetical_protein_plu0897_Photorhabdus_luminescens_subsp._laumondii_TTO1_1_1
GS258.0p8.filter.1126767027860
gi|256424167|ref|YP_003124820.1|amino_acid_adenylation_protein_Chitinophaga_pinensis_DSM_2588_3_6
acv1_C2
GS258.0p8.filter.1126766964586
gi|254877530|ref|ZP_05250240.1|non-ribosomal_peptide_synthetase_Francisella_philomiragia_subsp._philomiragia_ATCC_25015_1_1
GS264.0p8.filter.1121388860048.1
gi|333928020|ref|YP_004501599.1|amino_acid_adenylation_domain-containing_protein_Serratia_sp._AS12_2_3
gi|333928020|ref|YP_004501599.1|amino_acid_adenylation_domain-containing_protein_Serratia_sp._AS12_2_3
0.199
0.362
0.812
0.279
0.877
0.997
0.602
0.977
1.000
0.308
0.794
0.238
0.202
0.925
0.934
0.847
1.000
0.922
0.000
0.745
0.959
0.015
0.1
B
GS263.0p8.filter.1126724295716
GS263.0p8.filter.1126724282496
GS259.3p0.filter.1126690371803
GS258.0p8.filter.1126767058682
GS263.0p8.filter.1121370209914.1
GS257.3p0.filter.1126708574107
GS262.3p0.filter.1121390709124.1
GS260.0p8.filter.1121410930322.1
GS259.0p1.filter.1126714456603
GS262.0p1.filter.1126719201623
GS264.3p0.filter.1121380345270.1
Sare0363_1
SA02_0216_1
SA01_0366_1
Sare0353_1
SA01_0353_1
Sare0362_1
SA01_0363_1
SA01_8693_1
SA02_0215_1
Sare0363_2
cyclo1_C8_LCL
cyclo1_C10_LCL
cyclo1_C9_LCL
cyclo1_C11_LCL
cyclo1_C7_LCL
cyclo1_C6_LCL
cyclo1_C5_LCL
cyclo1_C4_LCL
cyclo1_C3_LCL
cyclo1_C2_LCL
cyclo1_C1_LCL
Stro4420_1
Sare4894_1
SP01_0005_1
GS257.0p8.filter.1126739704097
Stro2826_1
Stro2832_1
GS260.0p8.filter.1126683096769
0.586
0.655
0.930
0.423
0.757
0.801
0.323
0.680
0.659
0.986
0.908
0.932
0.922
0.950
1.000
0.822
0.432
0.333
0.783
0.838
1.000
0.869
0.237
0.870
0.751
0.088
0.789
0.932
0.1

Figure 4.9: Selected portion of the LCL group from the C domain CalCOFI phylogenetic tree. Green names represent CalCOFI C domains and red colored names are NaPDoS reference sequences (Appendix B). The sequence names in black are derived from the NCBI nr database but the last two numbers for each name are the domain number and the total number of C domains in the protein. In A) All NCBI based sequences are from genome sequencing projects in this tree. In B) no representatives from the nr C domain database are present all sequences are either reference sequences from NaPDoS in red (Appendix B) or CalCOFI KS domains. Branches are colored according to figure 4.8. Numbers associated with nodes are pseudo likelihood values generated by FastTree. Scale bar represents changes per site.

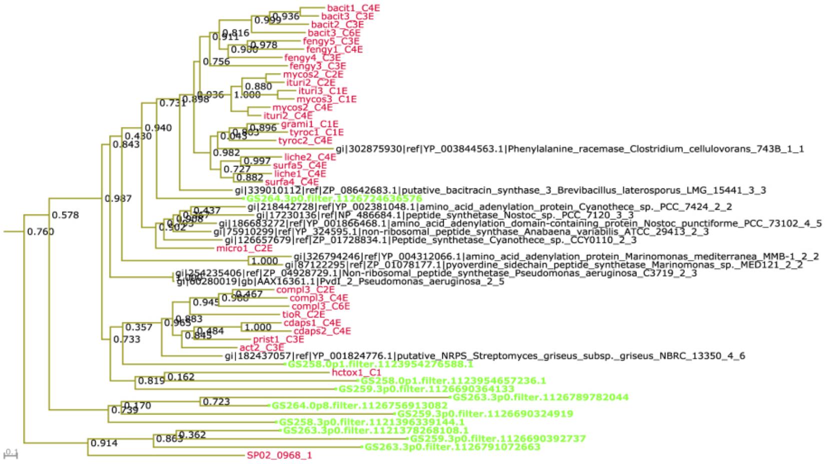

bacit1_C4E
bacit3_C3E
bacit2_C3E
bacit3_C6E
fengy5_C3E
fengy1_C4E
fengy4_C3E
fengy3_C3E
mycos2_C2E
ituri2_C2E
ituri3_C1E
mycos3_C1E
mycos2_C4E
ituri2_C4E
grami1_C1E
tyroc1_C1E
tyroc2_C4E
gi|302875930|ref|YP_003844563.1|Phenylalanine_racemase_Clostridium_cellulovorans_743B_1_1
liche2_C4E
surfa5_C4E
liche1_C4E
surfa4_C4E
gi|339010112|ref|ZP_08642683.1|putative_bacitracin_synthase_3_Brevibacillus_laterosporus_LMG_15441_3_3
GS264.3p0.filter.1126724636576
gi|218442728|ref|YP_002381048.1|amino_acid_adenylation_protein_Cyanothece_sp._PCC_7424_2_2
gi|17230136|ref|NP_486684.1|peptide_synthetase_Nostoc_sp._PCC_7120_3_3
gi|186683272|ref|YP_001866468.1|amino_acid_adenylation_domain-containing_protein_Nostoc_punctiforme_PCC_73102_4_5
gi|75910299|ref|YP_324595.1|non-ribosomal_peptide_synthase_Anabaena_variabilis_ATCC_29413_2_3
gi|126657679|ref|ZP_01728834.1|Peptide_synthetase_Cyanothece_sp._CCY0110_2_3
micro1_C2E
gi|326794246|ref|YP_004312066.1|amino_acid_adenylation_protein_Marinomonas_mediterranea_MMB-1_2_2
gi|87122295|ref|ZP_01078177.1|pyoverdine_sidechain_peptide_synthetase_Marinomonas_sp._MED121_2_2
gi|254235406|ref|ZP_04928729.1|Non-ribosomal_peptide_synthetase_Pseudomonas_aeruginosa_C3719_2_3
gi|60280019|gb|AAX16361.1|PvdI_2_Pseudomonas_aeruginosa_2_5
compl3_C2E
compl3_C4E
compl3_C6E
tioR_C2E
cdaps1_C4E
cdaps2_C4E
prist1_C3E
act2_C3E
gi|182437057|ref|YP_001824776.1|putative_NRPS_Streptomyces_griseus_subsp._griseus_NBRC_13350_4_6
GS258.0p1.filter.1123954276588.1
hctox1_C1
GS258.0p1.filter.1123954657236.1
GS259.3p0.filter.1126690364133
GS263.3p0.filter.1126789782044
GS264.0p8.filter.1126756913082
GS259.3p0.filter.1126690324919
GS258.3p0.filter.1121396339144.1
GS263.3p0.filter.1121378268108.1
GS259.3p0.filter.1126690392737
GS263.3p0.filter.1126791072663
SP02_0968_1
0.936
0.999
0.816
0.911
0.978
0.900
0.756
0.880
1.000
0.936
0.898
0.731
0.940
0.430
0.843
0.896
0.803
0.043
0.982
0.997
0.727
0.882
0.987
0.578
0.760
0.437
0.987
0.908
0.923
0.902
1.000
1.000
0.467
0.980
0.945
0.883
0.965
0.484
0.845
1.000
0.357
0.733
0.162
0.819
0.723
0.170
0.739
0.362
0.869
0.914
0.1

Figure 4.10: Selected portion of the epimerase group from the C domain CalCOFI phylogenetic tree. Green names represent CalCOFI C domains and red colored names are NaPDoS reference sequences (Appendix B). The sequence names in black are derived from the NCBI nr database but the last two numbers for each name are the domain number and the total number of C domains in the protein. Branches are colored according to figure 4.8. Numbers associated with nodes are pseudo likelihood values generated by FastTree. Scale bar represents changes per site.

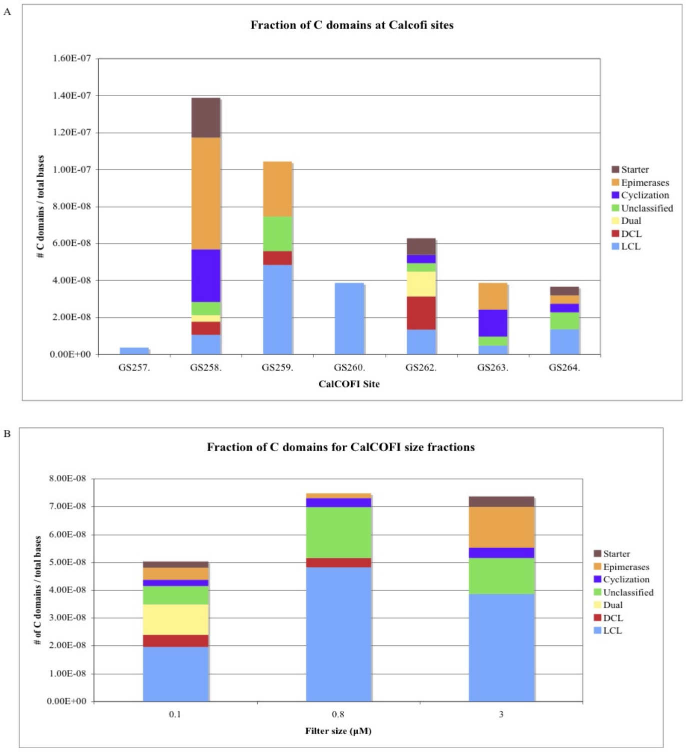


Figure 4.11: Chart of C domains for different CalCOFI sizes and sample. Bar graphs representing the fraction of C domains A) found at each sample site and B) in each size fraction. The numbers used to derive the values can be found in table 4.1 and table 4.2. Within each bar, the contribution of each domain class is colored according the labels next to the chart.

Tables

Table 4.1: Metadata for each site sample for metagenomes and metatranscriptomes.

Data set name	Sample site identification	CALCOFI location name	General location	Collection date	GPS coordinates	Filter size	Depth (m)	Temperature (°C)	Salinity
Dinoflagellate Bloom (cDNA)	1129899150649_0416109326-RLB-RL077-01-919_G2U4PMU01_EL1_RL077	93.26	Scripps Pier	4/16/10	32 57.0 N 117 178.0 W	20-µm plankton net	2	16.7	16.7
Dinoflagellate Bloom (cDNA)	1129899150649_0416109326-RLB-RL077-01-919_G2U4PMU02_EL1_RL077	93.26	Scripps Pier	4/16/10	32 57.0 N 117 18.0 W	20-µm plankton net	2	16.7	16.7
Dinoflagellate Bloom (cDNA)	1129899150649_0513109326-RLB-RL078-01-1083_G2U4PMU01_EL1_RL078	93.26	Scripps Pier	5/13/10	32 57.3 N 117 17.4 W	20-µm plankton net	2	17.5	17.5
Dinoflagellate Bloom (cDNA)	1129899150649_0513109326-RLB-RL078-01-1083_G2U4PMU02_EL1_RL078	93.26	Scripps Pier	5/13/10	32 57.3 N 117 17.4 W	20-µm plankton net	2	17.5	17.5
Dinoflagellate Bloom (cDNA)	1129899150649_0604109326-RLB-RL079-01-880_G2U4PMU01_EL1_RL079	93.26	Scripps Pier	6/4/10	32 57.1 N 117 18.0 W	20-µm plankton net	2	18.6	18.6
Dinoflagellate Bloom (cDNA)	1129899150649_0604109326-RLB-RL079-01-880_G2U4PMU02_EL1_RL079	93.26	Scripps Pier	6/4/10	32 57.076 N 117 18.0 W	20-µm plankton net	2	18.6	18.6
Dinoflagellate Bloom (cDNA)	1129899150649_060410IBMX-RLB-RL080-01-904_G2U4PMU01_EL1_RL080	IBMX	Coastal San Diego	6/4/10	32 46.0 N 117 24.2 W	20-µm plankton net	2	19.1	19.1
Dinoflagellate Bloom (cDNA)	1129899150649_060410IBMX-RLB-RL080-01-904_G2U4PMU02_EL1_RL080	IBMX	Coastal San Diego	6/4/10	32 46.0 N 117 24.2 W	20-µm plankton net	2	19.1	19.1
CalCOFI (DNA)	GS257	87 .40	nearshore	7/5/07	33 39.5 N 118 58.4 W	3.0-, 0.8-, 0.1- µm	2	18.64	33.744
CalCOFI (DNA)	GS258	87 .80	offshore	7/6/07	32 19.8 N 121 42.6 W	3.0-, 0.8-, 0.1- µm	2	14.71	33.511
CalCOFI (DNA)	GS259	83 .110	offshore	7/7/07	31 54.4 N 124 10.1 W	3.0-, 0.8-, 0.1- µm	2	17.25	33.362
CalCOFI (DNA)	GS260	83 .80	offshore	7/8/07	32 54.5 N 122 8.5 W	3.0-, 0.8-, 0.1- µm	2	14.68	33.155
CalCOFI (DNA)	GS262	80 .90	offshore	7/11/07	33 9.1 N 123 13.4 W	3.0-, 0.8-, 0.1- µm	2	17.4	33.095
CalCOFI (DNA)	GS263	77 .60	nearshore	7/13/07	34 43.4 N 121 33.2 W	3.0-, 0.8-, 0.1- µm	2	33.73	3.6
CalCOFI (DNA)	GS264	77 .49	nearshore	7/13/07	35 5.2 N 120 46.5 W	3.0-, 0.8-, 0.1- µm	2		
Antarctica (cDNA)	GS399	N/A	McMurdo Sound	11/21/09	77 40.38 S 166 25.43 E	0.1 µm	under sea Ice	-1.2	37.5
Antarctica (cDNA)	GS400	N/A	McMurdo Sound	11/18/09	77 39.521 S 166 10.082 E	0.1 µm	under sea Ice	-1.2	36
Antarctica (cDNA)	GS372	N/A	Ross Sea	01/30/09	77 40.457 S 166 0.117 E	0.1 µm	under sea Ice	-1.2	32
Antarctica (cDNA)	GS371	N/A	Ross Sea	1/28/09	77 41.7181 S 166 3.863 E	0.1 µm	under sea Ice	-1.2	33

Table 4.2: Sequence statistics for each metagenome.

					Metagenome statistics		
Data set name	Sequence type	Format analyzed	Sample identifiaction*		# of Sequences	# of Bases	# of Amino Acids
CalCOFI	metagenome	amino acids	GS257_0p1		269,326	86,843,484	28,947,828
CalCOFI	metagenome	amino acids	GS257_0p8		313,507	103,870,989	34,623,663
CalCOFI	metagenome	amino acids	GS257_3p0		239,615	81,195,939	27,065,313
CalCOFI	metagenome	amino acids	GS258_0p1		278,224	95,967,075	31,989,025
CalCOFI	metagenome	amino acids	GS258_0p8		281,792	90,844,818	30,281,606
CalCOFI	metagenome	amino acids	GS258_3p0		222,183	71,790,108	23,930,036
CalCOFI	metagenome	amino acids	GS259_0p1		254,039	59,829,201	19,943,067
CalCOFI	metagenome	amino acids	GS259_0p8		215,337	74,585,523	24,861,841
CalCOFI	metagenome	amino acids	GS259_3p0		251,315	83,990,124	27,996,708
CalCOFI	metagenome	amino acids	GS260_0p1		136,823	31,891,023	10,630,341
CalCOFI	metagenome	amino acids	GS260_0p8		370,659	119,684,679	39,894,893
CalCOFI	metagenome	amino acids	GS260_3p0		362,439	107,508,234	35,836,078
CalCOFI	metagenome	amino acids	GS262_0p1		149,035	35,610,729	11,870,243
CalCOFI	metagenome	amino acids	GS262_0p8		255,828	84,473,451	28,157,817
CalCOFI	metagenome	amino acids	GS262_3p0		233,874	76,195,425	25,398,475
CalCOFI	metagenome	amino acids	GS263_0p1		234,870	66,729,210	22,243,070
CalCOFI	metagenome	amino acids	GS263_0p8		212,647	67,950,375	22,650,125
CalCOFI	metagenome	amino acids	GS263_3p0		196,846	60,805,476	20,268,492
CalCOFI	metagenome	amino acids	GS264_0p1		292,405	80,137,011	26,712,337
CalCOFI	metagenome	amino acids	GS264_0p8		173,442	59,532,507	19,844,169
CalCOFI	metagenome	amino acids	GS264_3p0		185,577	60,955,695	20,318,565
				Total	5,129,783	1,600,391,076	533,463,692
Antarctica	metatranscriptome	DNA	GS371		181,347	44,370,090	N/A
Antarctica	metatranscriptome	DNA	GS372		171,962	48,004,990	N/A
Antarctica	metatranscriptome	DNA	GS399-ice		148,888	48,166,579	N/A
Antarctica	metatranscriptome	DNA	GS400-ice		186,665	56,756,358	N/A
				Total	688,862	197,298,017	
Dinoflagellate bloom	metatranscriptome	DNA	0416109326-RLB-RL077-01-919_G2U4PMU01_EL1_RL077		121,544	43,415,237	N/A
Dinoflagellate bloom	metatranscriptome	DNA	0416109326-RLB-RL077-01-919_G2U4PMU02_EL1_RL077		113,628	39,239,970	N/A
Dinoflagellate bloom	metatranscriptome	DNA	0513109326-RLB-RL078-01-1083_G2U4PMU01_EL1_RL078		127,866	46,378,011	N/A
Dinoflagellate bloom	metatranscriptome	DNA	0513109326-RLB-RL078-01-1083_G2U4PMU02_EL1_RL078		119,339	41,732,409	N/A
Dinoflagellate bloom	metatranscriptome	DNA	0604109326-RLB-RL079-01-880_G2U4PMU01_EL1_RL079		108,507	38,278,107	N/A
Dinoflagellate bloom	metatranscriptome	DNA	0604109326-RLB-RL079-01-880_G2U4PMU02_EL1_RL079		102,195	35,032,079	N/A
Dinoflagellate bloom	metatranscriptome	DNA	060410IBMX-RLB-RL080-01-904_G2U4PMU01_EL1_RL080		118,347	38,913,895	N/A
Dinoflagellate bloom	metatranscriptome	DNA	060410IBMX-RLB-RL080-01-904_G2U4PMU02_EL1_RL080		109,699	35,531,635	N/A
				Total	921,125	318,521,343	

N/A= not applicable

The sample identification is used in all tables and phylogenetic trees along with a unique sequence number. The CalCOFI IDs signify the sample site followed by the filter size given as 0p1 for example to signify 0.1 micron filter.

Table 4.3: Total sequences in the nr and reference dataset.

Data source	Sequence description	Total KS domains	Total C domains
NCBI	non-redundant proteins	17847	14448
NaPDoS	reference data set	197	258

Table 4.4: Number KS and C domains identified via blast and NaPDoS searches.

		Total KS domains		Total condensation domains	
Data set	Sample identifiaction	Initial BLAST *	NaPDoS result	Initial BLAST*	NAPDOS result
CalCOFI	GS257_0p1	264	263	4	0
CalCOFI	GS257_0p8	91	91	12	9
CalCOFI	GS257_3p0	139	138	12	7
CalCOFI	GS258_0p1	208	208	36	24
CalCOFI	GS258_0p8	127	126	28	18
CalCOFI	GS258_3p0	112	110	21	15
CalCOFI	GS259_0p1	164	163	5	4
CalCOFI	GS259_0p8	123	123	12	4
CalCOFI	GS259_3p0	65	65	20	12
CalCOFI	GS260_0p1	45	45	3	2
CalCOFI	GS260_0p8	179	176	2	2
CalCOFI	GS260_3p0	9	8	2	2
CalCOFI	GS262_0p1	86	83	14	10
CalCOFI	GS262_0p8	157	156	17	11
CalCOFI	GS262_3p0	88	88	12	11
CalCOFI	GS263_0p1	197	196	8	2
CalCOFI	GS263_0p8	148	148	21	15
CalCOFI	GS263_3p0	88	86	15	5
CalCOFI	GS264_0p1	303	298	9	5
CalCOFI	GS264_0p8	109	108	9	7
CalCOFI	GS264_3p0	72	71	39	29
	Total	2,774	2,750	301	194
		Initial BLAST**	NAPDOS result	Initial BLAST**	NAPDOS result***
Antarctica	GS371	15	15	5	1
Antarctica	GS372	12	12	6	2
Antarctica	GS399-ice	14	14	33	4
Antarctica	GS400-ice	27	27	33	7
	Total	68	68	80	14
		Initial BLAST**	NAPDOS result	Initial BLAST**	NAPDOS result***
Dinoflagellates	0416109326-RLB-RL077-01-919_G2U4PMU01_EL1_RL077	0	0	1	0
Dinoflagellates	0416109326-RLB-RL077-01-919_G2U4PMU02_EL1_RL077	3	3	0	0
Dinoflagellates	0513109326-RLB-RL078-01-1083_G2U4PMU01_EL1_RL078	9	9	2	2
Dinoflagellates	0513109326-RLB-RL078-01-1083_G2U4PMU02_EL1_RL078	7	6	2	1
Dinoflagellates	0604109326-RLB-RL079-01-880_G2U4PMU01_EL1_RL079	2	2	3	3
Dinoflagellates	0604109326-RLB-RL079-01-880_G2U4PMU02_EL1_RL079	1	1	3	1
Dinoflagellates	060410IBMX-RLB-RL080-01-904_G2U4PMU01_EL1_RL080	4	4	1	0
Dinoflagellates	060410IBMX-RLB-RL080-01-904_G2U4PMU02_EL1_RL080	3	3	0	0
	Total	29	28	12	7

* This is a blastp search of the reference KS or C domain dataset versus the Calcofi sequences

** This is a blastx search of the metatranscriptome versus the respective KS or C domains reference dataset

***Denotes unique sequence hits, due to matches in different frames actual number of hits is higher

Table 4.5: Results from CD-Hit. Shows the number of CalCOFI sequences for each NaPDoS based classification of KS domains after clustering and elimination based on length.

	Before elimination based on length			After elimination based on length			After phylogenetic classification
NaPDoS BLAST Classification	# of total clusters and singletons	# of clusters	# of singletons	# of clusters >124 aa	# of singletons > 124 aa	# of total clusters and singletons > 124 aa	# of clusters and singletons
FAS	1191	353	838	274	324	598	10**
typeII (KS-beta)	31	7	24	5	20	25	**11**
JamG, CurC	0	0	0	0	0	0	**2**
PUFA	54	10	44	8	25	33	72
enediyne	16	5	11	5	8	13	0
modular	390	23	367	18	258	276	337=(303)(24)(10)***
KS1	2	0	2	0	1	1	0
trans	112	8	104	6	78	84	**27**
iterative	33	4	29	4	19	23	**2**
hybridKS	22	3	19	1	15	16	**8**
KS (NaPDoS unclassified)	17	1	16	1	10	11	**2**
Unclassified	-	-	-	-	-	-	**11**
Total	1868	414	1454	322	758	1080	482
Total to Classify*	N/A	N/A	N/A	48	434	482	

reps=representatives
N/A= not applicable
*Total to classify excludes all FAS
** Does not include FAS classified by NaPDoS
*** Total modular =(eukaryotic) +(mixed)+ (bacterial)

Table 4.6: Results from CD-Hit. Shows the number of CalCOFI sequences for each NaPDoS based classification of condensation domains after clustering and elimination based on length.

	Before elimination based on length			After elimination based on length			
NaPDoS BLAST Classification	# of total clusters and singletons	# of clusters	# of singletons	# of clusters >124 aa	# of singletons > 124 aa	# of total clusters and singletons > 124 aa	# of total clusters and singletons after phylogenetic classification
C	1	0	1	0	1	1	0
Cyc	8	2	6	1	5	6	5
DCL	17	0	17	0	11	11	25
dual	5	0	5	0	3	3	6
Epim	11	3	8	2	6	8	11
LCL	117	7	110	5	74	79	59
Start	3	1	2	0	1	1	3
Total	162	13	149	8	101	109	109
Total to Classify	NA	NA	149	7	101	109	

Table 4.7: Top hit for KS and C domains in metatranscriptomes

<table>
<tr><th>Dataset ID</th><th>Dataset</th><th>Top BLAST hit in nr domain database*</th><th>Domain type</th><th># Hits</th><th>% Id**</th><th>Evalue**</th><th>NaPDoS classification</th><th>Alignment length</th></tr>
<tr><td>GS400-ice_0169732</td><td>Antarctica</td><td>gi|219129305|ref|XP_002184832.1| 3-oxoacyl-_acyl-carrier-protein_ synthase _Phaeodactylum tricornutum CCAP 1055/1_1_1</td><td>KS</td><td>17</td><td>59-92</td><td>1E-25-3E-69</td><td>FAS</td><td>x</td></tr>
<tr><td>GS372_146156</td><td>Antarctica</td><td>gi|224000768|ref|XP_002290056.1| 3-oxoacyl-synthase _Thalassiosira pseudonana CCMP1335_1_1</td><td>KS</td><td>16</td><td>55-82</td><td>3E-15-5E-70</td><td>FAS</td><td>x</td></tr>
<tr><td>GS371_076795</td><td>Antarctica</td><td>gi|219122229|ref|XP_002181453.1| 3-oxoacyl-_acyl-carrier-protein_ synthase _Phaeodactylum tricornutum CCAP 1055/1_1_1</td><td>KS</td><td>14</td><td>46-89</td><td>7E-14-7E-60</td><td>FAS</td><td>x</td></tr>
<tr><td>GS372_030280</td><td>Antarctica</td><td>gi|224013261|ref|XP_002295282.1| predicted protein _Thalassiosira pseudonana CCMP1335_1_1</td><td>KS</td><td>6</td><td>68-92</td><td>1E-13-1E-70</td><td>FAS</td><td>x</td></tr>
<tr><td>GS371_186573</td><td>Antarctica</td><td>gi|260061435|ref|YP_003194515.1| putative 3-oxoacyl-_acyl-carrier-protein_ synthase II _Robiginitalea biformata HTCC2501_1_1</td><td>KS</td><td>1</td><td>95.29</td><td>4.00E-57</td><td>FAS</td><td>85</td></tr>
<tr><td>GS371_030278</td><td>Antarctica</td><td>gi|90416593|ref|ZP_01224524.1| 3-oxoacyl-(acyl-carrier-protein) synthase _marine gamma proteobacterium HTCC2207_1_1</td><td>KS</td><td>1</td><td>90.23</td><td>3.00E-66</td><td>FAS</td><td>133</td></tr>
<tr><td>GS372_094992</td><td>Antarctica</td><td>gi|315444310|ref|YP_004077189.1| polyketide synthase family protein _Mycobacterium sp. Spyr1_1_1</td><td>KS</td><td>1</td><td>85.71</td><td>5.00E-06</td><td>modular</td><td>21</td></tr>
<tr><td>GS371_118649</td><td>Antarctica</td><td>gi|254476766|ref|ZP_05090152.1| 3-oxoacyl-(acyl-carrier-protein) synthase 2 _Ruegeria sp. R11_1_1</td><td>KS</td><td>1</td><td>83.33</td><td>3.00E-36</td><td>FAS</td><td>90</td></tr>
<tr><td>GS371_183267</td><td>Antarctica</td><td>gi|254785118|ref|YP_003072546.1| 3-oxoacyl-_acyl-carrier-protein_ synthase I _Teredinibacter turnerae T7901_1_1</td><td>KS</td><td>1</td><td>79.31</td><td>2.00E-10</td><td>FAS</td><td>29</td></tr>
<tr><td>GS371_024383</td><td>Antarctica</td><td>gi|255085212|ref|XP_002505037.1| predicted protein _Micromonas sp. RCC299_1_1</td><td>KS</td><td>1</td><td>77.3</td><td>3.00E-56</td><td>FAS</td><td>141</td></tr>
<tr><td>GS371_096098</td><td>Antarctica</td><td>gi|327403886|ref|YP_004344724.1| 3-oxoacyl-(acyl-carrier-protein) synthase II _Fluviicola taffensis DSM 16823_1_1</td><td>KS</td><td>1</td><td>76.92</td><td>6.00E-28</td><td>FAS</td><td>65</td></tr>
<tr><td>GS399-ice_0068950</td><td>Antarctica</td><td>gi|323453351|gb|EGB09223.1| hypothetical protein AURANDRAFT_25463 _Aureococcus anophagefferens_1_1</td><td>KS</td><td>1</td><td>74.03</td><td>3.00E-34</td><td>FAS</td><td>77</td></tr>
<tr><td>GS400-ice_0029375</td><td>Antarctica</td><td>gi|323454417|gb|EGB10287.1| hypothetical protein AURANDRAFT_71232 _Aureococcus anophagefferens_1_1</td><td>KS</td><td>1</td><td>73.68</td><td>6.00E-42</td><td>FAS</td><td>95</td></tr>
<tr><td>GS372_079517</td><td>Antarctica</td><td>gi|194476936|ref|YP_002049115.1| 3-oxoacyl-(acyl-carrier-protein) synthase II _Paulinella chromatophora_1_1</td><td>KS</td><td>1</td><td>67.44</td><td>9.00E-22</td><td>FAS</td><td>43</td></tr>
<tr><td>GS400-ice_0051088</td><td>Antarctica</td><td>gi|154685568|ref|YP_001420729.1| 3-oxoacyl-(acyl carrier protein) synthase II _Bacillus amyloliquefaciens FZB42_1_1</td><td>KS</td><td>1</td><td>64.47</td><td>1.00E-29</td><td>FAS</td><td>76</td></tr>
<tr><td>GS399-ice_0030617</td><td>Antarctica</td><td>gi|225165897|ref|ZP_03727666.1| 3-oxoacyl-(acyl-carrier-protein) synthase 2 _Opitutaceae bacterium TAV2_1_1</td><td>KS</td><td>1</td><td>62.5</td><td>1.00E-33</td><td>FAS</td><td>48</td></tr>
<tr><td>GS371_152716</td><td>Antarctica</td><td>gi|189424993|ref|YP_001952170.1| 3-oxoacyl-(acyl-carrier-protein) synthase 2 _Geobacter lovleyi SZ_1_1</td><td>KS</td><td>1</td><td>58.97</td><td>1.00E-08</td><td>FAS</td><td>39</td></tr>
<tr><td>GS371_058095</td><td>Antarctica</td><td>gi|296136825|ref|YP_003644067.1| 3-oxoacyl-(acyl-carrier-protein) synthase 2 _Thiomonas intermedia K12_1_1</td><td>KS</td><td>1</td><td>52.73</td><td>6.00E-11</td><td>FAS</td><td>55</td></tr>
<tr><td>GS399-ice_0107867</td><td>Antarctica</td><td>gi|291434987|ref|ZP_06574377.1| erythronolide synthase _Streptomyces ghanaensis ATCC 14672_1_1</td><td>KS</td><td>1</td><td>38.1</td><td>4.00E-06</td><td>modular</td><td>63</td></tr>
<tr><td>GS371_049689</td><td>Antarctica</td><td>gi|70731540|ref|YP_261281.1| peptide synthase [Pseudomonas fluorescens Pf-5]_3_4</td><td>C</td><td>1</td><td>92.39</td><td>1.00E-36</td><td>epim</td><td>92</td></tr>
<tr><td>GS371_064366</td><td>Antarctica</td><td>gi|295683941|gb|ADG27358.1| peptide synthetase [Streptomyces anulatus]_3_3</td><td>C</td><td>1</td><td>27.2</td><td>2.00E-04</td><td>epim</td><td>125</td></tr>
<tr><td>GS372_124713</td><td>Antarctica</td><td>gi|331013398|gb|EGH93454.1| non-ribosomal peptide synthetase [Pseudomonas syringae pv. tabaci ATCC 11528]_1_1</td><td>C</td><td>1</td><td>32.65</td><td>1.5</td><td>start</td><td>49</td></tr>
<tr><td>GS372_148674</td><td>Antarctica</td><td>gi|331013398|gb|EGH93454.1| non-ribosomal peptide synthetase [Pseudomonas syringae pv. tabaci ATCC 11528]_1_1</td><td>C</td><td>1</td><td>38.24</td><td>3.4</td><td>start</td><td>34</td></tr>
<tr><td>GS399-ice_0071749</td><td>Antarctica</td><td>gi|70729516|ref|YP_259254.1| nonribosomal peptide synthetase [Pseudomonas fluorescens Pf-5]_1_4</td><td>C</td><td>1</td><td>81.82</td><td>1.00E-38</td><td>dual</td><td>88</td></tr>
<tr><td>GS399-ice_0134765</td><td>Antarctica</td><td>gi|334838615|gb|EGM17328.1| peptide synthase [Pseudomonas aeruginosa 138244]_2_4</td><td>C</td><td>1</td><td>76.79</td><td>1.00E-45</td><td>LCL</td><td>56</td></tr>
<tr><td>GS399-ice_0141327</td><td>Antarctica</td><td>gi|70731540|ref|YP_261281.1| peptide synthase [Pseudomonas fluorescens Pf-5]_2_4</td><td>C</td><td>1</td><td>97.44</td><td>1.00E-62</td><td>LCL</td><td>117</td></tr>
<tr><td>GS399-ice_0143064</td><td>Antarctica</td><td>gi|226359672|ref|YP_002777450.1| non-ribosomal peptide synthetase [Rhodococcus opacus B4]_2_9</td><td>C</td><td>1</td><td>37.5</td><td>3.5</td><td>LCL</td><td>32</td></tr>
<tr><td>GS400-ice_0058051</td><td>Antarctica</td><td>gi|70731540|ref|YP_261281.1| peptide synthase [Pseudomonas fluorescens Pf-5]_3_4</td><td>C</td><td>1</td><td>92.17</td><td>3.00E-68</td><td>epim</td><td>115</td></tr>
<tr><td>GS400-ice_0116734</td><td>Antarctica</td><td>gi|70731540|ref|YP_261281.1| peptide synthase [Pseudomonas fluorescens Pf-5]_1_4</td><td>C</td><td>1</td><td>96.05</td><td>5.00E-29</td><td>C</td><td>76</td></tr>
<tr><td>GS400-ice_0133541</td><td>Antarctica</td><td>gi|70731449|ref|YP_261190.1| peptide synthase [Pseudomonas fluorescens Pf-5]_5_5</td><td>C</td><td>1</td><td>100</td><td>2.00E-39</td><td>DCL</td><td>78</td></tr>
<tr><td>GS400-ice_0137365</td><td>Antarctica</td><td>gi|70731449|ref|YP_261190.1| peptide synthase [Pseudomonas fluorescens Pf-5]_5_5</td><td>C</td><td>1</td><td>95.24</td><td>2.00E-29</td><td>DCL</td><td>63</td></tr>
<tr><td>GS400-ice_0179318</td><td>Antarctica</td><td>gi|341580261|gb|AAY93356.2| non-ribosomal peptide synthetase PvdI [Pseudomonas fluorescens Pf-5]_4_4</td><td>C</td><td>1</td><td>97.83</td><td>5.00E-29</td><td>DCL</td><td>46</td></tr>
<tr><td>GS400-ice_0185465</td><td>Antarctica</td><td>gi|70731450|ref|YP_261191.1| non-ribosomal peptide synthetase [Pseudomonas fluorescens Pf-5]_1_2</td><td>C</td><td>1</td><td>100</td><td>7.00E-21</td><td>LCL</td><td>39</td></tr>
<tr><td>G2U4PMU01D7Y9F</td><td>Dinoflagellate</td><td>gi|148536475|gb|ABQ85797.1| type I polyketide synthase-like protein KB2006 _Karenia brevis_1_1 Count</td><td>KS</td><td>4</td><td>47-64</td><td>1E-14-1E-44</td><td>modular</td><td>x</td></tr>
<tr><td>G2U4PMU01ARL26</td><td>Dinoflagellate</td><td>gi|148536481|gb|ABQ85800.1| type I polyketide synthase-like protein KB5361 _Karenia brevis_1_1 Count</td><td>KS</td><td>3</td><td>41-52</td><td>4E-20-2E-40</td><td>modular</td><td>x</td></tr>
<tr><td>G2U4PMU02GURJZ</td><td>Dinoflagellate</td><td>gi|148536485|gb|ABQ85802.1| type I polyketide synthase-like protein KB6736 _Karenia brevis_1_1 Count</td><td>KS</td><td>2</td><td>42-46</td><td>3E-15-3E-22</td><td>modular</td><td>x</td></tr>
<tr><td>G2U4PMU01C5NVD</td><td>Dinoflagellate</td><td>gi|302772915|ref|XP_002969875.1| hypothetical protein SELMODRAFT_171238 _Selaginella moellendorffii_ _1_1</td><td>KS</td><td>2</td><td>60</td><td>2.00E-30</td><td>FAS</td><td>x</td></tr>
<tr><td>G2U4PMU01ALEMP</td><td>Dinoflagellate</td><td>gi|302542518|ref|ZP_07294860.1| modular polyketide synthase _Streptomyces hygroscopicus ATCC 53653_ _2_2</td><td>KS</td><td>2</td><td>39</td><td>4E-6-3E-15</td><td>modular</td><td>x</td></tr>
<tr><td>G2U4PMU02IEWG2</td><td>Dinoflagellate</td><td>gi|62549357|gb|AAX86997.1| type I polyketide synthase-like _uncultured bacterium_1_1</td><td>KS</td><td>1</td><td>52.94</td><td>2.00E-15</td><td>modular</td><td>51</td></tr>
<tr><td>G2U4PMU02I22Y4</td><td>Dinoflagellate</td><td>gi|323454417|gb|EGB10287.1| hypothetical protein AURANDRAFT_71232 _Aureococcus anophagefferens_1_1</td><td>KS</td><td>1</td><td>67.44</td><td>2.00E-11</td><td>FAS</td><td>43</td></tr>
<tr><td>G2U4PMU02HLRDO</td><td>Dinoflagellate</td><td>gi|224013261|ref|XP_002295282.1| predicted protein _Thalassiosira pseudonana CCMP1335_ _1_1</td><td>KS</td><td>1</td><td>91.73</td><td>3.00E-67</td><td>FAS</td><td>133</td></tr>
<tr><td>G2U4PMU02GOVEW</td><td>Dinoflagellate</td><td>gi|307822825|ref|ZP_07653056.1| 6-deoxyerythronolide-B synthase _Methylobacter tundripaludum SV96_ _1_1</td><td>KS</td><td>1</td><td>40.2</td><td>7.00E-16</td><td>modular</td><td>102</td></tr>
<tr><td>G2U4PMU02GL41L</td><td>Dinoflagellate</td><td>gi|326326002|ref|YP_004250811.1| putative Beta-ketoacyl-ACP synthase _Vibrio nigripulchritudo_ _4_4</td><td>KS</td><td>1</td><td>45.28</td><td>4.00E-31</td><td>modular</td><td>159</td></tr>
<tr><td>G2U4PMU02GC0GK</td><td>Dinoflagellate</td><td>gi|158302786|dbj|BAF85839.1| modular polyketide synthase _Streptomyces cyaneogriseus subsp. noncyanogenus_3_4</td><td>KS</td><td>1</td><td>64.2</td><td>3.00E-27</td><td>modular</td><td>81</td></tr>
<tr><td>G2U4PMU02F294J</td><td>Dinoflagellate</td><td>gi|86134671|ref|ZP_01053253.1| 3-oxoacyl-(acyl-carrier-protein) synthase II _Polaribacter sp. MED152_ _1_1</td><td>KS</td><td>1</td><td>95.08</td><td>5.00E-78</td><td>FAS</td><td>122</td></tr>
<tr><td>G2U4PMU01EZD25</td><td>Dinoflagellate</td><td>gi|22138082|gb|AAM93417.1| polyketide synthase-like protein _Pseudopfiesteria shumwayae_1_1</td><td>KS</td><td>1</td><td>63.46</td><td>1.00E-31</td><td>modular</td><td>104</td></tr>
<tr><td>G2U4PMU01DWI4X</td><td>Dinoflagellate</td><td>gi|326433466|gb|EGD79036.1| MlnB _Salpingoeca sp. ATCC 50818_3_4</td><td>KS</td><td>1</td><td>58.24</td><td>1.00E-22</td><td>modular</td><td>91</td></tr>
<tr><td>G2U4PMU01DS05K</td><td>Dinoflagellate</td><td>gi|209877909|ref|XP_002140396.1| polyketide synthase _Cryptosporidium muris RN66_ _2_7</td><td>KS</td><td>1</td><td>43.21</td><td>5.00E-30</td><td>iterative</td><td>162</td></tr>
<tr><td>G2U4PMU01DLZ8C</td><td>Dinoflagellate</td><td>gi|326433466|gb|EGD79036.1| MlnB _Salpingoeca sp. ATCC 50818_2_4</td><td>KS</td><td>1</td><td>38.13</td><td>1.00E-13</td><td>modular</td><td>139</td></tr>
<tr><td>G2U4PMU01D9MJW</td><td>Dinoflagellate</td><td>gi|256787989|ref|ZP_05526420.1| 3-oxoacyl-(ACP) synthase II _Streptomyces lividans TK24_ _1_1</td><td>KS</td><td>1</td><td>49.02</td><td>3.00E-15</td><td>FAS</td><td>102</td></tr>
<tr><td>G2U4PMU01CQD04</td><td>Dinoflagellate</td><td>gi|225025714|ref|ZP_03714906.1| hypothetical protein EIKCOROL_02616 _Eikenella corrodens ATCC 23834_ _1_1</td><td>KS</td><td>1</td><td>66.27</td><td>3.00E-53</td><td>FAS</td><td>169</td></tr>
<tr><td>G2U4PMU01COUF8</td><td>Dinoflagellate</td><td>gi|209877909|ref|XP_002140396.1| polyketide synthase _Cryptosporidium muris RN66_ _6_7</td><td>KS</td><td>1</td><td>44.68</td><td>4.00E-15</td><td>hybridKS</td><td>94</td></tr>
<tr><td>G2U4PMU02IWHMS</td><td>Dinoflagellate</td><td>gi|94467032|dbj|BAE93722.1| type I polyketide synthase _Streptomyces sp. NRRL 11266_1_3</td><td>KS</td><td>1</td><td>32.63</td><td>3.00E-04</td><td>modular</td><td>95</td></tr>
<tr><td>G2U4PMU01BK5YT</td><td>Dinoflagellate</td><td>gi|77458076|ref|YP_347581.1| amino acid adenylation [Pseudomonas fluorescens Pf0-1]_1_1</td><td>C</td><td>1</td><td>36.14</td><td>3.00E-19</td><td>LCL</td><td>83</td></tr>
<tr><td>G2U4PMU01DIPED</td><td>Dinoflagellate</td><td>gi|88812177|ref|ZP_01127429.1| probable peptide synthetase protein [Nitrococcus mobilis Nb-231]_1_3</td><td>C</td><td>1</td><td>48.19</td><td>7.00E-22</td><td>LCL</td><td>166</td></tr>
<tr><td>G2U4PMU02HAG1S</td><td>Dinoflagellate</td><td>gi|88812177|ref|ZP_01127429.1| probable peptide synthetase protein [Nitrococcus mobilis Nb-231]_1_3</td><td>C</td><td>1</td><td>61.54</td><td>6.00E-14</td><td>LCL</td><td>65</td></tr>
<tr><td>G2U4PMU01AW7A2</td><td>Dinoflagellate</td><td>gi|220912507|ref|YP_002487816.1| amino acid adenylation protein [Arthrobacter chlorophenolicus A6]_1_3</td><td>C</td><td>1</td><td>50</td><td>0.055</td><td>LCL</td><td>36</td></tr>
<tr><td>G2U4PMU01EY08N</td><td>Dinoflagellate</td><td>gi|220912507|ref|YP_002487816.1| amino acid adenylation protein [Arthrobacter chlorophenolicus A6]_1_3</td><td>C</td><td>1</td><td>50</td><td>0.052</td><td>LCL</td><td>36</td></tr>
<tr><td>G2U4PMU01C990Q</td><td>Dinoflagellate</td><td>gi|254822516|ref|ZP_05227517.1| linear gramicidin synthetase subunit D [Mycobacterium intracellulare ATCC 13950]_4_5</td><td>C</td><td>1</td><td>57.89</td><td>1.9</td><td>Dual</td><td>19</td></tr>
<tr><td>G2U4PMU02GPXKR</td><td>Dinoflagellate</td><td>gi|220912507|ref|YP_002487816.1| amino acid adenylation protein [Arthrobacter chlorophenolicus A6]_1_3</td><td>C</td><td>1</td><td>40.48</td><td>0.74</td><td>LCL</td><td>42</td></tr>
</table>

x= for nr domain with mutltiple hits this information is not given

* names are genbank names associated with each sequence, the last two numbers for each name are the domain number and the total number of C or KS domains in the protein.

** sequences that are the top hit to multiple CalCOFI sequences have a range given and a representative ID is given

Chapter 5: Conclusions and future perspectives

Bacteria are no longer considered statistically or ecologically negligible. Many bacteria in the sea appear to be abundant and diverse; these are well represented in culture-dependent and independent based studies. However, there still remains an unknown fraction of slower growing and less abundant bacteria (e.g. actinomycetes). These rare bacteria must be included in our efforts to comprehensively understand microbes. Unfortunately, these slower growing bacteria are elusive even in metagenomics approaches. This dissertation presented three studies of bacteria and genes that are among the rare types in the ocean. The first study showed that the differences and similarities of two marine Actinobacteria are linked to secondary metabolism. The second study contributes to our knowledge of marine adaptation in the Gram-positive bacteria and shows evidence that they are adapted to the sea in fundamentally different way than Gram-negative bacteria. The final study demonstrates that there is an enormous potential to discover natural products in the seas but that potential seems to be hidden in eukaryotes and artifacts of DNA sequence generation. Bacterial type secondary metabolism genes are rare in metagenomic data relative to fatty acid synthases, phylogenetic markers and other housekeeping genes. Although bacterial genes related to secondary metabolism were not similar to anything observed in public databases signifying that there are natural products and potential cures to disease to be discovered.

The results in this dissertation contribute new insight about the evolution and ecology of *Salinispora*. A bioinformatics approach including controlled experiments has illuminated the differences and similarities of two species of *Salinispora*. The comparative genomics of *Salinispora* reveals adaptation at the species level and the genus level. In chapter 2 of this dissertation I was able to visualize on a gene-by-gene basis the differences and similarities of *Salinispora tropica* CNB-440 and *Salinispora arenicola* CNS-205. By manually curating the set of secondary metabolites and mobile genetic elements in each *Salinispora* genome, evidence was gained to show that the major functional types of genes that differ between the species are associated with secondary metabolism. The other main difference between the two species is the repertoire of mobile genetic elements they maintain, which happen to be located near most of the secondary metabolite gene clusters thus providing circumstantial evidence for how these clusters are horizontally transferred. Calculating a variety of metrics commonly used to test for HGT also supported horizontal gene transfer of secondary metabolites. Finally, the secondary metabolite genes were located in specific regions of each genome making them fit the general definition of a genomic island. The observation that secondary metabolites reside on genomic islands and are co-located with mobile genetic elements provides evidence that secondary metabolite genes are involved with ecological differentiation of the two species.

The prospects for the comparative genomics of *Salinispora* are bright. Currently approximately 100 *Salinispora* genomes are in the pipeline for sequencing. This large scale sequencing effort will further test the idea that secondary metabolites are species specific and allow people who want to understand the ecology and evolution of *Salinispora* to determine if the presence of specific gene clusters is truly species specific. One alternate hypothesis to species specificity is that secondary metabolite genes are derived from a local gene pool and provide adaptations to specific local conditions.

One of the major interests in *Salinispora* species is because of the observation that when seawater is replaced with deionized water no growth occurs. In Chapter 3, I took a bioinformatics approach to identify genes matching a specific set of criteria to identify genetic features related to the apparent seawater dependence of *Salinispora*. This approach identified that *Salinispora* has lost a mechanosensitive channel relative to closely related members of the Actinobacteria. Dr. Sergio Bucarey, a visiting professor from Chile, did a genetic experiment to test my hypothesis that *Salinispora* cannot survive osmotic down shock, because it does not have *mscL*. The results of Dr. Bucarey's experiment are the first to show physiological evidence in marine Actinobacteria that the lack of MscL prevents survival on DI water based media (Appendix A).

The physiological explanation for marine adaptation among actinomycetes is particularly useful because of the novel secondary metabolites that marine actinomycetes produce. The methods and genes from my study of marine adaptation can be used as a starting point to investigate why marine actinomycetes construct so many distinct natural products. Further systematic studies are needed to understand if marine actinomycetes truly produce significantly different secondary metabolites. Data from the third study showed results that are consistent with an ocean specific set of secondary metabolites relative to terrestrial habitats in metagenomes. The uniqueness of natural products from the sea may be related to adaptations to specific marine niches and unique marine biological targets. Alternatively, the unique chemical composition of seawater may be reflected in the types of natural products marine actinomycetes construct. For example, the abundance of chloride ions in natural products from the sea seems to be significantly higher then non-marine natural products. Future studies should test for this correlation and untangle ecological and chemical factors.

In response to the tedious task required to manually identify all secondary metabolite genes in *Salinispora* genomes, I created methods that automatically identify KS and C domains from natural product producing genes. This method was then adapted to become an online resource for people to analyze their own data. I applied this tool to study metagenomes and metatranscriptomes from marine plankton. The results of this research reveal several interesting things about KS and C domains in marine plankton. The abundance and distribution of natural product related gene sequences is not uniform and appears to be particularly enriched on larger particle sizes. The phylogeny of KS and C domains in metagenomes from plankton collected off the coast of California appears to contain known domain classes but represent novel groups within these classes perhaps representing biosynthetic pathways that make novel natural products. Finally, I show previously undiscovered diversity of modular KS domains from protists, which may represent an enormous potential for the discovery of novel natural products.

A by-product of the research on secondary metabolism in marine plankton was the creation of a database of all KS and C domains from nr. This dataset can serve as the basis for projects to manually curate and identify domains that have been linked to specific natural products in order to improve future predictions of the production of novel and known secondary metabolites.

These studies provide novel insight into an exciting group of newly discovered marine actinomycetes affectionately called *Salinispora*. The genome sequence analysis will be a resource as the search for new diversity of *Salinispora* continues. The NaPDoS tool will be of great use as genome and gene sequencing continues to become easier and cheaper. The great enigma of the research here is that *Salinispora* has yet to be found in any metagenome dataset. The ability to access *Salinispora* like species in metagenomes in the future may hold the key to further exploring "Neptune's medicine chest"(Balzar 2006).

None of this research would have been possible without understanding the facts of evolution. Darwin formulated his theory of natural selection without knowing how organisms inherit traits and was able to present a coherent argument. Ultimately the studies here can either be supported or refuted because of increased knowledge but the evidence from this dissertation can be considered a start to understand the comparative genomics of *Salinispora* and the abundance of secondary metabolites in marine communities.

Stephen J. Gould discussed the conundrum associated with the science of evolution (Gould 1990). It is not possible right now or perhaps it is that we are apart of the experiment and do not know it, to have a replicate samples of evolution. We can’t get back to the beginning, rerun the tape of time, and test how it would play out a second time. We cannot even observe directly how evolution occurred. We can however devise methods to provide evidence related to hypotheses of how evolution occurred. The methods we use are related to past discoveries and create progress towards a higher resolution picture of how evolution occurred. Currently it appears that metagenome and genome data has the potential to incredibly improve our understanding of evolution.

In conclusion, bioinformatics analyses have exhibited great utility. Bioinformatics studies are based on results of years of wet lab genetics and physiological experiments. Sequencing technology is outpacing Moores law and it is highly unlikely that genomic and metagenomic datasets will stop growing in size any time soon. What may have seemed inconceivable and in my eyes laughable, to sequence every living thing on the planet is becoming a more plausible goal each year. Assuming the technical difficulties of accessing the rare microbes are overcome and every gene that exists gets sequenced the next step will be to understand the function of every gene. The field of natural product research is dependent on knowing gene function and thus remains dependent on methods in biochemistry and genetics to properly identify new molecules and biosynthetic pathways. Yet, new high throughput methods are coming out that can quickly identify natural products (Kersten et al. 2011). In the future integration of automated sequencing, bioinformatics predictions like those from NaPDoS, and molecule detection will be able to search for predicted molecules.

References

Balzar J (2006) Neptune's Medicine Chest. Los Angeles Times. Los Angeles.

Gould SJ (1990) Wonderful life: the Burgess Shale and the nature of history: Norton.

Kersten RD, Yang Y-L, Xu Y, Cimermancic P, Nam S-J, Fenical W, Fischbach MA, Moore BS, Dorrestein PC (2011) A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nat Chem Biol 7(11): 794-802.

Appendix A: Genetic Complementation of the Obligate Marine Actinobacterium Salinispora tropica with the Large Mechanosensitive Channel Gene mscL Rescues Cells From Osmotic Downshock

Abstract

Marine actinomycetes in the genus *Salinispora* fail to grow when seawater is replaced with deionized water in complex growth media. While bioinformatic analyses have led to the identification of a number of candidate marine adaptation genes, there is currently no experimental evidence to support the genetic basis for the osmotic requirements associated with this taxon. One hypothesis is that the lineage specific loss of *mscL* is responsible for the failure to grow in DI water. The *mcsL* gene encodes a conserved trans-membrane protein that reduces turgor pressure under conditions of acute osmotic down-shock. In the present study, the *mscL* gene from a *Micromonospora* strain capable of growth on media prepared with DI water was transformed into *S. tropica* strain CNB-440. The single copy, chromosomal genetic complementation yielded a recombinant *Salinispora mscL*$^+$ strain that demonstrated an increased capacity to survive osmotic down-shock. The enhanced survival of the *S. tropica* transformant provides the first experimental genetic evidence that the loss of *mscL* is associated with the failure of *Salinispora* spp. to grow in low osmotic strength media.

Introduction

The obligate marine actinomycete genus *Salinispora* is comprised of the formally described species *S. tropica* and *S. arenicola (Maldonado et al. 2005)* and a third species for which the name "*S. pacifica"* has been proposed (Jensen and Mafnas 2006). The genus is broadly distributed in tropical and sub-tropical marine sediments (Jensen and Mafnas 2006) and is the source of a large number of structurally diverse secondary metabolites (Fenical and Jensen 2006) including the proteasome inhibitor salinosporamde A, which is in clinical trails as an anticancer agent (Fenical et al. 2009). *Salinispora* spp. produce a dense, non-fragmenting mycelium and non-motile spores that blacken the colony surface as is typical of the closely related genus *Micromonospora*. One of the unique characteristics of *Salinispora* spp., however, is that strains fail to grow when seawater is replaced with deionized (DI) water in complex growth media that lack added salts (Mincer et al. 2002; Maldonado et al. 2005).

Among Gram-negative marine bacteria, the requirement of seawater for growth has been linked to a specific sodium ion requirement (Oh et al. 1991). While a sodium requirement was originally reported for *Salinispora* spp., growth has subsequently been demonstrated with as little as 5 mM Na^+ if an appropriate osmotic environment is provided by the addition of alternative salts (Tsueng and Lam 2008). In addition, it was reported that *Salinispora* cells lyse in low ionic strength media (Tsueng and Lam 2008) suggesting they have poor tolerance for osmotic downshock. While the genetic basis for the failure of *Salinispora* strains to grow in low osmotic strength media has not been established, comparative genomics revealed a large family of highly duplicated polymorphic membrane proteins (PMPs) that were proposed to render cells unable to survive osmotic downshock (Penn et al. 2009). A more comprehensive bioinformatics analysis identified a larger pool of candidate marine adaptation genes and the lineage specific loss of *mscL* (see chapter 3), the product of which is a mechanosensitive channel that has been shown to alleviate cell lysis following osmotic downshock (Nakamaru et al. 1999).

Free-living microorganisms have developed robust mechanisms to maintain cell volume and integrity in response to changes in osmotic stress (Wood et al. 2001). These mechanisms include the accumulation of compatible solutes and mechanisms to release osmolytes under hypo-osmotic conditions. Mechanosensitive channels are present in a large variety of bacteria and thought to function as primary osmolyte release valves that reduce turgor pressure under conditions of osmotic downshock (Hoffmann et al. 2008). The mechanosensitive channel of large conductance (MscL) is nonselective in the ions and small molecules it transports and has been shown to open following osmotic downshock (Ajouz et al. 1998). Cells lacking MscL are thus unable to tolerate the transition from high to low osmotic conditions (Levina et al. 1999) as might be experienced in the transition from a marine to a non-marine environment.

The *E. coli mscL* gene was the first mechanosensitive channel to be cloned (Sukharev et al. 1994). Subsequent genetic experiments with the marine bacterium *Vibrio alginolyticus* revealed that the introduction of this gene alleviates cell lysis following osmotic downshock (Nakamaru et al. 1999). Similar functions were also demonstrated in the Gram-positive bacteria *Lactococcus lactis* (Folgering et al. 2005) and *B. subtilis* (Hoffmann et al. 2008).

Evidence that *Salinispora* spp. lack *mscL* coupled with the role of its protein product in relieving cell turgor pressure (Sukharev et al. 1997) led to the suggestion that the loss of this gene may account for the inability of *Salinispora* spp. to grow on complex media that lacks added salts (see chapter 3). In the experiments reported here, *S. tropica* strain CNB-440 was complemented with a copy of the *mscL* gene from a marine-derived *Micromonospora* strain (CNB-512) that was capable of growth on media prepared with DI water (Jensen et al. 1991). The resulting recombinant *Salinispora* strain displays enhanced survival following osmotic downshock. These results provide the first experimental evidence that the loss of *mscL* plays a major role in the failure of *Salinispora* strains to grow in low osmotic strength media.

Methods

Microorganisms

The type strain *S. tropica* CNB-440^T (accession number CP000667) (Maldonado et al. 2005) was chosen for complementation experiments based on an analysis of the genome sequence (Udwary et al. 2007), which did not contain the *MscL* gene (see chapter 3). *Micromonospora* sp. strain CNB-512 was used to complement CNB-440. It was isolated from a marine sediment sample and did not require seawater for growth (Jensen et al. 1991). Two exconjugants were generated from *S. tropica* CNB-440, one contained the recombinant plasmid pSET152::*mscl* and the other an empty plasmid. Strains CNB-440 and CNB-512 were grown in medium A1 (10 g starch, 4 g yeast extract, 2 g peptone, 1 liter natural seawater). *E. coli* was grown in Luria-Bertani (LB) medium (10 g Bacto tryptone, 5 g Bacto yeast extract, 5 g NaCl, 1 liter DI water). Descriptions of all *Salinispora*, *Micromonospora*, *and E.coli* strains and plasmids are presented in table A.1.

PCR analysis

Two sets of PCR primers were designed based on the *mscL* gene sequence (accession number NC_014815) obtained from the *Micromonospora* sp. L5 genome (accession number CP002399). One set *mscL-int*-F (5-TGACCTCCTCGCTGGGAGCC-3) and *mscL-int*-R (5-CGCGGTCGGCGTCGTCATC-3) amplifies a 320 bp internal fragment and the second set *mscL-ext*-F (5-GCCATCCGCGCCGGCGACCCG-3) and *mscL*-ext-R (5-GTCAGCGCGCGGCCGGGGGCTCC-3) amplifies 580 bp that includes the complete *mscL* gene and upstream flanking sequence that includes the promoter region. These primers were used to test for the presence of the *mscL* gene in a total of nine *Salinispora* strains (Table A.1) and to amplify the *mscL* gene sequence from *Micromonospora* strain CNB-512. *S. tropica* CNB-440, *S. arenicola* CNS-205, and "*S. pacifica"* CNT-133 were used as negative controls to verify that the primers were specific to *mscL*. Amplification was performed for 30 cycles (94°C denaturation for

30 s, 58°C annealing for 30 s, and 72°C extension for 1 min, followed by a 7 min extension at 72°C).

Cloning of the *Micromonospora mscL* gene

The *mscL* gene and flanking sequence (580 bp) was PCR amplified as described above from genomic DNA prepared from CNB-512 using the *mscL-ext* primers with restriction sites at the 5' ends, EcoRI-*mscL-ext*-F **(5-CTTGAATTC**AGCCGGTGCTTTTCTCGAAG-3) and XbaI-*mscL-ext*-R (**5-ATTCTAGA**GTCAGCGCGCGGCCGGGGGCTCC-3). The PCR product was purified, digested with the endonucleases EcoRI and XbaI, and ligated to the same sites of the Apra^{r} conjugative plasmid vector pSET152 (Bierman et al. 1992). The ligation mixture was electroporated into the *E. coli* host strain DH5α, plated on LB containing 50 µM apramacin, 0.5 mM isopropyl-D-thiogalactopyranoside (IPTG), and 40 µg/ml 5-bromo-4chloro-3-indolyl-D-galactopyranoside (X-Gal) at 37°C, and recombinants (white colonies) were screened by PCR using the same primers listed above. Plasmid DNA purified from one clone yielded an insert of the predicted size after digestion with EcoRI and XbaI and was subsequently sequenced verified. This plasmid, pSET152::*mscL*, was electroporated into the conjugative helper *E. coli* S17-1 (Simon et al. 1983) producing the strain *E. coli* S17-1/pSET152::*mscL*. Similar procedures were followed to generate a control plasmid that lacked the insert (pSET::empty).

Conjugation Assays

To conduct *E. coli/S. tropica* CNB-440 crosses, overnight LB cultures of the donor strains *E. coli* S17-1/pSET152::*mscL* and pSET::empty were grown for 4 h in 10 mL LB with 50 μg/ml apramycin. In parallel, *S. tropica* CNB-440 was grown in 30 mL A1 medium (70% seawater) for two days. One-half mL of the *E. coli* suspensions were then mixed with 0.5 mL of the *S. tropica* culture and the mixture spread onto A1 agar plates. After a 20h incubation at 33°C, the plates were overlaid with 1 mL of 2 mg/mL nalidixic acid to eliminate the *E. coli* donor strain and 1 mL of 4 mg/mL apramycin to select for *S. tropica* CNB-440 exconjugants. Exconjugants were visible after 2 weeks incubation at room temperature and individual colonies isolated onto A1 agar plates with 200 μg/ml apramacin and 100 μg/ml nalidixic acid (Lechner et al. 2011). In control experiments, plasmid insertion was highly stable even after three passages under non-selective conditions.

RNA isolation and *mscL*-specific RT-PCR

To isolate RNA, bacteria were grown for 5 days in medium A1 (70% sea water) at 27°C. Total RNA was extracted by TRIzol® reagent (Invitrogen, Carlsbad, CA) and treated with amplification grade, RNase-free DNase I (Gibco-BRL). Reverse transcription (RT) PCR was performed with 200 ng of DNase-treated RNA using a single-tube RT-PCR kit (Gibco-BRL). PCR amplification of the *mscL* gene was performed as previously described using the internal primer set. Genomic DNA served as a positive control, and DNase-treated RNA that had not been reverse transcribed was used as a negative control. Twenty-microliter aliquots were removed after 30 PCR cycles, stained with SYBR® Green, electrophoresed on a 1% agarose TBE gel, and analyzed using a Digital Science 120 system (Kodak).

Western blot MscL analysis

Membrane preparations followed previously described methods (Schnaitman 1971) with slight modification. Bacteria were grown for 5 days in 30 ml medium A1 with shaking (230 rpm, 27°C), chilled on ice, pelleted by centrifugation (7500 × *g*, 15 min, 4°C), resuspended in lysis buffer (10 mM Tris–HCl, pH 8, 10 mM $MgCl_2$), sonicated, and supplemented with 2 mM phenylmethylsulfonyl fluoride. Whole cells and debris were removed by low-speed centrifugation (5000 × *g*, 10 min) and total membrane fractions were obtained after 45 min of centrifugation at 13,000 × *g* at 4°C. Total membrane fractions were solubilized in 50 µl of Tris–HCl buffer (100 mM, pH 8) and 1% SDS. Proteins were separated by electrophoresis (12% SDS polyacrylamide gels), transferred to polyvinylidenedifluoride (PDVF) membranes, blocked for 1 hour in blocking buffer [phosphate-buffered saline (PBS), 5% nonfat milk with 3% bovine serum albumin] at room temperature. A rabbit polyclonal IgG antibody designed by Abgent Inc. (San Diego, Ca, USA) based on the *Micromonospora* L5 MscL immunogenic motif LDDVLGRRQEPPAPRC was then diluted 1:500 in blocking buffer and incubated overnight with the membrane at 4°C. Membranes were then incubated for 1 hour at room temperature with a 1:5000 dilution of IRDye®-conjugated goat anti-rabbit IgG (LI-COR® BIOSCIENCES) as a secondary antibody. Fluorescence was detected with a LI-COR Odyssey kit (LI-COR®BIOSCIENCES) and the membrane scanned using an Odyssey® CLx Infrared

Imaging System (LI-COR® BIOSCIENCES) operated in the 700/800 nm channel. The bands were analyzed using Odyssey® imaging software to quantify pixel intensity.

Growth estimates based on protein content

S. tropica CNB-440 and CNB-440 *mscL*$^+$ were grown in triplicate for 5 days in medium A1 (70% seawater) with apramycin (200 µg/ml), pelleted by centrifugation (7000 x *g*), washed twice with phosphate-buffered saline, and diluted 1:100 in PBS. Aliquots (200 µl) containing approximately 2x10^6 colony forming units (CFUs)/mL were inoculated into 100 ml medium A1 (70% sea water) and A1 prepared with DI water and allowed to grow for one week at 27°C while shaking at 230 rpm. Duplicate one ml subsamples were taken every 24h throughout the growth curve and assayed for total protein content using previously described methods (Makkar et al. 1982) and modifications (Meyers et al. 1998). In brief, the samples were centrifuged (13,800 × *g*) for 5 min. The pellets were washed by vortexing with 1 ml PBS (pH 7.0), centrifuged again as described above, and frozen (-20°C). For analysis, the pellets were re-suspended in 0.1 ml of 1 M NaOH, placed in boiling water for 10 min, neutralized by adding 0.02 ml of 5 M HCl, and the volumes adjusted to 1 ml by adding PBS. The samples were then centrifuged for 30 min and the absorbance of 0.8 ml measured at 230 and 260 nm using a Nanodrop 1000 spectrophotometer (Thermo Fisher Scientific). Protein concentration (µg/ml) was determined from the equation [Protein] = $(183 \times A_{230}) - (75.8 \times A_{260})$ (Makkar et al. 1982). The assay is linear over the range of 6-225 µg protein/ml (Makkar et al. 1982), and extracts from heavily turbid cultures were diluted in PBS to ensure that measurements remained within the linear range.

Effects of exposure to DI water on growth.

S. tropica CNB-440 and CNB-440 *mscL*$^{+}$ were grown in triplicate for 5 days in 30 ml A1 (70% seawater). The cells were pelleted, washed twice with DI water, and resuspended in 20 ml DI water without shaking at room temperature for various times from 1-72h. Aliquots (300 µl) were then spread plated onto medium A1 (70% seawater) and incubated at 30°C for two weeks. Growth was visually assessed.

Viability estimates

S. tropica CNB-440 and CNB-440 *mscL*$^{+}$ were grown in triplicate and exposed to DI water as described above. Live vs. dead cells were distinguished using the *Bac*Light LIVE/DEAD Bacterial Viability Kit (L7012, Life Technologies, Grand Island, NY) following the manufacture's instructions. In brief, equal volumes of dye components A and B were combined in a microfuge tube, mixed, and 3 µL added for each 1 mL of bacterial suspension analyzed. The suspensions were thoroughly mixed, incubated at room temperature in the dark for 15 min, and 5 µL placed between a glass microscope slide and 18 mm square coverslip. The samples were observed at 40x using an Olympus MVX10 fluorescence microscope (Olympus, Center Valley, PA) equipped with filter cube U-MCFPHQ/XL. Fluorescence associated with viable (green) and non-viable (red) cells was measured at 510-540 and 620-650 nm, respectively. Images of ten different fields were captured for each treatment using an

Olympus DC71 camera operated by DP Manager® software. The experiment was repeated three times for each strain.

In an effort to quantify cell viabillity, *S. tropica* CNB-440 and CNB-440 *mscL*$^+$ were cultured in triplicate for 5 days in 30 ml A1 (70% seawater). One half of each culture was heat-killed by boiling for 20 min and confirmed to be non-viable by plating on A1 agar (70% seawater). The suspensions of live and heat-killed cells were adjusted to an OD_{600} of 0.30 using a spectrophotometer (BioPhotometer, Eppendorf®). Live and dead bacterial suspensions (2 mL) were then prepared in ratios of 0:100, 10:90, 50:50, 90:10, 100:0 and stained as described above using the *Bac*Light LIVE/DEAD Bacterial Viability Kit to generate a standard curve of fluorescence vs. percent viable cells.

In parallel, *S. tropica* CNB-440 and CNB-440 *mscL*$^+$ were grown in triplicate for 5 days in 30 ml A1 (70% seawater). Cells were pelleted and washed as described above and soaked in DI water for 24 h. The cell suspensions were then adjusted to an OD_{600} of ca. 0.30 in 2 mL total volume, stained as described above, and 100 μL pipetted into separate wells of a 96-well, flat-bottomed, micro-titer plate. The plate was incubated at room temperature in the dark for 15 min after which the fluorescence emission at 500-700 nm was measured using a micro-titer plate reader (SpectraMax M2, Molecular Devices, Inc., Sunnyvale, CA) with the excitation wavelength set to 470 nm. The data were analyzed for each bacterial suspension by calculating the ratio of the integrated intensity of the green (510–540 nm) and red (620–650 nm) fluorescence emissions and plotting these values against the standard curve described above to estimate the percentage of live cells in the suspension.

Gadalidium experiments

Micromonospora strain CNB-512 was grown in triplicate for 5 days in 30 mL medium A1 (70% seawater), pelleted by centrifugation (7,000 x *g*), washed twice with PBS, and re-suspended in 10 mL DI water. Aliquots (200 µl) were inoculated into 100 ml medium A1 (70% sea water) with and without 1 mM gadolinium chloride and A1 (DI water) with and without 1 mM gadolinium chloride and allowed to grow for two week at 27°C while shaking at 230 rpm. Duplicate one ml subsamples were taken every 24h throughout the growth curve and assayed for total protein content using the method described above.

Results

PCR probing for the *mscL* gene

The *mscL* gene was not observed in the genome sequences of *S. tropica* strain CNB-440 (Figure A.1) or *S. arenicola* strain CNS-205. To determine if the absence of this gene is a common feature of the genus, we PCR probed for a 320 bp internal region and a 580 bp region that included the upstream *mscL* flaking sequence in three strains of *S. tropica*, *S. arenicola*, and "*S. pacifica"* (Table A.2). No PCR products were obtained from any of these nine strains while products of the predicted size and sequence were consistently amplified using both sets of primers and DNA templates prepared from three *Micromonospora* strains .

Genetic complementation and expression of *mscL* in *S. tropica*

The genera *Micromonospora* and *Salinispora* are closely related within the family Micromonosporarceae. Nonetheless, sequence differences even among closely related taxa can present formidable barriers to the construction of interspecific hybrids. To construct a *Salinispora* interspecies recombinant, we PCR amplified the *mscL* gene from *Micromonospora* strain CNB-512 using primers designed to amplify the complete gene and upstream promoter region (580 bp). This PCR product was then successfully ligated into the pSET152 conjugative plasmid and introduced into *E. coli* S17-1 as a donor strain (*E.coli* S17-1/pSET152::*mscL*) (Figure A.2A). Retrosequencing revealed that pSET152 integration occurred at strop_0483, one of three previously identified *S. tropica* pseudointegration sites (Lechner et al, in press). Following transformation and the selection of an apramycin resistant *S. tropica* exconjugant (*S. tropica mscL*$^{+}$), PCR amplification yielded a 580 bp product that was sequence verified as *mscL* (Figure A.2B). Furthermore, RT-PCR experiments revealed that *mscL* was expressed in the CNB-440 exconjugant (Figure A.2C). Thus, the first *Salinispora* interspecies genetic hybrid has been successfully constructed and the native *Micromonospora* promoter is active in a *Salinispora* genetic background.

Western blot analysis and MscL protein detection

To determine if the *mscL* transcripts were translated and the resulting protein incorporated into the cell membrane of *S. tropica mscL*$^{+}$, a polyclonal antibody targeting the *Micromonospora* CNB-512 MscL sequence was developed. Western blot analysis of membrane preparations derived from cultures of *S. tropica mscL*$^{+}$ revealed a specific, 15 kDa band that corresponds to MscL (Figure A.2D). This band was present in both the wild type *Micromonospora* strain CNB-512 and *S. tropica mscL*$^{+}$, however it was not observed in membrane preparations generated from the wild type *S. tropica* CNB-440 strain. MscL production in *Micromonospora* CNB-512 was standardized to 100% (16.66 pixels) and compared with two recombinant *S. tropica mscL*$^{+}$ strains. The fluorescence intensity of the hybridized probe was 6.89 and 6.75 pixels, corresponding to 41.5% and 40.66% of the positive control. These results demonstrate that MscL is incorporated into the *S. tropica* cell membrane albeit at reduced levels relative to the native *Micromonospora* strain.

Effect of osmotic downshock on *Salinispora* survival

Initial efforts to cultivate *S. tropica* CNB-440 *mscL*$^{+}$ revealed that this otherwise isogenic exconjugant, like the CNB-440 wild-type (WT) strain, failed to grow in complex media prepared with DI water (data not shown). Consequently, we used two different approaches to test for the effects of exposure to DI water on cell viability. The first test involved a visual examination of growth on A1 media prepared with seawater following exposure to DI water for 1-72h. The results provide clear and reproducible evidence that growth was reduced in a time-dependent fashion in the WT strain yet remained largely unchanged in the *S. tropica* CNB-440 *mscL*$^{+}$ exconjugant (Figure A.3). Given that *Salinispora* strains produce branching filaments, it was difficult to measure growth using traditional optical density or colony counting methods. For this reason, the effects of exposure to DI water on cell viability was further explored using the *Bac*Light LIVE/DEAD Bacterial Viability Kit. When grown in media prepared with seawater, cultures of the WT and *mscL*+ strains were dominated by viable cells (Figure A.4A, B). However, following a 24h exposure to DI water, green fluorescence was dramatically reduced in the WT strain indicating a lack of intact cell membranes (Figure A.4C). The intense red emission from the same sample indicates that most cellular membranes had been disrupted and supports prior observations that *Salinispora* strains lyse in low osmotic strength media (Tsueng and Lam 2008). Considerable green fluorescence is maintained in the *mscL*+ strain following exposure to DI water (Figure A.4D) suggesting that the introduction of this gene has made the cells less susceptible to lysis. In an effort to quantify viability using the *Bac*Light kit, the fluorescence emissions corresponding to various ratios of live and dead cells were measured (Figure A.5A). When plotted as the percentage of viable bacteria vs. the ratio of green to red fluorescence, a linear relationship was observed (Figure A.5B). Following a 24-hour exposure to DI water, the green/red fluorescence ratio for the wild type *S. tropica* CNB-440 strain corresponded to ca. 20% viable bacteria while the *mscL*+ exconjugant was greater than 80%. Thus it can be estimated that the introduction of the *mscL* gene increased viability by ca. 80%.

***MscL* chemical knock out**

Gadolinium chloride is a specific inhibitor of MscL function (Berrier et al. 1992). To test the hypothesis that MscL provides resistance to osmotic downshock, the marine derived *Micromonospora* strain CNB-512 was tested for growth on media prepared with DI water supplemented with 1mM $GaCl_2$. While this strain grew equally well on media prepared with seawater, DI water, and seawater supplemented with $GaCl_2$, growth as measured by total protein content was dramatically reduced when this compound was added to the medium prepared with DI water (Figure A.6A). Viability as measured using the *Bac*Light kit was also reduced dramatically when $GaCl_2$ was added to the medium prepared with DI water (Figure A.6B). Strain CNB-512 was capable of growth on $GaCl_2$ concentrations as high as 5 mM suggesting that compound toxicity was not a factor in the results. These experiments were repeated on two additional *Micromonospora* strains (Table A.1) and similar results were obtained (data not shown).

Discussion

The genus *Salinispora* is unique among marine-derived actinomycetes in that all species cultured to date fail to grow in low osmotic strength media. While comparative genomics has been used to identify a pool of candidate marine adaptation genes that may be associated with this phenotype (Penn et al. 2009), it has been proposed that the lineage specific loss of *mscL* plays a major role in the failure of *Salinispora* spp. to survive osmotic downshock (see chapter 3). The present study reports the first experimental evidence in support of this hypothesis.

The recently released *Micromonospora* L5 genome sequence (accession number CP002399) facilitated the design of two *mscL* specific primer sets that were used to successfully amplify this gene and the upstream promoter region from the marine-derived but non-seawater requiring *Micromonospora* strain CNB-512 (Jensen et al. 1991). *MscL* was not detected using either primer set in nine *Salinsipora* strains representing all three currently recognized species or in six *Salinispora* genome sequences (data not shown) supporting the proposal that the loss of this gene was a lineage-specific event. In future studies, a PCR assay targeting *mscL* may represent a quick approach to distinguish between *Salinispora* and *Micromonospora* strains, which are not readily resolved based on morphological features.

While a recently developed genetic system has been used to inactivate (Eustaquio et al. 2008) and reintroduce (Lechner et al., in press) genes in *Salinispora* spp., the results presented here represent the first use of the pSET152 conjugative plasmid to introduce a non-*Salinispora* gene into a *Salinispora* genetic background. Remarkably, only a small genetic cassette harboring the *mscL* open reading frame and the 100 base pair native promoter region was sufficient for the subsequent expression of this gene in *S. tropica* CNB-440 indicating that no additional species-specific factors are required. More importantly, a polyclonal antibody revealed that the gene product was associated with a membrane fraction of the CNB-440 *mscL*+ exconjugant providing evidence that it was incorporated into the cytoplasmic membrane as has been shown in similar experiments with *E. coli* (Sukharev et al. 1997).

Although the recombinant *Salinispora* *mscL*$^{+}$ strain expressed the MscL protein and it appears to have been incorporated into the cytoplasmic membrane, this in itself was not sufficient to facilitate growth in complex media prepared without added salts. There are a number of possible explanations for this including the relatively low levels of MscL expression relative to the parent *Micromonospora* strain CNB-512 (Figure A.2). Alternatively, other marine adaptation genes such as a highly duplicated family of polymorphic membrane proteins that appears to have been acquired from marine bacteria may contribute to the inhibitory effects of a low osmotic strength environment (Penn et al. 2009). Nevertheless, the introduction of the single copy *mscL* gene into *S. tropica* CNB-440 enhanced survival following osmotic downshock providing yet another example of the role of MscL in osmoadaptation. The chemical knockout of MscL function in *Micromonospora* CNB-512 using gadalidium further supports the role of this protein in surviving osmotic downshock. It is also of interest to note that *mscS* homologs detected in both *Salinispora* genomes do not appear to complement *mscL* function as has been observed in *E. coli* (Levina et al. 1999). These results provide the first experimental evidence that the loss of *mscL* is associated with the inability of *Salinispora* spp. to grow in complex media that lacks added salts. Although there are no known benefits associated with *mscL* loss in *Salinispora*, it may be an important factor that contributes to their reported requirement of seawater for growth.

The *Salinispora* 16S rRNA phylogeny reveals that it is closely related to a large number of non-marine actinomycete genera. Thus, it can be proposed that the environmental distribution of this lineage is the result of a secondary introduction into the marine environment. Given the consistent salinity of seawater, it would not be surprising if the loss of *mscL* had no effect on the ability of an ancestral *Salinispora* strain to survive in the marine environment. This loss likely occurred prior to speciation within the genus and may account for the fact that *Salinispora* strains have yet to be reported outside of the marine environment. It is of interest to note that no other marine-derived actinobacteria for which genome sequences are available lack *mscL* although many other marine bacteria are missing this gene. This may be due to the possibility that *Salinispora* spp. have been in the marine environment longer than other marine actinobacteria or simply reflect the stochastic nature of selectively neutral evolutionary events. It is intriguing to speculate that the random loss of a single gene may have resulted in the obligate marine distribution of this unusual actinomycete lineage.

Acknowledgements

Appendix A, in full, is submitted to Applied and Environmental Microbiology, 2012, Sergio Bucarey, Kevin Penn, Lauren Paul, William Fenical, Paul R. Jensen. The dissertation author conceived the methods and aided in parts of the analyses, which forms the basis of this publication.This research was supported by the National Institutes of Health grant R01-GM086261 to PRJ. SAB was supported by a STIPAS postdoctoral fellow from MECESUP 2 and the University of Chile. The authors thank Brad Moore and Anna Lechner for providing the pSET152 plasmid, transformation protocols, and E. coli strains.

References

Ajouz B, Berrier C, Garrigues A, Besnard M, Ghazi A (1998) Release of Thioredoxin via the Mechanosensitive Channel MscL during Osmotic Downshock of Escherichia coli Cells. Journal of Biological Chemistry 273(41): 26670-26674.

Berrier C, Coulombe A, Szabo I, Zoratti M, Ghazi A (1992) Gadolinium ion inhibits loss of metabolites induced by osmotic shock and large stretch-activated channels in bacteria. European Journal of Biochemistry 206(2): 559-565.

Bierman M, Logan R, O'Brien K, Seno ET, Nagaraja Rao R, Schoner BE (1992) Plasmid cloning vectors for the conjugal transfer of DNA from Escherichia coli to Streptomyces spp. Gene 116(1): 43-49.

Eustaquio AS, Pojer F, Noel JP, Moore BS (2008) Discovery and characterization of a marine bacterial SAM-dependent chlorinase. Nat Chem Biol 4(1): 69-74.

Fenical W, Jensen PR (2006) Developing a new resource for drug discovery: marine actinomycete bacteria. Nature Chemical Biology 2(12): 666-673.

Fenical W, Jensen PR, Palladino MA, Lam KS, Lloyd GK, Potts BC (2009) Discovery and development of the anticancer agent salinosporamide A (NPI-0052). Bioorganic &amp; Medicinal Chemistry 17(6): 2175-2180.

Folgering JHA, Moe PC, Schuurman-Wolters GK, Blount P, Poolman B (2005) Lactococcus lactis Uses MscL as Its Principal Mechanosensitive Channel. Journal of Biological Chemistry 280(10): 8784-8792.

Hoffmann T, Boiangiu C, Moses S, Bremer E (2008) Responses of Bacillus subtilis to Hypotonic Challenges: Physiological Contributions of Mechanosensitive Channels to Cellular Survival. Applied and Environmental Microbiology 74(8): 2454-2460.

Jensen PR, Mafnas C (2006) Biogeography of the marine actinomycete *Salinispora*. Environmental Microbiology 8(11): 1881-1888.

Jensen PR, Dwight R, Fenical W (1991) Distribution of actinomycetes in near-shore tropical marine sediments. Applied and Environmental Microbiology 57(4): 1102-1108.

Lechner A, Eust·quio AS, Gulder TAM, Hafner M, Moore BS (2011) Selective Overproduction of the Proteasome Inhibitor Salinosporamide A via Precursor Pathway Regulation. Chemistry & biology 18(12): 1527-1536.

Levina N, Totemeyer S, Stokes NR, Louis P, Jones MA, Booth IR (1999) Protection of Escherichia coli cells against extreme turgor by activation of MscS and MscL mechanosensitive channels: identification of genes required for MscS activity. EMBO J 18(7): 1730-1737.

Makkar HPS, Sharma OP, Dawra RK, Negi SS (1982) Simple Determination of Microbial Protein in Rumen Liquor. Journal of Dairy Science 65(11): 2170-2173.

Maldonado LA, Fenical W, Jensen PR, Kauffman CA, Mincer TJ, Ward AC, Bull AT, Goodfellow M (2005) *Salinispora arenicola* gen. nov., sp. nov. and *Salinispora tropica* sp. nov., obligate marine actinomycetes belonging to the family Micromonosporaceae. International Journal of Systematic and Evolutionary Microbiology 55(5): 1759-1766.

Meyers PR, Bourn WR, Steyn LM, van Helden PD, Beyers AD, Brown GD (1998) Novel Method for Rapid Measurement of Growth of Mycobacteria in Detergent-Free Media. Journal of Clinical Microbiology 36(9): 2752-2754.

Mincer TJ, Jensen PR, Kauffman CA, Fenical W (2002) Widespread and Persistent Populations of a Major New Marine Actinomycete Taxon in Ocean Sediments. Applied and Environmental Microbiology 68(10): 5005-5011.

Nakamaru Y, Takahashi Y, Unemoto T, Nakamura T (1999) Mechanosensitive channel functions to alleviate the cell lysis of marine bacterium, *Vibrio alginolyticus*, by osmotic downshock. FEBS Letters 444(2-3): 170-172.

Oh S, Kogure K, Ohwada K, Simidu U (1991) Correlation between Possession of a Respiration-Dependent Na+ Pump and Na+ Requirement for Growth of Marine Bacteria. Applied and Environmental Microbiology 57(6): 1844-1846.

Penn K, Jenkins C, Nett M, Udwary DW, Gontang EA, McGlinchey RP, Foster B, Lapidus A, Podell S, Allen EE, Moore BS, Jensen PR (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. ISME J 3(10): 1193-1203.

Schnaitman CA (1971) Solubilization of the Cytoplasmic Membrane of Escherichia coli by Triton X-100. Journal of Bacteriology 108(1): 545-552.

Simon R, Priefer U, Puhler A (1983) A Broad Host Range Mobilization System for In Vivo Genetic Engineering: Transposon Mutagenesis in Gram Negative Bacteria. Nat Biotech 1(9): 784-791.

Sukharev SI, Blount P, Martinac B, Kung aC (1997) Mechanosensitive Channels of *Escherichia coli*: The MscL Gene, Protein, and Activities. Annual Review of Physiology 59(1): 633-657.

Sukharev SI, Blount P, Martinac B, Blattner FR, Kung C (1994) A large-conductance mechanosensitive channel in E. coli encoded by mscL alone. Nature 368(6468): 265-268.

Tsueng G, Lam K (2008) A low-sodium-salt formulation for the fermentation of salinosporamides by *Salinispora tropica* strain NPS21184. Applied Microbiology and Biotechnology 78(5): 821-826.

Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W, Jensen PR, Moore BS (2007) Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. Proceedings of the National Academy of Sciences 104(25): 10376-10381.

Wood JM, Bremer E, Csonka LN, Kraemer R, Poolman B, van der Heide T, Smith LT (2001) Osmosensing and osmoregulatory compatible solute accumulation by bacteria. Comparative Biochemistry and Physiology - Part A: Molecular &amp; Integrative Physiology 130(3): 437-460.

Figures

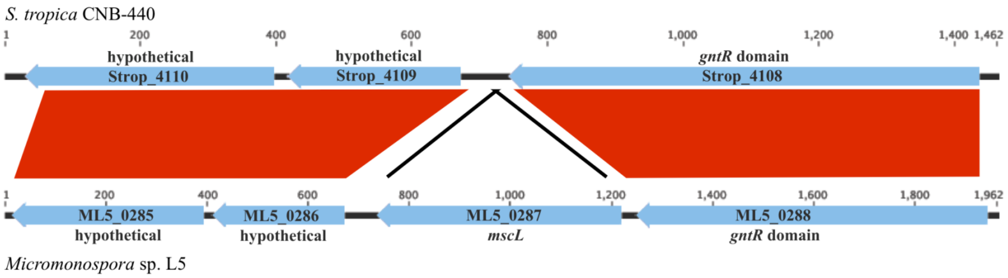


Figure A.1 Regional synteny plot of the *Micromonospora* L5 and *S. tropica* CNB-440 genomes. Red indicates syntenic regions. Gene numbers (locus tags) and Genbank annotations are listed. Tick marks represent base pairs.

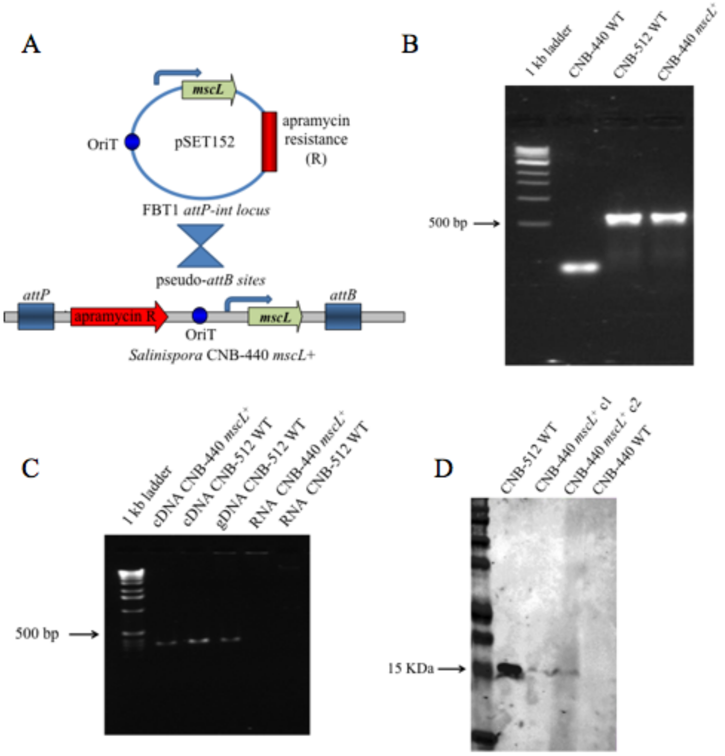


Figure A.2 Complementation experiments. (A) Diagram of the conjugation assay in which an *E. coli* donor strain harboring the *Micromonospora* CNB-512 *mscL* gene (S17-1/ pSET152::*mscL*) was used to introduce *mscL* into the recipient *S. tropica* CNB-440 strain. (B) PCR amplification of the *mscL* gene from *S. tropica* CNB-440 *mscL*+ and *Micromonospora* CNB-512 using the primer set EcoRI-*mscL*-*ext*-F/R (580 bp product). No appropriately sized product was observed from the CNB-440 WT strain. (C) PCR amplification of the *mscL* gene from cDNA generated from the CNB-440 *mscL*+ transformant and both cDNA and gDNA generated from *Micromonospora* strain CNB-512 using the primer set *mscL*-*int*-F/R (320 bp product). No products were observed from RNA controls. (D) Western blot analysis reveals the association of MscL with a membrane-enriched subcellular fraction as detected using an MscL specific polyclonal antibody. The arrow shows the expected size of the protein, which was detected in relatively low quantities in two CNB-440 *mscL*+ transformants relative to the CNB-512 WT.

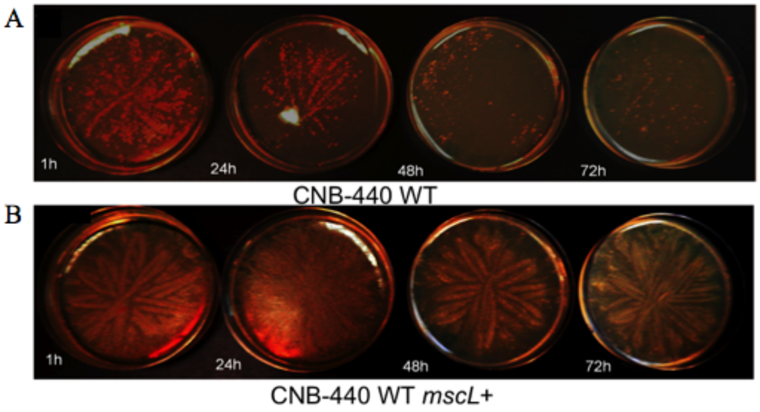


Figure A.3: Growth of the *S. tropica* strain CNB-440 wild type (WT) and *mscL*+ transformant after exposure to DI water. (A) The WT showed a negative visual growth response in relation to increased exposure to DI water from 1-72h prior to plating on media prepared with seawater. (B) The otherwise isogenic *mscL*+ plus transformant grew considerably better following DI exposure. A representative of three replicate experiments is shown.

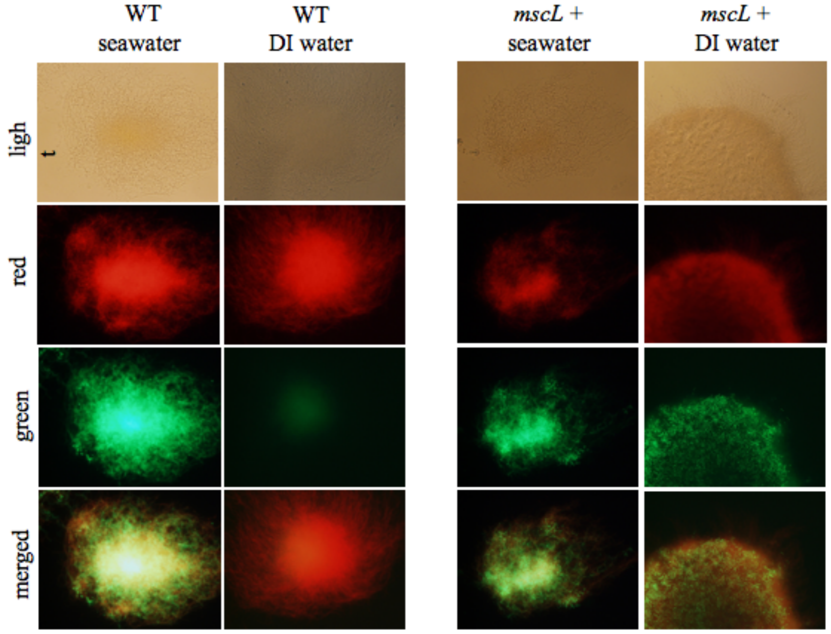


Figure A.4: Viability of *S. tropica* CNB-440 wild type (WT) and the *mscL*+ transformant as measured using the *Bac*Light Bacterial Viability Kit following exposure to seawater (control) or DI water for 24 h. Mycelial masses were viewed at 40x using bright field, red (620–650 nm) and green (510–540 nm) filters, and merged.

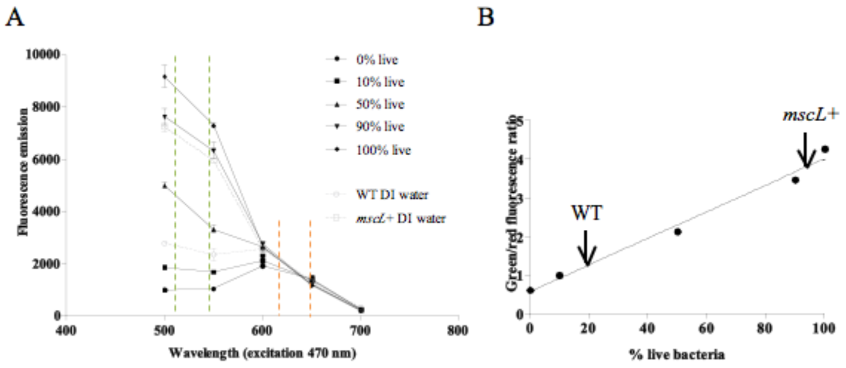


Figure A.5: Viability quantification of *S. tropica* CNB-440 using the *Bac*Light Bacterial Viability Kit. (A) Fluorescence emissions in the viable (green) and dead (red) wavelengths for different ratios of live and dead cells along with the WT strain and the *mscL*+ transformant following 24 h exposure to DI water. B) The integrated 510-540 nm (green) and 620-650 nm (red) fluorescence ratio for the WT following a 24-hour exposure to DI water corresponds to ca. 20% viable bacteria while the *mscL*+ exconjugant corresponds to greater than 80% viable bacteria. Average + STD for three replicate experiments plotted.

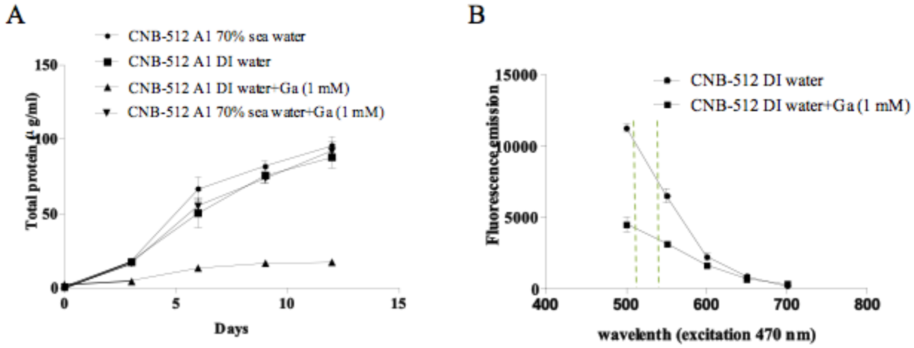


Figure A.6: Chemical knockout of *mscL* function. (A) Growth as measured by protein content was equal in *Micromonospora* strain CNB-512 grown in media prepared with 70% seawater, 100% DI water, and 100% DI water plus 1 mM gadalidium while growth in DI water with gadalidium was dramatically reduced. (B) Viability of *Micromonospora* strain CNB-512 grown in media prepared with 100% DI water with and without 1 mM gadalidium as measured using the *Bac*Light Bacterial Viability Kit. Green lines indicate wavelengths of viable fluorescence emmissions.

Tables

Table A.1: Bacterial strains and plasmids used in this study (16S rRNA accession numbers in parentheses)

Strain or plasmid	Genotype	Source
CNB-440	*S. tropica*	Bahamas (CP000667)
CNB-536	*S. tropica*	Bahamas (AY040618)
CNH-898	*S. tropica*	Bahamas (AY040622)
CNS-205	*S. arenicola*	Palau (NC_009953)
CNH-665	*S. arenicola*	Bahamas
CNS-325	*S. arenicola*	Palau (GU593973)
CNH-662	*S. arenicola*	Bahamas
CNT-133	*"S. pacifica"*	Fiji (HQ218996)
CNS-844	*"S. pacifica"*	Fiji (HQ642897)
CNT-131	*"S. pacifica"*	Fiji (HQ642896)
CNY-369	*S. tropica* pset152::*mscL*	This work
CNY-370	*S. tropica* pset152::*mscL*	This work
CNY-372	*S. tropica* pset152 empty	This work
CNB-512	*Micromonospora* sp.	Bahamas, AY040624
CNB-394	*Micromonospora* sp.	Bahamas, AY040625
CNX-434	*Micromonospora* sp.	Palmyra
DH5α	*E. coli* {*endA1hsdR17* (r_ m_) *supE44 thi-1 recA1 gyrA*(Nalr) *relA1* _(*lacZYA-argF*)*U169 deoR [_80_(lacZ)M15]*}	
S17-1	*E. coli* recA pro hsdR RP4-2-Tc::Mu-Km::Tn7	Simon *et al*., 1989
pTOPO	plasmid	Invitrogen®
pset152	plasmid	Bierman *et al*., 1992
pset152::mscL	plasmid	This work

Table A.2: PCR amplification of the *mscL* gene

		Growth		PCR product	
Species	Strain	Seawater	DI water	320 bp	580 bp
S. tropica	CNB-440	+++	---	no	no
S. tropica	CNB-536	+++	---	no	no
S. tropica	CNH-898	+++	---	no	no
S. arenicola	CNS205	+++	---	no	no
S. arenicola	CNH-665	+++	---	no	no
S. arenicola	CNS-325	+++	---	no	no
"*S. pacifica*"	CNT-133	+++	---	no	no
S. pacifica	CNS-844	+++	---	no	no
"*S. pacifica*"	CNT-131	+++	---	no	no
Micromonospora sp.	CNB-394	+++	+++	yes	yes
Micromonospora sp.	CNB-512	+++	+++	yes	yes
Micromonospora sp.	CNX-434	+++	+++	yes	yes

Appendix B: The Natural Product Domain Seeker NaPDoS: a Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity

Abstract

New bioinformatic tools are needed to analyze the growing volume of DNA sequence data. This is especially true in the case of secondary metabolite biosynthesis, where the highly repetitive nature of the associated genes creates major challenges for accurate sequence assembly and analysis. Here we introduce the web tool Natural Product Domain Seeker (NaPDoS), which provides an automated method to assess the secondary metabolite biosynthetic gene diversity and novelty of strains or environments. NaPDoS analyses are based on the phylogenetic relationships of sequence tags derived from polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) genes, respectively. The sequence tags correspond to PKS-derived ketosynthase domains and NRPS-derived condensation domains and are compared to an internal database of experimentally characterized biosynthetic genes. NaPDoS provides a rapid mechanism to extract and classify ketosynthase and condensation domains from PCR products, genomes, and metagenomic datasets. Close database matches provide a mechanism to infer the generalized structures of secondary metabolites while new phylogenetic lineages provide targets for the discovery of new enzyme architectures or mechanisms of secondary metabolite assembly. Here we outline the main features of NaPDoS and test it on four draft genome sequences and two metagenomic datasets. The results provide a rapid method to assess secondary

metabolite biosynthetic gene diversity and richness in organisms or environments and a mechanism to identify genes that may be associated with uncharacterized biochemistry.

Introduction

Genome sequencing has revealed that the secondary metabolite potential of even well studied bacteria has been severely underestimated (Bentley et al. 2002; Ikeda et al. 2003). This revelation has led to an explosion of interest in genome mining as an approach to natural product discovery (Lautru et al. 2005; Hornung et al. 2007; Udwary et al. 2007; Challis 2008; Winter et al.; Eustáquio et al. 2011). Considering that natural products remain one of the primary sources of therapeutic agents (Baker et al. 2007; Newman and Cragg 2007), sequence analysis provides opportunities to identify strains with the greatest genetic potential to yield novel secondary metabolites prior to chemical analysis and thus increase the rate and efficiency with which new drug leads are discovered. In addition, community or metagenomic analyses can be used to identify environments with the greatest secondary metabolite potential and to address ecological questions related to secondary metabolism. To capitalize on these opportunities, it is critical that new bioinformatics tools be developed to handle the massive influx of sequence data that is being generated from next generation sequencing technologies (McPherson 2009).

Polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs) are large enzyme families that account for many clinically important pharmaceutical agents. These enzymes employ complimentary strategies to sequentially construct a diverse array of natural products from relatively simple carboxylic acid and amino acid building blocks using an assembly line process (Finking and Marahiel 2004; Hertweck 2009). The molecular architectures of PKS and NRPS genes have been reviewed in detail and minimally consist of activation (AT or A), thiolation (ACP or PCP), and condensation (KS or C) domains, respectively (Shen 2003; Lautru and Challis 2004; Weissman 2004; Sieber and Marahiel 2005; Fischbach and Walsh 2006). These genes are among the largest found in microbial genomes and can include highly repetitive modules that create considerable challenges to accurate assembly and subsequent bioinformatic analysis (Udwary et al. 2007).

When the challenges associated with PKS and NRPS gene assembly can be overcome, a number of effective bioinformatics tools have been developed for domain parsing (Ansari et al. 2004; Rausch et al. 2005) and domain string analysis (Starcevic et al. 2008; Yadav et al. 2009). In cases of modular type I PKSs and NRPSs where domain strings follow the "co-linearity rule" such that substrates are incorporated and processed following the precise domain organization observed in the pathway, bioinformatics has been used to make accurate structural predictions about the metabolic products of those pathways (McAlpine et al. 2005). However, the increasing number of exceptions to co-linearity, such as module skipping and stuttering (Moss et al. 2004), create limitations for precise, sequence-based structure prediction. The bioinformatic tools currently available for secondary metabolism have been reviewed (Bachmann et al. 2009; Jenke-Kodama and Dittmann 2009a) and are complemented by the recent release of antiSMASH, which has the capacity to accurately identify and provide detailed sequence analysis of gene clusters associated with all known secondary metabolite chemical classes (Medema et al. 2011). While all of these tools have useful applications, NaPDoS employs a phylogeny based classification system that can be used to quantify and distinguish KS and C domain types from a variety of datasets including the incomplete genome assemblies typically obtained using next generation sequencing technologies. These specific domains were selected because they are highly conserved and have proven to be among the most informative in a phylogenetic context (Rausch et al. 2007; Nguyen et al. 2008).

Phylogenomics provides a useful approach to infer gene function based on phylogenetic relationships as opposed to sequence similarities (Eisen 1998; Eisen and Fraser 2003). While the evolutionary histories of PKS and NRPS genes are largely uninformative due to their size and complexity, KS and C domain phylogenies reveal highly supported clustering patterns. These patterns have been used to distinguish type II PKSs associated with spore pigment and antibiotic biosynthesis (Metsä-Ketelä et al. 1999), type I modular and hybrid PKSs (Moffitt and Neilan 2003), and subsequently to identify many different PKSs types (Jenke-Kodama et al. 2005). KS phylogeny has also been used to predict pathway associations (Ginolhac et al. 2005; Jenke-Kodama and Dittmann 2009a) and, in some cases, the secondary metabolic products of those pathways (Gontang et al. 2007; Nguyen et al. 2008; Freel et al. 2011). Phylogenetics has also been used to successfully identify PKS sequences from complex metagenomic datasets (Foerstner et al. 2008). Likewise, C domain phylogeny clearly delineates functional subtypes as opposed to species relationships (Roongsawang et al. 2005) and has been used to identify new functional classes, such as the “starter” C domain (Rausch et al. 2007). Taken together, the established phylogenetic relationships of KS and C domains provide an effective framework within which to assess secondary metabolite gene richness and diversity and to identify new functional classes that may be associated with uncharacterized biosynthetic mechanisms.

Here we introduce the web tool Natural Product Domain Seeker (NaPDoS), which extracts and rapidly classifies KS and C domains from a wide range of sequence data. The results can be used to assess the potential for PKS and NRPS secondary metabolite biosynthesis in organisms or environments and to identify new phylogenetic lineages, which can subsequently be investigated as a source of new mechanistic biochemistry. We tested NaPDoS on four draft bacterial genome sequences and two metagenomic datasets. The results reveal a remarkable level of secondary metabolite gene diversity among closely related strains and provide a mechanism to assess secondary metabolism from poorly assembled genomic data.

Methods

Reference database. KS and C domains were extracted from select PKS and NRPS genes associated with experimentally characterized biosynthetic pathways using the online program NRPS-PKS (http://www.nii.res.in/searchall.html) (Ansari et al. 2004; Yadav et al. 2009). The pathways selected include representatives of the currently known enzyme architectures and functions associated with type I and II PKSs and NRPSs and thus this database is not meant to be comprehensive. The biochemical function and enzyme architecture of each domain was manually confirmed by analysis of the associated domain string and secondary metabolic product. Based on these results, each sequence was preliminarily assigned to a domain class. The compound produced by the associated pathway, the literature reference including PubMed ID, and the gene accession number was also recorded for each domain.

Sequence alignment and phylogeny. The amino acid sequences of all reference KS and C domains were aligned using either MUSCLE (Edgar 2004) or ClustalX (version 1.83) (Thompson et al. 1997) with the BLOSUM 62 protein weight matrix. The alignments were manually adjusted using Mesquite (Maddison and Maddison 2010). Maximum likelihood, parsimony, and neighbor-joining phylogenetic trees were constructed using the "a la carte" mode at the Phylogeny.fr website (http://www.phylogeny.fr/) (Dereeper et al. 2008). Final maximum likelihood trees were constructed from the reference data set with the program PHYML (Guindon and Gascuel 2003). Final domain classifications were made based on the phylogenetic relationships observed in these trees.

NaPDoS and Webportal. The NaPDoS web portal identifies candidate KS and C domains through a combination of hidden markov model (HMM) searches and the basic local alignment search tool (BLAST) algorithm (Altschul et al. 1990) optimized for query input type as shown in figure B.1. PCR products or coding sequences (CDS) in nucleotide or amino acid format are analyzed directly by local BLASTX or BLASTP searches against the manually curated reference database of experimentally verified KS and C domains described above. This BLAST-based approach proved more effective than HMM models in detecting the target domains from short query sequences. Genomic sequences (including contigs, incomplete drafts, or complete genomes) and metagenomic nucleotide data sets are first pre-screened to obtain rough coordinates for KS and C domains using the KS domain HMM developed by Yadav and co-workers (Yadav et al. 2009) and the PFAM C domain model PF00668 (Finn et al. 2008). The resulting candidate domains are then subjected to BLAST analyses using the same manually curated reference database as described above.

BLAST results are linked to a back-end MySQL relational database via CGI-scripting to retrieve and report domain classification and related pathway information. Query sequences are trimmed according to their BLAST match coordinates by a custom Perl script then aligned to each other and their database matches using MUSCLE (Edgar 2004). Trimmed sequences can be downloaded along with best BLAST matches in FASTA or MSF aligned format. Finally, trimmed and aligned candidate KS and C sequences plus BLAST matches can be inserted into a phylogenetic tree generated from the reference database using FastTree to estimate maximum likelihood (Guindon and Gascuel 2003). Newick format output from FastTree is converted to SVG format graphic images using the Newick-Utilities program (Junier and Zdobnov 2010). NaPDoS does not employ any stand-alone software that was created specifically for its operation but instead employs pre-existing and publically available programs as described above.

Draft genomes and metagenomes. Draft genome sequences of *S. arenicola* strain CNH-643 (accession number PRJNA84391), *S. arenicola* strain CNT-088 (accession number PRJNA84269), *"S. pacifica"* strain CNS-143 (accession number

PRJNA84389), and *"S. pacifica"* strain CNT-133 (accession number PRJNA84271) were obtained at 8X coverage at the J. Craig Venter Institute using 454 GS FLX pyrosequencing and 0.5X Sanger sequencing as previously described (Goldberg et al. 2006) based on an estimated genome size of 5.6 Mb. The sequence data were assembled using the Newbler Assembler with the mapping option (Margulies et al. 2005). *S. arenicola* strains were mapped onto the complete *S. arenicola* strain CNS-205 genome and the *S. pacifica* strains were mapped to the complete *S. tropica* CNB-440 genome (Penn et al. 2009) while any unmapped sequence data was assembled de novo. The four draft *Salinispora* genomes were mined for KS and C domains using NaPDoS with default settings. The metagenomic datasets (whale fall, AAFZ00000000, AAFY00000000, AAGA00000000 and Minnesota farm soil, AAFX00000000,(Tringe et al. 2005)) were mined using default HMM settings (e^{-5}) and the resulting sequences further subjected to a loose BLAST analysis with an e-value cut-off of 1 to obtain more precise coordinates and assign initial domain classifications.

Results

The Natural Product Domain Seeker (NaPDoS). The publically available web tool NaPDoS (http://npdomainseeker.ucsd.edu/) was created to detect and classify KS and C domains in nucleotide and amino acid sequence data. The query data can be PCR amplicons, genes, contigs, genomes, or metagenomes. The current query size limits are <30 MB and <50,000 individual sequences. The website provides a detailed tutorial on the use of this tool, which is implemented using a web interface (Table B.2) that follows the bioinformatic pipeline shown in table B.1. Query sequences are BLASTed against the reference database, which currently contains 459 KS and 190 C domains derived from 66 PKS, 20 NRPS, 8 PKS/NRPS hybrid, and 5 fatty-acid synthase (FAS) biosynthetic pathways. These sequences can be downloaded from the website and encompass all major classes of type I and II KS and C domains currently described in the literature (Nguyen et al. 2008; Ridley et al. 2008; Hertweck 2009; Jenke-Kodama and Dittmann 2009b). This manually curated database will be updated periodically as new modular architectures and biochemical features are discovered for each domain type.

The primary output for all analyses includes the query identification, best database match, percent identity, alignment length, e-value, and product and classification of the biosynthetic pathway associated with the best match. KS and C domain sequences derived from the input data can then be output in raw format or aligned with the best BLAST matches. A NaPDoS independent BLAST of the output domain sequence(s) against the NCBI nr database is also highly recommended to check for matches that do not occur in the reference database.

To generate a final classification for each domain sequence, it is highly recommended to construct a phylogenetic tree, especially in cases where the percent sequence identity to the top database match is low. If that option is chosen, a profile alignment is generated in which the query sequences are incorporated into a carefully curated reference alignment generated from the sequences in the reference database. This alignment is then used to create a phylogenetic tree, which needs to be manually interpreted to establish a final classification for each sequence. Interpreting sequences in the context of a phylogenetic tree is particularly important given that the NaPDoS pipeline is intentionally set to low stringency in an effort to detect all possible KS and C domains. Thus, homologs not involved in secondary metabolism such KSs associated with fatty acid biosynthesis are regularly detected. These sequences can readily be classified in the phylogenetic tree.

Domain classification. KS and C domain phylogenies form the basis of the NaPDoS classification system (Table B.3). KS domains clade based on biochemical function and enzyme architecture, which are described in table B.1. In some cases, e.g. enediynes, these clades are also predictive of structural motifs associated with the secondary metabolites produced. The KS phylogeny clearly delineates type I and II PKSs (Table B.3A). The shared ancestry reported between type II PKS and FAS sequences (Jenke-Kodama et al. 2005) is clearly maintained in this tree. The vast majority of the reference sequences fall into the type I PKS clade. This clade can be further resolved into seven classes, which are not always monophyletic. This polyphyly reflects the complex evolutionary histories of the different classes such as the *trans*-AT KSs, which evolved by extensive HGT and exploit considerably greater modular architectures than the *cis*-AT group, which has largely evolved by gene duplication (Piel 2010). However, all of these lineages are highly supported in the tree (likelihood values 0.7-1.0) and largely agree with previous phylogenetic studies (Jenke-Kodama and Dittmann 2005; Rausch et al. 2007).

In the case of C domains, the sequences generally clade based on substrate specificity, i.e. the stereochemistry of the amino acids incorporated, and the subsequent tailoring reactions they perform (Table B.3B). Eight clades are identified in the tree of which six are functionally characterized. The characterized clades are comprised of LCL domains, which catalyze a peptide bond between two L-amino acids, DCL domains, which link an L-amino acid to a growing peptide ending with a D-amino acid, starter C domains, which acylate the first amino acid with a β-hydroxy-carboxylic acid, cyclization domains, which catalyze both peptide bond formation and the subsequent cyclization of cysteine, serine or threonine residues, epimerization (E) domains, which switch the chirality of the last amino acid in the growing peptide, and dual E/C domains, which catalyze both epimerization and condensation reactions. These six functional classes are well supported in the tree and largely monophyletic. Two experimentally uncharacterized clades are identified in the tree, one of which has been conditionally assigned the name "modified AA" (Table B.3B). This clade contains domains from the bleomycin and microcystin pathways. Although the biochemical function of these domains has not been experimentally defined, they appear to be involved in the modification of the incorporated amino acid, for example the dehydration of serine to dehydroalanine (Du et al. 2000; Tillett et al. 2000). C domains in the second functionally uncharacterized clade have been conditionally assigned the name "hybrid C". The three sequences in this clade (micro5, ituri1, and mycos1) are each located downstream of an aminotransferase domain and appear to be involved in the condensation of an amino acid to an aminated polyketide resulting in a hybrid PKS/NRPS secondary metabolite. The phylogenetic relationships of the KS and C domains in the reference dataset form the basis of the NaPDoS classification system and provide a framework within which new clades and biochemical functions can be discovered.

Genome analyses. As a positive control, NaPDoS was used to analyze the genome sequence of *Streptomyces avermitilis* strain MA-4680. This analysis revealed 67 KS and 15 C domains (Table B.2), which encompass all of the PKS, NRPS, and hybrid PKS/NRPS gene clusters that were reported to contain these domains (Nett et al. 2009). NaPDoS also correctly identified all of the KS and C domains in the complete genome sequences of *S. tropica* (strain CNB-440) and *S. arenicola* (strain CNS-205) (Penn et al. 2009). NaPDoS was then tested on four draft *Salinispora* genome sequences. These low coverage drafts were generated using 454 technology and yielded poor assemblies and a large number of contigs (Table B.3). There was no evidence that any biosynthetic gene clusters had been completely assembled based on the analysis of flanking regions and comparisons with pathways that appeared common with the CNB-440 and CNS-205 sequences (Penn et al. 2009). None-the-less, NaPDoS successfully detected 18-36 KS domains and 5-14 C domains in the un-annotated FASTA files generated for each of the four draft genomes (Table B.3). More than half (56%) of these sequences showed no significant BLAST matches to domains associated with biochemically characterized biosynthetic genes and thus could not be linked to specific secondary metabolic products. More significantly, 8 KS and 9 C domains detected in the four draft sequences were not observed in the two closed *Salinispora* genomes (Table B.4). These sequences (KS7-14 and C5-13) cover a broad range of domain classes and indicate considerable new biosynthetic potential among a group of closely related strains. Two C domains fell into the “Modified AA” clade, which has yet to be experimentally characterized. Given that the upstream A domain specifies serine in both cases, it can be predicted that this domain results in the incorporation of dehydrated serine (ie., dehyroalanine) into the non-ribosomal peptide. This hypothesis has not yet been tested, but is supported by the reference sequences in this clade, which perform similar dehydration reactions.

Interestingly, two KS domains with close matches (89% and 94%) to those associated with the biosynthesis of salinosporamide A (Eustáquio et al. 2011) were observed in "*S. pacifica*" strain CNT-133. This was unexpected given that compounds in this series had previously been reported exclusively from *S. tropica* (Jensen et al. 2007). This observation subsequently led to the discovery of a new compound in the salinosporamide series (Eustáquio et al. 2011) and a rare window into pathway evolution in two closely related bacterial species (Freel et al. 2011). Furthermore, a KS domain that shares close sequence identity with domains involved in the biosynthesis of tylactone in *Streptomyces fradiae* (Cundliffe et al. 2001) was detected in strain CNH-643 (Table B.4). Subsequent chemical studies revealed the production of several new tylactone derivatives by this strain (unpublished data). The same four draft genome sequences were also analyzed using antiSMASH (Medema et al. 2011), a sophisticated pipeline that can make structure predictions for a diverse range of secondary metabolic pathways. While antiSMASH worked well on the two complete *Salinispora* genomes, NaPDoS consistently detected more KS domains in the draft genomes (Table B.5). While this is not surprising given that NaPDoS is specifically designed for this purpose, it nonetheless highlights the value of the sequence tag approach when working with draft genome sequences that contain many unassembled contigs.

Metagenomic analyses. NaPDoS was further tested on metagenomic data sets generated from a Minnesota farm soil and whale fall (Tringe et al. 2005). While the numbers of KS domains detected in both datasets are similar (Table B.6), removing redundant sequences reveals a higher diversity of KS domains in the soil sample. The majority of the whale fall KS domains were classified as FASs suggesting they are associated with fatty acid biosynthesis. In contrast, nearly half of the KS domains detected in the Minnesota farm soil appear to be involved in secondary metabolite biosynthesis. These results are in agreement with a previous study in which these datasets were manually screened for type I PKSs (Foerstner et al. 2008). All of the sequences shared <70% identity to the reference database or NCBI BLAST matches associated with experimentally characterized pathways and thus no predictions could be made about the structures of the potential secondary metabolic products. None-the-less, the majority of the KS domains detected could be rapidly classified by NaPDoS. The incorporation of these domains into a phylogenetic tree containing the reference sequences led to the reclassification of some and the prediction that others are functionally distinct from KS domains (Tables B.7 and B.8). These sequences were likely detected due to the low stringency at which the NaPDoS BLAST analyses were performed on the meta-data and is a positive indication that the KS analysis was comprehensive. The reclassification of some sequences emphasizes the importance of incorporating phylogeny into the analyses.

Discussion

Rapid advances in DNA sequencing technologies are providing unprecedented opportunities to incorporate DNA sequence data into the natural product discovery process. The effective use of this information requires bioinformatic tools that can rapidly analyze large datasets in the context of a wide array of complex biosynthetic paradigms. While a number of excellent bioinformatic tools targeting secondary metabolism have been developed (Starcevic et al. 2008; Bachmann et al. 2009; Medema et al. 2011), they are largely predicated on accurate gene and operon assembly, something that has proven challenging to obtain given the modular and highly repetitive nature of many genes involved in secondary metabolism (Udwary et al. 2007). This challenge can become especially problematic in the case of metagenomic analyses of complex microbial communities.

The Natural Product Domain Seeker (NaPDoS) is a web-based bioinformatic tool that was developed to detect and classify KS and C domains from a wide variety of sequence data. The use of domain sequence tags as proxies for the biosynthetic genes in which they reside is based on the well established and highly informative phylogentic relationships they maintain. These relationships form the foundation of the NaPDoS classification system and provide a rapid mechanism to delineate secondary metabolite biosynthetic gene richness and diversity within a genome or environmental sample. Sequence tags as short as 600 base pairs can be effectively analyzed using NAPDoS and thus minimum coverage, next generation sequence assemblies are well suited for this tool. The resulting estimates of biosynthetic potential can be used to guide more extensive sequencing efforts or targeted operon assembly. In cases where query sequences closely match domains derived from experimentally characterized biosynthetic pathways (eg., >90% sequence identity), it may even be possible to make accurate predictions about the structural class of the secondary metabolite(s) produced (Gontang et al. 2010; Freel et al. 2011). The low stringency of the HMM searches and the ability to adjust the internal BLAST parameters provides opportunities to detect more highly divergent KS and C domains associated with secondary metabolism as well as domains that are not associated with secondary metabolism (e.g. fatty acid biosynthesis) and thus all results should be carefully scrutinized. As the number of experimentally characterized biosynthetic pathways increases, this approach will provide an increasingly effective method to "de-replicate", i.e. to identify strains that have the greatest potential to produce known compounds.

There is ample evidence that the mechanistic diversity of polyketide and non-ribosomal peptide assembly is considerably greater than originally anticipated (Shen 2003; Wenzel and Müller 2005), and thus it can be expected that the NaPDoS classification system will need to evolve as new phylogenetic lineages are linked to specific biochemical functions and enzyme architectures. There is considerable preliminary evidence that the classes defined here will be further delineated once more experimentally characterized sequence data is obtained. For example, the current KS1 clade includes traditional starter KSs (KSQ) as well as domains from the salinosporamide (stro1024) and jamaicamide (JamE) pathways, which are involved in the incorporation of unusual extender units (Edwards et al. 2004; Udwary et al. 2007). Likewise, the Type II clade includes deeply branching KS domains derived from CurC and JamG that are predicted to be involved in decarboxylation as opposed to condensation reactions (Chang et al. 2004; Edwards et al. 2004). A third example is the Iterative (a) class, which include traditional iterative KSs as well as those involved in the biosynthesis of polycyclic tetramate macrolactams (Blodgett et al. 2010). Finally, the *trans*-AT (b) clade is comprised of KS sequences derived from what appears to be an evolutionarily independent lineage of *trans*-AT sequences as well as genes associated with *beta*-branching (Blodgett et al. 2010; Piel 2010). Despite the potential oversimplification of the current classification system, it provides a useful method to estimate the numbers and functional types of biosynthetic genes present in complex data sets.

Despite poor assembly, a large number and diversity of KS and C domains were detected among the four draft *Salinispora* genome sequences. Seventeen of these domains were not observed in either of the two complete *Salinispora* genomes providing evidence of the considerable biosynthetic variability that may occur among closely related strains. In addition, two C domains fell into the “Modified AA” clade, a lineage whose biochemical function has yet to be experimentally characterized. While the metagenomic datasets revealed similar total numbers of KS domains, the classification of those domains revealed dramatic differences in functional types. Analyses such as these provide insight into the potential significance of secondary chemistry in mediating population and community dynamics while at the same time identifying environments that can be prioritized for secondary metabolite discovery efforts.

Traditional natural product discovery paradigms have become increasingly inefficient (Li and Vederas 2009) and are rapidly moving towards approaches that capitalize on access to DNA sequence data (Davies 2010). NaPDoS is a publically available bioinformatic tool that capitalizes on the well-established phylogenetic relationships of KS and C domains. It provides a rapid method to make informed interpretations of secondary metabolism based on small sequence tags extracted from a variety of data types including poorly assembled, next generation datasets. A major application of NaPDoS is the exploration of sequence space and the identification of new domain lineages, which have a high probability of being associated with new mechanisms of secondary metabolite biosynthesis. Prioritizing these lineages for experimental characterization will facilitate the discovery of new biochemistry and represents a rationale approach to secondary metabolite discovery.

At present, NaPDoS is optimized for the identification and classification of bacterial PKS and NRPS genes. Nonetheless, it is possible for NaPDoS to identify eukaryotic KS and C domains given their shared evolutionary history with prokaryotic homologs. The results obtained for non-bacterial sequences should be interpreted with caution however, as the reference database has not been adequately populated with these sequences to provide a robust classification system. Future plans include the expansion of NaPDoS to include additional eukaryotic sequences and subgroups with the FAS and PUFA lineages, the later of which were recently shown to cluster phylogenetically based on functional type (Shulse and Allen 2011). Additional goals are to include type III PKSs, which were originally found in plants but are now known to occur in a wide range of bacteria (Moore et al., 2001). These PKSs are distantly related to types I and II and thus will require a separate alignment and analysis pipeline. The inclusion of additional secondary metabolite families, such as terpenoids, alkaloids, and ribosomal peptides, are also conceivable.

Acknowledgements

Appendix B, in part, contains methods, data and results that is submitted to PLoS ONE, 2012, Nadine Ziemert, Sheila Podell, Kevin Penn, Eric E. Allen, Paul R. Jensen. The dissertation author conceived the methods and aided in parts of the analyses, which forms the basis of this publication. This work was supported by the National Institutes of Health grant numbers RO1GM086261, UO1TW0007401, and RO1GM085770 to PRJ and a grant from the German Research Foundation (DFG 1325/1-1) to NZ.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215(3): 403-410.

Ansari MZ, Yadav G, Gokhale RS, Mohanty D (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Research 32(suppl 2): W405-W413.

Bachmann BO, Ravel J, David AH (2009) Chapter 8 Methods for In Silico Prediction of Microbial Polyketide and Nonribosomal Peptide Biosynthetic Pathways from DNA Sequence Data. Methods in Enzymology: Academic Press. pp. 181-217.

Baker DD, Chu M, Oza U, Rajgarhia V (2007) The value of natural products to future pharmaceutical discovery. Natural Product Reports 24(6).

Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang CH, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabbinowitsch E, Rajandream MA, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, Hopwood DA (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417(6885): 141-147.

Blodgett JAV, Oh D-C, Cao S, Currie CR, Kolter R, Clardy J (2010) Common biosynthetic origins for polycyclic tetramate macrolactams from phylogenetically diverse bacteria. Proceedings of the National Academy of Sciences.

Challis GL (2008) Genome Mining for Novel Natural Product Discovery. Journal of Medicinal Chemistry 51(9): 2618-2628.

Chang Z, Sitachitta N, Rossi JV, Roberts MA, Flatt PM, Jia J, Sherman DH, Gerwick WH (2004) Biosynthetic Pathway and Gene Cluster Analysis of Curacin A, an Antitubulin Natural Product from the Tropical Marine Cyanobacterium Lyngbya majuscula. Journal of Natural Products 67(8): 1356-1367.

Cundliffe E, Bate N, Butler A, Fish S, Gandecha A, Merson-Davies L (2001) The tylosin-biosynthetic genes of Streptomyces fradiae. Antonie van Leeuwenhoek 79(3): 229-234.

Davies J (2010) How to discover new antibiotics: harvesting the parvome. Current Opinion in Chemical Biology 15(1): 5-10.

Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Research 36(suppl 2): W465-W469.

Du L, Sánchez C, Chen M, Edwards DJ, Shen B (2000) The biosynthetic gene cluster for the antitumor drug bleomycin from Streptomyces verticillus ATCC15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase. Chemistry &amp; Biology 7(8): 623-642.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32(5): 1792-1797.

Edwards DJ, Marquez BL, Nogle LM, McPhail K, Goeger DE, Roberts MA, Gerwick WH (2004) Structure and Biosynthesis of the Jamaicamides, New Mixed Polyketide-Peptide Neurotoxins from the Marine Cyanobacterium Lyngbya majuscula. Chemistry &amp; Biology 11(6): 817-833.

Eisen JA (1998) A phylogenomic study of the MutS family of proteins. Nucleic Acids Research 26(18): 4291-4300.

Eisen JA, Fraser CM (2003) Phylogenomics: Intersection of Evolution and Genomics. Science 300(5626): 1706-1707.

Eustáquio AS, Nam S-J, Penn K, Lechner A, Wilson MC, Fenical W, Jensen PR, Moore BS (2011) The Discovery of Salinosporamide K from the Marine Bacterium "Salinispora pacifica" by Genome Mining Gives Insight into Pathway Evolution. ChemBioChem 12(1): 61-64.

Finking R, Marahiel MA (2004) Biosynthesis of nonribosomal peptides. Annual Review of Microbiology 58: 453 - 488.

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz H-R, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, Bateman A (2008) The Pfam protein families database. Nucleic Acids Research 36(suppl 1): D281-D288.

Fischbach MA, Walsh CT (2006) Assembly-Line Enzymology for Polyketide and Nonribosomal Peptide Antibiotics: Logic, Machinery, and Mechanisms. Chemical Reviews 106(8): 3468-3496.

Foerstner KU, Doerks T, Creevey CJ, Doerks A, Bork P (2008) A Computational Screen for Type I Polyketide Synthases in Metagenomics Shotgun Data. PLoS ONE 3(10): e3515.

Freel KC, Nam S-J, Fenical W, Jensen PR (2011) Evolution of Secondary Metabolite Genes in Three Closely Related Marine Actinomycete Species. Applied and Environmental Microbiology 77(20): 7261-7270.

Ginolhac Al, Jarrin C, Robe P, Perri√®re G, Vogel TM, Simonet P, Nalin R (2005) Type I Polyketide Synthases May Have Evolved Through Horizontal Gene Transfer. Journal of Molecular Evolution 60(6): 716-725.

Goldberg SMD, Johnson J, Busam D, Feldblyum T, Ferriera S, Friedman R, Halpern A, Khouri H, Kravitz SA, Lauro FM, Li K, Rogers Y-H, Strausberg R, Sutton G, Tallon L, Thomas T, Venter E, Frazier M, Venter JC (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proceedings of the National Academy of Sciences 103(30): 11240-11245.

Gontang EA, Fenical W, Jensen PR (2007) Phylogenetic Diversity of Gram-Positive Bacteria Cultured from Marine Sediments. Applied and Environmental Microbiology 73(10): 3272-3282.

Gontang EA, Gaudencio SP, Fenical W, Jensen PR (2010) Sequence-Based Analysis of Secondary-Metabolite Biosynthesis in Marine Actinobacteria. Applied and Environmental Microbiology 76(8): 2487-2499.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52(5): 696 - 704.

Hertweck C (2009) The Biosynthetic Logic of Polyketide Diversity. Angewandte Chemie International Edition 48(26): 4688-4716.

Hornung A, Bertazzo M, Dziarnowski A, Schneider K, Welzel K, Wohlert S-E, Holzenkämpfer M, Nicholson GJ, Bechthold A, Süssmuth RD, Vente A, Pelzer S (2007) A Genomic Screening Approach to the Structure-Guided Identification of Drug Candidates from Natural Sources. ChemBioChem 8(7): 757-766.

Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, Omura S (2003) Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. Nat Biotech 21(5): 526-531.

Jenke-Kodama H, Dittmann E (2005) Combinatorial polyketide biosynthesis at higher stage. Mol Syst Biol 1.

Jenke-Kodama H, Dittmann E (2009a) Bioinformatic perspectives on NRPS/PKS megasynthases: Advances and challenges. Natural Product Reports 26(7).

Jenke-Kodama H, Dittmann E (2009b) Evolution of metabolic diversity: Insights from microbial polyketide synthases. Phytochemistry 70(15,Äì16): 1858-1866.

Jenke-Kodama H, Sandmann A, Müller R, Dittmann E (2005) Evolutionary Implications of Bacterial Polyketide Synthases. Molecular Biology and Evolution 22(10): 2027-2039.

Jensen PR, Williams PG, Oh D-C, Zeigler L, Fenical W (2007) Species-Specific Secondary Metabolite Production in Marine Actinomycetes of the Genus *Salinispora*. Applied and Environmental Microbiology 73(4): 1146-1152.

Junier T, Zdobnov EM (2010) The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. Bioinformatics 26(13): 1669-1670.

Lautru S, Challis GL (2004) Substrate recognition by nonribosomal peptide synthetase multi-enzymes. Microbiology 150(Pt 6): 1629 - 1636.

Lautru S, Deeth RJ, Bailey LM, Challis GL (2005) Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. Nat Chem Biol 1(5): 265-269.

Li JWH, Vederas JC (2009) Drug Discovery and Natural Products: End of an Era or an Endless Frontier? Science 325(5937): 161-165.

Maddison WP, Maddison D (2010) Mesquite: a modular system for evolutionary analysis. Version 2.73.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057): 376-380.

McAlpine JB, Bachmann BO, Piraee M, Tremblay S, Alarco A-M, Zazopoulos E, Farnet CM (2005) Microbial Genomics as a Guide to Drug Discovery and Structural Elucidation: ECO-02301, a Novel Antifungal Agent, as an Example. Journal of Natural Products 68(4): 493-496.

McPherson JD (2009) Next-generation gap. Nat Meth 6(11s): S2-S5.

Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Research 39(suppl 2): W339-W346.

Metsä-Ketelä M, Salo V, Halo L, Hautala A, Hakala J, Mäntsälä P, Ylihonko K (1999) An efficient approach for screening minimal PKS genes from Streptomyces. FEMS Microbiology Letters 180(1): 1-6.

Moffitt MC, Neilan BA (2003) Evolutionary Affiliations Within the Superfamily of Ketosynthases Reflect Complex Pathway Associations. Journal of Molecular Evolution 56(4): 446-457.

Moss SJ, Martin CJ, Wilkinson B (2004) Loss of co-linearity by modular polyketide synthases: a mechanism for the evolution of chemical diversity. Natural Product Reports 21(5).

Nett M, Gulder TAM, Kale AJ, Hughes CC, Moore BS (2009) Function-Oriented Biosynthesis of β-Lactone Proteasome Inhibitors in Salinispora tropica. Journal of Medicinal Chemistry 52(19): 6163-6167.

Newman DJ, Cragg GM (2007) Natural Products as Sources of New Drugs over the Last 25 Years. Journal of Natural Products 70(3): 461-477.

Nguyen T, Ishida K, Jenke-Kodama H, Dittmann E, Gurgui C, Hochmuth T, Taudien S, Platzer M, Hertweck C, Piel J (2008) Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. Nat Biotech 26(2): 225-233.

Penn K, Jenkins C, Nett M, Udwary DW, Gontang EA, McGlinchey RP, Foster B, Lapidus A, Podell S, Allen EE, Moore BS, Jensen PR (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. ISME J 3(10): 1193-1203.

Piel J (2010) Biosynthesis of polyketides by trans-AT polyketide synthases. Natural Product Reports 27(7).

Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). Nucleic Acids Research 33(18): 5799 - 5808.

Rausch C, Hoof I, Weber T, Wohlleben W, Huson D (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. BMC Evolutionary Biology 7(1): 78.

Ridley CP, Lee HY, Khosla C (2008) Evolution of polyketide synthases in bacteria. Proceedings of the National Academy of Sciences 105(12): 4595-4600.

Roongsawang N, Lim SP, Washio K, Takano K, Kanaya S, Morikawa M (2005) Phylogenetic analysis of condensation domains in the nonribosomal peptide synthetases. FEMS Microbiology Letters 252(1): 143-151.

Shen B (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. Current Opinion in Chemical Biology 7(2): 285-295.

Shulse CN, Allen EE (2011) Widespread Occurrence of Secondary Lipid Biosynthesis Potential in Microbial Lineages. PLoS ONE 6(5): e20146.

Sieber SA, Marahiel MA (2005) Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics. Chem Rev 105(2): 715 - 738.

Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Research 36(21): 6882-6892.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X Windows Interface: Flexible Strategies for Multiple Sequence Alignment Aided by Quality Analysis Tools. Nucleic Acids Research 25(24): 4876-4882.

Tillett D, Dittmann E, Erhard M, von Dohren H, Borner T, Neilan BA (2000) Structural organization of microcystin biosynthesis in Microcystis aeruginosa PCC7806: an integrated peptide-polyketide synthetase system. Chem Biol 7(10): 753 - 764.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM (2005) Comparative Metagenomics of Microbial Communities. Science 308(5721): 554-557.

Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W, Jensen PR, Moore BS (2007) Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. Proceedings of the National Academy of Sciences 104(25): 10376-10381.

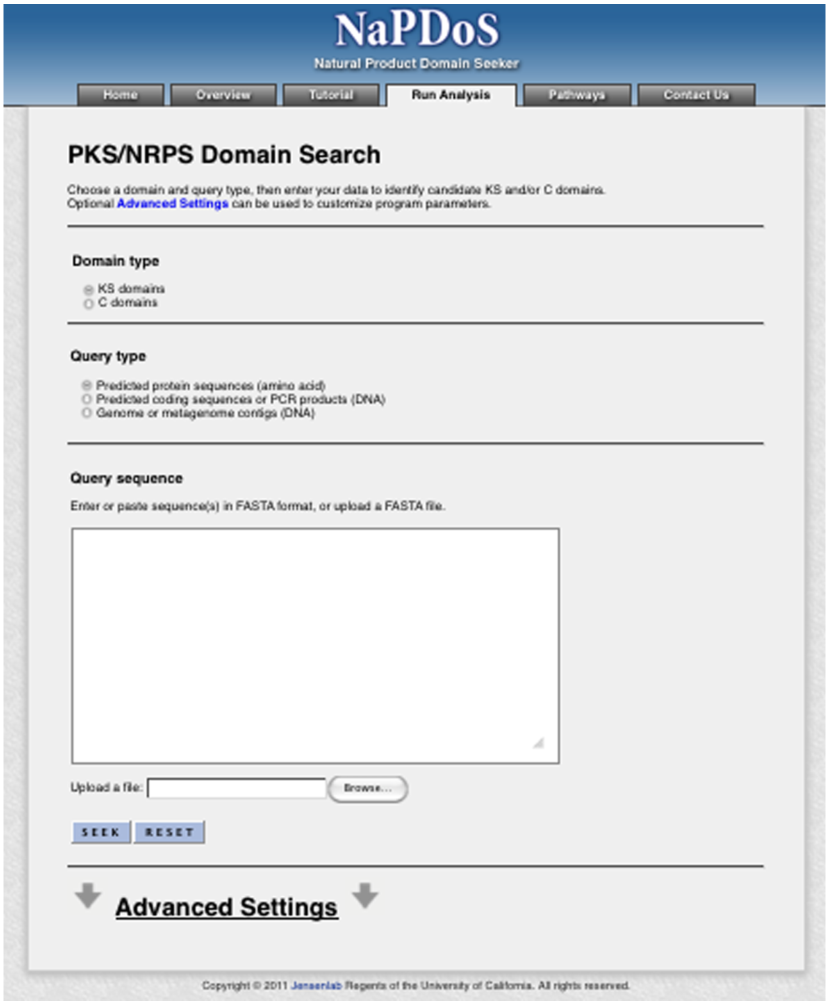
Weissman KJ (2004) Polyketide biosynthesis: understanding and exploiting modularity. Philosophical Transactions of the Royal Society of London Series A: Mathematical, Physical and Engineering Sciences 362(1825): 2671-2690.

Wenzel SC, Müller R (2005) Formation of novel secondary metabolites by bacterial multimodular assembly lines: deviations from textbook biosynthetic logic. Current Opinion in Chemical Biology 9(5): 447-458.

Winter JM, Behnken S, Hertweck C (2010) Genomics-inspired discovery of natural products. Current Opinion in Chemical Biology 15(1): 22-31.

Yadav G, Gokhale RS, Mohanty D (2009) Towards Prediction of Metabolic Products of Polyketide Synthases: An *In Silico* Analysis. PLoS Comput Biol 5(4): e1000351.

Figures



PKS/NRPS Domain Search

Choose a domain and query type, then enter your data to identify candidate KS and/or C domains.
Optional Advanced Settings can be used to customize program parameters.

Domain type

● KS domains
○ C domains

Query type

● Predicted protein sequences (amino acid)
○ Predicted coding sequences or PCR products (DNA)
○ Genome or metagenome contigs (DNA)

Query sequence

Enter or paste sequence(s) in FASTA format, or upload a FASTA file.

Upload a file:

Copyright © 2011 Jensenlab Regents of the University of California. All rights reserved.

Figure B.1: NaPDoS bioinformatic pipeline. The web interface to this pipeline is divided 3 consecutive steps. Nucleic acid sequences are translated into predicted amino acids and genomic sequences are screened using Hidden Markov Models (HMM). For protein and small nucleic acid sequences a BLAST search is performed against curated reference database examples to identify matches to known PKS/NRPS pathways. Selected candidate sequences plus the BLAST results are trimmed and inserted into a manually curated reference alignment, keeping the original reference alignment intact. This alignment is used to build a tree.

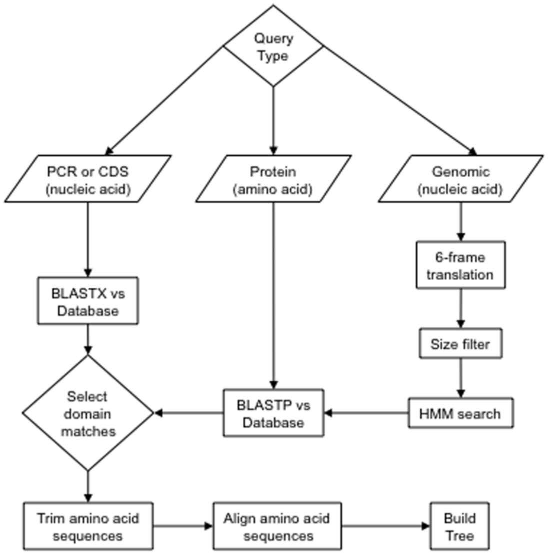


Figure B.2: Screen shot of the NaPDoS webpage.

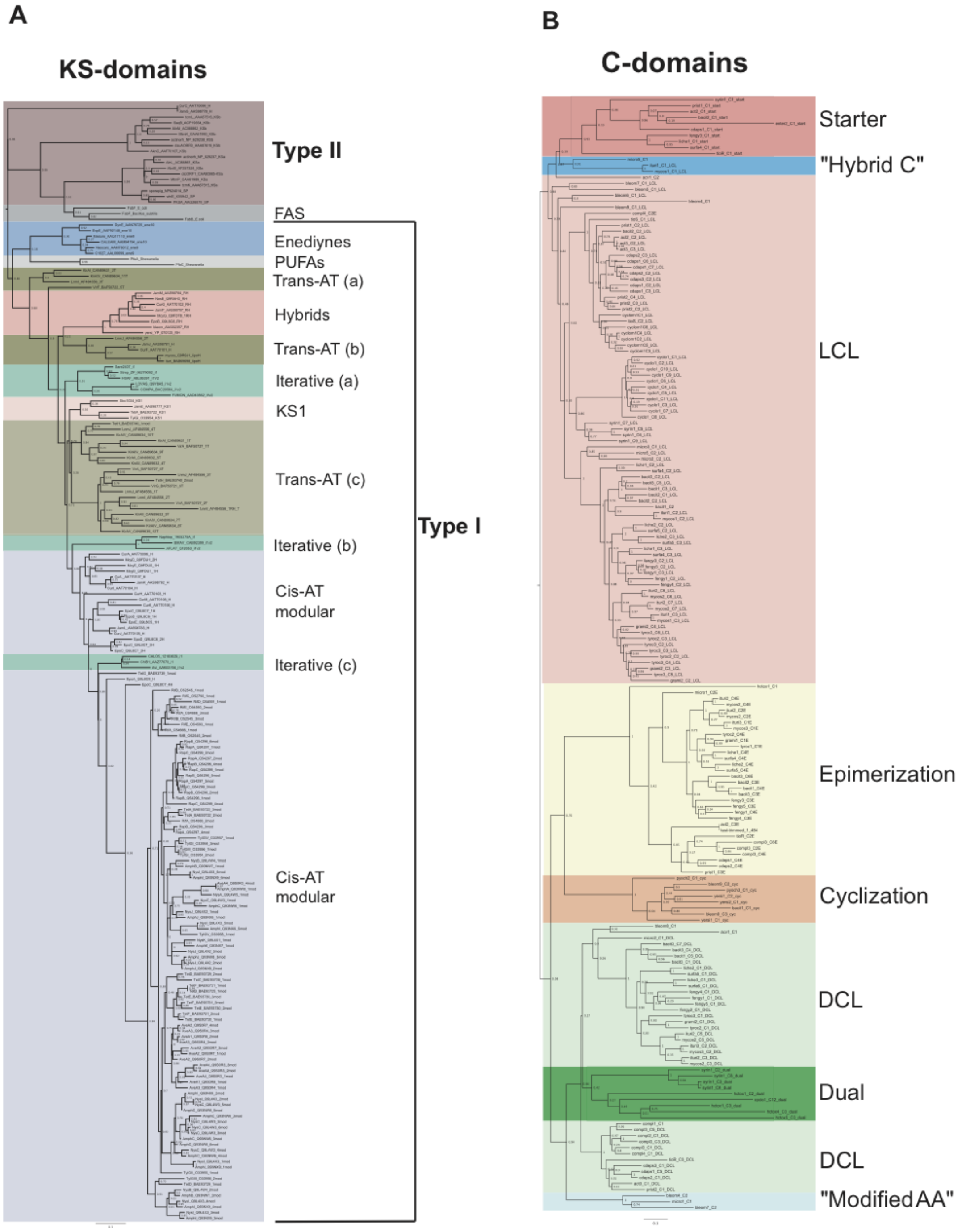


Figure B.3: Phylogeny based domain classification. A) KS domain phylogeny. Polyphyletic groups are distinguished by letters. B) C domain phylogeny.

Tables

Table B.1: KS domain classification.

Type	Class	Description	Product (example)
I	Enediyne	Iteratively acting, builds typical 9 or 10 membered enediynes.	Enediyne (calicheamicin)
	Trans-AT	Module lacks cognate AT domain; this activity is provided instead by a discrete protein encoded in *trans*.	Polyketide/macrolide (leinamycin)
	Cis-AT	Multi-domain module that includes AT domain.	Polyketide/macrolide (erythromycin)
	Hybrid	Catalyzes a condensation reaction between an amino acid and an acyl extender unit in a NRPS/PKS pathway.	Peptide-polyketide (microcystin)
	Iterative	Domain is used multiple times in a cyclic fashion.	Polycyclic polyketide (aflatoxin)
	PUFA	Produces long chain fatty acids containing more than one double bond.	Polyunsaturated fatty acid (omega-3-fatty-acid)
	KS1	Occurs in the first module of multimodular genes, includes typical starter KSs (KSQ) as well as KS domains that incorporate unusual precursors.	Polyketide, peptide-polyketide (salinosporamide)
II	Type II	Each domain occurs on a discrete protein.	Aromatic polyketide (actinorhodin)
	FAS	Involved in fatty acid biosynthesis (eg., FabB and FabF from bacteria).	Fatty acid (palmitic acid)

Table B.2: NaPDoS derived KS and C domains from the *S. avermitilis* MA-4680 genome.

Domain #	NaPDoS database match	% ID	Length	e-value	NaPDoS pathway	Domain classification	Locus tag[a]	Predicted compound
KS1	AlnL_ACI88861_KSa	43	326	2.00E-51	alnumycin	type II	SAV_2292	fatty acid
KS2	AlnM_ACI88862_KSb	47	407	8.00E-81	alnumycin	type II	SAV__2839	spore pigment
KS3	CALO5_12183629_i1	48	429	8.00E-102	calicheamicin	Iterative	SAV_2893	oligomycin
KS4	AlnL_ACI88861_KSa	37	372	3.00E-43	alnumycin	type II	SAV_2944	fatty acid
KS5	AveA1_Q9S0R8_1mod	100	226	2.00E-127	avermectin	Modular	SAV_943	avermectin
KS6	AveA1_Q9S0R8_2mod	100	222	1.00E-126	avermectin	Modular	SAV_943	avermectin
KS7	AveA2_Q9S0R7_1mod	100	223	6.00E-122	avermectin	Modular	SAV_943	avermectin
KS8	AveA2_Q9S0R7_2mod	100	222	5.00E-126	avermectin	Modular	SAV_943	avermectin
KS9	AveA2_Q9S0R7_3mod	100	222	2.00E-126	avermectin	Modular	SAV_943	avermectin
KS10	AveA2_Q9S0R7_4mod	100	224	4.00E-120	avermectin	Modular	SAV_943	avermectin
KS11	sporepig_NP824014_SP	100	239	8.00E-141	spore pigment	type II	SAV_2838	spore pigment
KS12	Strep_ZP_06279092_i1	46	427	2.00E-90	unknown	Iterative	SAV_1249	PK-NRP hybrid
KS13	Avi_AAK83194_i1v2	48	436	2.00E-110	avilamycin	Iterative	SAV_2892	oligomycin
KS14	Avi_AAK83194_i1v2	51	436	4.00E-113	avilamycin	Iterative	SAV_2892	oligomycin
KS15	HSAF_ABL86391_i1V2	39	462	1.00E-80	HSAF	Iterative	SAV_100	polyketide
KS16	Stro2795_1	54	211	5.00E-62	ST Sid 3	KS	SAV_3665	polyketide
KS17	Avi_AAK83194_i1v2	48	436	7.00E-94	avilamycin	Iterative	SAV_2898	oligomycin
KS18	Avi_AAK83194_i1v2	51	436	2.00E-104	avilamycin	Iterative	SAV_2898	oligomycin
KS19	Avi_AAK83194_i1v2	47	441	8.00E-102	avilamycin	Iterative	SAV_2898	oligomycin
KS20	Avi_AAK83194_i1v2	52	436	7.00E-109	avilamycin	Iterative	SAV_2864	oligomycin
KS21	Avi_AAK83194_i1v2	50	436	2.00E-97	avilamycin	Iterative	SAV_2864	oligomycin
KS22	Avi_AAK83194_i1v2	48	436	2.00E-102	avilamycin	Iterative	SAV_2864	oligomycin
KS23	AlnL_ACI88861_KSa	66	366	2.00E-137	alnumycin	type II	SAV_2376	polyketide
KS24	bleom_AAG02357_RH	51	428	1.00E-104	bleomycin	Hybrid	SAV_845	NRP
KS25	Avi_AAK83194_i1v2	47	436	1.00E-105	avilamycin	Iterative	SAV_416	filipin
KS26	Avi_AAK83194_i1v2	50	436	4.00E-110	avilamycin	Iterative	SAV_416	filipin
KS27	Avi_AAK83194_i1v2	49	436	3.00E-113	avilamycin	Iterative	SAV_416	filipin
KS28	Avi_AAK83194_i1v2	50	436	4.00E-112	avilamycin	Iterative	SAV_416	filipin
KS29	Avi_AAK83194_i1v2	48	436	3.00E-106	avilamycin	Iterative	SAV_416	filipin
KS30	Avi_AAK83194_i1v2	49	436	3.00E-116	avilamycin	Iterative	SAV_416	filipin
KS31	Stro3381_1	63	237	5.00E-74	unknown	FAS	SAV_5785	fatty acid
KS32	KirAIV_CAN89634_11T	44	438	8.00E-84	kirromycin	trans-AT	SAV_7362	polyketide
KS33	KirAIV_CAN89634_11T	38	472	9.00E-66	kirromycin	trans-AT	SAV_7361	polyketide
KS34	VirF_BAF50722_5T	38	208	3.00E-26	virginiamycin	trans-AT	SAV_3667	polyketide
KS35	AlnL_ACI88861_KSa	36	276	2.00E-22	alnumycin	type II	SAV_3660	polyketide
KS36	Strep_ZP_06279092_i1	46	430	1.00E-102	unknown	iterative	SAV_7184	polyketide
KS37	HSAF_ABL86391_i1V2	48	427	1.00E-95	HSAF	iterative	SAV_2899	oligomycin
KS38	Avi_AAK83194_i1v2	51	436	2.00E-114	avilamycin	iterative	SAV_2899	oligomycin
KS39	CALO5_12183629_i1	51	427	6.00E-109	calicheamicin	Iterative	SAV_2899	oligomycin
KS40	Avi_AAK83194_i1v2	51	435	3.00E-113	avilamycin	Iterative	SAV_2899	oligomycin
KS41	Avi_AAK83194_i1v2	51	436	4.00E-111	avilamycin	Iterative	SAV_2899	oligomycin
KS42	Avi_AAK83194_i1v2	51	435	3.00E-112	avilamycin	Iterative	SAV_2899	oligomycin
KS43	Avi_AAK83194_i1v2	50	436	1.00E-96	avilamycin	Iterative	SAV_1551	polyketide
KS44	CALO5_12183629_i1	50	428	7.00E-113	calicheamicin	Iterative	SAV_1551	polyketide
KS45	Avi_AAK83194_i1v2	50	438	5.00E-114	avilamycin	Iterative	SAV_410	filipin
KS46	AlnL_ACI88861_KSa	35	142	4.00E-08	alnumycin	type II	SAV_3663	aromatic polyketide
KS47	Avi_AAK83194_i1v2	48	436	2.00E-102	avilamycin	Iterative	SAV_2895	oligomycin
KS48	Avi_AAK83194_i1v2	47	447	3.00E-103	avilamycin	Iterative	SAV_2895	oligomycin
KS50	AlnM_ACI88862_KSb	53	405	3.00E-104	alnumycin	type II	SAV_2375	polyketide
KS51	KirAI_CAN89631_1T	46	425	6.00E-84	kirromycin	trans-AT	SAV_2368	polyketide
KS52	Avi_AAK83194_i1v2	52	438	2.00E-112	avilamycin	iterative	SAV_2368	polyketide
KS53	Avi_AAK83194_i1v2	50	436	5.00E-108	avilamycin	iterative	SAV_2368	polyketide
KS54	CALO5_12183629_i1	39	426	3.00E-52	calicheamicin	Iterative	SAV_2281	polyketide
KS55	AveA4_Q9S0R3_1mod	100	222	8.00E-118	avermectin	Modular	SAV_943	avermectin
KS56	AveA4_Q9S0R3_2mod	100	222	6.00E-125	avermectin	Modular	SAV_943	avermectin
KS57	AveA4_Q9S0R3_3mod	100	222	1.00E-125	avermectin	Modular	SAV_943	avermectin
KS58	AveA3_Q9S0R4_1mod	100	222	7.00E-104	avermectin	modular	SAV_943	avermectin
KS59	AveA3_Q9S0R4_2mod	100	223	1.00E-126	avermectin	modular	SAV_943	avermectin
KS60	AveA3_Q9S0R4_3mod	100	222	3.00E-126	avermectin	modular	SAV_943	avermectin
KS61	HSAF_ABL86391_i1V2	49	424	2.00E-111	HSAF	iterative	SAV_419	filipin
KS62	Avi_AAK83194_i1v2	49	435	2.00E-106	avilamycin	iterative	SAV_419	filipin
KS63	Avi_AAK83194_i1v2	48	436	2.00E-106	avilamycin	iterative	SAV_419	filipin
KS64	Avi_AAK83194_i1v2	49	435	3.00E-116	avilamycin	iterative	SAV_419	filipin
KS65	Avi_AAK83194_i1v2	50	437	3.00E-112	avilamycin	iterative	SAV_419	filipin
KS66	Avi_AAK83194_i1v2	48	436	5.00E-113	avilamycin	iterative	SAV_415	filipin
KS67	Avi_AAK83194_i1v2	48	437	2.00E-111	avilamycin	iterative	SAV_415	filipin
C1	cyclo1_C7_LCL	27	295	5.00E-17	cyclosporin	LCL	SAV_859	NRP
C2	act3_C3_LCL	39	192	1.00E-22	actinomycin	LCL	SAV_869	NRP
C3	syrin1_C6_LCL	32	300	5.00E-29	syringomycin	LCL	SAV_857	NRP
C4	ituri1_C3_LCL	27	245	2.00E-15	iturin	LCL	SAV_1551	polyketide
C5	bacil2_C1_start	47	293	5.00E-77	bacillibactin	starter	SAV_603	NRP
C6	syrin1_C6_LCL	44	298	4.00E-60	syringomycin	LCL	SAV_3643	NRP
C7	micro1_C1	36	302	2.00E-51	microcystin	Mod.AA	SAV_3197	NRP
C8	syrin1_C6_LCL	40	298	7.00E-56	syringomycin	LCL	SAV_3159	NRP
C9	act3_C3_LCL	49	295	1.00E-63	actinomycin	LCL	SAV_865	NRP
C10	syrin1_C9_LCL	38	303	1.00E-48	syringomycin	LCL	SAV_852	NRP
C11	micro3_C1_LCL	28	220	3.00E-17	microcystin	LCL	SAV_847	NRP
C12	Sare2407_1	33	295	2.00E-31	pksnrps2	LCL	SAV_3647	NRP
C13	cdaps2_C2_LCL	47	306	5.00E-60	Ca-dependent antibiotic	LCL	SAV_3642	NRP
C14	micro1_C1	35	293	1.00E-34	microcystin	Mod.AA	SAV_3642	NRP
C15	micro1_C1	34	293	2.00E-36	microcystin	Mod.AA	SAV_3642	NRP

a) as defined in the *S. avermitilis* MA-4680 genome sequence.

Table B.3: NaPDoS results for six *Salinispora* genomes.

Species	Strain	Size (Mb)	Contigs	KS	KS class[a]							C	C class[b]				
				Total	Ene	II	Cis	Iter	Hyb	KS1	FA	Total	LCL	Cyc	Starter	DCL	Mod
S. arenicola	CNS-205	5.1	1	33	2	4	20	3	1	1	3	24	20	3	0	0	1
S. tropica	CNB-440	5.7	1	28	2	8	12	0	2	1	3	16	8	7	1	0	0
S. arenicola	CNT-088	5.4	2304	32	2	1	21	4	2	1	1	16	13	2	0	0	1
S. arenicola	CNH-643	4.8	3823	29	1	1	21	1	2	1	2	9	6	1	0	1	1
"S. pacifica"	CNT-133	4.5	5214	32	1	4	19	1	1	2	4	6	6	0	0	0	0
"S. pacifica"	CNS-143	4.1	5260	25	1	1	18	0	3	2	0	7	3	2	1	1	0

a) Ene = enediyne, II = type II, cis = *cis*-AT modular, Iter = iterative, Hyb = hybrid, FA = fatty acid.

b) Cyc = cyclization, Mod = "modified amino acid".

Table B.4: KS and C domains detected in four draft *Salinispora* genomes.

Pathway name[a]	Domain classification	Predicted compound[b]	*S. arenicola* CNS-205	*S. tropica* CNB-440	*S. arenicola* CNT-088	*S. arenicola* CNH-643	*S. pacifica* CNT-133	*S. pacifica* CNS-143
PKS1A	enediyne	9 membered enediyne	X	-	X	X	-	-
PKS2	type II	polyketide	X	-	-	-	-	-
Rif	modular	rifamycin and saliniketals	X	-	X	X	-	-
PKS3A	iterative	calicheamicin-related fragment A	X	-	X	X	-	-
Sid1	hybrid	yersiniabactin related siderophore	X	X	X	X	-	-
PKS3B	enediyne	calicheamicin-related fragment B	X	-	X	-	-	-
PKS4	type II	aromatic polyketide	X	X	X	X	X	-
PKSNRPS2	modular	ND SApksnrps2	X	-	X	-	-	-
PKS5	modular	macrolide	X	-	X	X	-	-
lym	modular	**lymphostin**	X	X	X	X	-	X
pks1C	iterative	kedarcin related fragment C	X	-	X	-	-	-
STpks1	enediyne	10 membered enediyne STpks1	-	X	-	-	-	-
sal	KS1, hybrid	salinisporamide	-	X	-	-	X	-
STPKS2	type II	glycosylated decaketide	-	X	-	-	X	1
spo	enediyne	**sporolide**	-	X	-	-	X	-
slm	modular	**salinilactam**	-	X	-	-	X	X
cya	enediyne	**cyanosporaside**	-	-	-	-	-	X
STSid3	type II	dihydroaeruginoic acid related siderophore	-	X	-	-	-	-
tyl	modular	**tylactone**	-	-	-	-	X	-
fa	fatty acid	fatty acid	X	X	X	X	X	-
PKS7	modular	polyketide	-	-	-	X	X	X
PKS8	hybrid	NRP/PK hybrid	-	-	-	X	-	-
PKS9	modular	polyketide	-	-	-	-	-	X
PKS10	fatty acid	fatty acid	-	-	-	-	X	-
PKS11	iterative	polyketide	-	-	-	-	X	-
PKS12	modular	polyketide	-	-	-	-	X	-
PKS13	KS1	polyketide	-	-	-	-	-	-
PKS14	KS1	polyketide	-	-	-	-	-	X
PKS15	hybrid	NRP/PK hybrid	-	-	-	-	-	X
PKS16	modular	polyketide	-	-	-	-	-	X
PKS17	fatty acid	fatty acid	-	-	-	X	X	-
PKS18	hybrid	NRP/PK hybrid	-	-	-	-	-	X
PKS19	modular	FD-891-like	-	-	-	-	-	X
PKS20	modular	polyketide	-	-	-	-	-	X
PKS21	hybrid	NRP/PK hybrid	-	-	-	-	-	X
NRPS 1	LCL, modified AA	pentapeptide	X	-	X	X	-	-
Sid1	cyclization	yersiniabactin-related	X	X	X	X	-	-
PKS1B	LCL	kedarcidin-related	X	-	X	X	-	-
PKSNRPS2	LCL	polyketide/non-ribosomal peptide	X	-	X	-	-	-
NRPS2	LCL	tetrapeptide	X	-	X	-	-	-
NRPS3	LCL	dipeptide	-	X	-	-	-	-
Cym	LCL	**cyclomarin**	X	-	-	-	-	-
NRPS4	LCL	tetrapeptide	X	X	X	-	X	X
Sal	LCL	**salinosporamide**	-	X	-	-	X	-
Sid3	LCL	dihydroaeruginoic-acid related	-	X	-	-	-	-
Sid4	cyclization, LCL	coelibactin-related siderophore	-	X	-	-	-	-
Spo	LCL	**sporolide**	-	X	-	-	-	-
NRPS5	LCL	NRP	-	-	X	-	-	-
NRPS6	LCL	NRP	-	-	X	-	-	-
NRPS7	LCL	NRP	-	-	X	-	-	-
NRPS8	DCL	NRP	-	-	-	-	-	X
NRPS9	LCL	NRP	-	-	-	-	X	-
NRPS10	LCL	NRP	-	-	-	-	X	-
NRPS11	LCL	NRP	-	-	-	-	X	-
NRPS12	cyclization	NRP	-	-	-	-	-	X
NRPS13	cyclization	NRP	-	-	-	-	-	X
NRPS14	LCL	NRP	-	-	-	X	-	-
NRPS15	starter	NRP	-	-	-	-	-	X
NRPS16	DCL	NRP	-	-	-	X	-	-

a) Pathway names and associated compounds are as previously reported (Penn et al., 2009). In cases of <90% sequence identity to an experimentally characterized pathway, domains were given PKS and NRPS numbers.

b) Compounds in bold have been isolated from at least one of the strains.

Table B.5: NaPDoS and antiSMASH-derived KS and C domains.

		KS domains		C domains	
Species	Strain	antiSMASH	NaPDoS[a]	antiSMASH	NaPDoS[b]
S. arenicola	CNH-643	27	34	16	15
S. arenicola	CNT-088	25	30	13	14
S. pacifica	CNS-143	10	16	10	9
S. pacifica	CNT-133	7	17	7	8

[a] KS domains associated with fatty acid biosynthesis were manually removed from the NaPDoS totals as antiSMASH does this automatically.

[b] The NaPDoS C domain cut-off was set to 100 amino acids to be more comparable with antiSMASH.

Table B.6: NaPDoS KS results for metagenomic data sets.

			Class								
Dataset	Total KS domains	Distinct KS domains	Fatty acid	Type II	Hybrid	Modular	*Trans*-AT	KS1	Iterative	PUFA	Non-KS
Whale fall	129	42	27	1	1	0	1	0	0	1	11
Farm soil	128	127	43	15	11	20	8	4	4	0	22

Table B.7: KS domains detected in the whale fall metagenomic data set.

Query KS		NaPDoS database match					
KS	Domain class	Database name	Percent identity	Align length	e-value	Pathway product	Domain class
1	non KS	PfaA_Shewanella_PUFA	44	203	2.00E-52	polyunsaturated fatty acid	PUFA
2	FAS	FabF_Bacillus_FAS	47	309	6.00E-69	fatty acid synthesis	FAS
3	FAS	FabB_Ecoli_FAS	72	240	9.00E-98	fatty acid synthesis	FAS
4	FAS	LnmJ_AF484556_4T	41	94	2.00E-14	leinamycin	trans
5	FAS	FabB_Ecoli_FAS	42	280	2.00E-53	fatty acid synthesis	FAS
6	non KS	PfaA_Shewanella_PUFA	88	255	8.00E-139	polyunsaturated fatty acid	PUFA
7	PUFA	PfaC_Shewanella_PUFA	36	140	9.00E-24	polyunsaturated fatty acid	PUFA
8	FAS	FabF_Ecoli_FAS	38	267	3.00E-22	fatty acid synthesis	FAS
9	type II	VicB_BAD08358_1KSB	36	108	5.00E-08	vicenistatin	modular
10	non KS	FabF_Bacillus_FAS	33	320	1.00E-35	fatty acid synthesis	FAS
11	FAS	FabF_Bacillus_FAS	53	254	5.00E-71	fatty acid synthesis	FAS
12	non KS	FabF_Bacillus_FAS	37	131	9.00E-16	fatty acid synthesis	FAS
13	non KS	FabF_Bacillus_FAS	50	236	9.00E-60	fatty acid synthesis	FAS
14	FAS	FabF_Bacillus_FAS	58	210	1.00E-54	fatty acid synthesis	FAS
15	non KS	FabF_Bacillus_FAS	51	230	4.00E-56	fatty acid synthesis	FAS
16	hybrid	bleom_AAG02357_H	54	197	2.00E-57	bleomycin	hybrid
17	FAS	FabF_Bacillus_FAS	28	104	2.00E-07	fatty acid synthesis	FAS
18	FAS	FabB_Ecoli_FAS	77	193	4.00E-85	fatty acid synthesis	FAS
19	non KS	Nostoc_glycolipid_PUFA	39	136	6.00E-21	heterocyst glycolipid	PUFA
20	FAS	FabF_Bacillus_FAS	55	134	1.00E-39	fatty acid synthesis	FAS
21	FAS	FabB_Ecoli_FAS	64	214	1.00E-71	fatty acid synthesis	FAS
22	FAS	FabB_Ecoli_FAS	36	143	1.00E-14	fatty acid synthesis	FAS
23	non KS	bleom_AAG02357_H	65	104	3.00E-39	bleomycin	hybrid
24	FAS	FabF_Bacillus_FAS	41	248	2.00E-44	fatty acid synthesis	FAS
25	FAS	FabF_Bacillus_FAS	37	155	9.00E-30	fatty acid synthesis	FAS
26	FAS	FabF_Bacillus_FAS	34	190	3.00E-23	fatty acid synthesis	FAS
27	FAS	FabF_Bacillus_FAS	56	214	1.00E-52	fatty acid synthesis	FAS
28	FAS	KirAII_CAN89632_4T	36	121	9.00E-10	kirromycin	trans
29	FAS	FabB_Ecoli_FAS	72	193	1.00E-75	fatty acid synthesis	FAS
30	non KS	FabF_Bacillus_FAS	43	154	1.00E-25	fatty acid synthesis	FAS
31	FAS	FabF_Bacillus_FAS	54	255	3.00E-63	fatty acid synthesis	FAS
32	FAS	FabF_Bacillus_FAS	49	185	2.00E-43	fatty acid synthesis	FAS
33	FAS	FabF_Bacillus_FAS	57	244	4.00E-79	fatty acid synthesis	FAS
34	FAS	FabF_Bacillus_FAS	50	125	2.00E-31	fatty acid synthesis	FAS
35	FAS	FabB_Ecoli_FAS	29	210	4.00E-16	fatty acid synthesis	FAS
36	non KS	FabF_Bacillus_FAS	44	186	1.00E-37	fatty acid synthesis	FAS
37	non KS	FabF_Ecoli_FAS	46	212	2.00E-44	fatty acid synthesis	FAS
38	FAS	FabF_Bacillus_FAS	33	244	3.00E-20	fatty acid synthesis	FAS
39	FAS	FabF_Bacillus_FAS	52	157	3.00E-46	fatty acid synthesis	FAS
40	FAS	FabF_Bacillus_FAS	51	262	1.00E-65	fatty acid synthesis	FAS
41	FAS	FabF_Bacillus_FAS	57	210	3.00E-65	fatty acid synthesis	FAS
42	trans	mycos_Q9R9J1_T	43	287	1.00E-59	mycosubtilin	trans

Table B.8: KS domains detected in the Minnesota farm soil data set.

Query KS		NaPDoS database match					
KS	Domain class	Database name	Percent identity	Align length	e-value	Pathway product	Domain class
1	FAS	Strep_ZP_06279092_i	54	87	6.00E-22	unknown	iterative
2	FAS	FabF_Bacillus_FAS	55	288	8.00E-95	fatty acid synthesis	FAS
3	modular	Avi_AAK83194_i	55	254	3.00E-73	avilamycin	iterative
4	FAS	FabF_Bacillus_FAS	58	280	1.00E-97	fatty acid synthesis	FAS
5	non KS	MxaC_Q93TW9_3KSB	37	63	2.00E-06	myxalamid	modular
6	typeII	FabF_Bacillus_FAS	35	364	8.00E-42	fatty acid synthesis	FAS
7	iterative	CALO5_12183629_i	55	209	1.00E-45	calicheamicin	iterative
8	hybrid	bleom_AAG02357_H	55	221	5.00E-67	bleomycin	hybrid
9	non KS	FabF_Bacillus_FAS	49	179	5.00E-40	fatty acid synthesis	FAS
10	modular	Avi_AAK83194_i	52	206	6.00E-58	avilamycin	iterative
11	typeII	FabF_Bacillus_FAS	33	304	6.00E-37	fatty acid synthesis	FAS
12	non KS	yersi_YP_070123_H	36	100	2.00E-15	yersiniabactin	hybrid
13	KS1	HSAF_ABL86391_i	53	258	7.00E-73	HSAF	iterative
14	typeII	FabF_Bacillus_FAS	43	210	9.00E-44	fatty acid synthesis	FAS
15	FAS	FabF_Bacillus_FAS	54	257	2.00E-66	fatty acid synthesis	FAS
16	trans	LnmJ_AF484556_2T	52	296	1.00E-85	leinamycin	trans
17	typeII	FabF_Bacillus_FAS	54	234	9.00E-74	fatty acid synthesis	FAS
18	FAS	FabF_Bacillus_FAS	54	240	8.00E-76	fatty acid synthesis	FAS
19	trans	LnmJ_AF484556_2T	50	346	3.00E-94	leinamycin	trans
20	FAS	FabF_Streptomyces_FAS	34	273	2.00E-31	fatty acid synthesis	FAS
21	trans	LnmJ_AF484556_1T	61	283	5.00E-88	leinamycin	trans
22	typeII	FabF_Bacillus_FAS	38	276	7.00E-48	fatty acid synthesis	FAS
23	typeII	FabF_Bacillus_FAS	27	327	3.00E-22	fatty acid synthesis	FAS
24	FAS	FabF_Ecoli_FAS	55	362	4.00E-101	fatty acid synthesis	FAS
25	non KS	LnmI_AF484556_2T	45	110	1.00E-26	leinamycin	trans
26	FAS	FabF_Bacillus_FAS	43	243	2.00E-41	fatty acid synthesis	FAS
27	modular	CALO5_12183629_i	58	144	2.00E-45	calicheamicin	iterative
28	non KS	KirAI_CAN89631_2T	47	73	2.00E-13	kirromycin	trans
29	FAS	AknB_AF257324_KSa	40	330	1.00E-50	aclacinomycin	typeII
30	non KS	Avi_AAK83194_i	48	153	2.00E-34	avilamycin	iterative
31	non KS	FabF_Ecoli_FAS	53	110	4.00E-27	fatty acid synthesis	FAS
32	FAS	FabF_Bacillus_FAS	36	303	9.00E-52	fatty acid synthesis	FAS
33	modular	HSAF_ABL86391_i	50	321	1.00E-87	HSAF	iterative
34	FAS	FabF_Bacillus_FAS	59	137	6.00E-48	fatty acid synthesis	FAS
35	modular	HSAF_ABL86391_i	51	245	4.00E-60	HSAF	iterative
36	FAS	FabF_Ecoli_FAS	55	268	1.00E-87	fatty acid synthesis	FAS
37	KS1	HSAF_ABL86391_i	50	306	2.00E-84	HSAF	iterative
38	iterative	Strep_ZP_06279092_i	55	219	2.00E-64	unknown	iterative
39	typeII	FabF_Bacillus_FAS	44	323	9.00E-64	fatty acid synthesis	FAS
40	FAS	FabF_Bacillus_FAS	60	201	2.00E-72	fatty acid synthesis	FAS
41	typeII	FabF_Streptomyces_FAS	39	143	4.00E-20	fatty acid synthesis	FAS
42	modular	JamK_AAS98782_mod	62	227	8.00E-83	jamaicamide	modular
43	non KS	LnmJ_AF484556_3T	48	86	9.00E-13	leinamycin	trans
44	trans	VirA_BAF50727_4T	45	317	5.00E-60	virginiamycin	trans
45	FAS	FabF_Bacillus_FAS	36	242	8.00E-35	fatty acid synthesis	FAS
46	typeII	FabF_Bacillus_FAS	36	221	3.00E-28	fatty acid synthesis	FAS
47	typeII	FabF_Bacillus_FAS	47	183	4.00E-41	fatty acid synthesis	FAS
48	non KS	FabF_Ecoli_FAS	48	173	3.00E-41	fatty acid synthesis	FAS
49	modular	HSAF_ABL86391_i	49	205	1.00E-50	HSAF	iterative
50	modular	StiH_Q8RJX9_1KSB	56	151	9.00E-37	stigmatellin	modular
51	non KS	bleom_AAG02357_H	33	166	2.00E-18	bleomycin	hybrid
52	FAS	FabB_Ecoli_FAS	64	284	5.00E-105	fatty acid synthesis	FAS
53	modular	CALO5_12183629_i	49	336	3.00E-82	calicheamicin	iterative
54	KS1	KirAII_CAN89632_5T	52	190	1.00E-52	kirromycin	trans
55	non KS	bleom_AAG02357_H	40	126	2.00E-27	bleomycin	hybrid
56	FAS	FabF_Bacillus_FAS	37	260	4.00E-43	fatty acid synthesis	FAS
57	modular	HSAF_ABL86391_i	51	290	2.00E-83	HSAF	iterative
58	FAS	FabF_Bacillus_FAS	46	181	5.00E-43	fatty acid synthesis	FAS
59	hybrid	bleom_AAG02357_H	63	283	2.00E-92	bleomycin	hybrid
60	FAS	FabF_Bacillus_FAS	57	144	1.00E-45	fatty acid synthesis	FAS
61	hybrid	bleom_AAG02357_H	54	293	9.00E-76	bleomycin	hybrid
62	modular	Strep_ZP_06279092_i	48	244	5.00E-64	unknown	iterative
63	FAS	FabF_Bacillus_FAS	67	243	3.00E-89	fatty acid synthesis	FAS
64	iterative	CALO5_12183629_i	53	276	3.00E-57	calicheamicin	iterative
65	modular	KirAIV_CAN89634_10T	49	134	8.00E-28	kirromycin	trans

Table B.8 (continued)

Query KS		NaPDoS database match					
KS	Domain class	Database name	Percent identity	Align length	e-value	Pathway product	Domain class
66	KS1	HSAF_ABL86391_i	55	202	4.00E-58	HSAF	iterative
67	FAS	FabF_Bacillus_FAS	54	255	4.00E-73	fatty acid synthesis	FAS
68	hybrid	bleom_AAG02357_H	55	356	7.00E-105	bleomycin	hybrid
69	FAS	FabF_Ecoli_FAS	44	162	1.00E-33	fatty acid synthesis	FAS
70	FAS	FabF_Bacillus_FAS	59	228	2.00E-61	fatty acid synthesis	FAS
71	trans	VirA_BAF50727_4T	50	230	7.00E-59	virginiamycin	trans
72	modular	CALO5_12183629_i	55	287	5.00E-77	calicheamicin	iterative
73	trans	KirAIV_CAN89634_7T	49	259	8.00E-66	kirromycin	trans
74	modular	COMPA_BAC20564_i	41	252	1.00E-59	compactin	iterative
75	hybrid	yersi_YP_070123_H	57	92	1.00E-25	yersiniabactin	hybrid
76	trans	LnmJ_AF484556_4T	58	259	7.00E-88	leinamycin	trans
77	FAS	FabF_Ecoli_FAS	54	200	5.00E-50	fatty acid synthesis	FAS
78	FAS	KirAIV_CAN89634_11T	41	150	6.00E-29	kirromycin	trans
79	non KS	Nostoc_glycolipid_PUFA	50	105	2.00E-24	heterocyst glycolipid	PUFA
80	modular	CALO5_12183629_i	50	309	2.00E-80	calicheamicin	iterative
81	FAS	FabF_Ecoli_FAS	76	248	3.00E-100	fatty acid synthesis	FAS
82	typeII	FabF_Bacillus_FAS	38	299	3.00E-34	fatty acid synthesis	FAS
83	typeII	FabF_Bacillus_FAS	35	252	1.00E-33	fatty acid synthesis	FAS
84	hybrid	yersi_YP_070123_H	51	348	4.00E-103	yersiniabactin	hybrid
85	modular	Strep_ZP_06279092_i	50	218	1.00E-57	unknown	iterative
86	typeII	FabF_Bacillus_FAS	38	211	6.00E-35	fatty acid synthesis	FAS
87	FAS	FabF_Bacillus_FAS	54	245	3.00E-72	fatty acid synthesis	FAS
88	non KS	FabF_Bacillus_FAS	48	169	3.00E-36	fatty acid synthesis	FAS
89	FAS	FabF_Bacillus_FAS	50	240	4.00E-62	fatty acid synthesis	FAS
90	hybrid	bleom_AAG02357_H	61	266	2.00E-83	bleomycin	hybrid
91	non KS	HSAF_ABL86391_i	49	80	1.00E-20	HSAF	iterative
92	hybrid	bleom_AAG02357_H	61	287	2.00E-90	bleomycin	hybrid
93	FAS	FabF_Bacillus_FAS	33	152	1.00E-19	fatty acid synthesis	FAS
94	FAS	FabB_Ecoli_FAS	63	259	5.00E-88	fatty acid synthesis	FAS
95	hybrid	bleom_AAG02357_H	60	287	2.00E-89	bleomycin	hybrid
96	iterative	CALO5_12183629_i	60	213	4.00E-68	calicheamicin	iterative
97	typeII	FabF_Bacillus_FAS	39	270	1.00E-44	fatty acid synthesis	FAS
98	non KS	pfaA_omega3_PUFA	65	134	5.00E-44	omega3_FA	PUFA
99	FAS	FabF_Bacillus_FAS	68	130	1.00E-52	fatty acid synthesis	FAS
100	modular	bleom_AAG02357_H	59	59	1.00E-19	bleomycin	hybrid
101	non KS	FabF_Bacillus_FAS	36	170	3.00E-25	fatty acid synthesis	FAS
102	FAS	FabF_Bacillus_FAS	56	178	2.00E-53	fatty acid synthesis	FAS
103	hybrid	bleom_AAG02357_H	54	299	6.00E-90	bleomycin	hybrid
104	modular	CALO5_12183629_i	49	265	2.00E-65	calicheamicin	iterative
105	non KS	MerB_ABJ97438_2KSB	42	76	3.00E-07	meridamycin	modular
106	FAS	bleom_AAG02357_H	66	79	2.00E-25	bleomycin	hybrid
107	FAS	FabF_Bacillus_FAS	54	267	7.00E-75	fatty acid synthesis	FAS
108	non KS	FabF_Bacillus_FAS	48	125	5.00E-31	fatty acid synthesis	FAS
109	FAS	FabF_Bacillus_FAS	52	231	8.00E-66	fatty acid synthesis	FAS
110	typeII	FabF_Bacillus_FAS	46	195	3.00E-38	fatty acid synthesis	FAS
111	FAS	FabF_Bacillus_FAS	47	220	1.00E-48	fatty acid synthesis	FAS
112	FAS	FabF_Bacillus_FAS	59	257	2.00E-88	fatty acid synthesis	FAS
113	FAS	FabF_Bacillus_FAS	50	296	1.00E-76	fatty acid synthesis	FAS
114	FAS	FabF_Ecoli_FAS	45	252	3.00E-48	fatty acid synthesis	FAS
115	FAS	FabF_Ecoli_FAS	60	270	7.00E-92	fatty acid synthesis	FAS
116	FAS	FabF_Bacillus_FAS	62	191	4.00E-44	fatty acid synthesis	FAS
117	non KS	HSAF_ABL86391_i	43	81	8.00E-16	HSAF	iterative
118	trans	KirAII_CAN89632_5T	34	157	5.00E-14	kirromycin	trans
119	FAS	FabF_Bacillus_FAS	40	265	6.00E-49	fatty acid synthesis	FAS
120	non KS	FabF_Bacillus_FAS	38	210	1.00E-36	fatty acid synthesis	FAS
121	FAS	FabF_Bacillus_FAS	46	162	6.00E-41	fatty acid synthesis	FAS
122	FAS	FabF_Bacillus_FAS	51	250	5.00E-65	fatty acid synthesis	FAS
123	hybrid	bleom_AAG02357_H	58	262	3.00E-81	bleomycin	hybrid
124	modular	KirAII_CAN89632_5T	51	223	1.00E-61	kirromycin	trans
125	modular	HSAF_ABL86391_i	51	264	2.00E-75	HSAF	iterative
126	non KS	FabF_Bacillus_FAS	53	211	7.00E-50	fatty acid synthesis	FAS
127	non KS	VirA_BAF50727_4T	46	81	3.00E-17	virginiamycin	trans