

UNIVERSITY OF CALIFORNIA SAN DIEGO

**Population genomics and the basis of species delineations in the marine
actinomycete *Salinispora***

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Marine Biology

by

Krystle L. Chavarria

Committee in charge:

Professor Paul R. Jensen, Chair
Professor Eric E. Allen
Professor Douglas Bartlett
Professor Ronald S. Burton
Professor Susan S. Golden

2018

Copyright
Krystle L. Chavarria, 2018
All rights reserved.

The dissertation of Krystle L. Chavarria is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2018

DEDICATION

To my parents

who instilled in me a drive to continue learning and never to fear failure
as long as I tried my best.

EPIGRAPH

There are opportunities
even in the most difficult moments.

— *Wangari Maathai*

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Epigraph	v
Table of Contents	vi
List of Figures	ix
List of Tables	xi
Acknowledgements	xii
Vita	xiv
Abstract of the Dissertation	xvi
Chapter 1	
Introduction	1
1.1 Bacterial Species Concepts	2
1.2 Actinobacteria	4
1.3 <i>Salinispora</i>	5
1.4 Comparative Genomics	6
1.5 Differential Expression	7
1.6 Lateral Gene Transfer	8
1.7 Molecular Clock	9
Chapter 2	
Population Genomics of <i>Salinispora</i>	10
2.1 Abstract	10
2.2 Introduction	11
2.2.1 <i>Salinispora</i>	13
2.3 Materials and Methods	15
2.3.1 Strain Selection	15
2.3.2 DNA Extraction Protocol	15
2.3.3 Genome Sequencing	16
2.3.4 Project 1 (CNB-440 and CNS-205)	17
2.3.5 Project 2 (Six Fiji Strains)	17
2.3.6 Project 3 (111 Genome Project)	18
2.3.7 Computational Analysis	19
2.4 Results	20
2.4.1 Sequencing and Genome Statistics	20
2.4.2 Ortholog Analysis and <i>Salinispora</i> Pangenome	21

	2.4.3 Genus and Species COGs	26
	2.5 Discussion	29
Chapter 3	Species-specific functional traits between two species in the genus <i>Salinispora</i>	32
	3.1 Abstract	32
	3.2 Introduction	33
	3.3 Materials and Methods	37
	3.3.1 Chitinase Gene Identification	37
	3.3.2 Colloidal Chitin Preparation	37
	3.3.3 Strain Cultivation for Transcriptome Study	38
	3.3.4 Transcriptome Analyses	39
	3.3.5 Differential Gene Expression Analysis	39
	3.4 Results	41
	3.4.1 Chitinase Genes Identified	41
	3.4.2 Chitinase expression	41
	3.4.3 Global Differential Expression	44
	3.5 Discussion	61
	3.5.1 Chitinase	61
	3.5.2 Transcriptomics	64
	3.5.3 KEGG Pathways	65
Chapter 4	Lateral Gene Transfer Dynamics and the <i>Salinispora</i> Molecular Clock	70
	4.1 Abstract	70
	4.2 Introduction	71
	4.3 Methods	74
	4.3.1 Genome Annotation	74
	4.3.2 Molecular Clock Analyses	74
	4.3.3 Biosynthetic Gene Cluster Likelihood Analysis	76
	4.3.4 Lateral Gene Transfer Analysis	76
	4.4 Results	77
	4.4.1 Molecular Clock Results	77
	4.4.2 DarkHorse Results	84
	4.5 Discussion	89
Chapter 5	Final Remarks	95
References		99
Appendix A	<i>Salinispora</i> strain metadata	117
Appendix B	Species-specific Core Genes for <i>S. arenicola</i> and <i>S. tropica</i>	121

Appendix C	ANOVA and Tukey's Post Hoc Test for Chitinase Gene Expression	128
Appendix D	Differential Expression of <i>Salinispora</i> : Exponential Phase	131
Appendix E	Differential Expression of <i>Salinispora</i> : Stationary Phase	144
Appendix F	KEGG Maps	150
Appendix G	GenProp0799	185
Appendix H	Biosynthetic Gene Cluster Presence/Absence Matrix	189
Appendix I	Contributed Publications	194

LIST OF FIGURES

Figure 2.1:	Distribution and number of strains sequenced for each species. Pie chart indicates total number of strains sequenced for all species.	21
Figure 2.2:	Number of ortholog groups based on FastOrtho analysis.	22
Figure 2.3:	Rarefaction curves for the number of ortholog groups as sample size of genomes sequenced increases. Curves account for ortholog accumulation and decumulation for the entire genus as well as by species. Blue shading indicates standard error.	24
Figure 2.4:	Histogram of the number of ortholog groups found across all genomes. There are 2603 ortholog groups in the 119 strains sequenced (core genome). The number of ortholog groups found in only 2 genomes is comparable at 2108.	25
Figure 2.5:	Clusters of Orthologous Groups (COGs) as percentage of genome averaged across all 119 <i>Salinispora</i> genomes.	27
Figure 2.6:	Average distribution of major COG categories for <i>S. arenicola</i> and <i>S. tropica</i> genomes.	28
Figure 2.7:	Average distribution of COG categories for the <i>S. arenicola</i> and <i>S. tropica</i> core genomes.	29
Figure 3.1:	Chitinase ortholog groups identified in the genus <i>Salinispora</i>	42
Figure 3.2:	Demonstration of the ability of <i>Salinispora</i> to utilize colloidal chitin as a nutrient source on agar plates containing 0.4% chitin and agarose. Upper two plates are <i>S. tropica</i> , lower two are <i>S. arenicola</i>	43
Figure 3.3:	Average zone of clearing in mm of media with colloidal chitin as the sole carbon and nitrogen source for <i>S. tropica</i> and <i>S. arenicola</i> after 16 days of growth.	44
Figure 3.4:	Expression levels of 3 <i>Salinispora</i> strains for 4 different chitinase ortholog groups sampled at exponential and stationary phase.	45
Figure 3.5:	RNAseq library size for each experimental replicate.	46
Figure 3.6:	Squared Pearson's correlation matrix of exponential and stationary phase for each strain.	47
Figure 3.7:	Heatmap of Top Hits	48
Figure 3.8:	MAplot of log fold change during exponential growth for pairwise comparison of strains	50
Figure 3.9:	MAplot of log fold change during stationary growth for pairwise comparison of strains	51
Figure 3.10:	Differential expression of pathways in exponential phase growth	53
Figure 3.11:	Differential expression of pathways in stationary phase growth	54
Figure 3.12:	ABC Transporters - Exponential Phase	56
Figure 3.13:	Homologous Recombination - Exponential Phase	57
Figure 3.14:	Oxidative Phosphorylation - Exponential Phase	58
Figure 3.15:	Oxidative Phosphorylation - Stationary Phase	59

Figure 3.16: Glycine Serine Threonine Metabolism - Stationary Phase	60
Figure 4.1: Phylogenomic <i>Salinispora</i> species tree	78
Figure 4.2: Bacterial tree of life phylogeny	79
Figure 4.3: The <i>Salinispora</i> molecular clock	83
Figure 4.4: 16S rRNA phylogeny by genus for Actinobacterial DarkHorse hits .	85
Figure 4.5: Pie chart representing the 225 genes likely laterally transferred from taxa outside the Actinobacteria clade	87
Figure F.1: KEGG Map key of genes found in biosynthetic pathways of a genome	151
Figure F.2: 2-Oxocarboxylic Acid Metabolism - Exponential Phase	152
Figure F.3: Bacterial Secretion System - Exponential Phase	153
Figure F.4: Biosynthesis of Amino Acids - Exponential Phase	154
Figure F.5: Biotin Metabolism - Exponential Phase	155
Figure F.6: Butanoate Metabolism - Exponential Phase	156
Figure F.7: Carbon Fixation Pathways Prokaryotes - Exponential Phase	157
Figure F.8: Carbon Metabolism - Exponential Phase	158
Figure F.9: Cysteine Methionine Metabolism - Exponential Phase	159
Figure F.10: Galactose Metabolism - Exponential Phase	160
Figure F.11: Methane Metabolism - Exponential Phase	161
Figure F.12: Monobactam Biosynthesis - Exponential Phase	162
Figure F.13: Novobiocin Biosynthesis - Exponential Phase	163
Figure F.14: Pantothenate CoA Biosynthesis - Exponential Phase	164
Figure F.15: Phenylalanine Metabolism - Exponential Phase	165
Figure F.16: Phenylalanine Tyrosine Tryptophan Biosynthesis - Exponential Phase	166
Figure F.17: Porphyrin Chlorophyll Metabolism - Exponential Phase	167
Figure F.18: Propanoate Metabolism - Exponential Phase	168
Figure F.19: Protein Export - Exponential Phase	169
Figure F.20: Purine Metabolism - Exponential Phase	170
Figure F.21: Pyrimidine Metabolism - Exponential Phase	171
Figure F.22: Pyruvate Metabolism - Exponential Phase	172
Figure F.23: Sulfur Metabolism - Exponential Phase	173
Figure F.24: Taurine Hypotaurine Metabolism - Exponential Phase	174
Figure F.25: Tryptophan Metabolism - Exponential Phase	175
Figure F.26: Tyrosine Metabolism - Exponential Phase	176
Figure F.27: Alanine Aspartate Glutamate Metabolism - Stationary Phase	177
Figure F.28: Biosynthesis of Amino Acids - Stationary Phase	178
Figure F.29: Carbon Metabolism - Stationary Phase	179
Figure F.30: Glycerophospholipid Metabolism - Stationary Phase	180
Figure F.31: Glyoxylate Dicarboxylate Metabolism - Stationary Phase	181
Figure F.32: Oxidative Phosphorylation - Stationary Phase	182
Figure F.33: Phenylalanine Metabolism - Stationary Phase	183
Figure F.34: Thiamine Metabolism - Stationary Phase	184

LIST OF TABLES

Table 2.1:	Averaged genome statistics for the genus <i>Salinispora</i> and each species.	21
Table 3.1:	MA-Plot of number of genes found showing differential expression between each pair-wise comparison.	52
Table 4.1:	List of most common genera from which <i>Salinispora</i> has acquired genes based on DarkHorse predictions.	86
Table 4.2:	Average number of lateral gene transfer events from Actinobacterial and non-Actinobacterial strains into <i>Salinispora</i> species.	88
Table C.1:	One-way ANOVA of chitinase gene expression between strains for both exponential and stationary phase growth. Significance values: (***) $p < 0.001$, (**) $p < 0.01$, (*) $p < 0.05$, (.) $p < .1$	129
Table C.2:	Tukey's HSD post hoc test to determine pairwise significance of differential expression between strains.	130

ACKNOWLEDGEMENTS

Just don't get an ulcer. These were the words that brought me a surprising amount of encouragement from Paul Jensen as I pushed through the final months of my PhD. Unconventional advice, yet effective. I would like to thank Paul for supporting me throughout my time at SIO. His understanding and accommodating attitude towards the amount of time I dedicated spent away from the lab and towards my family alleviated much of the stress I would have otherwise felt.

Thank you to my committee. Eric, I still remember my first meeting with you at SIO during Open House. Any apprehension or fatigue I was feeling after a long run of back to back meetings was alleviated after half an hour talking about how kombucha was the next big marketable fermented beverage. Doug, it was a pleasure TAing your class. I appreciate your patience and approachable disposition. Ron, thank you for taking the time to make students feel like they could approach you with problems, from CUPS lunches to reaching out when you suspected there were issues with how research was going. Susan, thank you for setting a phenomenal example of how to set expectations and keep everyone on task.

Thank you to the instrumental individuals in both the SIO Grad Office and the Center for Marine Biotechnology and Biomedicine. Thanks especially to Beth Masek who, upon retiring, left us lost and appreciative of how she kept our department functioning. Thank you to Gilbert Bretado, Maureen McGreevy, Maureen McCormack, Shelly Weisel, Iliana Perez, and to Amy Butros who kept us students plied with candy.

My gratitude goes out to the rotating members of my lab with whom I've had the pleasure to work, play, and commiserate. These individuals enriched my experience at SIO and made coming to work a pleasure: Alyssa Demko, Dulce Guillén Matús, Natalie Millán-Aguíñaga, Gregory Amos, Anne-Catrin Letzel, Henrique Ramalho Machado, Leesa Klau, Julia Busch, Kaitlin Creamer, Eun Ju Choi, and Michelle Schorn.

Thank you to Juan Ugalde, Sheila Podell, and Greg Rouse for your generous contributions to my thesis. I would have been lost without your patience and guidance.

To others in the SIO community who have become like family, thank you to: PO Family/Bad Luck Bananas - Julia Fiedler, Amy Van Cise, Sean Crosby, Kate Furby. Scripps Motorcycle Club - Peter Kannberg, Dave Myers Outside-of-lab support - Mariela Brooks, Raymond Ku, Ben Reineman, Amy Waterhouse, Maggie Johnson, Rich Walsh. Babies - Ingrid, Baz, Leo, Atzín, and Amayrú.

Thank you especially to my family for all their support over the years. My parents shaped the person I am today and instilled in me a value system that is full of perspective. Thanks to Cal for keeping me company on my journey through higher education. And thanks to Josh for making our Venn diagram more 'Venn-y'.

Chapter 2 is coauthored with Millán-Aguñaga N, JA Ugalde, and PR Jensen. The dissertation author was the primary investigator and author of this chapter.

Chapter 3 is coauthored with Amos GCA, and PR Jensen. The dissertation author was the primary investigator and author of this chapter.

Chapter 4 is coauthored with Podell S, and PR Jensen. The dissertation author was the primary investigator and author of this chapter.

VITA

2003 - 2008	Bachelor of Science in Molecular Environmental Biology University of California Berkeley. Berkeley, CA
2008 - 2011	Research Associate, Lawrence Berkeley National Laboratory Department of Energy. Berkeley, CA
2011 - 2014	Master of Science in Marine Biology Scripps Institution of Oceanography University of California San Diego. La Jolla, CA
2011 - 2018	Doctor of Philosophy in Marine Biology Scripps Institution of Oceanography University of California San Diego. La Jolla, CA

PUBLICATIONS

Millán-Aguiñaga N, KL Chavarría, JA Ugalde, A-C Letzel, GW Rouse, PR Jensen. Phylogenomic Insight into *Salinispora* (Bacteria, Actinobacteria) Species Designations. *Nature Scientific Reports* 7, 3564. 2017.

Ziemert N, A Lechner, M Wietz, N Millán-Aguiñaga, KL Chavarría, PR Jensen. Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proceedings of the National Academy of Sciences* 111, E1130-E1139. 2014.

Jensen PR, KL Chavarría, W Fenical, BS Moore, N Ziemert. Challenges and triumphs to genomics-based natural product discovery. *Journal of Industrial Microbiology & Biotechnology* 41:203-209. 2014.

Hu P, S Borglin, NA Kamennaya, L Chen, H Park, L Mahoney, A Kijac, G Shan, KL Chavarría, C Zhang, NWT Quinn, D Wemmer, H-Y Holman, C Jansson. Metabolic phenotyping of the cyanobacterium *Synechocystis* 6803 engineered for production of alkanes and free fatty acids. *Applied Energy* 102:850-859. 2013.

Bell RC, JB MacKenzie, MJ Hickerson, KL Chavarría, M Cunningham, S Williams, C Moritz. Comparative multi-locus phylogeography confirms multiple vicariance events in co-distributed rainforest frogs. *Proceedings of the Royal Society B: Biological Sciences* 279:991-999. 2012.

Mackelprang, R, MP Waldrop, KM DeAngelis, MM David, KL Chavarría, SJ Blazewicz, EM Rubin, JK Jansson. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 480(368-371). 2011.

Lu, Z, Y Deng, JD Van Nostrand, Z He, J Voordeckers, A Zhou, YJ Lee, OU Mason, EA Dubinsky, KL Chavarría, LM Tom, JL Fortney, R Lamendella, JK Jansson, P Dhaeseleer, TC Hazen, J Zhou. Microbial gene functions enriched in the Deepwater Horizon deep-sea oil plume. *ISME Journal* 6 (2), 451-460. 2011.

Chourey, K, JK Jansson, N VerBerkmoes, M Shah, KL Chavarría, L Tom, EL Brodie, RL Hettich. A Direct Cellular Lysis/Protein Extraction Protocol for Soil Metaproteomics. *Journal of Proteome Research* 9(12):6615-22. 2010.

Hazen, TC, EA Dubinsky, TZ DeSantis, GL Andersen, YM Piceno, N Singh, JK Jansson, A Probst, SE Borglin, JL Fortney, WT Stringfellow, M Bill, MS Conrad, LM Tom, KL Chavarría, TR Alusi, R Lamendella, DC Joyner, C Spier, J Baelum, M Auer, ML Zemla, R Chakraborty, EL Sonnenthal, P Dhaeseleer, HN Holman, S Osman, Z Lu, JD Van Nostrand, Y Deng, J Zhou, OU Mason. Deep-sea oil plume enriches indigenous oil-degrading bacteria. *Science* 330(6001):204-8. 2010.

ABSTRACT OF THE DISSERTATION

Population genomics and the basis of species delineations in the marine actinomycete *Salinispora*

by

Krystle L. Chavarria

Doctor of Philosophy in Marine Biology

University of California San Diego, 2018

Professor Paul R. Jensen, Chair

The genus *Salinispora* is a marine actinomycete that is known for producing an assortment of secondary metabolites with anticancer and antibiotic properties. The genus is comprised of three species, *S. arenicola*, *S. pacifica*, and *S. tropica*, and are closely related based on 16S rRNA gene similarity. The goal of this dissertation is to use a population-scale comparative genomics approach to study the evolutionary diversity of *Salinispora*. Bacterial population genomics allows for the study of genome-wide patterns of sequence variation between closely related species. Aided by the declining costs of Next Generation Sequencing, whole genome analyses have become more

commonplace and accessible. Genomes from 119 *Salinispora* strains representing all 3 species and 11 different geographic locations were sequenced. Ortholog analyses of these genomes reveal the pangenome of the genus and species-specific gene pools are identified, illuminating the composition of their function. Transcriptomics analyses are incorporated into identify differential gene expression as an additional way to identify significant differences between two species. Specifically, chitinase genes as well as genes included in the species core genome are investigated. Bioinformatic predictions suggest *S. tropica* has the ability to better cope with osmotic stress and may be more affected by nutrient or oxidative stress while *S. arenicola* has more energetic needs during stationary phase growth potentially due to costly secondary metabolism. Comparative genomics allows us to identify gene content differences between related bacteria and many of these differences can be attributed to lateral gene transfer. This dissertation also examines this type of exchange of genetic information and the introduction of genes and gene clusters from neighboring bacteria which may have conferred an evolutionary advantage. For the first time a molecular clock is also presented for the genus, providing a new temporal framework with which to understand how genetic information moves across species and strains both at the gene and biosynthetic cluster level.

Chapter 1

Introduction

Research Outline

The overarching goal of this thesis is to use a population-scale comparative genomics approach to study the evolutionary diversity of the marine actinomycete genus *Salinispora*. To achieve this goal, I employed comparative genomics, bioinformatics, transcriptomic and phylogenomic techniques. This thesis is divided into three main research sections:

In chapter 2, I investigate *Salinispora* population genomics using comparative genomics techniques. A large dataset of 119 *Salinispora* genome sequences representing three named species from 11 localities across the globe are the subject of this study. The pangenome, or suite of genes found across the entire genus is identified. Genomic features of *Salinispora* based on bioinformatic functional predictions are presented. Genes specific to the co-occurring species *S. arenicola* and *S. tropica* are probed to look for traits that would differentiate them as species.

Chapter 3 seeks to further define species-specific traits by incorporating transcriptomic analyses to address how shared genes may be differentially expressed under

the same culturing conditions. Gene expression data is used to identify significant differences between two species to determine if differences in the expression of shared genes, rather than just gene content, may help explain how two species differ. Specifically, chitinase genes as well as genes included in the species core genome are investigated for differential expression.

Chapter 4 looks at the role of lateral gene transfer in *Salinispora* evolution. Horizontally transferred genes were identified bioinformatically and their likely taxonomic origins analyzed. An additional goal of this chapter was to establish a molecular clock for the genus to create a temporal framework for *Salinispora* evolution.

Chapter 5 is an overview of the significant findings brought forth by this dissertation and general conclusions from each research chapter. It also further expands into potential future directions for each project.

1.1 Bacterial Species Concepts

Defining species for any group of organisms has been a controversial undertaking for centuries. Humans have an innate desire to categorize living organisms into distinct groups in order to study them more effectively. The endeavor to formally describe species has spanned several hundred years, from Linnaeus' development of binomial Latin taxonomic nomenclature in the mid-18th century (Knapp et al., 2007), to Darwin's contribution to the theory of the 'origin of species' (Darwin, 1859). In the mid-1900s, Ernst Mayr defined species as groups of actually or potentially interbreeding natural populations which are reproductively isolated from other such groups in what is described as the Biological Species Concept (Mayr, 1942). The Morphological Species Concept determined species by the most distinguishable and distinctive means. That is, the smallest natural populations permanently separated from each other by a

distinct discontinuity (Aldhebiani, 2018). The Ecological Species Concept describes a species as a lineage which occupies an adaptive zone that is minimally different from that of any other lineage in its range and evolves separately from all lineages outside its range (Van Valen, 1976). The Evolutionary Species Concept describes species as a single lineage of ancestor-descendant populations of organisms maintaining its identity from other such lineages with its own evolutionary tendencies and historical fate (Simpson, 1951). Bacteria are particularly challenging to group into species (Cohan, 2002). In the general sense, species comprise individuals that are phenotypically and, therefore, ecologically more similar to each other than to other species (Gevers et al., 2005; Cohan and Koeppel, 2008). This concept was developed to include asexual organisms to which the biological species concept could not be applied. Sexual reproduction was thought to make determining species in eukaryotes more convoluted and that bacterial species were more straightforward due to asexual reproduction and inheritance of genes vertically. This of course was proven not to be the case as lateral gene transfer was discovered, further complicating the establishment of species concepts as they pertain to prokaryotes. Prokaryotic species concepts are now focused on the interpretation of phylogenetic diversity (Cohan and Koeppel, 2008). While the list of species concepts is long, there is a general consensus that species should be a cohesive group and their diversity should be limited by an evolutionary force (de Queiroz, 2005). Determining what these forces are, however, is highly contentious and it is suggested that microbial diversity should be ordered into ecologically and genetically cohesive units (Shapiro and Polz, 2015).

1.2 Actinobacteria

The Actinobacteria are a phylum of Gram-positive bacteria. They occupy a wide range of niches from soil to marine sediments, and occur as pathogens as well as symbionts in plant roots (Stach and Bull, 2005; Manivasagan et al., 2013; Doroghazi and Metcalf, 2013). It represents one of the largest taxonomic groups with 18 major lineages. Actinobacteria have characteristically high G+C content in their DNA, ranging from 51% in some corynebacteria to more than 70% in *Streptomyces* and *Frankia* (Ventura et al., 2007). Bacteria in the order Actinomycetales, commonly called actinomycetes, include many taxa that form filaments and produce spores. Actinomycetes are prolific producers of natural products with bioactivity that has been harnessed for pharmaceutical purposes. In fact, 65%-70% of current antibiotics originate from actinomycetes (Bérdy, 2005). Notably, the acquisition of foreign genetic material through lateral gene transfer is widespread within and between *Streptomyces* species and is a hallmark of the genus (Doroghazi and Buckley, 2010). Actinobacteria exhibit diverse physiological and metabolic properties including the production of a wide variety of secondary metabolites, small molecules that are not essential for basic cellular function and survival (Schrempf, 2001). Notably, many of these metabolites are potent antibiotics (Lechevalier and Lechevalier, 1967). This attribute has turned *Streptomyces* species into the primary antibiotic-producing organisms harnessed by the pharmaceutical industry (Bérdy, 2005). Because of their diverse antibiotic production, the first sequenced actinomycete genomes were of strains of *Streptomyces coelicolor* and *S. avermitilis* (Bentley et al., 2002; Ikeda et al., 2003). These genomes revealed a large number of genes predicted to encode the biosynthesis of secondary metabolites. This led to the field of genome mining and efforts to find the products of orphan biosynthetic gene clusters (Challis, 2008).

1.3 *Salinispora*

The level of investigation into marine Actinobacteria pales in comparison to their terrestrial counterparts (Manivasagan et al., 2013). However, indigenous marine taxa began to be described (Bull and Stach, 2007) and noted for the production of a large number of bioactive secondary metabolites (Jensen et al., 2013; Fenical and Jensen, 2014; Jensen et al., 2015a). A model genus for natural product discovery and species concepts has been found in the obligate marine actinomycete genus *Salinispora* (Mincer et al., 2002). *Salinispora* is a Gram-positive genus with high G+C content (~69%). Its cultivation was first reported in 1991 as part of a study addressing actinomycete distributions in marine sediments (Jensen et al., 1991). The distribution of *Salinispora* is quite cosmopolitan and has been reported from sediment collected from tropical and sub-tropical latitudes around the globe (Jensen et al., 2005; Jensen and Mafnas, 2006). It has been isolated from a wide range of depths, from less than a meter to depths as great as 1100m (Mincer et al., 2005). The genus is comprised of three named species: *S. arenicola*, *S. tropica*, and *S. pacifica* (Maldonado et al., 2005; Ahmed et al., 2013). Morphologically, the genus is characterized as having orange colonies with black spores. They fail to grow when seawater is replaced by deionized water in the growth medium (Penn and Jensen, 2012). The three species are closely related and their phylogeny is poorly resolved using the conventional and highly conserved 16S rRNA marker gene. A recent phylogenomic study suggests that *S. pacifica* should be divided into at least six additional species (Millán-Aguiñaga et al., 2017). *Salinispora* dedicates a surprisingly large proportion of its genome, ~10%, to secondary metabolite production (Penn et al., 2009). It produces a diverse suite of secondary metabolites that include anticancer and antibiotic compounds. One compound in phase II clinical trials for multiple myeloma is salinosporamide A (Feling et al., 2003; Williams et al., 2005; Jensen et al., 2007; Fenical

et al., 2009). While natural product discovery was the impetus for a major sequencing project across all three species from various locations, a rather large dataset of 119 whole genomes provided opportunities to address bacterial species concepts through comparative genomics.

1.4 Comparative Genomics

Comparative genomics provides a powerful opportunity to identify genetic similarities and differences among bacteria. The advent of Next Generation Sequencing has decreased sequencing costs so profoundly that it has outpaced Moores Law (Muir et al., 2016). This has made whole genome sequencing more readily accessible and provided new opportunities for those of us studying comparative genomics. It has already come to pass that the bottleneck in comparative genomics is not sequence acquisition but the ability to process the vast amounts of data now available. We have stepped into an unprecedented era in genomics, which can now be used to illuminate genomic features across closely related taxa and to study a wide variety of bacteria from pathogenic to free-living forms (Reno et al., 2009; Remenant et al., 2010). Newly sequenced bacterial genomes are usually analyzed by comparison with previously characterized genomes.

The genomic era has brought with it amazing opportunities to study bacterial evolution at the molecular level. Bacterial relatedness was once determined by denaturing the DNA of two bacteria and measuring the amount that hybridized when mixed. DNA-DNA hybridization values of $>70\%$ was arbitrarily used to assign species designations (Goris et al., 2007; Laird et al., 1969). The use of a single gene, the small subunit of the 16S rRNA was then used for phylogenetic studies due to its highly conserved nature (Woese and Fox, 1977). This, however, did not provide the necessary resolution to study closely related taxa, so instead of a single gene, many genes were

used in a method called MultiLocus Sequence Analysis (MLSA) (Maiden et al., 1998). Although this increased the representative portion of the genome that was analyzed, it was the availability of whole genome sequencing that allowed researchers to see the full complement of genes, known as the pangenome (Vernikos et al., 2015), and apply phylogenomic approaches and measures such as ANI to species designations (Goris et al., 2007).

1.5 Differential Expression

The suite of genomic tools has expanded to allow scientists to gather whole transcriptome data through RNAseq, the isolation and sequencing of the total complement of RNAs in a given sample. This provides the opportunity to assess the relationships between genetic and phenotypic similarity. Previous studies have highlighted the need to combine genetic diversity and ecology to approach the species definition problem (Fraser et al., 2009). The importance of adding ecological theory has become apparent as studies have identified a striking level of versatility in gene expression among closely related taxa under the same growth conditions. The ability for two individuals with the same genes to exhibit varying expression levels has been identified previously in the bacterial genus *Shewanella* (Fredrickson et al., 2008). Notably, more differences were found in expression levels than at the genome level, suggesting that gene regulation and expression levels should constitute another important parameter for species descriptions (Konstantinidis et al., 2009). A global gene expression analysis of *E. coli* strains highlights the ecological relevance of differential gene regulation and its role in the diversification of a model species (Vital et al., 2014). It remains unclear how differential gene expression should be incorporated into species concepts. This field is largely unexplored and understanding the link between gene regulation and phylogeny under

different culture conditions is necessary in order to understand the role gene expression plays in bacterial diversification.

1.6 Lateral Gene Transfer

In addition to leaving a faint fossil record, the evolution of microbial life has also left a tangled phylogenetic signal due to lateral gene transfer (LGT). LGT is the mechanism by which genetic material is acquired, potentially from distant relatives, and has long made reconstructing the history of life a difficult endeavor (Doolittle, 1999). LGT was first described in the 1940s in microorganisms (Lederberg and Tatum, 1946). It became widely recognized for its significance in the 1950s when multidrug resistance patterns emerged worldwide (Davies, 1996). Since then, methods to identify LGT have improved and revealed the surprising extent to which this strategy for genetic exchange has impacted the variation in viral, prokaryotic, and eukaryotic gene content (Soucy et al., 2015). Bacteria often exchange genes via LGT and distantly related groups may appear more closely related to one another than they actually are, potentially confounding the ability to discern genetic cohesion in a particular group (Doolittle and Papke, 2006). Additionally, the rate at which LGT occurs is not uniform and recombination rates are higher among closely related groups (Hanage et al., 2006). The availability of genome sequences has allowed scientists to measure and compare the total amount of laterally transferred sequences between diverse bacterial genomes. There is evidence that a sizeable fraction of some bacterial genomes have been acquired from other species, upwards of 17% of the chromosome of *Synechocystis* PCC6803 and between 10% and 16% of *E. coli* (Ochman et al., 2000).

1.7 Molecular Clock

During the last five decades, the molecular clock hypothesis has provided an indispensable tool for building evolutionary timescales. Molecular clocks have revolutionized evolutionary biology by providing a framework for estimating the times of divergence of populations and species, the diversification of gene families and the origin of sequence variations (Doolittle et al., 1996). As DNA sequencing has progressed, the use of molecular clocks has increased providing profound insight into the temporal diversification of species (Battistuzzi et al., 2004). *Salinispora* is a model genus for studying species concepts by investigating dynamics at the population-level. The availability of a large dataset of 119 genome sequences provides an exceptional opportunity to use comparative genomics to study the drivers of fine-scale phylogenetic diversity. These analyses provide opportunities to explore the *Salinispora* pangenome and identify species-specific genes. The comparison of global gene expression in two *Salinispora* species has provided evidence that the differential expression of shared genes may be as important as differential gene content when comparing closely related species. And finally, contextualizing the effect of lateral gene transfer provides insights into the evolutionary history of the genus. The following chapters provide a first glimpse into a large-scale comparative analysis of species that comprise a closely related genus in order to study species concepts, but much remains to be learned.

Chapter 2

Population Genomics of *Salinispora*

2.1 Abstract

Comparative bacterial genomics aims to study closely related taxa. This chapter aims to analyze the marine actinomycete genus, *Salinispora*, at the population level. A large dataset of 119 *Salinispora* genomes, including representatives from each of the three named species, was analyzed in this study. The pangenome, the entire set of genes that comprise the genus, was determined. Genome strains for each species were rarefied to determine if the pangenome had been captured. These sets of genes offer glimpses of *Salinispora* evolution through different lenses. The pangenome consists of the core genome, or genes found in all sequenced strains, and the flexible genome, or set of genes found in a few strains to all genomes of a species. This study identifies these gene pools and looks more closely at the flexible genome to determine whether two of the co-occurring species, *S. arenicola* and *S. tropica*, have species-defining gene pools. Based on Cluster of Orthologous Group (COG) annotations, *S. arenicola* has more genes associated with metabolism while *S. tropica* specific genes are associated with cellular processing and signaling in *S. tropica*. This appears to be congruent with observations

that *S. arenicola* devotes a larger portion of its genome to secondary metabolism whereas *S. tropica* employs a life history tradeoff that includes a relative reduction in secondary metabolism and more genomic resources dedicated towards faster growth.

2.2 Introduction

As sequencing technology becomes faster, cheaper, and more widely accessible across many platforms, the rate at which whole bacterial genome sequences have become available is unprecedented. Newly sequenced bacterial genomes are usually analyzed by comparing sequences to previously studied and characterized genomes. These comparisons can be used to determine functional differences and relationships between genes as well as identify genes that are novel, rapidly evolving, or introduced through horizontal gene transfer events (Carver et al., 2005; Elnitski et al., 2010; Hallin et al., 2008). Comparative genomics can be used to provide an overview of genomic features across closely related taxa and has been employed in studies that include pathogenic to free-living environmental bacteria (Reno et al., 2009; Remenant et al., 2010; Qin et al., 2013). One of the most surprising results of these analyses is the level of genomic diversity observed among bacteria that form closely related clades in phylogenetic trees (Haubold and Wiehe, 2004; Gan et al., 2013).

Previous comparative genomic studies on marine phytoplankton have shown that *Prochlorococcus* species are capable of occupying the entire euphotic zone due mainly to its microdiversity. Different subgroups have adapted to grow under various optimal light conditions (Moore et al., 1995; Johnson et al., 2006). This example of niche-partitioning shows that isolates from deeper in the water column can grow at substantially lower light intensities whereas those isolated from the surface have adapted to high-light conditions (Scanlan and West, 2002; Ahlgren et al., 2006). *Prochlorococcus*

has become one of the best models for the assignment of ecological function to lineages recognized in phylogenetic trees. Establishing these types of linkages between phylogeny and function remains one of the fundamental goals in microbial ecology.

It is becoming increasingly apparent that microorganisms have the potential to adapt to and utilize biotic and abiotic resources under any environmental conditions, diversifying extensively within local populations and as a function of distance between them. There are different mechanisms that can drive diversification among bacteria. One is allopatry, in which case diversification is driven by geographic isolation. One of the most notable examples of this is found within the genus *Sulfolobus*. These thermoacidophilic archaea are found in volcanic springs. In order to successfully migrate, propagules must remain viable over long distance transport and enter small islands of geothermal habitat. The species *S. islandicus* has been shown to be a model for allopatric diversification in prokaryotes with a restricted geographic distribution and a short distance dispersal capacity (Martiny et al., 2006). Sharply contrasting these results, *S. acidocaldarius*, a species within the same genus, shows evidence of rapid gene flow despite a severe discontinuity in habitat, and despite being geographically diverse, proved to have nearly identical genomes (Mao and Grogan, 2012).

A fundamental aspect of comparative genomics is the construction of accurate ortholog groups. Orthologs are genes in different species that arise from a common ancestor and contrast paralogs, another type of homolog, which are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution while paralogs are free to evolve new functions (Fang et al., 2010). Understanding this distinction allows one to more precisely describe the evolution of a genome and to understand the function of the genes they contain (Shapiro et al., 2012). Protein coding regions have a phylogenetic history that can lead to insights into diversification or conservation of function. Investigating this phylogenetic history in a model taxonomic

group of closely related bacterial strains, in this case the genus *Salinispora*, can serve to answer fundamentally important questions in the field of evolutionary microbiology.

2.2.1 *Salinispora*

The genus *Salinispora* was first described over a decade ago as the first obligate marine genus within the order Actinomycetales. It was originally designated as the MAR1 16S phylogenetic clade and belongs to the family *Micromonosporaceae* (Maldonado et al., 2005). *Salinispora* is comprised of G+C rich (~69%) Gram positive bacteria belonging to the Phylum Actinobacteria and is nested within the Order Actinomycetales, a group capable of producing a wide array of secondary metabolites (Jensen et al., 2015a). Found in marine sediment, *Salinispora* was discovered to be the first actinomycete to require a combination of salts found in seawater for growth (Jensen et al., 1991), and serves as a model organism for natural product discovery. Its global distribution is restricted to warmer tropical and subtropical latitudes (Mincer et al., 2002).

Currently, the genus *Salinispora* consists of three named species: *S. arenicola*, *S. tropica*, and *S. pacifica* (Jensen and Mafnas, 2006; Maldonado et al., 2005). Complete genome sequencing of *S. arenicola* strain CNS-205 and *S. tropica* strain CNB-440 revealed that a large percentage of the genome (~10%) is dedicated to natural product production (Penn et al., 2009). At the time of its description, *S. tropica* was determined to be the source of the sporolides, a type of halogenated macrolide (Buchanan et al., 2005). When *Salinispora pacifica* was discovered, it was shown to be the source of two additional structurally novel compounds (Oh et al., 2006). *Salinispora* has been shown to be capable of producing interesting molecules such as salinosporamide A, a highly cytotoxic proteasome inhibitor currently in phase II clinical trials for the treatment of multiple myeloma. These are just a few examples of natural products produced by this genus with potential for many more (Jensen et al., 2015b).

Using *Salinispora* as a model organism, I had the opportunity to explore how closely related organisms differ at the genome level. Early *Salinispora* studies have shown that *S. arenicola* has been cultured from all locations sampled and previous research has posited that the co-occurrence of *S. arenicola* with both *S. tropica* and *S. pacifica* suggests that ecological differentiation as opposed to geographic isolation was driving speciation within the genus (Jensen and Mafnas, 2006). I had the opportunity to look at a system wherein the same species can be found in multiple locations, globally, and different species can be found in the same location. I looked more deeply into this system through whole genome sequencing. Early genetic studies once took snapshots of small pieces of a taxons genetic information in the form of 16S rRNA, covering only $\sim 0.07\%$ of a genome. This genetic resolution has progressively expanded into larger snippets of coverage as MultiLocus Sequence Analysis (MLSA) accounting to on average $\sim 0.2\%$ of a genome. Today we now have the ability to observe and unlock an entire pangenome with realistic expectations of capturing 100% of a genome (Vernikos et al., 2015).

The goal of this chapter is to investigate the pangenome of the genus *Salinispora* by interrogating various gene pools. The distribution of a gene, whether it be present in a single strain, all strains of a species, at a single location, or across all species provides insight into the evolutionary history of the gene and how important it is to the function of the organisms that possess it. The genes shared by all *Salinispora spp.* become fixed long ago in their evolutionary history and unify *Salinispora* as a genus. Species-specific genes are likely to confer ecological adaptations and help define ecological differences between co-occurring species. An additional goal of this chapter is to determine a genetic basis for the delineation of *S. arenicola* and *S. tropica*. One of the major trends in the data is that half of the genes in a representative *Salinispora* genome belongs to the core, and are therefore essential in defining the genus. The remainder consist of

genes that vary in their distribution. These genes are species-specific and provide potential ecological advantages showing patterns of species specificity, demonstrating a genome's state of flux, collecting and purging genes as they encounter them.

2.3 Materials and Methods

2.3.1 Strain Selection

Genomic DNA was extracted from a total of 119 *Salinispora* strains covering a diverse geographic and phylogenetic distribution. Strains were isolated from marine sediments that were acquired between June 1989 and 2012. Sediment collection depths ranged from 1 meter to 700 meters. Strains originated from 11 localities: Fiji, Bahamas, Caribbean, Palau, Sea of Cortez, Guam, Hawaii, Palmyra, Red Sea, Madeira, and Puerto Vallarta and represent all three species in the genus *Salinispora* (Appendix A).

2.3.2 DNA Extraction Protocol

Bacterial isolates from glycerol stocks were grown in 100 ml A1 broth (10g starch, 4g yeast, 2g peptone, 1 liter 75% Instant Ocean (Aquarium Systems, Mentor, OH)). Liquid cultures were grown for 5-7 days at 30°C while shaking at 240 rpm. Prior to cell harvesting, 10% glycerol stocks were prepared and stored at -80°C. An agar plate of A1 media was streaked with cells to ensure purity. The remaining cells were separated into two equal-volume aliquots and centrifuged at 10,000 rpm for 5 minutes. The supernatant was discarded and the pellet was frozen at -20°C.

Prior to extraction, frozen pure cell pellets were resuspended in TE buffer to OD₆₀₀ ~ 1 (ca. 5 mL). Samples were extracted using a modified phenol-chloroform-CTAB (hexadecyltrimethylammonium bromide) protocol. To initiate cellular lysis, 150

μl of 100 mg/ml lysozyme was added. In order to remove bacterial RNA, 5 μl of 100 mg/ml RNase A was added to the mixture and incubated for 80 min in a water bath at 37°C. After incubation, 50 μl of 20 mg/ml Proteinase K and 500 μl of 10% SDS were added and tubes mixed by inversion and incubated at 55°C overnight.

The following day, 1.5 ml of 5M NaCl and 1 ml of CTAB/NaCl (10 g CTAB added to dissolved 4.1 g NaCl in 80 ml water at 65°C adjusted to 100 ml) were added, mixed and incubated for 10 min at 65°C. After incubating, tubes were placed on ice for approximately 30 min and 4 ml phenol:chloroform:isoamylalcohol (25:24:1 saturated with 10mM Tris, pH 8, 1mM EDTA) was added and tubes mixed gently by inversion. Samples were centrifuged for 10 min at 10,000 rpm and 4°C. The aqueous layer was then transferred to a fresh tube using a wide-bore pipette tip and 4 ml chloroform added and gently mixed. Samples were again centrifuged for 10 min at 10,000 rpm at 4°C. The aqueous layer was then transferred to a fresh tube using a wide-bore pipette tip and 0.6 volume isopropanol was added and mixed by inversion. Samples were centrifuged a final time for 10 min at 10,000 rpm and 4°C. The supernatant was pipetted off and the DNA pellet was rinsed in ice-cold 70% ethanol and subsequently air-dried. The DNA pellet was then dissolved in 500 μl TE buffer at 4°C overnight.

Genomic DNA was quantified using size and mass standards provided by the Joint Genome Institute. Standards and DNA were simultaneously run on 1% TAE agarose gels to ensure high molecular weight and quality before sequencing.

2.3.3 Genome Sequencing

Genome sequencing was conducted by the U.S. Department of Energy (DOE) Joint Genome Institute (JGI) as part of the Community Science Program (CSP) (<http://jgi.doe.gov/user-program-info/community-science-program/>) using Sanger and Illumina sequencing technology. *Salinispora* genomes involved in this study were se-

quenced in three distinct CSP sequencing projects. IMG genome ID and NCBI taxon numbers can be found in Appendix Table A. Genomic data is available from the Integrated Microbial Genomes (IMG) database (<http://img.jgi.doe.gov>)

2.3.4 Project 1 (CNB-440 and CNS-205)

The sequencing and annotation of *S. arenicola* CNS-205 and *S. tropica* CNB-440 were as previously reported for *S. tropica* (Udwary et al., 2007). These two strains were sequenced using Sanger technology and are closed genomes.

2.3.5 Project 2 (Six Fiji Strains)

The draft genomes of six *Salinispora* strains were generated at the JGI using Illumina technology (Bennett, 2004). These strains have been assigned DSM numbers from the German Collection of Microorganisms and Cell Cultures (DSMZ) that correspond to CN numbers taken from our lab's culture collection and are as follows: DSM45543 - CNS863 (*S. pacifica*), DSM45544 - CNS960 (*S. pacifica*), DSM45545 - CNS991 (*S. arenicola*), DSM45547 - CNT138 (*S. pacifica*), DSM45548 - CNT148 (*S. pacifica*), DSM45549 - CNT150 (*S. pacifica*). For these genomes, both an Illumina short-insert and long-insert paired-end library were constructed.

All general aspects of library construction and sequencing performed at the JGI can be found at <http://www.jgi.doe.gov/>. The initial draft data were assembled with All-paths, version r39750, and the consensus was computationally shredded into 10 Kbp overlapping fake reads (shreds). The Illumina draft data were also assembled with Velvet, version 1.1.05 (Zerbino and Birney, 2008), and the consensus sequences were computationally shredded into 1.5 Kbp overlapping fake reads (shreds). The Illumina draft data were assembled again with Velvet using the shreds from the first Velvet assembly to

guide the next assembly. The consensus from the second VELVET assembly was shredded into 1.5 Kbp overlapping fake reads. The fake reads from the Allpaths assembly and both Velvet assemblies and a subset of the Illumina CLIP paired-end reads were assembled using parallel phrap, version 4.24 (High Performance Software, LLC). Possible mis-assemblies were corrected with manual editing in Consed (Ewing and Green, 1998; Ewing et al., 1998; Gordon et al., 1998). Gap closure was accomplished using repeat resolution software (Wei Gu, unpublished), and sequencing of bridging PCR fragments with Sanger and/or PacBio technologies (unpublished, Cliff Han). For improved high-quality draft and noncontiguous finished projects, one round of manual/wet lab finishing was completed. Primer walks, shatter libraries, and/or subsequent PCR reads were also included for a finished project.

2.3.6 Project 3 (111 Genome Project)

The draft genomes of *Salinispora spp.* were generated at the DOE Joint Genome Institute (JGI) using Illumina technology (Bennett, 2004). An Illumina Std shotgun library was constructed and sequenced using the Illumina HiSeq 2000 platform. The data consisted of Illumina 8kbp long-mate pair library coverage at 100x coverage as well as Illumina Standard library coverage at 200x coverage. Library sequences were subjected to JGI's quality control standards. Contamination and artifacts were removed and identification was verified. All general aspects of library construction and sequencing performed at the JGI can be found at <http://www.jgi.doe.gov>. All raw Illumina sequence data was passed through DUK (Mingkun L, Copeland A, Han J. DUK, unpublished, 2011) a filtering program developed at JGI, which removes known Illumina sequencing and library preparation artifacts. The following steps were then performed for assembly: (1) filtered Illumina reads were assembled using Velvet (version 1.1.04) (Zerbino and Birney, 2008), (2) 1-3 Kbp simulated paired end reads were created from Velvet contigs

using wgsim (<https://github.com/lh3/wgsim>), (3) Illumina reads were assembled with simulated read pairs using Allpaths-LG (version r41043) (Gnerre et al., 2011; Butler et al., 2008). Parameters for assembly steps were: 1) Velvet (velveth: 63 -shortPaired and velvetg: -very clean yes -export- Filtered yes -min contig lgth 500 -scaffolding no -cov cutoff 10, 2) wgsim (-e 0 -1 100 -2 100 -r 0 -R 0 -X 0) 3) Allpaths-LG (PrepareAllpathsInputs: PHRED 64=1 PLOIDY=1 FRAG COVERAGE=125 JUMP COVERAGE=25 LONG JUMP COV=50, RunAllpathsLG: THREADS=8 RUN=std shredpairs TARGETS=standard VAPI WARN ONLY=True OVERWRITE=True).

Genes were identified using Prodigal (Hyatt et al., 2010), followed by a round of manual curation using GenePRIMP (Pati et al., 2010) for finished genomes and draft genomes in fewer than 10 scaffolds. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) non-redundant database, UniProt, TIGRFam, Pfam, KEGG, COG, and InterPro databases. The tRNAScanSE tool (Lowe and Eddy, 1997) was used to find tRNA genes, whereas ribosomal RNA genes were found by searches against models of the ribosomal RNA genes built from SILVA (Pruesse et al., 2007). Other noncoding RNAs such as the RNA components of the protein secretion complex and the RNase P were identified by searching the genome for the corresponding Rfam profiles using INFERNAL (<http://infernal.janelia.org>). Additional gene prediction analysis and manual functional annotation was performed within the Integrated Microbial Genomes (IMG) platform (<http://img.jgi.doe.gov>) developed by the Joint Genome Institute, Walnut Creek, CA, USA (Markowitz et al., 2009).

2.3.7 Computational Analysis

Genomes from 119 strains (12 *S. tropica*, 62 *S. arenicola* and 45 *S. pacifica*) were analyzed using the program FastOrtho (<http://enews.patricbrc.org/fastortho>), a reimple-

mentation of the program OrthoMCL (Li et al., 2003), to identify clusters of protein coding genes (Orthologous Groups (OGs)). This program performs an all-vs-all comparison of amino acids, followed by a clustering step (percent match cutoff=70, e-value cutoff=1e-05, and inflation index (I)=1.5). The results were then processed using a series of python scripts (<https://github.com/juanu/MicroCompGenomics>), developed in collaboration with Juan Ugalde, to generate a matrix of orthologous groups detected among all the genomes, and the number of gene copies in each group. This matrix was used to extract the core and flexible genomes at the genus level. OGs were associated with COG (Clusters of Orthologous Groups) numbers, allowing classification into functional categories, by following a majority-rule of the annotation of the individual genes that are part of each OG. Rarefaction curves for all OGs in the *Salinispora* pangenome were generated using the 'vegan' package in R.

2.4 Results

2.4.1 Sequencing and Genome Statistics

Out of 119 strains sequenced, *S. arenicola* accounted for 62 strains, *S. pacifica* accounted for 45 strains, and *S. tropica* accounted for 12 strains (Figure 2.1). The average genome size for the genus was 5.57 Mbp, however this number varied at the species level. Notably, *S. arenicola* genomes were on average larger than *S. tropica* and *S. pacifica* by approximately 400 Kbp and 300 Kbp, respectively. As would be expected, this larger genome equates to more genes in *S. arenicola*. The average GC content for each species ranged from 69.2-69.8% for all genomes. The average scaffold size reflects how well the entire genome was sequenced. While having a closed genome eliminates sequence gaps in downstream analyses, these 'permanent drafts' are of excellent quality given the sequencing effort involved (Table 2.1).

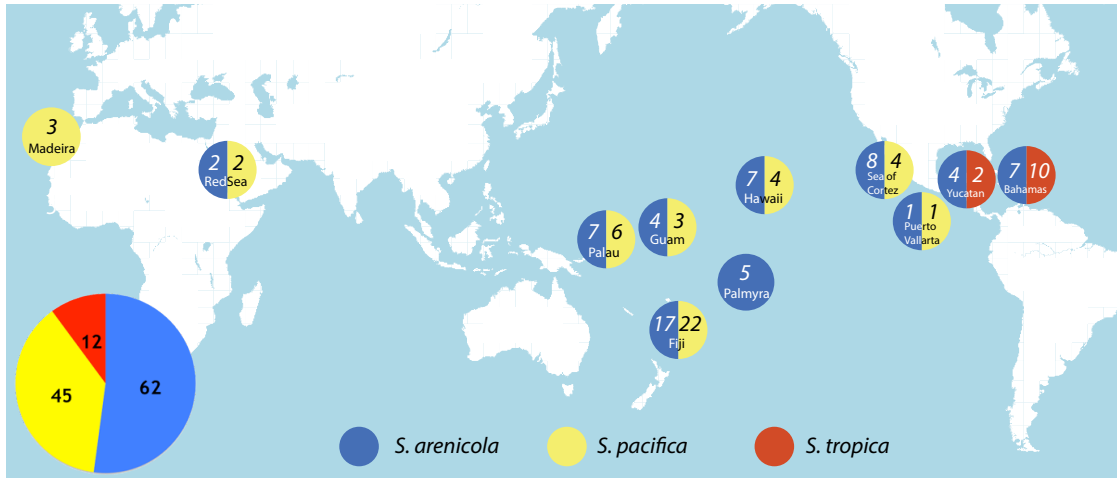


Figure 2.1: Distribution and number of strains sequenced for each species. Pie chart indicates total number of strains sequenced for all species.

Taxa	Average Genome Size (Mbps)	Average Gene Count	Average Scaffold Count	Average GC Content (%)
<i>Salinispora</i>	5.57	5148	85	69.7
<i>S. arenicola</i>	5.74	5234	80	69.8
<i>S. pacifica</i>	5.42	5079	90	69.9
<i>S. tropica</i>	5.31	4959	89	69.2

Table 2.1: Averaged genome statistics for the genus *Salinispora* and each species.

2.4.2 Ortholog Analysis and *Salinispora* Pangenome

The FastOrtho analysis revealed a pangenome comprised of the 18,492 orthologous groups (OGs). This gene pool encompasses all of the genetic diversity detected in the 119 genomes sequenced. Of these OGs, 5,176 were found in all three species (Figure 2.2a), but not necessarily in all strains within each species.

The Venn diagram reveals large gene pools observed only in *S. arenicola* and *S. tropica* (Figure 2.2a). This corresponds with the reduced phylogenetic diversity observed in *S. tropica*, but is inconsistent with the relatively large levels of diversity detected within *S. pacifica* relative to *S. arenicola*. This discrepancy may well be an artifact of uneven sampling of the two species sample size. The species-specific gene pools in-

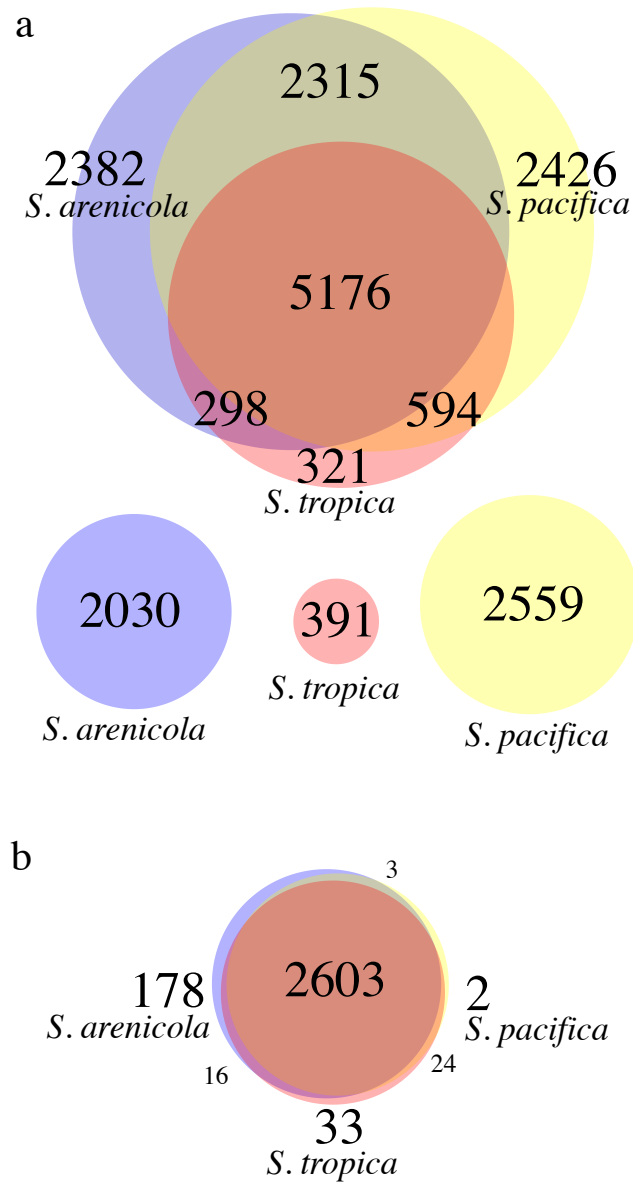


Figure 2.2: Number of ortholog groups based on FastOrtho analysis. **a**, all 18,492 genes in the *Salinispora* pangenome separated into a Venn diagram (to scale) of gene pools by species (upper). The lower three circles indicate the number of genes found in only one strain of each species (singletons). **b**, Venn diagram showing the relationships among the core gene pools identified for each species. These are the genes that are found in all strains of each species.

clude OGs that occur in anywhere from one strain to all strains within each species. Surprisingly, many of the OGs that were only observed in one species were only observed in one strain (singletons) (Figure 2.2a). This result provided the first hint of the extraordinary genetic diversity that can be observed among closely related *Salinispora* strains.

Shared OGs between two species are especially interesting when accounting for biogeography. *S. arenicola* co-occurs with both *S. pacifica* and *S. tropica*. FastOrtho results identified 2,315 and 298 ortholog groups, respectively, that are shared by these species pairs. It is expected that important, species-specific traits would be found exclusively in all strains of a given species. A second Venn diagram was created that included only those OGs that were observed in all strains of each species (Figure 2.2b). Of these, the vast majority (2,603 OGs) was found in all 119 strains (Figure 2.2b). This core genome unifies *Salinispora* as a genus and accounts for 28% of the entire pangenome.

Another surprising result from the FastOrtho analyses was the relatively few OGs that occurred exclusively in all strains of each *Salinispora spp.* (Figure 2.2b). These OGs are expected to define the genetic basis for differences among the species. These species-specific OGs ranged from 178 in *S. arenicola* to only two in *S. pacifica*. The lack of genetic coherence within *S. pacifica* supports a developing hypothesis that the clade encompassed by this species more appropriately represents an amalgam of species. This hypothesis was supported in a recent paper in which it was suggested that the *S. pacifica* clade more appropriately represents seven different species (Millán-Aguíñaga et al., 2017).

Singleton pools are also of interest as they provide a look into more recent evolutionary history and suggest strains have sampled genetic information from other taxa in their environment via horizontal gene transfer (HGT). These genes are by definition strain-specific and are explored further in Chapter 4. It is interesting to note how high

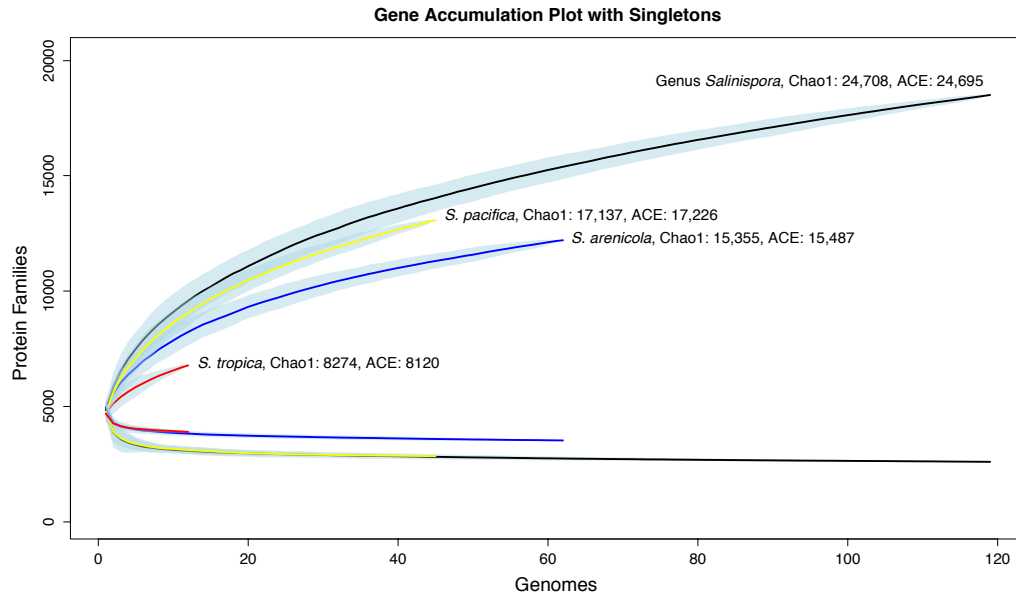


Figure 2.3: Rarefaction curves for the number of ortholog groups as sample size of genomes sequenced increases. Curves account for ortholog accumulation and decumulation for the entire genus as well as by species. Blue shading indicates standard error.

these numbers are, especially in relation to the species-specific gene pools. A large proportion of these genes, however, are of unknown function.

Rarefaction curves help to assess how effectively a group of organisms has been sampled. In the case of rarefying pangenomes, the potential number of ortholog groups for a taxon to reach saturation can be determined (Figure 2.3). Despite the *Salinispora* pangenome having more than 18,000 genes, Chao1 and ACE indices suggest an uncaptured gene pool of more than 6,000 additional genes. When broken down by species, this trend still holds true and all rarefaction/accumulation curves do not appear to be approaching an asymptote or a flattening of the curve to indicate saturation. The decumulation curves at the bottom of (Figure 2.3) depict the core genome and do appear to stabilize, suggesting this analysis has captured the basis of what represents a core bacterium in the genus *Salinispora*.

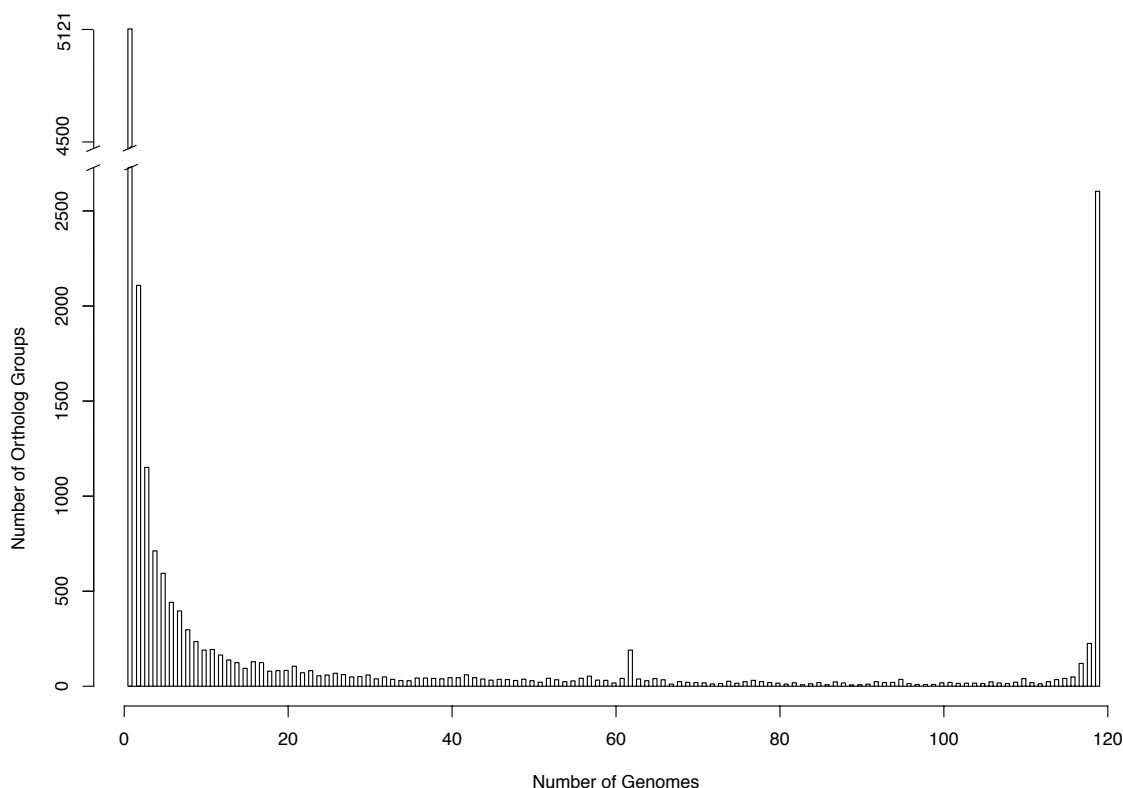


Figure 2.4: Histogram of the number of ortholog groups found across all genomes. There are 2603 ortholog groups in the 119 strains sequenced (core genome). The number of ortholog groups found in only 2 genomes is comparable at 2108.

The occurrence of OGs across the 119 genomes is represented in a histogram (Figure 2.4). The core genome comprising the 2,603 OGs found in all 119 strains is represented on the right end of the graph. These genes include housekeeping genes necessary for basic cellular function and notably is missing one particular gene *mscL*, that has been attributed to the obligate marine nature and osmotic requirements necessary for the genus to grow (Penn and Jensen, 2012; Bucarey et al., 2012). As the number of genomes decreases, the number of OGs drops precipitously with a notable spike at genome 62, which corresponds to the number of *S. arenicola* strains sequenced and the 178 OGs detected in all of those strains.

The number of OGs rises again as the number of genomes decreases suggesting there are, proportionately, many more OGs that are present in only a handful of genomes.

This spike in the left portion of the graph provides further support for the concept that a large proportion of the genetic diversity observed in this genus is distributed among only a few strains.

2.4.3 Genus and Species COGs

Looking specifically at annotation and Clusters of Orthologous Groups (COGs), I was able to see how the average *Salinispora* genome is broken down by putative function (Figure 2.5). COGs relating to energy production as well as transport and metabolism constitute a significant part of each genome. Unfortunately, a larger part of the genome, ~38%, is comprised of genes with unknown function or general prediction functions only. To assess species-specific functional traits, genes found in all strains of *S. arenicola* and *S. tropica* were analyzed further. *S. pacifica* was omitted due to only two genes in all *S. pacifica* species, one of which is annotated as function unknown. COG function for the 178 genes found in all *S. arenicola* and the 33 genes found in all *S. tropica* were analyzed (Appendix B and Figure 2.6).

Only three of the species-specific genes found in *S. arenicola* were found in a biosynthetic gene cluster specific to this species. This pathway was *terp1* and the genes are predicted to encode a class I diterpene synthase, cytochrome P450, and class II terpene cyclase. The *S. tropica* core had 15% more genes relating to cellular processing and signaling. COGs for information, storage and processing were nearly equal for the two species. Metabolism COGs, however, were 10% higher in *S. arenicola* than *S. tropica*. A large percentage of the species-specific core (37-45%) had COG annotation of poorly characterized. Upon closer examination of each COG category, *S. tropica* had three times as many species-specific genes for cell wall/membrane/envelope biogenesis, cell cycle control/cell division/chromosome partitioning, replication/recombination repair, and coenzyme transport/metabolism than *S. arenicola* (Figure 2.7). *S. arenicola*, on

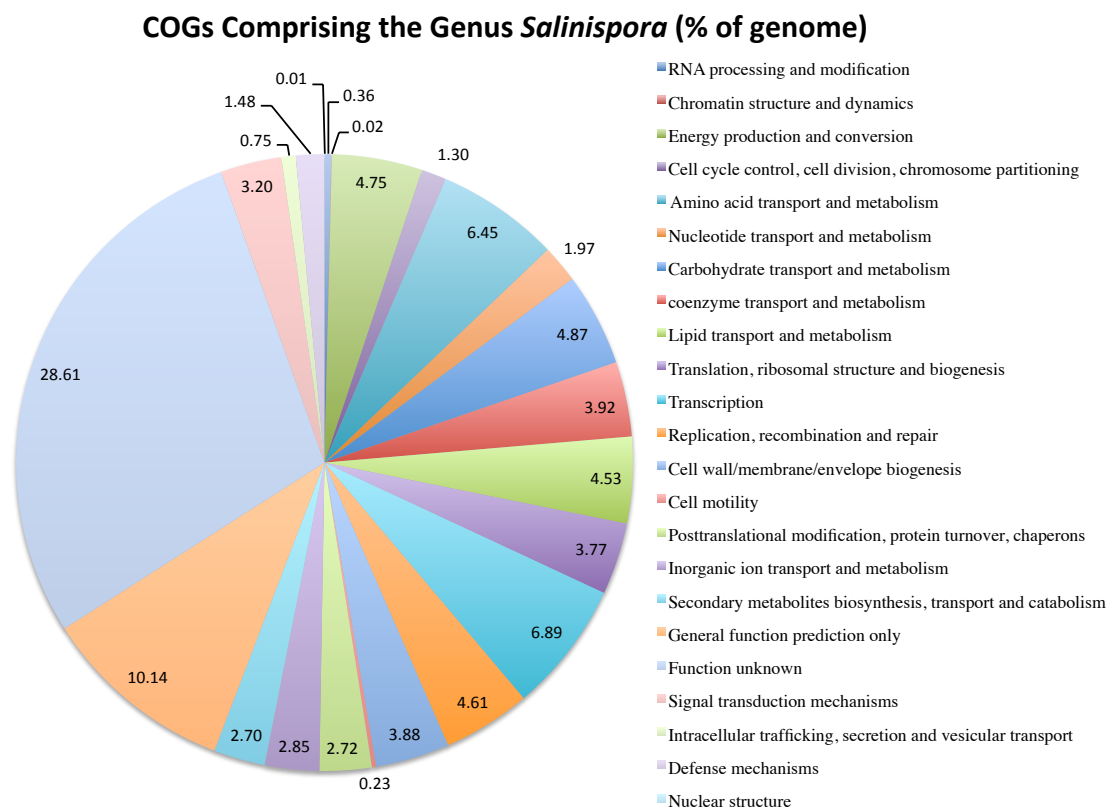


Figure 2.5: Clusters of Orthologous Groups (COGs) as percentage of genome averaged across all 119 *Salinispora* genomes.

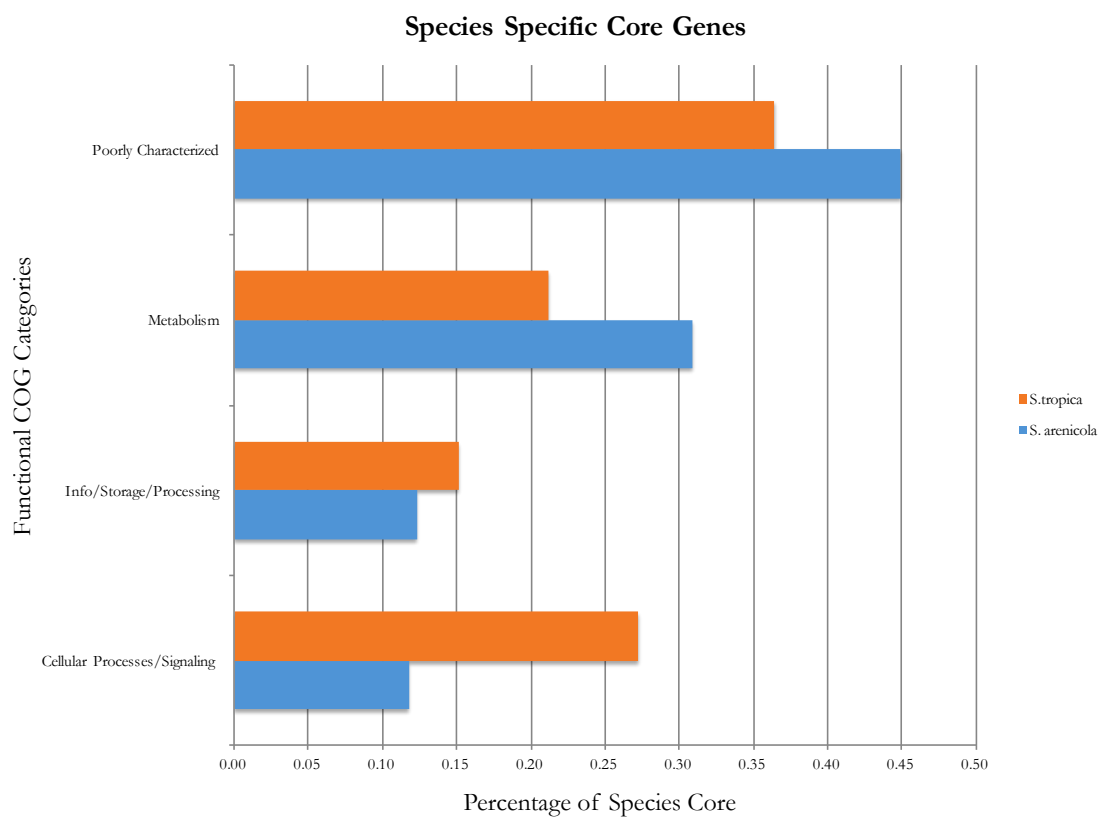


Figure 2.6: Average distribution of major COG categories for *S. arenicola* and *S. tropica* genomes.

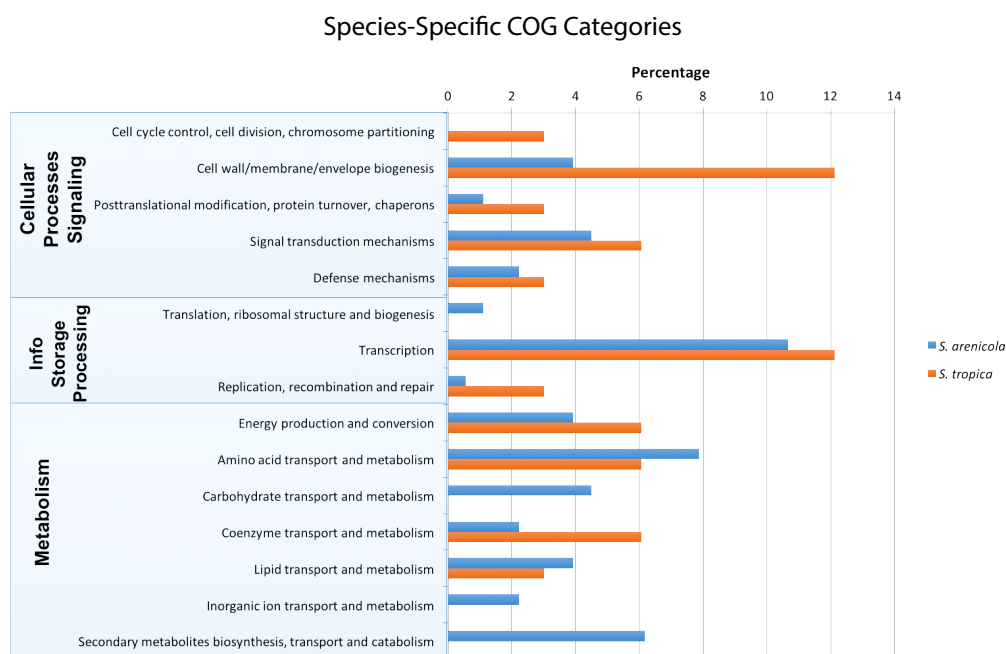


Figure 2.7: Average distribution of COG categories for the *S. arenicola* and *S. tropica* core genomes.

the other hand, had species-specific genes not found in *S. tropica*. These were involved in inorganic ion transport and metabolism (2%), carbohydrate transport and metabolism (4%), and secondary metabolite biosynthesis, transport and catabolism (6%).

2.5 Discussion

This study investigated the pangenome of the marine actinomycete genus *Salinispora* with the aim of determining the genetic basis for species specificity. The three species that comprise *Salinispora* vary in genome size with *S. arenicola* having, on average, the largest genome. *S. tropica* has the smallest genome, and *S. pacifica*'s genome sizes fall between the other two. The pangenome of *Salinispora* is comprised of more

than 18,000 genes yet rarefaction curves suggest that there is potential for the discovery of thousands more. Isolation and sequencing of more *Salinispora* strains would very quickly add many more genes to the pangenome. The division of the pangenome into a core and flexible component can allow for the identification of genes that either unify *Salinispora* as a genus or differentiate the three species.

The core genome encodes basic functions and contains many housekeeping genes as well as genes for translation, replication and energy production (Bentley, 2009). Of note is the absence of a gene from all *Salinispora* genomes. The apparent loss of a large conductance mechano-sensitive channel (*mscL*) gene previously identified as missing in two closed *Salinispora* genomes (Penn and Jensen, 2012), appears to be likewise absent in the rest of the genomes sequenced. This gene, however is found in its most closely related genus, *Micromonospora*. The *mscL* gene occurs in other Actinobacterial genomes and provides a mechanism to survive osmotic downshock (Sukharev et al., 1997; Roberts, 2005). The absence of this gene has been the predominant factor attributed to *Salinispora*'s inability to survive in media with low osmotic strength (Bucarey et al., 2012) and supports the hypothesis that gene loss contributes to the obligate marine nature of this genus (Penn and Jensen, 2012).

An additional aim of this study was to identify the genetic basis for differences between *S. arenicola* and *S. tropica*. The species-specific core genomes help answer this question. Differences in cellular processes between the two species show a larger proportion of species-specific genes devoted to cellular processes and signaling in *S. tropica*. Genes associated with metabolism appear to be more represented in *S. arenicola*. These genetic differences support results from recent observations that *S. arenicola* can produce more secondary metabolites than *S. tropica*, while *S. tropica* grows faster. This may be evidence of tradeoffs of energy investment as competitive strategies (Patin et al., 2016). Because differences in secondary metabolite production have been identi-

fied among *Salinispora* species, genes from secondary metabolite biosynthetic clusters in *Salinispora* were compared to this gene pool. Only three genes were found. These genes are a part of the terp pathway with no known associated compound. A substantial proportion of the species-specific genes are of unknown function. Although many OGs are function unknown, these genes could hold the key to what defines the taxa as evidenced by the recent construction of the smallest viable bacterial cell, of which 32% of the genes necessary for viability are annotated 'function unknown' (Hutchison et al., 2016).

We are still at the forefront of determining how to best delineate species. Asking this question using closely related taxa within the genus *Salinispora* has shown that a surprisingly few number of genes were specific to each particular species. While genomic differences may not be fully apparent, the manner in which these genes are transcribed may provide more insight into how these species differ, and in fact, it has been suggested that species level descriptions should take into account another parameter, gene regulation and expression (Konstantinidis et al., 2009).

Chapter 2 is coauthored with Millán-Aguiñaga N, JA Ugalde, and PR Jensen. The dissertation author was the primary investigator and author of this chapter.

Chapter 3

Species-specific functional traits between two species in the genus *Salinispora*

3.1 Abstract

This chapter seeks to further define species-specific traits through transcriptomics. It has been suggested that comparative genomics solely at the genome level may limit the differences found between species and that gene regulation and expression should constitute another important parameter for species descriptions. Transcriptome data for three strains, two *S. arenicola* and one *S. tropica* under the same laboratory conditions were analyzed during both exponential and stationary phase. Differential expression was identified in chitinase genes shared between the two species. Growth on colloidal chitin demonstrated larger zones of clearing for *S. tropica* than *S. arenicola* suggesting *S. tropica* is better able to utilize the most abundant biopolymer in the ocean. Shared genes between the two species were identified and analyzed for differential ex-

pression to look for traits that differentiate them as species. Bioinformatic predictions suggest *S. tropica* has the ability to better cope with osmotic stress and may be more affected by nutrient or oxidative stress while *S. arenicola* has more energetic needs during stationary phase growth potentially due to costly secondary metabolism.

3.2 Introduction

In what capacity do genotypic differences correlate to phenotypic difference? How are these differences reflected in how we describe bacterial species? These questions have historically been difficult to answer. Bacterial species concepts are controversial. Delineating species has long been a contentious issue. Many studies focus on how phylogenetic diversity can be interpreted in the context of species (Cohan and Koeppel, 2008). Prokaryotic species concepts address the evolutionary forces that give rise to discrete clusters observed in phylogenetic trees. However, this is complicated by the issue that bacteria exchange genes through horizontal gene transfer, which can blur species boundaries further (Koonin et al., 2001). Species definitions address the criteria for which these clusters are delimited (Doolittle and Zhaxybayeva, 2009a). As scientists, we desire a species definition to clearly adhere to a set of stable rules that govern when two organisms are similar enough to be given the same name. However, this is likely to be unattainable as no single set of rules will apply to a group as diverse as bacteria. Thus, a species concept rooted in ecological and evolutionary theory is likely to be more meaningful.

Powerful genomic tools to investigate genetic diversity are readily available and inexpensive, allowing scientists a detailed look at relationships between genotype and predicted phenotype. Past studies have underscored the need to combine both genetic diversity and ecology to define species thus providing a firm foundation for bacterial

species concepts (Fraser et al., 2009). The notion that two individuals with the same genes have varying expression levels has been identified in several bacterial taxa. For example, rates of evolution of different cellular functions have been identified in the genus *Shewanella*. A systems-level analysis has shown a remarkable level of versatility in this genus (Fredrickson et al., 2008). Expression differed among these organisms even when grown under the same conditions. In fact, more differences were found in expression levels than those at the genome level, suggesting that similarity in gene regulation and expression should constitute another important parameter for species descriptions. It has been found that similarities in expressed pathways are determined by genetic relatedness, distinct ecological adaptation or a combination of the two (Doolittle and Zhaxybayeva, 2009b; Konstantinidis et al., 2009).

Previous studies in *Vibrio* show distinct populations where one is free-living, scavenging nutrients in the water column and the other is particle associated and capable of producing biofilms in order to attach to nutrient particles (Doolittle and Papke, 2006; Yawata et al., 2014). Theory suggests co-occurring bacteria should be ecologically distinct otherwise one would out-compete the other. This is further exemplified in closely related *Leptospirillum* groups. Researchers addressed the relationship between gene content and ecological divergence. They were able to determine that one genotypic group was an early colonizer while the other group proliferated in later successional stages. Across each subset of groups, only ~15% of genes were unique to each genotype and were involved in niche partitioning showing how subtle genetic variations can lead to distinct ecological strategies (Deneff et al., 2010). Certain adaptive traits have also been shown to limit dispersal. Latitudinal gradients have limited dispersal of *Streptomyces* sister-taxa due to thermal trait adaptation, restricting gene flow across climate regimes (Choudoir and Buckley, 2018). Studying marine bacteria adds another dimension to dispersion. While temperature decreases with latitude, it also decreases

with depth.

When looking at the genus *Salinispora*, the different pools of genes I have identified can be used to help answer questions about what evolutionary processes and ecological traits make *S. arenicola* different from *S. tropica*. However, genes that differentiate the two species have not shown a clear association between genotypic analyses and function. *S. arenicola* and *S. tropica* have been isolated from the same sites in the Caribbean and the Bahamas. Perhaps one species is an early colonizer while the other takes advantage of a later successional stage? How are these very closely related species able to cohabitate and what are the tradeoffs? Much of what we know about *Salinispora* pertains to secondary metabolism and not as much about habitat and nutrient utilization. Experimental differences in growth rate have been observed for the two species and a current hypothesis is that *S. arenicola* appears to invest in the production of secondary metabolites early in development as a trade-off to growth rate (Patin et al., 2016). Whereas *S. tropica* is able to grow relatively quickly, potentially capable of establishing itself in the environment more readily, giving itself a competitive advantage. This however may come at the expense of having a smaller genome and producing fewer secondary metabolites (Fraser et al., 2009).

The description of the type strains for each *Salinispora* species provided limited insight into primary metabolism, however much more can be learned. These species descriptions have identified some phenotypic differences. In particular, *S. arenicola* can use arbutin, L-proline, (+)-D-salicin, L-threonine and L-tyrosine as sole carbon sources for energy and growth, but not (+)-D-galactose or inulin and can grow in the presence of rifampicin (25 mg/ml). *S. tropica* can use (+)-D- Galactose and inulin as sole carbon sources for energy and growth but does not grow in the presence of rifampicin (25 mg/ml) (Maldonado et al., 2005). Although some phenotypic differences have been identified, I have chosen to utilize a transcriptomics approach in order to study differ-

ential expression of genes in a quantitative manner rather than the ability to utilize a particular nutrient source. *Salinispora* grows in a clumping manner and it is no easy task to homogenize. This could lead to false negatives in phenotypic experiments and, indeed, this has been the case in Biolog (Biolog Inc, Hayward CA) assays conducted in the lab previously.

I first use chitinase genes to exemplify that genes shared by two *Salinispora* species are differentially expressed. The ability to metabolize chitin has potential ecological implications due to its abundance as a biopolymer in the marine environment and as a carbon and nitrogen source. It also has been shown to be implicated in secondary metabolite production in *Streptomyces spp.* (Rigali et al., 2008).

The second part of this chapter utilizes transcriptomics data that has previously been used to identify expression levels of secondary metabolite biosynthetic gene clusters in strains grown under standard laboratory conditions (Amos et al., 2017). This dataset had not been used to address the expression of primary metabolism genes in *Salinispora*. In order to determine differences between these *Salinispora* species, genes found in all strains were studied and assessed for differential expression. The majority of these genes will be involved in some type of primary metabolism. I was also able to determine if they are involved in a previously characterized metabolic pathway. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a collection of databases used for bioinformatic studies in various omics fields and contains maps representing experimental knowledge on metabolism and various other functions of the cell and organism (www.kegg.jp). Because these pathways have been so thoroughly studied in other organisms, we can gain a greater insight into how metabolism differs in our organism of interest, *Salinispora*.

In the previous chapter, I identified ortholog groups that define the pangenome of *Salinispora*, the entire suite of genes that have thus far been sequenced across the

entire genus. The aim of this chapter is to investigate the metabolic potential of two species of *Salinispora* in order to find species-specific phenotypes that would have not been uncovered using traditional comparative genomics techniques.

3.3 Materials and Methods

3.3.1 Chitinase Gene Identification

Chitinase genes were identified based on annotation using IMG/ER ([http://img-jgi.doe.gov](http://img.jgi.doe.gov)). The ortholog group for each chitinase gene was determined from the FastOrtho analysis of Chapter 2 (<http://enews.patricbrc.org/fastortho>). In order to prevent overlooking genes that may have not been annotated correctly as chitinases, BLAST searches of representatives from each chitinase ortholog group were performed against the National Center for Biotechnology Information (NCBI) non-redundant database.

3.3.2 Colloidal Chitin Preparation

Practical-grade chitin was purified using the following protocol to be used as a carbon/nitrogen source in bacterial growth media. 400 mL of cold 12N HCl was added to 10 g chitin from shrimp shells (Sigma C7170) in a 1 L beaker and stirred for approximately 10 minutes. The beaker was placed in a 37°C water bath and stirred every 5 minutes for 30 minutes using a glass rod until the solution became homogenous. The solution was poured into a 6L Erlenmeyer flask and 4 L dH₂O was added and stirred completely. The flask was sealed with parafilm and placed at 4°C overnight for chitin precipitation. The following day, the supernatant was removed using a vacuum pump and the pH of the chitin precipitate was adjusted using NaOH pellets to pH 7. Chitin concentration was determined by drying 10 mL of the chitin solution overnight at 70°C

and taking a dry weight measurement. The colloidal chitin solution was added to a final concentration of 0.4% to agar growth media in plates. Two strains each of *S. arenicola*, (CNS-205 and DSM45545) and *S. tropica* (CNB-440 and CNR-699) were grown to early exponential phase. Cellular mass was normalized using packed cell volume and homogenized using a pestle and 1.5 mL Eppendorf tube. Four replicates of 40 μ L of the homogenate were pipetted onto plates of agar containing 0.4% colloidal chitin media. Cultures were grown for 16 days. Chitin metabolism was determined as the radial length of zones of clearing around colonies.

3.3.3 Strain Cultivation for Transcriptome Study

Growth curves were generated for each of the three strains (CNR-699 was not included in the transcriptome study) to establish time points for transcriptome and metabolome analyses (Amos et al., 2017). These strains were *S. tropica* CNB-440, *S. arenicola* CNS-205, and *S. arenicola* DSM45545. Starter cultures were inoculated from frozen stocks into 50 mL of medium A1 (10 g/L starch, 4 g/L yeast extract, 2 g/L peptone, and 1 L of 0.2- μ m filtered seawater) in 250 mL flasks and grown for 5 d at 25 °C with shaking at 160 rpm (New Brunswick Innova 2300). Starter cultures (1 mL) were then used to inoculate each strain into triplicate flasks containing 50 mL of A1 and glass beads to reduce clumping. Optical density (600 nm) was monitored at 24-h intervals, with three readings averaged for each replicate culture at each time point. Based on the results of the growth curves, 96 h and 216 h were selected as time points for the transcriptome and metabolome analyses.

3.3.4 Transcriptome Analyses

At each time point, RNA was extracted from 5 mL of culture following an acid phenol/chloroform/isoamyl alcohol procedure (Nieselt et al., 2010). RNA was sent to the US Department of Energy Joint Genome Institute (JGI) for sequencing, quality control, and read mapping as previously described (Letzel et al., 2017). One of three replicates of DSM45545 failed JGI's QC criteria due to degraded RNA and another replicate could not be obtained due to the synchronized nature of the original transcriptomics experiment. In brief, Illumina HiSeq 2500 sequencing generated $>3 \times 10^7$ paired-end reads (100 bp) per replicate. Using BBMap (<https://sourceforge.net/projects/bbmap/>), raw reads were evaluated for artifact sequences by kmer matching (kmer = 25). Quality trimming was performed using the phred trimming method set at Q10 (Ewing and Green, 1998), with reads under 45 bases removed. Raw reads were aligned to their respective reference genome using Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2010). FeatureCounts was used to generate raw gene counts (Liao et al., 2014). Mapped reads were visualized using BamView in Artemis (Rutherford et al., 2000). The number of reads per kilobase of transcript per million mapped reads (RPKM) was used to normalize raw data in Artemis (Mortazavi et al., 2008). Expression levels were derived from average values calculated for key biosynthetic genes. These included polyketide synthases, nonribosomal peptide synthetase, terpene synthase, precursor peptide (bacteriocin), and LanM (lantibiotic) genes. Additional genes associated with key biosynthetic operons were checked to confirm the expression levels.

3.3.5 Differential Gene Expression Analysis

The RPKM values for chitinase genes were analyzed using a one-way ANOVA to determine significant differential expression between strains at both exponential and

stationary growth phases. A Tukey's HSD post hoc test was then run to determine pairwise statistical differential expression between strains.

Genome-wide differential expression for the three strains was analyzed using T-REx, a transcriptome analysis webserver for RNA-seq expression data (de Jong et al., 2015). The RPKM values generated in the previous section were analyzed in a pairwise comparison between strains for both exponential and stationary phase. The pipeline requires raw RNA expression level data for analysis. The input files were divided into five categories. The first was a 'Gene Counts' input file containing RPKM values for every ortholog Salin Group. Salin Groups for this analysis were limited to only those found as a single copy in each of the three genomes sequenced. The second input file, 'Factors File', defines the factors that were used to describe the experiments and the replicates. This file consists of 'Experiments' broken down by each replicate and by 'Strain'. The third input file, 'Contrasts File', defined which comparisons should be made between the various experimental conditions. The fourth input file, 'Annotation File', attributed an annotation for each Salin Group as well as its gene length. Gene expression data were first normalized and the statistical method of RNA-seq analysis R-package EdgeR and DEseq was implemented (Robinson et al., 2010; Anders and Huber, 2010; Love et al., 2014). In order to calculate p-values for differential expression, the dispersion model of EdgeR was employed. The output of the T-REx analysis was filtered for gene expression of logFC between -1.0 and 1.0 for both *S. arenicola* strains, CNS-205 and DSM45545. This subset of genes was then filtered against the *S. tropica* strain, CNB-440 for gene expression of logFC less than -1.0 and greater than 1.0. All genes with a p-value greater than 0.05 were disregarded. T-REx uses two predefined cutoff values to generate 2 lists of differentially expressed genes; TopHits (fold change ≥ 2 and a p-value ≤ 0.05) and HighFold (fold change ≥ 5 and a p-value ≤ 0.01).

3.4 Results

3.4.1 Chitinase Genes Identified

A bioinformatic survey of chitinase genes across 119 genomes representing all three named *Salinispora* species resulted in four distinct chitinase ortholog groups (Figure 3.1). These ortholog groups, Salin3372, Salin984, Salin5528, and Salin4659 were present in varying degrees by species. *S. tropica* genomes had all four chitinase genes present in all strains sequenced. The five *S. arenicola* strains from PM were the only other strains to have all four genes. Salin984 was identified in all strains of *Salinispora*. Salin5528 was not observed in any *S. pacifica* and only six *S. arenicola*, one strain from the Bahamas (CNB-527) and all strains from Palmyra. This gene was also pseudogenized in eight *S. arenicola* strains from various geographic locations. Salin4659 was identified in three *S. pacifica* genomes and 34 of 62 *S. arenicola* strains from various locations.

Two strains of *Salinispora arenicola* and *S. tropica* were grown in four replicates on media containing colloidal chitin as the sole carbon and nitrogen source. All four strains were able to degrade chitin as evidenced by a zone of colloidal chitin clearing (Figure 3.2). The extent of clearing differed between strains. On average, *S. tropica* had a clearing zone of $3.9\text{mm} \pm 0.1\text{mm}$ while *S. arenicola* had an average clearing zone of $1.8\text{mm} \pm 0.2\text{mm}$ (Figure 3.3), less than half of *S. tropica*.

3.4.2 Chitinase expression

Chitinase expression values collected during exponential and stationary phase growth were analyzed for three strains: *S. tropica* CNB-440, *S. arenicola* CNS-205, and *S. arenicola* DSM45545. The reads per kilobase of transcript per million mapped reads (RPKM) values for all four chitinase genes were compared based on growth phase

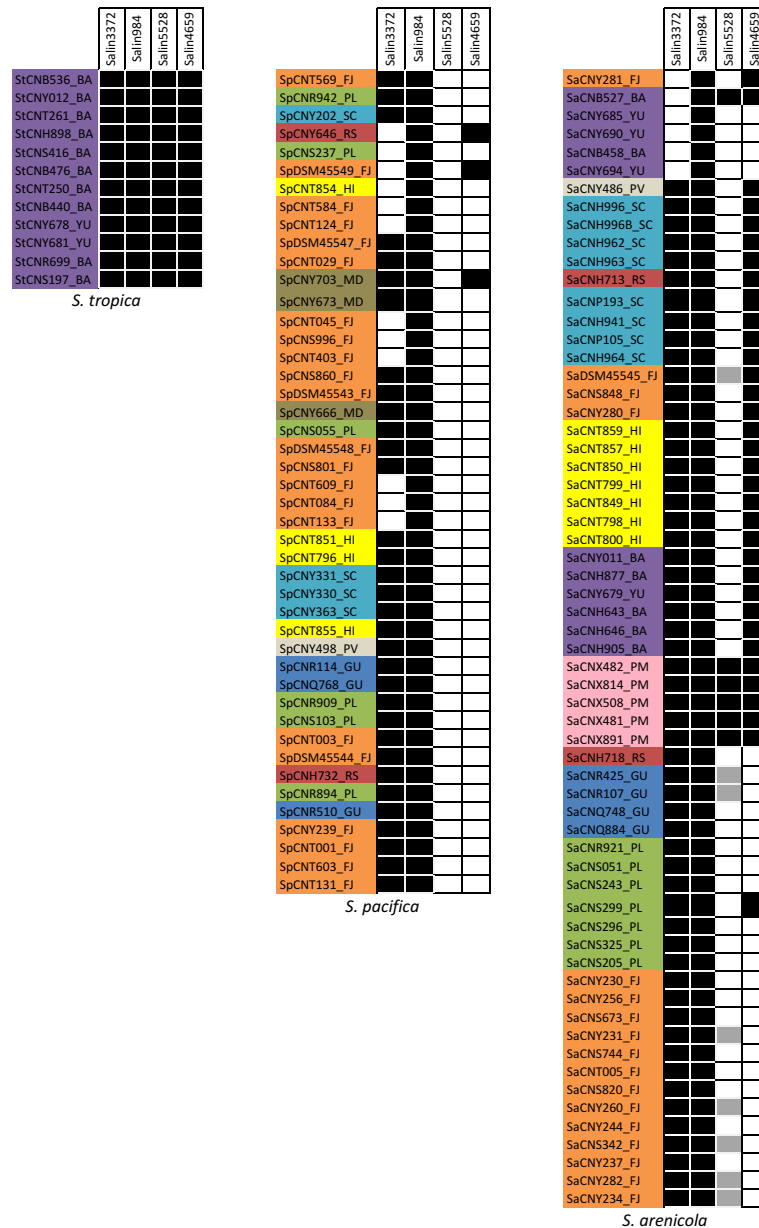


Figure 3.1: Four chitinase ortholog groups were identified in the genus *Salinispora*. All 4 genes were found in all *S. tropica* strains (first column), and occurred varying in *S. pacifica* and *S. arenicola* (second and third columns, respectively). Ortholog group Salin984 was identified in all 119 genomes. Cell color of strain distinguishes collection location. BA Bahamas (purple), YU Yucatan (purple), FJ Fiji (orange), PL Palau (green), SC - Sea of Cortez (light blue), RS - Red Sea (red), HI Hawaii (yellow), MD Madeira (brown), PV - Puerto Vallarta (tan), GU Guam (dark blue), PM Palmyra (pink). Strain order correlates to position on *Salinispora* phylogenetic tree. Black cells indicate presence of chitinase gene in genome. Grey cells indicate pseudogenes.

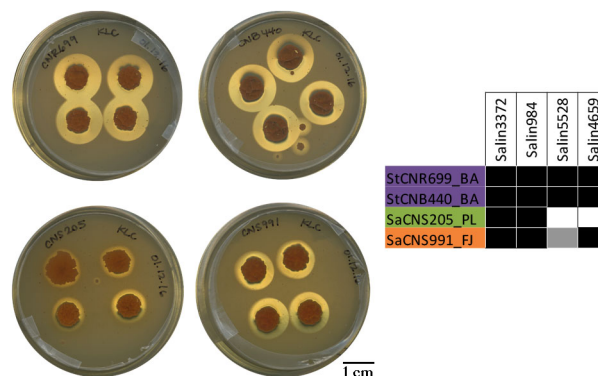


Figure 3.2: Demonstration of the ability of *Salinispora* to utilize colloidal chitin as a nutrient source on agar plates containing 0.4% chitin and agarose. Upper two plates are *S. tropica*, lower two are *S. arenicola*. Note that CNS-991 is also known as DSM45545. Black cells indicate presence of chitinase gene in genome. Grey cells indicate pseudo-genes.

(Figure 3.4). According to (Amos et al., 2017), the baseline to distinguish between silent and expressed biosynthetic gene clusters was established at 27.1 RPKM. Although chitinase expression levels appear to be less than 27.1 RPKM for all strains except CNB-440 at stationary phase for Salin984 and Salin5528 and exponential phase for Salin984, there is still evidence of chitinase activity through clearing zones of colloidal chitin on solid media (Figure 3.2). Note that CNS-991 is also known as DSM45545. A one-way ANOVA was used to determine if there were significant differences in expression between strains at each growth phase (Table C.1). Expression levels for chitinase genes were significantly different between strains for Salin3372 during exponential phase ($p < 0.001$), Salin984 during stationary phase ($p < 0.001$), Salin984 during exponential phase ($p < 0.05$).

A Tukey's HSD post hoc test was used to determine pair-wise significance between strains (Table C.2). Letters below bars in Figure 3.4 denote significant differences between strains. Capital letters denote results from one-way ANOVA of exponential phase, while lower-case letters denote stationary phase. There appeared to be no species-specific signal in expression except for Salin984 during stationary phase. Salin5528 was

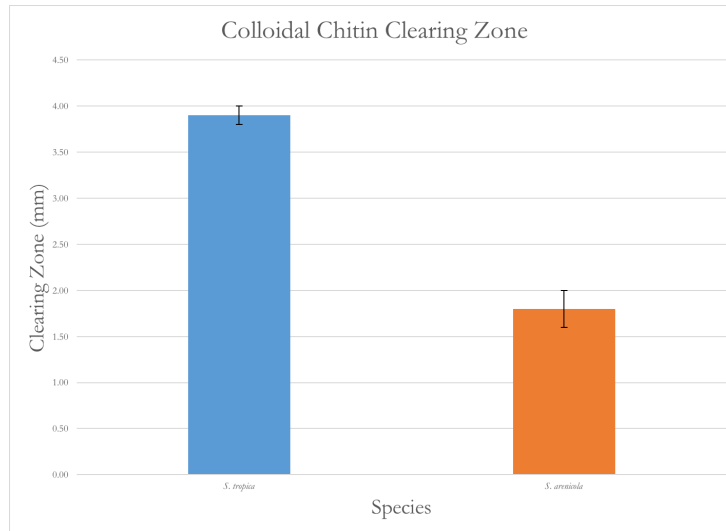


Figure 3.3: Average zone of clearing in mm of media with colloidal chitin as the sole carbon and nitrogen source for *S. tropica* and *S. arenicola* after 16 days of growth.

present in *S. tropica* and shows expression higher than the 27.1 RPKM threshold during stationary phase. Bioinformatic analysis shows that Salin5528 is present in the genome of *S. arenicola* DSM45545 as a pseudogene and is not expressed. It is not present in CNS-205. Salin4659 is also not present in CNS-205 and shows very low expression levels for the remaining two strains.

3.4.3 Global Differential Expression

I next used the T-REx pipeline to perform global analyses on the RNA-seq gene expression data derived from the two *S. arenicola* and one *S. tropica* strains. A total of 3365 Salin Groups were found in single copy in all three strains with available transcriptomics data. Library sizes for every experimental replicate largely showed good coverage of over 100,000 reads for experiments done in triplicate. Lack of replication mostly seemed to affect *S. arenicola* DSM45545 during exponential phase (Figure 3.5). Library reads for these samples averaged 10,000 reads fewer than the other experiments at 90,000 reads (Figure 3.5A).

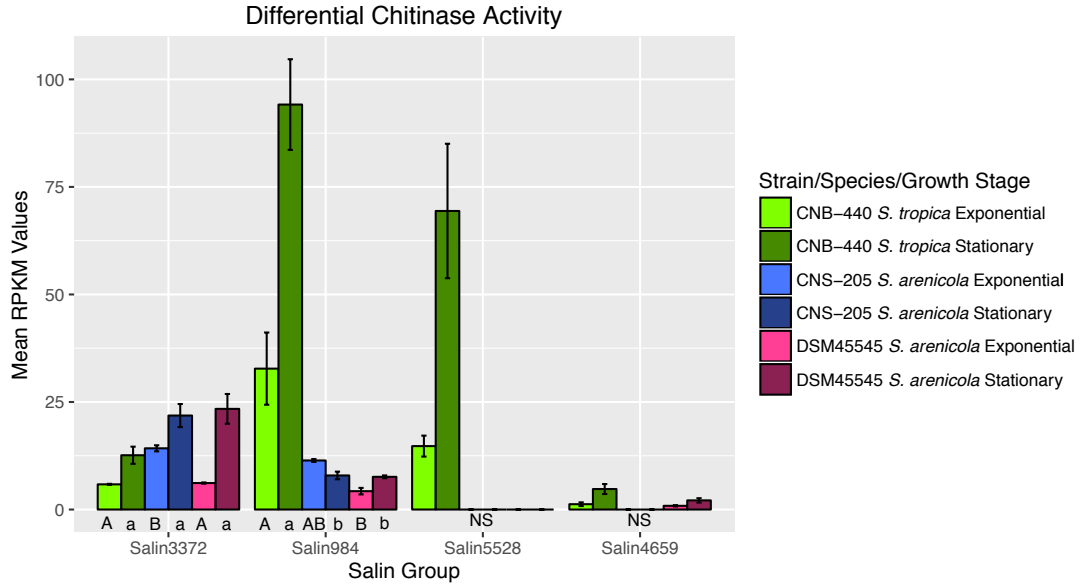


Figure 3.4: Expression levels of 3 *Salinispora* strains for 4 different chitinase ortholog groups sampled at exponential and stationary phase. Salin5528 and Salin4659 are not found in CNS-205. NS stands for *Not Significant*

A squared Pearson's correlation matrix of exponential and stationary phase for each strain was made (Figure 3.6). The scale is from light blue (max = 1.00) to dark blue (min = 0.00) indicating high to low correlation, respectively. *Salinispora arenicola* strains appear to be more closely correlated to each other (0.90) than to *S. tropica* (CNS-205: 0.73 and DSM45545: 0.75). A more in-depth look at a heatmap of top hits with fold change > 2 and $p < 0.5$ (Figure 3.7) shows a species-specific signal based on color intensity. Genes that are more differentially expressed are more intensely orange or blue for downregulated or upregulated genes, respectively. Intensity appears increased for interspecies pairwise comparisons.

An MA-plot shows log-fold change (M-values are the log2 ratio of level counts for each gene) against the log-mean (A-values are the average level counts for each gene). Genes with similar expression levels in two samples will appear around the horizontal line $y = 0$. A log2 transformed gene expression between -1 and 1 is indicated by a blue hash-marked band on the plot and genes are denoted by black dots. A normalized

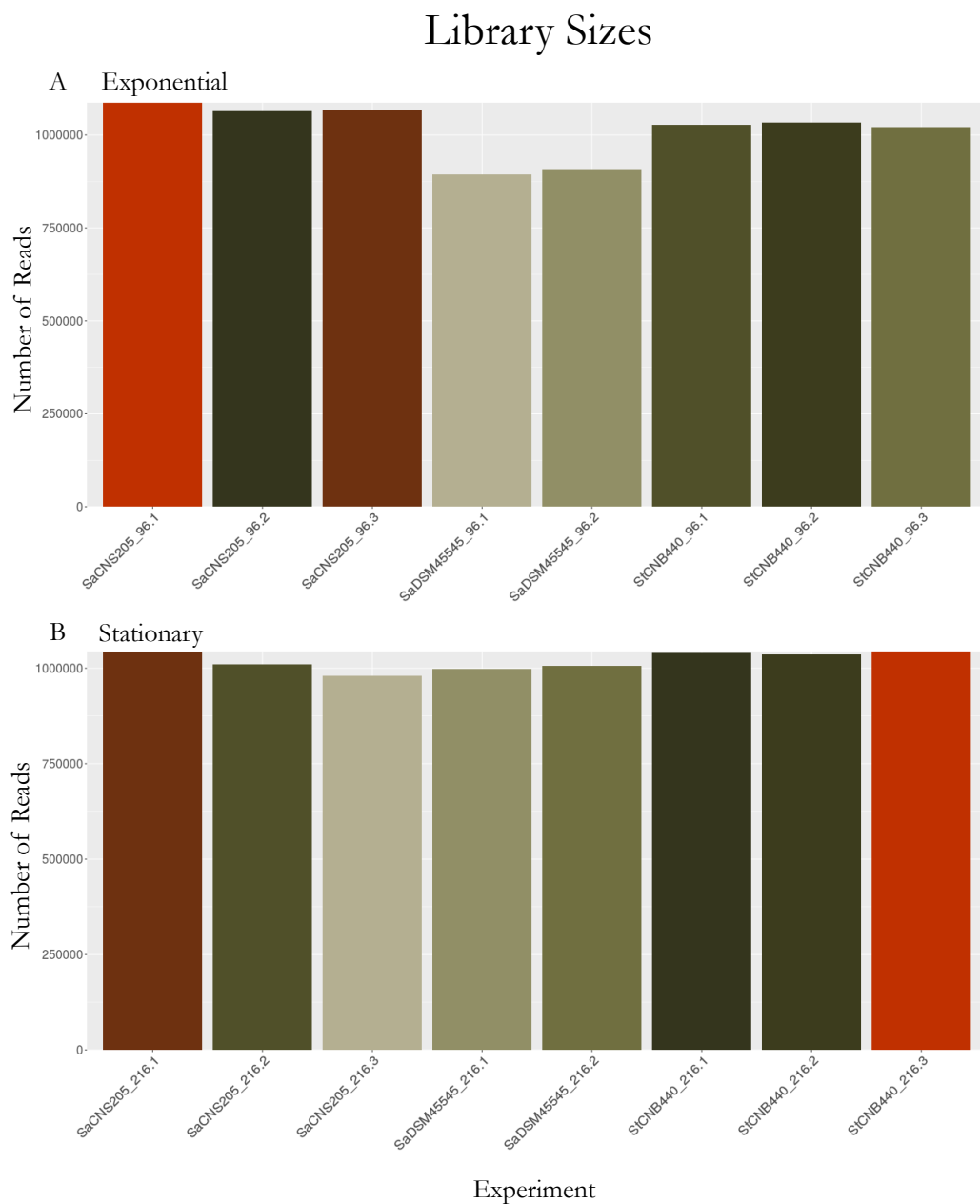


Figure 3.5: RNAseq library size for each experimental replicate. Exponential growth phase replicates include 96 and replicate number to denote a 96-hour timepoint. Stationary phase replicates include 216 and replicate number to denote a 216-hour timepoint. Panel **A** shows results for exponential phase experiments. Panel **B** shows results for stationary phase experiments.

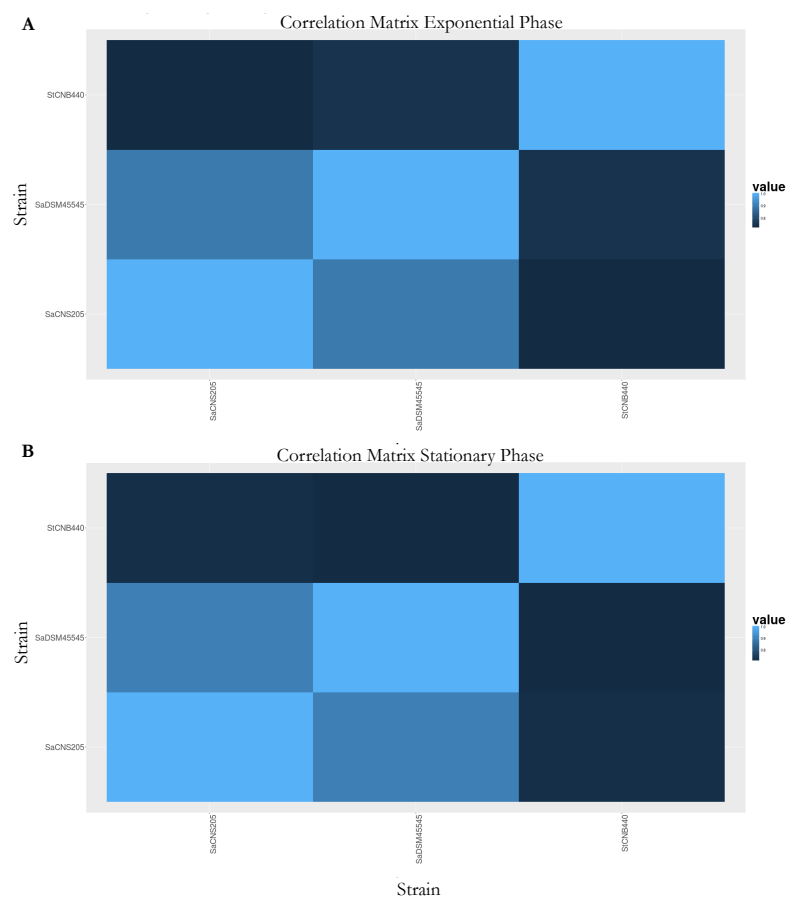


Figure 3.6: Squared Pearson's correlation matrix of exponential and stationary phase for each strain. The scale is from light blue (max = 1.00) to dark blue (min = 0.00) indicating high to low correlation, respectively. Panel **A** shows results for exponential phase experiments. Panel **B** shows results for stationary phase experiments.

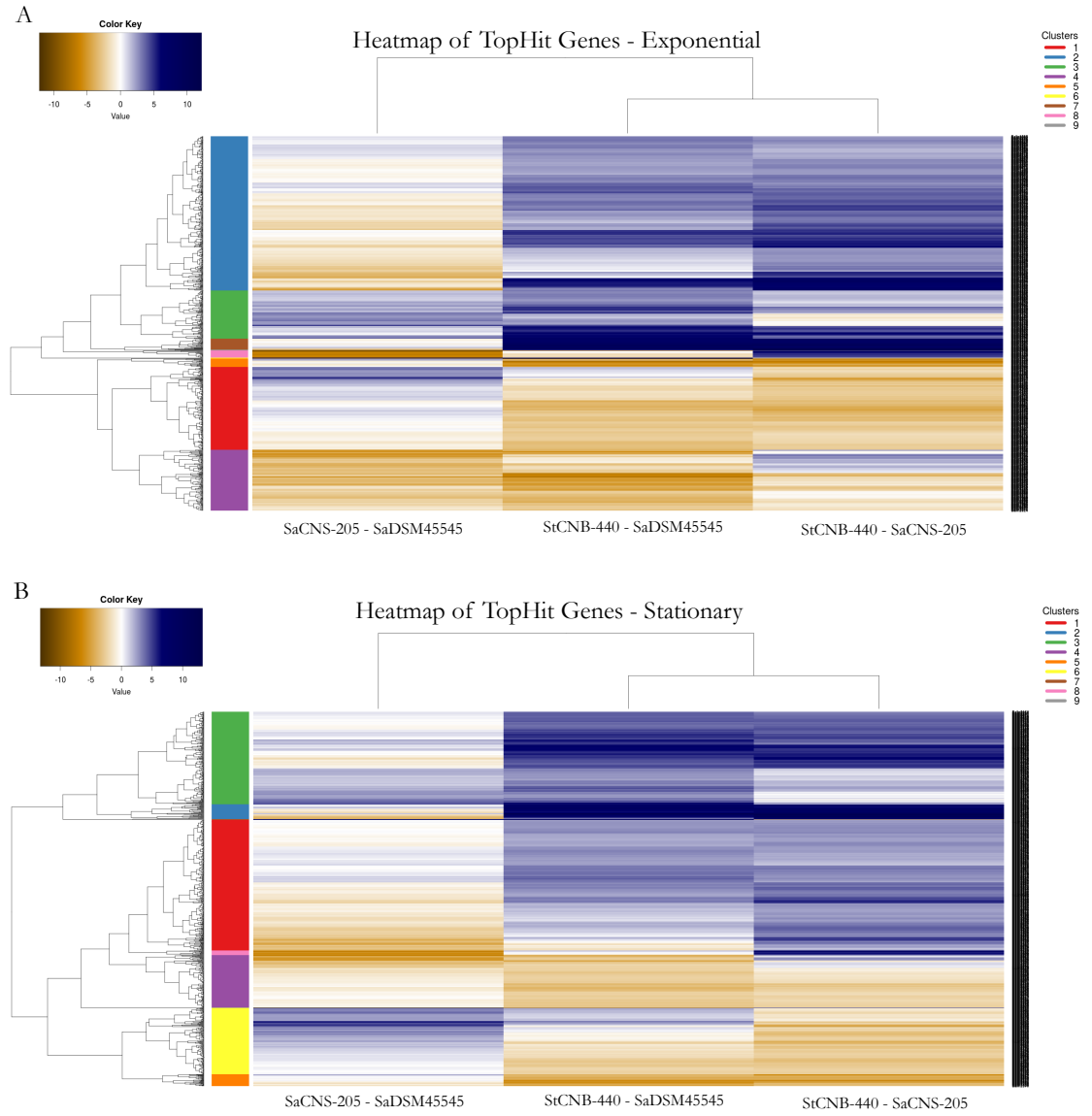


Figure 3.7: Heatmap of Top Hits (Fold change >2 and $p < 0.05$). Genes and comparisons are hierarchical clustered as indicated by the dendrogram on the left and top of the heatmap. Blue indicates that the first genome in the pairwise comparison has a positive fold change versus the second genome. Orange indicates a negative fold change. Panel **A** shows results for exponential phase experiments. Panel **B** shows results for stationary phase experiments.

expression value above \log_2 of 1 is indicative of doubling of expression and a value of -1 is considered a reduction to 50%. For each pairwise comparison of strains, genes with a log normalized value within the blue band fall into a region that is considered to not be biologically relevant with regards to differential expression (Friedman et al., 2006).

In order to attribute expression patterns to a particular species, genes were sorted based on \log_2 normalized values. In *S. arenicola* pairwise comparisons, genes with \log_2 fold change values between -1 and 1 were assigned to that species. Pairwise comparisons between *S. arenicola* and *S. tropica* that indicated genes with \log_2 fold change values above 1 or below -1 were then assigned as being differentially expressed between species. MA-plots for all pairwise comparisons at exponential phase (Figure 3.8) and stationary phase (Figure 3.9) show gene distribution based on mean expression levels. Top panels for each growth phase (Figure 3.8A and 3.9A) show a comparison between both *S. arenicola* strains and genes cluster more tightly to the blue hash-marked band indicating more similar gene expression.

Table 3.1 shows the number of genes found inside and outside of this blue band. At exponential phase, 72% of genes are similarly expressed between the two *S. arenicola* strains. However, 50% and 52% are differentially expressed when DSM45545 and CNS-205 are compared to *S. tropica* CNB-440, respectively. Similarly, at stationary phase, *S. arenicola* strains show 71% of genes with very low levels of differential expression. However, 40% and 52% are differentially expressed when DSM45545 and CNS-205 are compared to *S. tropica* CNB-440, respectively.

Looking more closely at differentially expressed genes between *S. tropica* and *S. arenicola*, 215 genes were differentially expressed during exponential phase. Of these genes, 91 were upregulated and 124 were downregulated in *S. tropica* versus *S. arenicola* (Appendix D). Expression during stationary phase showed that only 75 genes were differentially expressed. Of these genes, a similar trend showed fewer genes upregulated

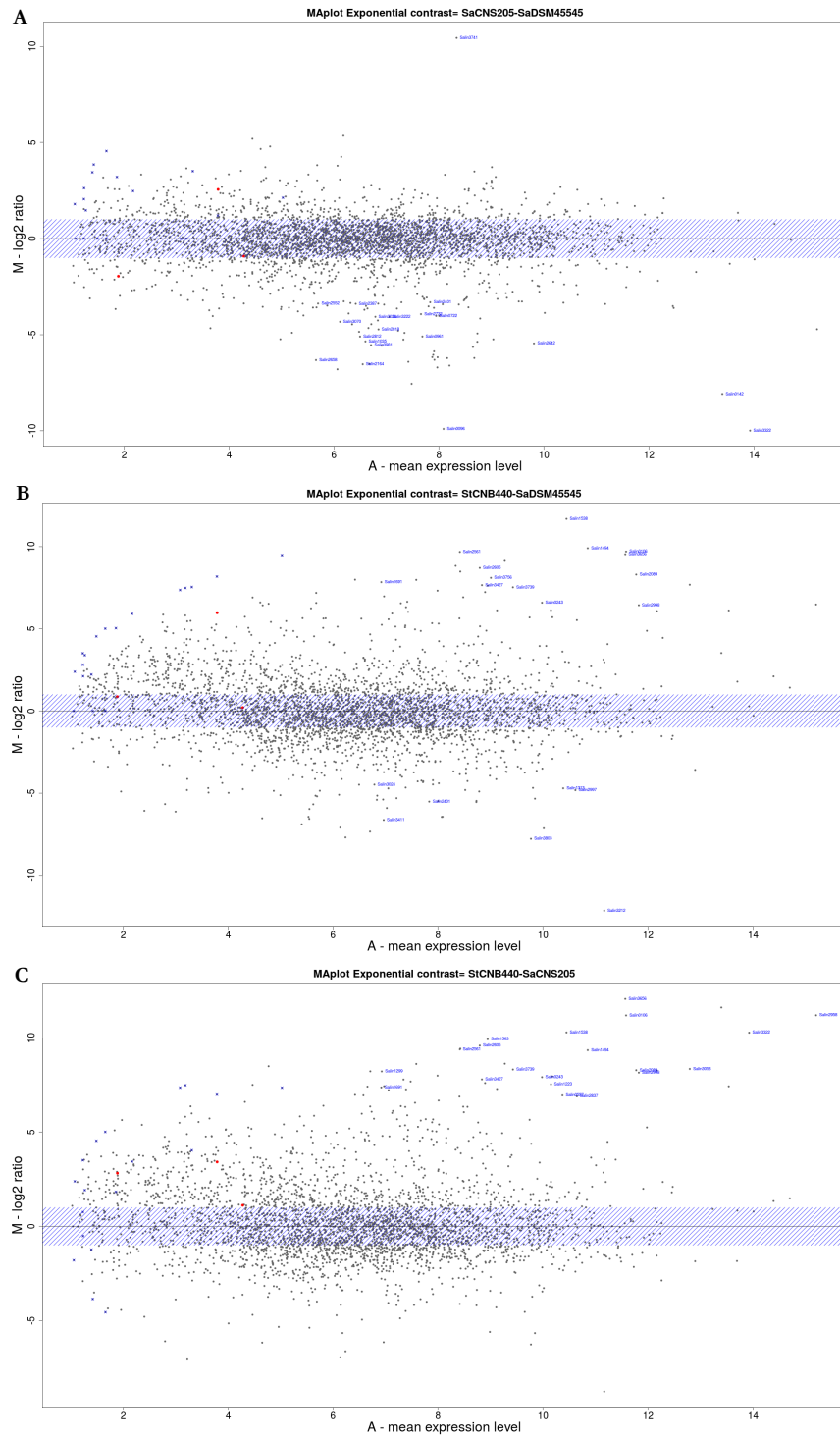


Figure 3.8: MAplot of log fold change during exponential growth for pairwise comparison of strains. Genes with similar expression levels in two samples will appear around the horizontal line $y = 0$. Panel A shows SaCNS205 vs SaDSM45545. Panel B shows StCNB440 vs SaDSM45545. Panel C shows StCNB440 vs SaCNS205.

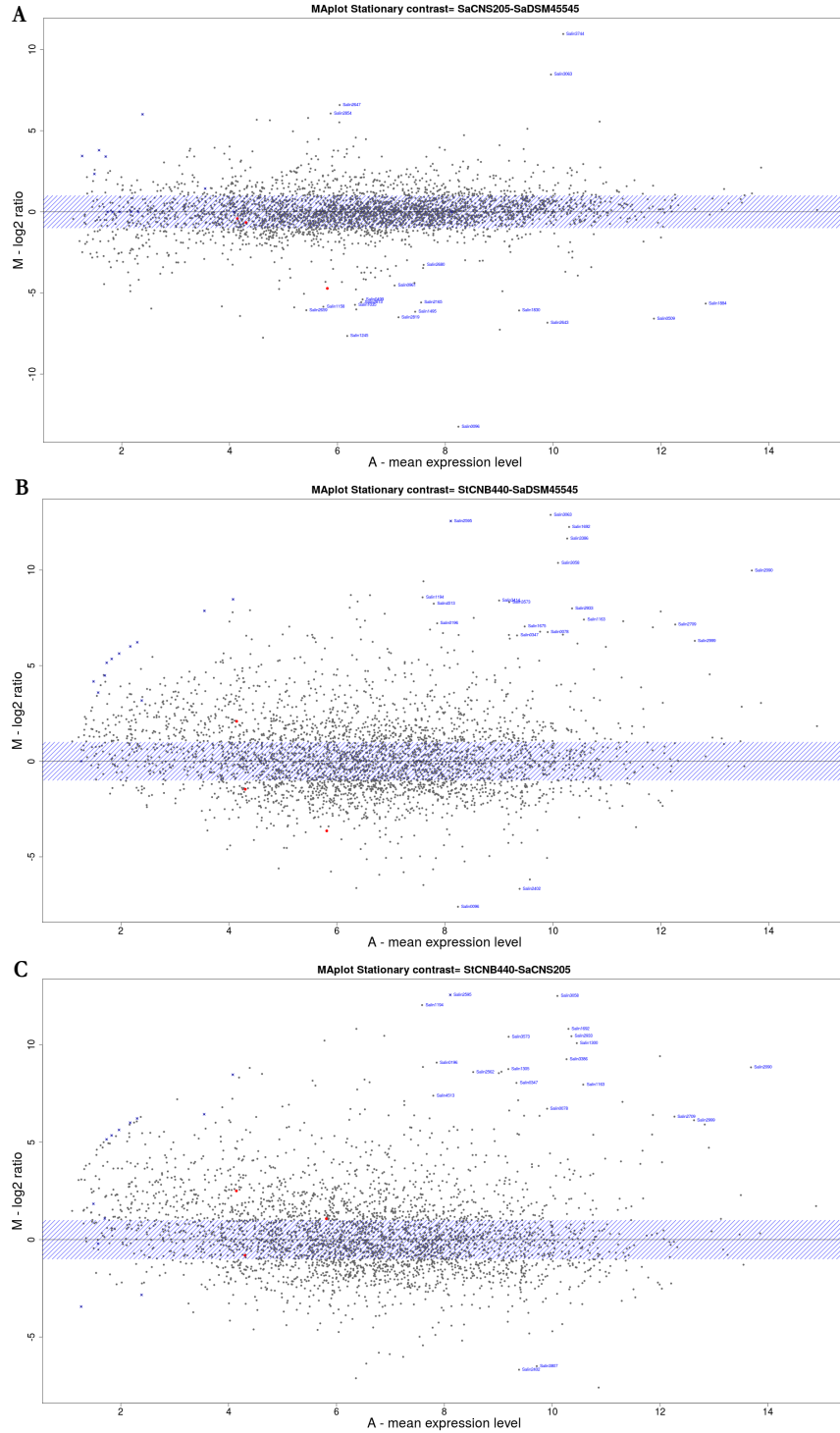


Figure 3.9: MAplot of log fold change during stationary growth for pairwise comparison of strains. Genes with similar expression levels in two samples will appear around the horizontal line $y = 0$. Panel **A** shows SaCNS205 vs SaDSM45545. Panel **B** shows StCNB440 vs SaDSM45545. Panel **C** shows StCNB440 vs SaCNS205.

Table 3.1: MA-Plot of number of genes found showing differential expression between each pair-wise comparison.

MA-Plot Comparison	-1 <# genes <1	-1 <% genes <1	-1 ># genes >1	-1 >% genes >1
Exponential				
SaCNS-205 vs SaDSM45545	2427	72	938	28
StCNB-440 vs SaDSM45545	1685	50	1680	50
StCNB-440 vs SaCNS-205	1765	52	1600	48
Stationary				
SaCNS-205 vs SaDSM45545	2393	71	972	29
StCNB-440 vs SaDSM45545	1340	40	2025	60
StCNB-440 vs SaCNS-205	1571	47	1794	53

(21) and downregulated (54) in *S. tropica* (Appendix E).

Due in large part to many of the differentially expressed gene products being function unknown, genes with annotations were given priority. Furthermore, solitary genes provide less insight into biological function, thus I prioritized annotated genes to determine if they were part of a KEGG pathway. Specifically, I searched for at least two genes that were differentially expressed and associated with the same KEGG pathway. In total, I identified 37 pathways where these criteria were met. Of these, 29 KEGG pathways and 56 genes showed evidence of differential expression between species during exponential phase. Twenty-one of these genes were involved in KEGG pathways upregulated in *S. tropica* and 35 in pathways in which they were downregulated (Figure 3.10). Pathways which had genes only upregulated during exponential phase in *S. tropica* were: carbon fixation pathways in prokaryotes, citrate cycle (TCA cycle), methane metabolism, oxidative phosphorylation, phenylalanine metabolism, propanoate metabolism, pyruvate metabolism, taurine and hypotaurine metabolism, and tyrosine metabolism. Pathways which had genes only downregulated during exponential phase in *S. tropica* were: 2-oxocarboxylic acid metabolism, biotin metabolism, cysteine and methionine metabolism, homologous recombination, monobactam biosynthesis, porphyrin and chlorophyll metabolism, and pyrimidine metabolism (Appendix F).

Downregulated					KEGG Pathway	Upregulated				
			Salin 0791	Salin 2010	2-Oxocarboxylic acid metabolism					
Salin 2428	Salin 0925	Salin 2088	Salin 1375		ABC transporters	Salin 2153				
			Salin 1815		Bacterial secretion system	Salin 0677				
Salin 0288	Salin 1396	Salin 0791	Salin 2010		Biosynthesis of amino acids	Salin 1276	Salin 2675			
		Salin 1535	Salin 1002		Biotin metabolism					
			Salin 1859		Butanoate metabolism	Salin 1526	Salin 1605	Salin 1302		
					Carbon fixation pathways in prokaryotes	Salin 0915	Salin 1988	Salin 1526	Salin 1605	Salin 1302
			Salin 1396		Carbon metabolism	Salin 0915	Salin 1526	Salin 2675	Salin 1605	Salin 1302
					Citrate cycle (TCA cycle)	Salin 1526	Salin 1605	Salin 1302		
Salin 1396	Salin 2010				Cysteine and methionine metabolism					
	Salin 1103				Galactose metabolism	Salin 1604				
Salin 0774	Salin 1616				Homologous recombination					
					Methane metabolism	Salin 0915	Salin 1988			
Salin 3172	Salin 2010				Monobactam biosynthesis					
	Salin 0288				Novobiocin biosynthesis	Salin 1276				
					Oxidative phosphorylation	Salin 1526	Salin 1605	Salin 1302		
Salin 0654	Salin 0791				Pantothenate and CoA biosynthesis	Salin 0850				
					Phenylalanine metabolism	Salin 1424	Salin 1276	Salin 0254		
		Salin 0288			Phenylalanine, tyrosine and tryptophan biosynthesis	Salin 1276				
Salin 2814	Salin 0431	Salin 1091			Porphyrin and chlorophyll metabolism					
					Propanoate metabolism	Salin 0915	Salin 1988			
		Salin 1815			Protein export	Salin 0677				
		Salin 1634			Purine metabolism	Salin 2675				
Salin 0679	Salin 0428	Salin 0512	Salin 1652	Salin 1338	Pyrimidine metabolism					
				Salin 2518						
					Pyruvate metabolism	Salin 0915	Salin 1988			
	Salin 1854	Salin 1396			Sulfur metabolism	Salin 3870				
					Taurine and hypotaurine metabolism	Salin 0915	Salin 1988			
		Salin 2107			Tryptophan metabolism	Salin 0254	Salin 2435			
					Tyrosine metabolism	Salin 1424	Salin 1276			

Figure 3.10: Differential expression of pathways in exponential phase growth with more than one gene differentially expressed between *S. tropica* and *S. arenicola*. Red cell color indicates genes down-regulated in *S. tropica* CNB-440. Green cell color indicates genes up-regulated in *S. tropica* CNB-440.

Downregulated				KEGG Pathway	Upregulated	
			Salin 0567	Alanine, aspartate and glutamate metabolism	Salin 0331	
Salin 1241	Salin 0575	Salin 0767	Salin 1377	Biosynthesis of amino acids	Salin 0331	
Salin 0793	Salin 2839	Salin 0767	Salin 1377	Carbon metabolism	Salin 1302	
			Salin 1246	Glycerophospholipid metabolism	Salin 2343	
Salin 0793	Salin 1246	Salin 0767	Salin 1377	Glycine, serine and threonine metabolism		
			Salin 0793	Glyoxylate and dicarboxylate metabolism	Salin 0392	
		Salin 0783	Salin 0527	Oxidative phosphorylation	Salin 1302	
				Phenylalanine metabolism	Salin 0254	Salin 1383
		Salin 1097	Salin 0699	Thiamine metabolism		

Figure 3.11: Differential expression of pathways in stationary phase growth with more than one gene differentially expressed between *S. tropica* and *S. arenicola*. Red cell color indicates genes down-regulated in *S. tropica* CNB-440. Green cell color indicates genes up-regulated in *S. tropica* CNB-440.

During stationary phase, 27 genes were differentially expressed in nine KEGG pathways. Eight genes were involved in pathways upregulated in *S. tropica* and 19 genes in pathways in which they were downregulated (Figure 3.11). The only pathway which had genes only upregulated during stationary phase in *S. tropica* were: phenylalanine metabolism. Pathways which had genes only downregulated during stationary phase in *S. tropica* were: glycine, serine and threonine metabolism and thiamine metabolism (Appendix F)

Three interesting KEGG pathways were found to have genes with species-specific differential expression during exponential phase growth. These pathways were ABC transporters, Homologous recombination, and Oxidative Phosphorylation. The ABC transporters pathway (Figure 3.12) consists of more than one type of transporter. Op-uBB is part of an osmotically regulated binding protein-dependent transport system, specifically, a periplasmic substrate-binding fusion protein (Schuster et al., 2016). Tran-

scription of the gene that encodes this protein is downregulated in *S. tropica* compared to *S. arenicola*. Similarly downregulated genes encode FhuB in the iron-complex transport system and ZnuA and ZnuC involved in the zinc transport system. BioY is encoded by a gene that is upregulated in *S. tropica* and is involved in biotin transmembrane transport. These products are notable because they are essential cofactors for enzymes in key metabolic pathways.

The homologous recombination pathway (Figure 3.13) shows downregulation in *S. tropica* for genes encoding RecO and PriA. RecO is a DNA double strand break repair and homologous recombination factor. PriA is a DNA replication priming protein required for homologous recombination and double strand break repair.

The oxidative phosphorylation pathway (Figure 3.14) shows expression levels of genes upregulated across almost the entire succinate dehydrogenase enzyme complex. With the exception of SdhD, genes encoding SdhA, SdhB, and SdhC were upregulated in *S. tropica*.

Interestingly only one gene involved in the succinate dehydrogenase complex, *sdhC*, remained upregulated in *S. tropica* during stationary phase (Figure 3.15). Two other genes, however, were downregulated. These were genes encoding NuoA, part of the inner membrane component of NADH dehydrogenase, and CoxB, a promoter involved in cytochrome c oxidase.

Three genes are downregulated in the glycine, serine, threonine metabolism pathway (Figure 3.16). These genes are EC 5.4.2.12 - phosphoglycerate mutase, CDP-diacylglycerol-serine O-phosphatidyltransferase 2.7.8.8, and EC 1.4.4.2 - glycine dehydrogenase (aminomethyl-transferring).

While there are many more pathways with genes that are differentially expressed, some do not demonstrate species specificity. Pathways with two or more genes showing differential expression between species can be referenced in Appendix F.

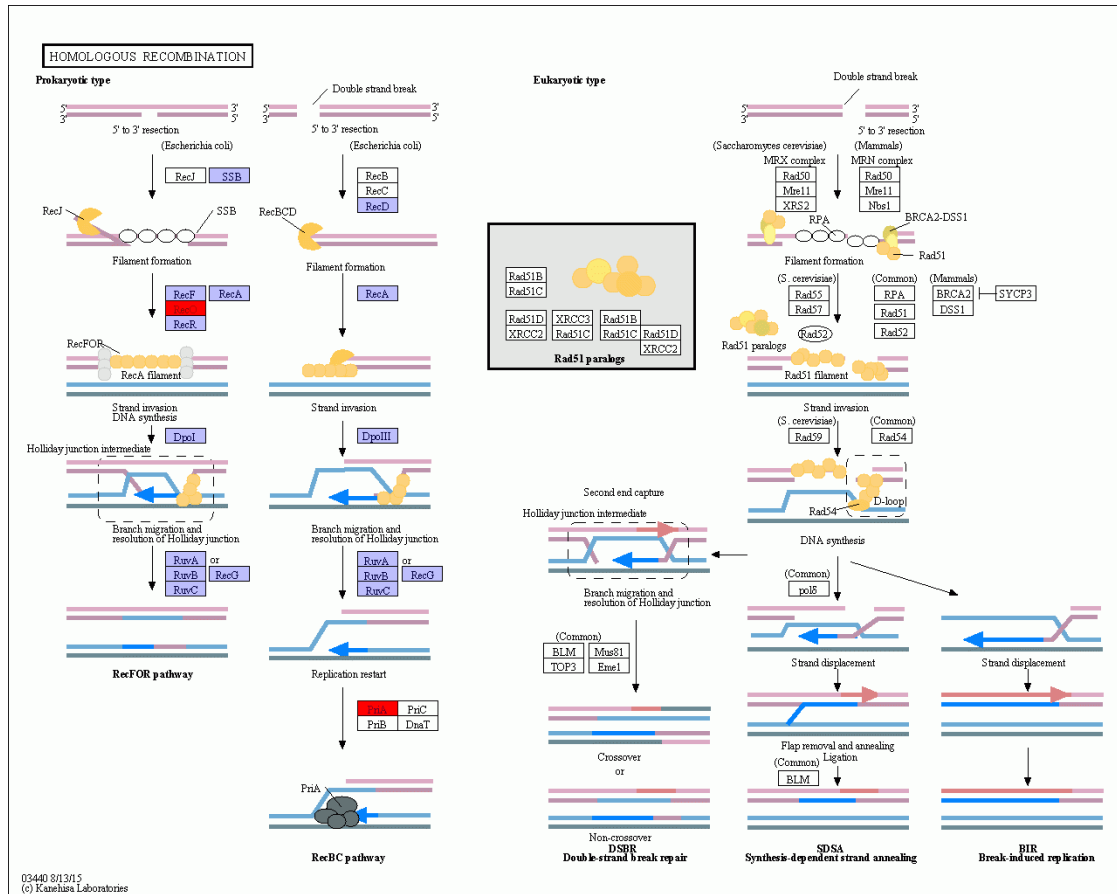


Figure 3.13: Differential expression of genes during exponential phase growth found in pathway: Homologous Recombination in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545. Red cell color indicates genes down-regulated in *S. tropica* CNB-440. Green cell color indicates genes up-regulated in *S. tropica* CNB-440. Yellow cell color indicates positional cluster genes. Purple cell color indicates other genes found in *S. tropica* CNB-440.

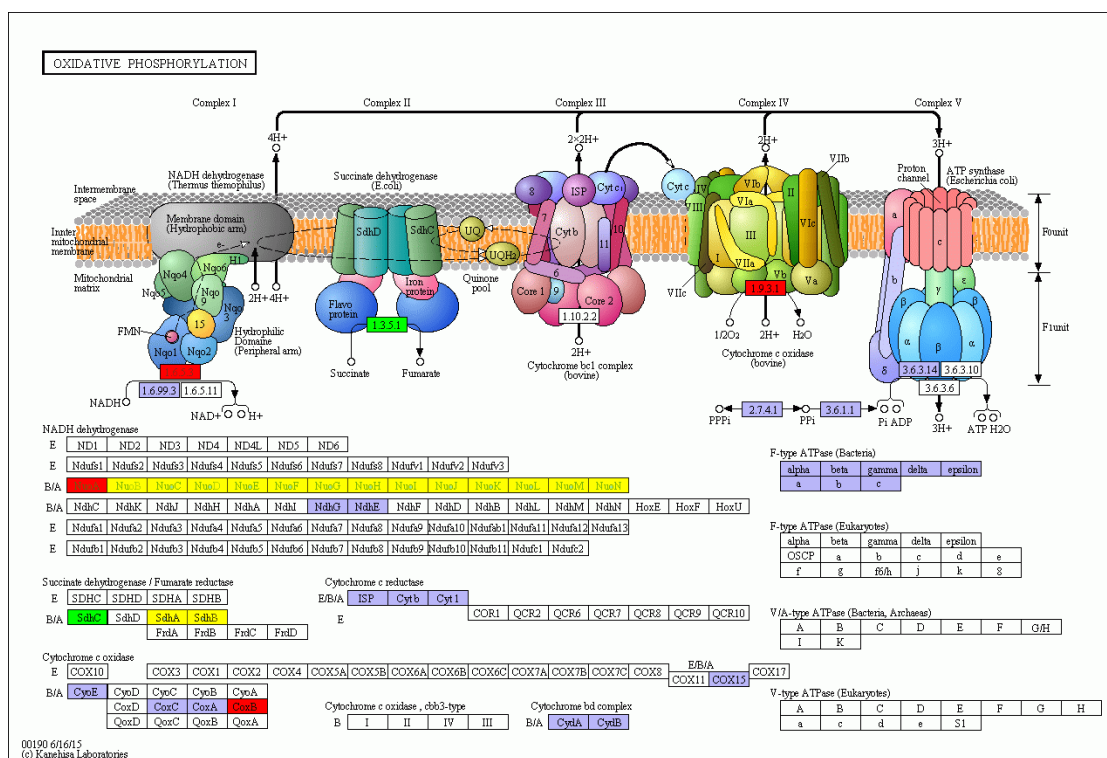


Figure 3.15: Differential expression of genes during stationary phase growth found in pathway: Oxidative Phosphorylation in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545. Red cell color indicates genes down-regulated in *S. tropica* CNB-440. Green cell color indicates genes up-regulated in *S. tropica* CNB-440. Yellow cell color indicates positional cluster genes. Purple cell color indicates other genes found in *S. tropica* CNB-440.

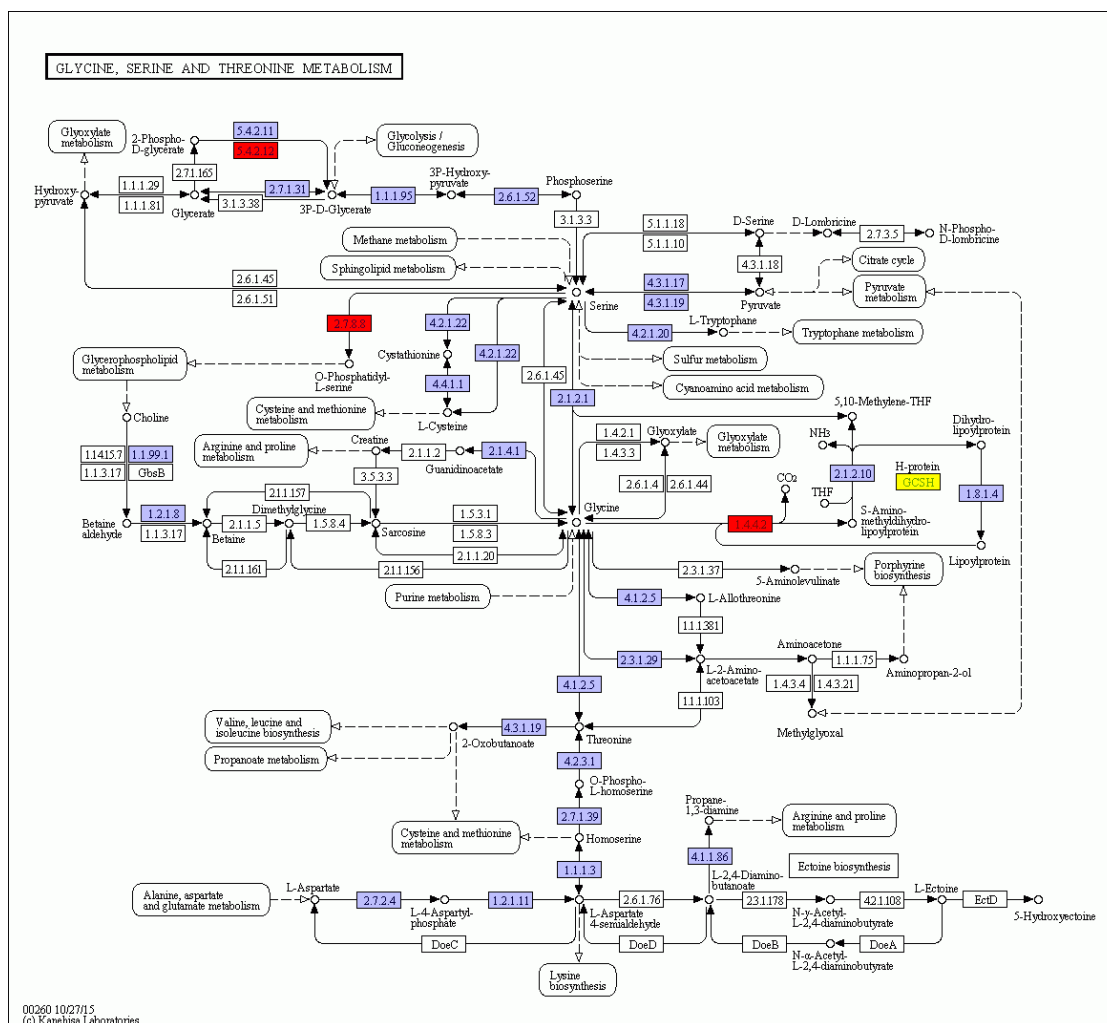


Figure 3.16: Differential expression of genes during stationary phase growth found in pathway: Glycine Serine Threonine Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

3.5 Discussion

Salinispora has been studied extensively for secondary metabolites (Jensen, 2016) and only recently have genomes become available for comparative genomic studies. Even more recently RNAseq data from transcriptomics studies have allowed researchers in our lab to better understand how genes are expressed. We have gained insight into secondary metabolism expression and shed light on which individual genes in biosynthetic gene clusters are switched on or off under standard laboratory conditions (Amos et al., 2017), however until now, no one has investigated the expression of primary metabolism genes in this genus. It is especially important to understand how species within a particular genus may be utilizing their genetic potential in differing ways, not just whether a gene is present or absent. This adds a layer of complexity to how we understand species concepts in the context of bacteria.

3.5.1 Chitinase

Looking at chitinase genes in particular has revealed an interesting example of how two species may share a particular gene yet express them at different levels. Four different chitinase genes were identified in at least two families. All four chitinase genes were present in all 12 *S. tropica* strains suggesting that *S. tropica* may have a greater capacity for chitin metabolism. This is supported by the sizes of the zones of clearing detected for each species. The ability to metabolize chitin has great implications as a carbon and nitrogen source given chitin is the most abundant polymer in the ocean and second most abundant on earth, exceeded only by cellulose (Aluwihare et al., 2005; Jeuniaux and Voss-Foucart, 1991). Based on their primary structures, chitinases can be allocated to the GH (Glycoside Hydrolase) families 18 and 19. Family 18 chitinases occur in viruses, bacteria, archaea and eukaryotes, while family 19 chitinases are mainly asso-

ciated with plants, though more than a decade ago they have been shown to be associated with bacteria, specifically, in *Streptomyces* species (Frederiksen et al., 2013; Itoh et al., 2002; Funkhouser and Aronson, 2007). One ortholog group (Salin984) was present in all 119 strains. Salin984 and Salin3372 belong to the family 18 chitinases. Family 18 chitinases are more common and are capable of degrading alpha-chitin. Salin5528, however, belongs to the rarer family 19 chitinases. Family 19 chitinases are able to degrade beta-chitin. The chitinase family of Salin4659 has not yet been determined. The presence of Salin5528 and Salin4659 show a seemingly random distribution, with some hints of geographic specificity (Salin5528 are found in all Palmyra strains) however this needs to be investigated more fully. The widespread occurrence of chitinase activity and the ubiquity of chitinase genes in marine bacteria (Cottrell and Kirchman, 2000) indicate that chitin-like biopolymers are important substrates in the marine environment. The increased abundance of chitinase genes in *S. tropica* may also facilitate the faster growth rate of *S. tropica* compared to *S. arenicola* (Patin et al., 2016) either through the ability to hydrolyze chitin as a nitrogen and carbon source or actively breaking down its own cell wall as cells rapidly multiply during exponential phase growth.

Several *S. arenicola* strains (denoted by grey cells) in Salin5528 are pseudogenes. Pseudogenes are commonly found in genomes and are homologous to functional genes but are unlikely to be functional due to genetic defects such as mutations or deletions and hint at likely removal from the genome sometime in the future (Holt et al., 2009). It has been reported that loss of function by pseudogenization may play a role in bacterial evolution (Tutar, 2012; Valenzuela et al., 2015; Ortega et al., 2016). There appears to be no clear pattern to pseudogenization in these *S. arenicola* strains and their imminent departure from the genome may have been a chance capture.

Despite no chitin present in the media during transcriptomic experiments (strains were grown in A1 media), chitinase genes showed various levels of expression, espe-

cially in stationary phase of *S. tropica*. It is possible this gene is constitutively transcribed in stationary phase rather than in response to outside cues or environmental factors. Sporulation may be a trigger initiating the degradation of n-acetylglucosamine, a chitin monomer found in the cell wall. This has been an observation made in fungi (van Munster et al., 2013) and may be a mechanism to support the faster growth rates observed in this species. While expression levels in *S. arenicola* were generally low, perhaps the sheer number of chitinase genes has created a compounding effect, whereby an evident zone of clearing indicating chitin metabolism is still notable.

Chitin metabolism also has ties to the production of secondary metabolites including antibiotics and anti-tumor agents in another actinomycete, the soil and marine-dwelling genus *Streptomyces*. This genus has been extensively studied and a transcriptional regulator in the GntR-family, DasR, has been shown to regulate antibiotic production, pigment biosynthesis, and morphological development (Liao et al., 2015). The *dasABC* gene cluster adjacent to *dasR* encodes a novel ABC transporter for the uptake of chitin in *Streptomyces coelicolor* A3(2). This expression is induced by a monomer of chitin, n-acetylglucosamine (NAG) (Saito et al., 2007). This monomer, in fact, has been linked to a major checkpoint for the onset of secondary metabolism. Rigali et al. (2008) proposed that a signaling cascade, from nutrient sensing to development, and antibiotic production, involves NAG. High concentrations of NAG is suggested to mimic the accumulation of NAG after autolytic degradation of the vegetative mycelium initiating a major checkpoint for the onset of secondary metabolism. A DasR homolog has been bioinformatically investigated yielding candidates outside the scope of this study, however the propensity of *Salinispora* to metabolize chitin could indicate greater implications in secondary metabolite production in this genus. The percentage of orphan gene clusters (defined as not having an assigned product) in *Salinispora* is incredibly high. Approximately 85% of the identified pathways are orphan. Of these, 50% are silent,

meaning that gene expression has not been detected. It is currently unknown what regulates these pathways, however studies of more closely related taxonomic groups may enable us to link expression of primary metabolite gene clusters and pathways to uncovering ways to unlock secondary metabolism potential.

3.5.2 Transcriptomics

Transcriptomics data revealed notable differences based on growth phase and species. Library sizes for each experimental time-point appear uniform with the exception of SaDSM45545 (exponential phase) due likely to the missing third replicate that did not pass QC during sequencing (Figure 3.5A). Nonetheless, the number of reads were fairly high given the scope of the study. Although there were only three strains to compare, the correlation matrix (Figure 3.6) shows that the two *S. arenicola* strains are more closely correlated to each other than either one was to *S. tropica*.

This species-specific pattern was resolved further at the gene level. The heatmap in Figure 3.7 shows genes with differential expression levels greater than two-fold and $p < 0.5$. Hierarchical clustering organizes genes but does not lead to clusters. A k-means clustering was used to find the number of clusters in the dendrogram. Unsurprisingly, an intraspecies pairwise comparison of both *S. arenicola* strains in both panels shows a heatmap with a less intense signal coloration than an interspecies comparison with *S. tropica*. Note brighter blue and orange intensities in the right two columns for both exponential and stationary phase indicating larger fold-change in expression.

Alternatively, a way to visualize similarities between intraspecific gene expression is using MA-plots. The closer the clustering of genes along the x-axis, the more similarly they are expressed. For both exponential and stationary growth phases, both *S. arenicola* comparisons show a tighter cluster of genes and were 20-30% more similar to each other than to *S. tropica* (Figure 3.8A and 3.9A). Genes that fell within the two-fold

differential expression range were used to attribute species-specific expression.

Forty-nine more genes were differentially expressed during exponential phase than stationary. This is not entirely surprising considering exponential phase growth involves more metabolic activity and growth. However, it has been accepted that stationary phase is the common phase for bacterial survival in the environment due to the ability of cells to grow once again when conditions become ideal (Kolter et al., 1993; Navarro Llorens et al., 2010). Long-term viability of bacteria depends on defensive measures against numerous stressors and includes strategies such as spore formation or entry into stationary phase (Ishihama, 1997). In *Vibrio parahaemolyticus*, stationary phase cells showed greater resistance to stressors such as thermal or oxidative stress than in exponential phase (Koga and Takumi, 1995; Koga et al., 1999). In *E. coli*, exponential phase gene expression studies have shown that ~1000 genes are inducibly expressed while they are considerably repressed in stationary cells. However a set of 50-100 genes was induced upon entering stationary phase (Ishihama, 1997). These were genes that were especially involved in stress response and long-term survival. Conversely, genes responsible for transcription and translation, nucleotide biosynthesis, aerobic metabolism and cell processes were down regulated (Chang et al., 2002). Reeve et al. (1984) found that upon entering stationary phase, *E. coli* growth rate decrease was accompanied by an 80% reduction in protein synthesis compared with exponential growth. This is thought to occur as cells reach a high cell density and with the onset of starvation as cell size and metabolism rates decrease (Kolter et al., 1993).

3.5.3 KEGG Pathways

The ABC transporters pathway (Figure 3.12) Opu transport system has been well studied in *Bacillus subtilis* and is part of an osmotically regulated binding protein-dependent transport system, specifically a periplasmic substrate-binding fusion protein

(Schuster et al., 2016). *B. subtilis* can take advantage of a wide spectrum of osmoprotectants via their import through their Opu transport system (Hahne et al., 2010; Du et al., 2011). The gene that encodes this protein is downregulated in *S. tropica* compared to *S. arenicola*, conversely this can be interpreted to mean that the gene is upregulated in *S. arenicola*. This could mean that *S. arenicola* experienced more salt or osmotic stress during exponential phase growth or *S. tropica* experienced less.

Another downregulated gene encodes FhuB in the ferric iron-complex transport system. Unsurprisingly, iron acquisition and storage systems are regulated in response to iron availability. This regulation is mediated by the homodimeric repressor protein, Fur, which employs ferrous iron as a co-repressor (Hantke, 2001). There is evidence that the Fe²⁺-Fur complex also represses genes (*cyoA*, *flbB*, *fumC*, *gpmA*, *metH*, *nohB*, *purR*, and *sodA*) involved in non-iron metabolic pathways including respiration, flagella chemotaxis, the TCA cycle, glycolysis, methionine biosynthesis, phage-DNA packaging, purine metabolism, and redox-stress resistance and consequently can be considered a global regulator (pathways can be found in Appendix F) (Stojiljkovic et al., 1994; Park and Gunsalus, 1995; Touati, 1988).

ZnuA is the periplasmic component of the zinc transporter ZnuABC which captures Zn(II) and delivers it to ZnuB. ZnuA plays a role in zinc homeostasis (Petrarca et al., 2010). ZnuABC is activated in several bacteria in response to zinc deficiency. ZnuC is the membrane permease involved in the zinc transport system (Ammendola et al., 2007; Campoy et al., 2002). The downregulation of these genes related to zinc transport in *S. tropica* may indicate that it is not as crucial to its growth as it is for *S. arenicola*. BioY is encoded by a gene that is upregulated in *S. tropica* and is involved in biotin transmembrane transport. Biotin is a water-soluble b-vitamin also called vitamin B7 and may be required by *S. tropica* for growth.

The homologous recombination pathway is mediated by *rec* genes (Figure 3.13)

and shows downregulation in *S. tropica* for genes encoding RecO. RecO is a DNA double strand break repair and homologous recombination factor (Redfield, 2001; Vos, 2009). Likewise, the gene encoding PriA is downregulated and is a DNA replication priming protein required for homologous recombination and double strand break repair (Kogoma et al., 1996). The downregulation of these genes in *S. tropica* may contribute to less incorporation of lateral gene transfer and the clonal nature of the species.

The expression levels of genes associated with the succinate dehydrogenase enzyme complex of the oxidative phosphorylation pathway (Figure 3.14) were upregulated. With the exception of *sdhD*, genes encoding *SdhA*, *SdhB*, and *SdhC* were upregulated in *S. tropica*. *Sdh* is the only enzyme that participates in both the citric acid cycle and the electron transport chain. Succinate dehydrogenase (EC 1.3.99.1) is part of the nonoxidative branch of the TCA cycle and is directly linked to the respiratory chain. This enzyme complex catalyzes the oxidation of succinate to fumarate, donating FADH₂ for oxidative phosphorylation. It consists of three subunits: membrane-bound cytochrome b₅₅₈ (*SdhC*), a flavoprotein containing an FAD binding site (*SdhA*), and an iron-sulfur protein showing a binding region signature of the 4Fe-4S type (*SdhB*) (Hederstedt and Rutberg, 1980, 1981). The upregulation of these genes has proven to be advantageous in *Staphylococcus aureus* under biofilm conditions when nutrients and oxygen may be limiting (Gaupp et al., 2010). Although *Salinispora* does not produce biofilms, *S. tropica* does grow faster than *S. arenicola* potentially producing microenvironments of low nutrients and oxygen. Upregulating the succinate dehydrogenase genes might enhance survival under less than ideal conditions.

Interestingly, only one gene involved in the succinate dehydrogenase complex, *sdhC*, remained upregulated in *S. tropica* during stationary phase (Figure 3.15). Two other genes, however, were downregulated. One of these genes encodes NuoA, part of the inner membrane component of NADH dehydrogenase. NuoA is one of 14 different

protein subunits also known as complex I. Complex I of the respiratory chain connects energy currencies by using NADH produced during nutrient breakdown to generate a proton motive force, which is subsequently used for ATP synthesis (Spero et al., 2015). The other downregulated gene encodes CoxB, a promoter involved in cytochrome c oxidase. The *coxB* promoter depends on a stationary phase sigma factor, RpoS (Osamura et al., 2017). These genes are downregulated in *S. tropica* meaning that they are conversely upregulated in *S. arenicola*. Upregulation of genes necessary for energy acquisition may mean that *S. arenicola* has greater energetic needs during stationary phase, potentially for pathways involved in secondary metabolite production.

Three genes were downregulated in *S. tropica* in the glycine, serine, threonine metabolism pathway (Figure 3.16). These genes are EC 5.4.2.12 - phosphoglycerate mutase, EC 2.7.8.8 - CDP-diacylglycerol-serine O-phosphatidyltransferase, and EC 1.4.4.2 - glycine dehydrogenase (aminomethyl-transferring). One study showed that glucose acted as a potentiator for activity of kanamycin in order to effectively kill the multidrug resistant pathogenic bacterium *Edwardsiella piscicida* through the activation of the TCA cycle (Ye et al., 2018). Glucose significantly altered eight amino acids, however glycine, serine and threonine showed the strongest efficacy. Furthermore, succinate dehydrogenase activity increased as well as proton motive force (PMF). Inhibitors that disrupted PMF also abolished potentiation (Ye et al., 2018). The upregulation of genes in the glycine, serine, threonine metabolism pathway in *S. arenicola* may be indicative of a pathway that could be harnessed in the future for antibiotic potentiation in the species starting with the addition of glucose to the media in which it is cultured.

In summary, the species-specific traits I identified can be broken down as follows. Experimental evidence has shown that chitinase gene number and expression levels are increased in *S. tropica* when compared to *S. arenicola*. Based on bioinformatic predictions, *S. tropica* has the ability to better cope with osmotic stress. Iron and

zinc are less crucial for *S. tropica* growth; however, biotin may be critical for growth. *S. tropica* undergoes less homologous recombination and this may contribute to its clonal nature. *S. tropica* may be more affected by nutrient or oxidative stress. *S. arenicola* has more energetic needs during stationary phase potentially because of costly secondary metabolite production and has shown upregulation of pathways that have implications for the potentiation of antibiotics.

Linking species-specific expression to genotype adds a layer of complexity to species definitions, yet provides valuable insights into how bacterial species differ from each other. Although this study was limited by the number of strains analyzed, species-specific signals were detected, and it was a worthwhile endeavor to explore more fully how *Salinispora tropica* and *S. arenicola* differed from one another. This study could be enhanced further still by increasing the sample size and adding more RNAseq data from representatives from each species. By the same token, varying growth conditions to better mimic environmental conditions would add more ecological relevance to this study. Another limitation, which remains in almost all bioinformatic studies, is that many genes in a genome have yet to be experimentally characterized leaving their physiological roles unknown. In chapter 2, the goal was to look for a genetic basis for species delineation between *S. arenicola* and *S. tropica*. The nature of that study relied heavily on an orthologous gene comparison that would concede a less than ideal level of resolution for a genus with species so closely related to one another. Because of this, having transcriptome data in hand becomes a powerful resource in diving below the surface of conventional comparative genomics and brings us one step closer to understanding microbial species and populations.

Chapter 3 is coauthored with Amos GCA, and PR Jensen. The dissertation author was the primary investigator and author of this chapter.

Chapter 4

Lateral Gene Transfer Dynamics and the *Salinispora* Molecular Clock

4.1 Abstract

Determining how bacterial diversity is created is challenging especially due to the surprising mechanisms by which genes are acquired. Lateral gene transfer (LGT) is a means by which foreign DNA can become incorporated into a bacterial chromosome. The goal of this chapter is to examine how LGT has affected the diversity of the genus *Salinispora* by querying the pangenome for evidence of foreign genetic elements. Analysis shows that the largest proportion of LGT events occurred within the phylum Actinobacteria and that 75% of those genes originated from only four genera. Many of these genes are annotated to encode mobile genetic elements, ABC transporters, and secondary metabolites. For the first time a molecular clock is also presented for the genus, providing a new temporal framework with which to understand how genetic information moves across species and strains both at the gene and biosynthetic cluster level.

4.2 Introduction

Bacterial diversity formed over billions of years of evolution, proliferating into the furthest imaginable reaches on Earth (Hoehler and Jørgensen, 2013) and occupying every possible metabolic niche (Rinke et al., 2013). They are the pillars of life that have engineered the foundation of our environment. (Gibbons and Gilbert, 2015). Their role on this planet is not insignificant and many are major ecological players in their respective environments (Cohan and Koeppel, 2008). Bacteria have been known to be major contributors in the marine environment and the ecological roles have been well characterized (Cohan and Koeppel, 2008; Azam et al., 1983; Fraser et al., 2009). Bacterial evolution over such a long expanse of time has undoubtedly resulted in an extraordinarily complex history. Determining how bacteria have created so much diversity is a challenging endeavor complicated by the surprising means by which their genomes are able to acquire genes.

Until the 1940s, bacteria were believed to be clonal and incapable of exchanging genetic information because they did not reproduce sexually (Bobay et al., 2015). Instead, scientists discovered that genetic exchange was an underexplored driver of bacterial evolution (Daubin and Szöllősi, 2016). Various mechanisms by which their ability to gain genetic information began with the discovery of conjugative plasmids (Lederberg and Tatum, 1946), then transformation (Thomas and Nielsen, 2005; Avery, 1944) and transduction (Zinder and Lederberg, 1952). After the discovery of antibiotic resistance (Davies, 1994) and virulence phenotypes (Ochman et al., 2000), it became evident that non-homologous lateral gene transfer (LGT) was indeed a major driver of bacterial evolution.

In the previous two chapters, I have attempted to show that comparative genomics allows us to identify gene content differences between related bacteria. Many

of these differences can be attributed to LGT and this mechanism should also be examined when studying bacterial evolution. On larger scales, gene transfers occur more often within phyla than between (Beiko et al., 2005). And in some cases, certain phyla are more biased towards particular groups (Andam and Gogarten, 2011). Upon closer inspection at the population level, genomes exhibiting greater dissimilarity have shown reduced rates of recombination (Whitaker et al., 2003; Cadillo-Quiroz et al., 2012; Lerat et al., 2005). This widespread exchange of genetic information has led to the suggestion that bacterial species do not actually exist as discernibly distinct units (Boucher et al., 2003). Bacterial evolutionary histories are suggested to look more like a web of life rather than a tree. Upon closer inspection at the population level, genomes exhibiting greater dissimilarity have shown reduced rates of recombination (Gogarten et al., 2002; Soucy et al., 2015; Ge et al., 2005; Retchless and Lawrence, 2010).

The goal of this chapter is to take a broader view of the evolutionary history of the genus *Salinispora*. *Salinispora* does not exist in isolation, but in complex microbial communities that create opportunities for genetic exchange. Many genes have been acquired from neighboring bacteria via lateral gene transfer at some point in their evolutionary history. The aim is to identify genes that were acquired and subsequently maintained, suggesting they confer a selective advantage. Additionally, putting these transfer events onto a phylogeny with a molecular clock provides new perspective on when speciation events occurred and the potential genetic drivers. This dissertation has identified the entire suite of genes that have thus far been sequenced, the pangenome, and focused primarily on the genes that have the potential to differentiate species. These analyses targeted genes that were either species-specific or shared genes that were differentially expressed between species.

Here I extend these analyses by identifying genes that were observed in a single genome. This group, singleton genes, are unique to individual genomes and is ana-

lyzed here. I also mapped biosynthetic gene cluster acquisition events onto the species phylogeny and dated these events using a molecular clock. A recent study of the actinobacterial genus *Streptomyces* investigated how LGT shaped the evolution of this ubiquitous and medically important taxon (McDonald and Currie, 2017). Using a molecular clock, they estimated that the genus *Streptomyces* is ~ 380 million years old. They determined that the acquisition and retention of genes through LGT was quite rare in this lineage. McDonald and Currie also noted that in contrast to *Salinispora*, most biosynthesis clusters were composed of genes from multiple sources rather than a single full-operon transfer event. They suggest that *Streptomyces* should not actually be considered 'closely related' despite being categorized as a genus. Therefore, their analysis only provides insight into LGT dynamics at the intermediate-scale, something more akin to a bacterial family. Ergo, in order to investigate these processes at the population-scale, closely related bacterial strains should be sampled.

In *Salinispora*, many criteria that make investigating population-scale processes more meaningful are met. Unlike the >550 species in the genus *Streptomyces* that span soil, sediment, and seawater environments, *Salinispora* consists of three closely related species (99% 16S rRNA sequence identity) and are found solely in the marine environment associated with sediments. Not only does this make the genus an ideal candidate for population-scale investigation, but the availability of 118 genomes across all species lends itself to a robust dataset. The potential for more named species (Millán-Aguinaga et al., 2017) means we can look at LGT across physiologically similar organisms that we know are diverging. *Salinispora*'s secondary metabolite biosynthetic gene clusters have been well characterized (Jensen et al., 2015a; Jensen, 2016; Letzel et al., 2017) and creating a molecular clock for the genus allows patterns over evolutionary time scales to be quantified. Additionally, identifying the functional annotation of laterally transferred genes provides insight into which types of genes are in some way beneficial to the strain,

enough to retain it in its genome.

4.3 Methods

4.3.1 Genome Annotation

Whole genomes were sequenced, assembled, and annotated according to the methods described in Chapter 2. One genome, *S. arenicola* CNY-281, was determined to be contaminated and removed from the dataset for the analyses described in this chapter.

4.3.2 Molecular Clock Analyses

Two different phylogenies were generated using Reltime in the Mega-CC 7 package distribution to approximate divergence times (Tamura et al., 2012; Kumar et al., 2016). The first was a tree of 55 taxa representing the bacterial tree of life and included 20 strains representing the three *Salinispora* species and all 16S sequence types identified to date (Figure 4.2), genera representing the *Salinispora* closest phylogenetic neighbors, as well as taxa used as calibration points. The second was a *Salinispora* phylogeny (Figure 4.3) containing the full set of 118 genomes.

Both multilocus phylogenies were generated using TIGRFAM annotated proteins. The 109 full TIGRFAM proteins in the core bacterial protein set GenProp0799 were used as the molecular clock data set (Appendix G). This set of genes includes those which are generally found exactly one to a bacterial genome and tend to exhibit little to no lateral gene transfer events. The protein sequences with the top HMMER bitscore for each protein family in each genome were concatenated and aligned using the MUSCLE plug-in (Edgar, 2004) in Geneious (v. 5.5.8 <https://www.geneious.com>). Protest3

(v. 3.4.2) (Darriba et al., 2011) was used to find the best-fit model for protein evolution for the bacterial tree of life phylogeny. Amino acid sequences were used instead of nucleotide sequences due to the complications aligning genes from taxonomically diverse genomes. A phylogeny was then generated with RAxML (Stamatakis, 2014) using the PROTGAMMABLOSUM62 substitution model and 100 rapid bootstraps of the final alignment. The *Salinispora* species phylogeny incorporated 118 genome sequences. As an outgroup, *Micromonospora aurantica* was used rather than *Verrucosispora maris* because of the availability of gene sequences with homology to GenProp0799. Nucleotide sequences of all 109 genes from GenProp0799 were concatenated (117,671 bp) and aligned for all 119 strains including the outgroup, *Micromonospora*. jModelTest (v. 2.1.10) (Darriba et al., 2012; Guindon and Gascuel, 2003) was used to determine the appropriate nucleotide substitution model. A phylogeny was then generated with RAxML using the GTRgamma substitution model and 100 rapid bootstraps of the final alignment.

The Time Tree Tool in Reltime can be used for calculating relative and absolute divergence times for all branching points in the tree. Because there is no global calibration rate, clocks for any given data set must be calibrated. The Many Clocks algorithm was used in Reltime and the analysis was set to Estimate Divergence Times (ML). Three calibration points were used as approximate time intervals for the evolution of Cyanobacteria (2,500 to 3,500 million years ago) (Brocks et al., 1999; Garvin et al., 2009), the divergence of *Salmonella* and *Escherichia* (50 to 150 million years ago) (Ochman and Wilson, 1987), and the origin of bacteria (3,500 to 3,800 million years ago) (Mojzsis et al., 1996; Rosing, 1999). The RelTime algorithm touts the ability to use a single calibration point in order to correctly infer the divergence rate of the other taxa with an error rate of less than 20%. The confidence intervals for the divergence of *S. arenicola* from *S. tropica* and *S. pacifica* were used to calibrate the molecular clock

analysis of the *Salinispora*-specific phylogeny.

4.3.3 Biosynthetic Gene Cluster Likelihood Analysis

Biosynthetic gene clusters (BGCs) identified from *Salinispora* strains (Letzel et al., 2017) were incorporated onto the species time tree. A likelihood analysis of gains and losses of BGCs was run for all clusters found in >10 strains even if an associated compound has not yet been identified. Mesquite (v 3.40, <http://www.mesquiteproject.org>) was used to conduct the likelihood analysis using a presence/absence matrix of BGCs (Appendix H) and mapping onto the *Salinispora* time tree. The presence/absence matrix was supplied as standard categorical data. The last row of the matrix provides the number of steps determined by Mesquite for each BGC. A trace character history analysis was run for each BGC as 'parsimony ancestral states'. Each gain/loss event is denoted with arrows at a node to represent a recent common ancestor, or at an individual strain.

4.3.4 Lateral Gene Transfer Analysis

A genome-wide prediction of horizontal gene transfer was conducted for all 118 genomes using DarkHorse (Podell and Gaasterland, 2007). DarkHorse uses an automated pipeline that rapidly identifies and ranks phylogenetically atypical proteins and selects potential ortholog matches from a reference database of amino acid sequences. The analysis identifies the taxonomy of the closest Genbank nr match to each protein, excluding all matches to the genus *Salinispora* (Genbank database version October 2015). For each individual protein, only matches with a bitscore within 10% of the closest non-self hit were considered for assigning taxonomy (DarkHorse program filter threshold setting of 0.1). The BLAST cutoff threshold was an e-value of 1e-5, and alignment length >70% of both query and subject sequences. The pipeline employs a

Lineage Probability Index (LPI) that is inversely proportional to the phylogenetic distance between database match sequences and the query genome. Match organisms at similar phylogenetic distances receive similar LPI scores regardless of their database abundance. This feature is helpful in compensating for database bias in number of sequences associated with different taxonomic groups. A phylogenetic tree was generated using 16S sequences for each Actinobacteria hit using the Integrated Microbial Genomes and Microbiomes tree generator (<https://img.jgi.doe.gov/>)

4.4 Results

4.4.1 Molecular Clock Results

A bacterial tree of life phylogeny was generated using GenProp0799 gene set (Figure 4.2). Calibration points were used as approximate time intervals for the evolution of Cyanobacteria (2,500 to 3,500 million years ago), *Salmonella* and *Escherichia coli* (50 to 150 million years ago), and the origin of bacteria (3,500 to 3,800 million years ago). Twenty representative *Salinispora* strains are shown within colored boxes. Eleven strains represent *S. arenicola*, two strains represent *S. tropica*, and seven strains come from *S. pacifica* (Figure 4.1). These strains broadly cover the *Salinispora* phylogenetic tree and represent various sequence types for each species as well as sampling sites.

Based on the molecular clock analysis, *Salinispora* diverged much more recently than the *Streptomyces* lineage. *Salinispora* appears to have diverged from *Verrucosispora maris* approximately 68.68 million years ago (mya). This occurred roughly near the time of the Cretaceous-Tertiary (K-T) extinction event leading to the sudden mass extinction of three-quarters of plant and animals species approximately 66 mya (Renne et al., 2013). *S. arenicola* diverged from sister species *S. tropica* and *S. pacifica* ap-

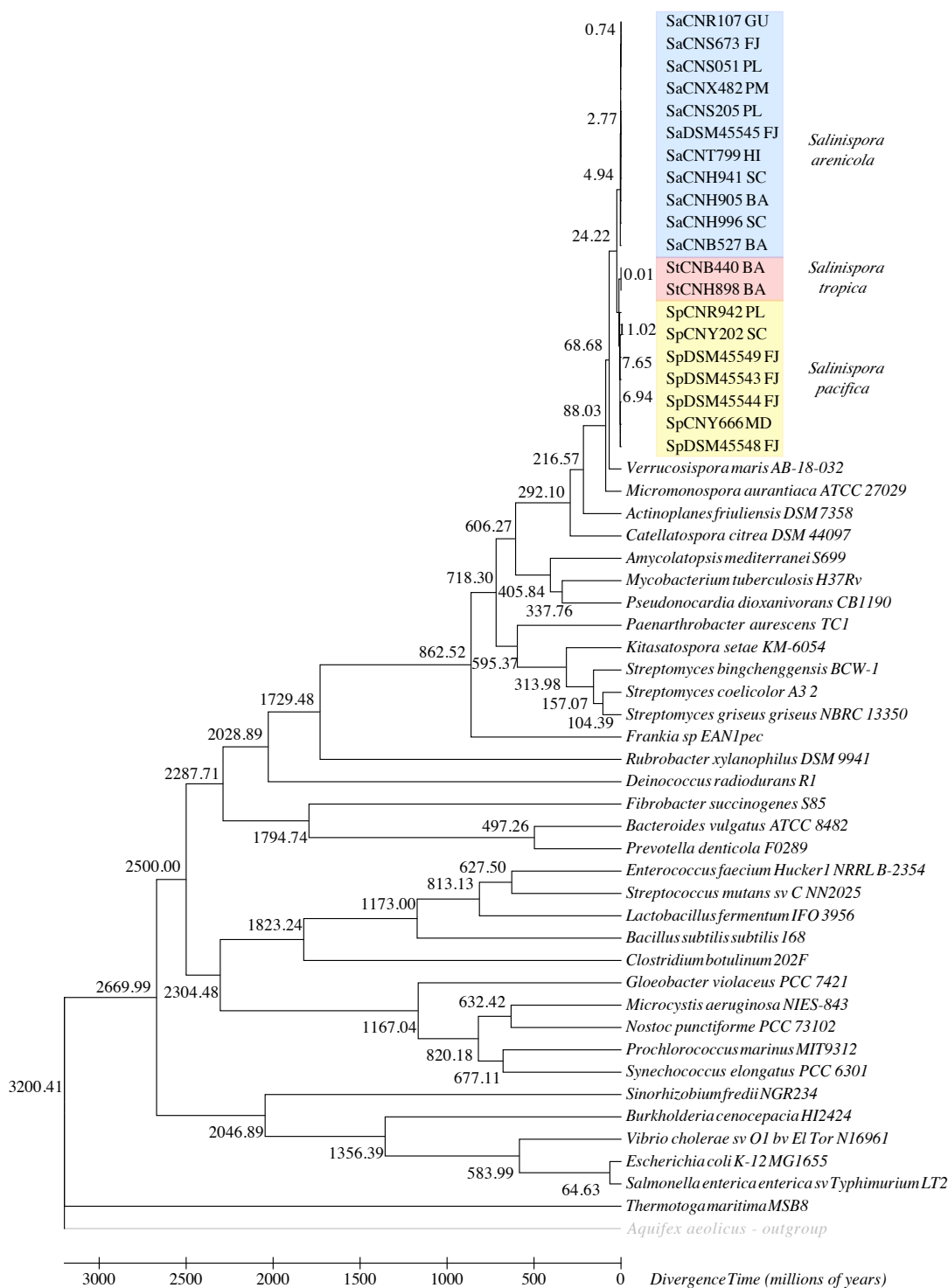


Figure 4.2: Bacterial tree of life phylogeny generated using GenProp0799 gene set. Calibration points were used as approximate time intervals for the evolution of Cyanobacteria (2,500 to 3,500 million years ago), Salmonella and Escherichia (50 to 150 million years ago), and the origin of bacteria (3,500 to 3,800 million years ago). Twenty representative *Salinispora* strains are shown in colored boxes. Numbers at nodes denote estimated time of divergence according to RelTime analysis.

proximately 24.22 mya. This corresponds to the Oligocene-Miocene transition (OMT) approximately 23 mya and is associated with a transient global cooling event (Beddow et al., 2016). The sister-species *S. tropica* and *S. pacifica* diverged about 11.02 mya corresponding to the mid-Miocene. Diversification within *S. arenicola* appears to have occurred more recently in evolutionary history, starting approximately 4.94 mya.

A more refined *Salinispora* molecular clock was generated by calibrating at the node where *S. arenicola* diverged from *S. tropica* and *S. pacifica* (24 mya). This clock predicts that *S. tropica* and *S. pacifica* diverged 12.2 mya and that diversification within *S. arenicola* initiated 5.3 mya (Figure 4.3). Thus, it remains unknown why in the course of *Salinispora* evolution *S. arenicola*, which diverged from *S. tropica* and *S. pacifica* 24 mya, only began to diverge relatively recently. Possible explanations include a selective sweep in *S. arenicola* and biases associated with strain cultivation.

This genus-level time tree produces similar divergence time estimates as the dates for the bacterial tree of life. Each clade associated with a named species is shown in a different color. *S. pacifica* strain names are colored based on suggested ANI species designations (Millán-Aguíñaga et al., 2017). In total 48 BGCs were identified in >10 *Salinispora* strains. The likelihood analysis identified the most parsimonious gain and loss events for each BGC and these results are denoted with arrows pointing to various points of the tree either at a node or one particular strain. Some of these BGCs have been characterized and have a known product. Those with no known product are classified by the class of biosynthetic enzyme or type of compound they are predicted to encode. These include PKS (polyketide synthase), NRPS (nonribosomal peptide synthetase), Terp (terpene), Bac (bacteriocin). This schematic is not comprehensive for the 178 BGCs identified by (Letzel et al., 2017), but instead includes the 48 most abundant and well-characterized BGCs. The complete species tree with all gain and loss events for every strain can be found in Figure 4.3.

The *mscL* gene was lost in all *Salinispora* but present in closely related genera including *Micromonospora*, *Verrucosispora*, and *Actinoplanes* which have also been isolated from marine sediment. The loss of this gene was previously proposed as a marine adaptation based on two genomes (Penn and Jensen, 2012; Bucarey et al., 2012). These results provide strong support for the loss of this gene at the genus level.

Four BGCs were either inherited from a common *Salinispora* ancestor or were acquired at the time the genus split from *Verrucosispora* some 68.7 mya and remain conserved in all 119 genome sequences. These BGCs are sioxanthin (*sio*), responsible for the orange cell pigment, and PKS4, Bac2, and aminocyclitol, all of which have yet to be linked to their products. There are clearly strong but yet to be defined selective pressures maintaining these BGCs. Five additional BGCs were also inherited or acquired around this time yet show evidence of more recent loss events in some strains. These BGCs encode desferrioxamine (*des*), lymphostin (*lym*), and salinipostin, while NRPS4 and Sid 1/2 remain uncharacterized. Thirteen BGCs were acquired by *S. arenicola* around the time it diverged (24 mya). They encode the biosynthesis of the potent antibiotic rifamycin (*rif*), the potent cytotoxic staurosporine (*sta*), along with 11 uncharacterized BGCs. Four of these BGCs (*sta*, Lan2, NRPS1, and PKS2) have also been observed in *S. pacifica*, with the likelihood analyses predicting they were acquired independently (with *S. arenicola* as a potential source). Interestingly, *S. pacifica* has never been shown to produce staurosporine despite 16 strains having the *sta* cluster. PKS1A and PKS5 clusters show multiple loss events within the *S. arenicola* clade. The Palmyra specific *S. arenicola* clade (PM) acquired the *slc*, *sid5*, and PKSNRPS2 BGCs while losing *des* and PKS5. Additionally, a larger *S. arenicola* clade with strains representing many geographic locations (Red Sea, Bahamas, Yucatan, Sea of Cortez, Fiji, and Hawaii), acquired PKS7, PKS15, PKS16, and NRPS14 while losing PKS1C. The Sea of Cortez specific clade acquired the *sal*, Lan1 and Lan9 BGCs while losing PKS5.

S. tropica and *S. pacifica* are sister species and share two BGCs, *lom* and Bac4, that were introduced into a recent common ancestor and largely maintained among strains belonging to both species. Six additional BGCs were acquired by *S. tropica* around the time it diversified from *S. pacifica*. These were *slm*, *sal*, NRPS3, *sid3*, *sid4*, and Lan9. Of the 12 *S. tropica* strains sequenced, 10 strains acquired the *spo* BGC and two strains acquired *cya*. Lan9 was lost in eight strains. *S. pacifica* has shown more diversity at the genome level than the other two species and this is also reflected in the gain and loss of BGCs in the clade. The likelihood analysis predicts that a modified version of the *sal* pathway was acquired seven times by *S. pacifica* and shown to produce salinosporamide K (Freel et al., 2011; Eustáquio et al., 2010). A recent *S. pacifica* diversity study has identified seven different clades within the species, and these clades are numbered and colored by strain in Figure 4.3.

The divergence of *S. pacifica* from *S. tropica* is associated with the gain of NRPS20 and the loss of *sid1/2* in *S. pacifica*. Within this species, additional clade-specific gains and losses were also observed. Clade 2 acquired PKS2, Lan2, PKS16, *sid3*, and betalactam but lost Bac4 and salinipostin. Clade 3 and 4 gained the *sta* and *cya* clusters. Clade 4 also gained PKS25 but lost *lom* and NRPS20. Clade 5 acquired *sta*, *cya*, and PKS25 and lost *lom*. Within this clade, five strains have lost the *lym* BGC while gaining *sal*, betalactame, and NRPS2. Clade 6, which is represented by strain CNY-666, shows the acquisition of NRPS19 and NRPS27 and the loss of *des*. Clade 7 is notable for a large number of BGC acquisitions in strain CNS-055. These BGCs are salinipostin, NRPS1, NRPS4, PKS16, PKS19, *sid3*, and *sid4*. Clade 8 shows a gain of the PKS16 while subclades show the acquisition of *cya*, *slc*, *sid5*, *terp6*, PKS19, NRPS19 and the loss of *des*. One subclade shows the gain of the *spo*, *sal*, and betalactam BGC and the loss of NRPS20. Individual strains likewise show gains and losses of BGCs but will not be addressed in this study.

4.4.2 DarkHorse Results

In total 4,980 proteins were identified by DarkHorse. Of these, 2,918 or 58.6% had non-self matches that met the minimum initial BLAST cutoff threshold. Non-self matches were divided into genes likely to be horizontally transferred from taxa in the phylum Actinobacteria (2,692 genes) and genes likely originating from taxa outside the Actinobacteria (225 genes). The remaining 2,062 genes with no Genbank hits are likely to be either highly original (mutated or re-arranged), mis-assembled, or pseudogenes in the process of degeneration.

Figure 4.4 shows a 16S rRNA, polar tree phylogeny for the Actinobacterial genera identified by DarkHorse as hits for foreign proteins. Within this 16S tree is a donut chart showing a total of 93 Actinobacterial genera and the proportion of genera represented by these hits. Of these genes, 73% (1965/2692) were represented by only four genera: *Micromonospora*, *Streptomyces*, *Verrucosipora*, and *Actinoplanes*. These genes are colored in the phylogeny and *Salinispora* is shown enlarged and in bold. For the sake of brevity, Table 4.1 shows the most common genera from which *Salinispora* has acquired genes based on DarkHorse predictions. Of the 2,692 Actinobacterial gene hits, 1,499 genes (56%) are annotated as hypothetical proteins. Transposases, mobile elements and integrases comprise 275 genes (10%). A total of 85 genes were transcriptional regulators and 47 were ABC transporters. With respect to secondary metabolism, 16 NRPS genes were found with closest matches to the following genera: *Saccharothrix*, *Micromonospora*, *Kitasatospora*, *Alloactinosynnema*, and *Actinokineospora*. Two PKS genes were also found and likely laterally transferred from *Streptomyces* and *Micromonospora*. Seven multidrug transporters had close hits to the following genera: *Alloactinosynnema*, *Micromonospora*, *Saccharothrix*, *Streptomyces*, and *Nonomuraea*.

Of the 225 genes originating from taxa outside the Actinobacteria, greater than 50% are predicted to originate from the phyla Proteobacteria and Firmicutes (Figure

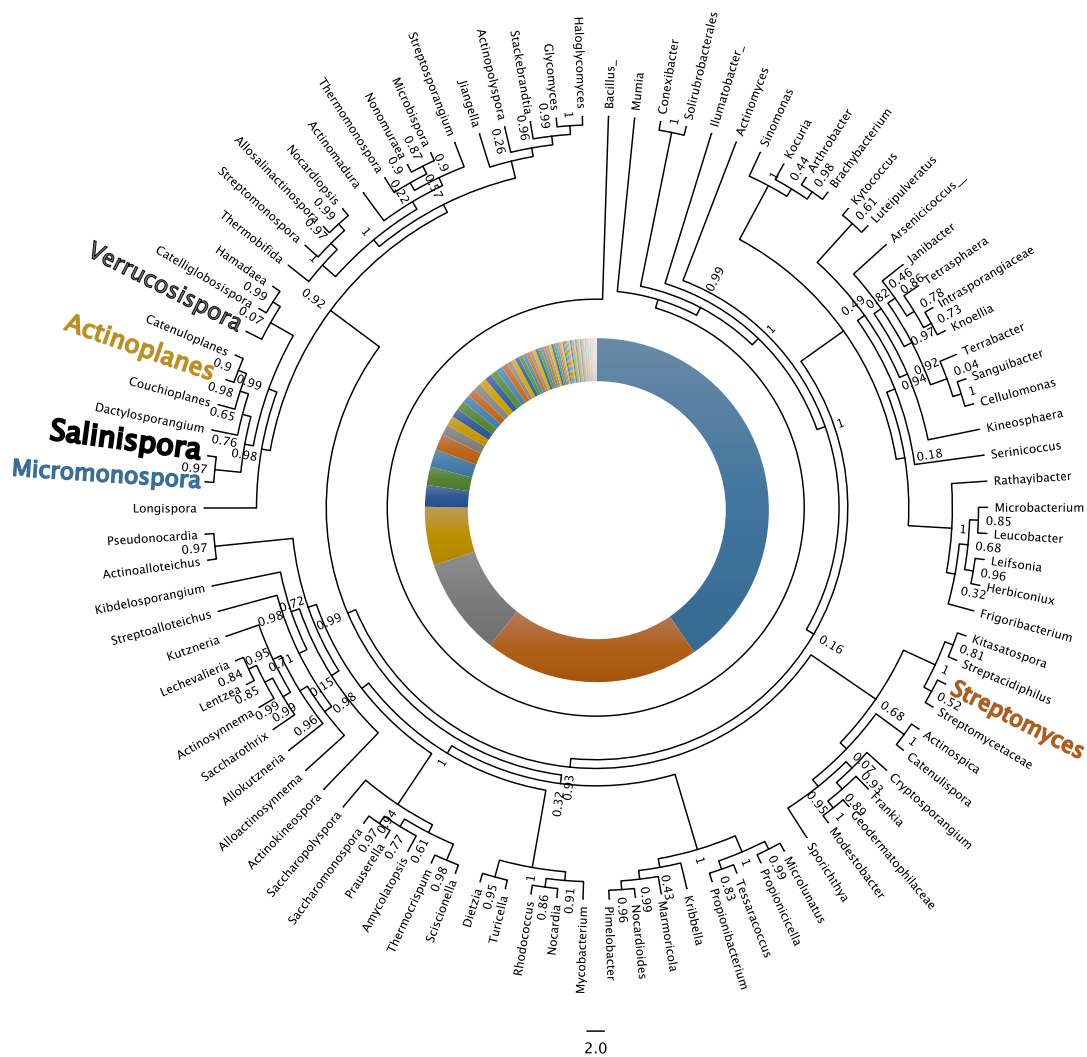


Figure 4.4: 16S rRNA phylogeny by genus for Actinobacterial DarkHorse hits. Colors of inner donut chart represent different genera, 73% of genes transferred (1965/2692) are represented by four genera: *Micromonospora*, *Streptomyces*, *Verrucosisspora*, and *Actinoplanes*. These genes are colored in the phylogeny and *Salinispora* is enlarged and in bold.

Table 4.1: List of most common genera from which *Salinispora* has acquired genes based on DarkHorse predictions.

Genus	Count
<i>Micromonospora</i>	1055
<i>Streptomyces</i>	527
<i>Verrucosispora</i>	239
<i>Actinoplanes</i>	144
<i>Frankia</i>	53
<i>Alloactinosynnema</i>	45
<i>Mycobacterium</i>	44
<i>Nocardia</i>	39
<i>Kitasatospora</i>	26
<i>Nonomuraea</i>	24
<i>Rhodococcus</i>	22
<i>Pseudonocardia</i>	22

4.5). A majority of the 225 genes (52%) were annotated as hypothetical proteins. Fourteen were annotated as spore germination proteins from *Bacillus*. Six gene hits for cell wall anchor proteins likely originated from *Listeria*, *Enterobacter*, and *Streptococcus*. Again, for secondary metabolite related genes, six NRPS genes were acquired from the genera *Bacillus*, *Chroococcales*, *Methylobacterium*, and *Bradyrhizobium*. Finally, one PKS gene was found to come from *Burkholderia*.

Lateral gene transfer events were averaged per genome by species both for Actinobacterial and non-Actinobacterial hits (Table 4.2). The average number of LGT events per genome from Actinobacteria was 16.6 for *S. arenicola*, 32.1 for *S. pacifica*, and 14.3 for *S. tropica*. *S. pacifica* had on average twice as many laterally transferred genes in its genome. The average number of LGT events per genome from non-Actinobacterial taxa was 1.8 for *S. arenicola*, 2.9 for *S. pacifica*, and 3.7 for *S. tropica*.

Another interesting result to come from the non-Actinobacteria DarkHorse analysis is the ability to identify phage infection. One strain in particular, *S. pacifica* CNS-055, which was identified in the previous section for having a disproportionately large

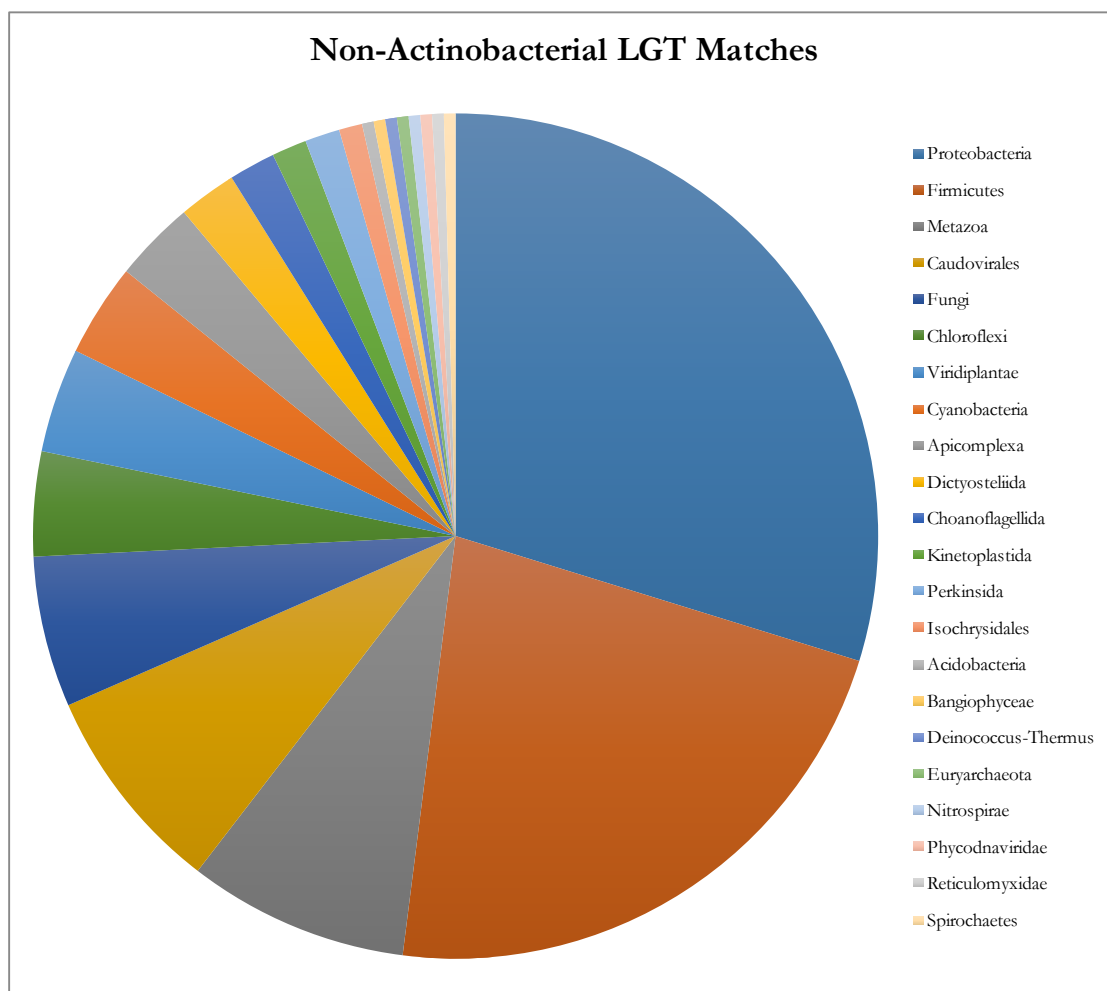


Figure 4.5: Pie chart representing the 225 genes likely laterally transferred from taxa outside the Actinobacteria clade

Table 4.2: Average number of lateral gene transfer events from Actinobacterial and non-Actinobacterial strains into *Salinispora* species.

	<i>S. arenicola</i>	<i>S. pacifica</i>	<i>S. tropica</i>
Average Actino LGT per Genome	16.6	32.1	14.3
Average NonActino LGT per Genome	1.8	2.9	3.7

number of BGC acquisitions, had 10 gene hits from a *Streptomyces* phage originating from the order Caudovirales, family Siphoviridae. These *Streptomyces* phage isolates are: Hydra, Danzina, Verse, Aaronocolus, Amela, Caliburn, and Sujidade. While most genomes only show one or a few instances of phage infection, this strain had the most hits and is indicative of a major phage infection.

4.5 Discussion

Calibrating the *Salinispora* molecular clock provides a new perspective on the time frames within which the three named species have been on independent evolutionary trajectories. Layering LGT events onto the calibrated phylogeny provides a temporal framework with which to understand how genetic information moves across species and strains both at the gene and biosynthetic gene cluster level. Those who study actinobacteria have compared *Salinispora* with the genus *Streptomyces* due to their similarly prolific secondary metabolite production and ability to swap genes promiscuously (Maldonado et al., 2009; Doolittle and Zhaxybayeva, 2009b; Bull et al., 2005). However, *Streptomyces* is a much more ancient and diverse lineage than *Salinispora* and studying LGT dynamics in a genus of over 550 species at best provides insight into intermediate-scale dynamics (McDonald and Currie, 2017). In contrast, access to a large number of closely related *Salinispora* genomes provides the opportunity to study population-scale LGT dynamics.

I was able to generate a molecular clock for *Salinispora* by creating a robust phylogenetic tree of conserved genes not under selection pressure as well as a tree placing the genus within the context of other bacterial lineages. This bacterial species tree was congruent with trees made for *Streptomyces* (McDonald and Currie, 2017). It should be noted that both *Salinispora* and *Streptomyces* are spore formers, and spore dormancy might affect molecular clock calibrations. However, it has been found that rates of molecular evolution in bacteria are relatively constant despite spore dormancy (Maughan, 2007).

I determined that *Salinispora* is a much younger lineage than *Streptomyces*. The emergence of *Salinispora* 68 mya interestingly corresponds with a mass extinction event. These extinction events were turning points in biotic evolution. The Cretaceous-

Tertiary (K-T) mass extinction event 66 mya was responsible for the sudden mass extinction of up to 75% of the plants and animals on Earth (Renne et al., 2013). While microbes do not readily leave a fossil record, some things are known about microbiota at the time. This extinction event represents a dramatic turnover in the fossil record for certain calcareous nanoplankton that formed the calcium deposits for which the Cretaceous was named. There is a marked turnover of calcareous nanoplankton at the species level (Pospichal, 1996; Bown, 2005). Statistically-based analyses of marine losses at the time suggest that decreases in diversity were caused by a sharp increase in extinctions rather than a decrease in speciation (Bambach et al., 2004).

It is unclear if a geologic event led to the speciation events within the genus some 24.22 mya. However, potential cooling associated with a large-scale Antarctic ice sheet expansion may have been a driver of *Salinispora* speciation during the Oligocene-Miocene transition approximately 23 mya (Beddow et al., 2016). Perhaps optimal growth temperatures for each species is more specific than we currently understand and there is a phylogenetic conservatism of thermal traits that limits dispersal as it does in *Streptomyces* sister-taxa (Choudoir and Buckley, 2018). The emergence of the genus *Salinispora* also coincides with the loss of the *mscL* gene. This gene encodes a large conductance mechano-sensitive channel that plays a role in osmotic adaptation and provides a mechanism to survive osmotic down shock. It has been suggested that the loss of this gene has relegated *Salinispora* to the marine environment (Penn and Jensen, 2012; Bucarey et al., 2012). Close relatives of *Salinispora*, which include the genera *Micromonospora*, *Verrucosispora* and *Actinoplanes*, have been isolated from marine sediments. These taxa, however, have maintained the *mscL* gene in their genomes. Members of the common ancestor of *Salinispora* introduced to the sea likely lost *mscL* since it was not needed. This prevented *Salinispora* from returning to land, and thus reduced recombination rates leading to an evolutionary independent ecotype.

The *Salinispora* molecular clock places the many gains and losses of biosynthetic gene clusters over evolutionary history. It is interesting to note that 178 BGCs have been identified and only 28 of these have been linked to their products (Letzel et al., 2017). Of these, rifamycin is in the strictest sense species-specific, having been observed in all *S. arenicola* strains sequenced to date. However, the rifamycin pathway is also found in the Actinobacterial genus *Amycolatopsis*. Given that it is not observed in the sister genus *Micromonospora*, it was likely acquired by *S. arenicola* around the time of its divergence (Figure 4.3). Similarly, many other gene clusters detected in *Salinispora* (e.g., staurosporine, enterocin, and lymphostin) have been observed in distantly related Actinobacteria such as the genus *Streptomyces*.

In total 48 BGCs were identified in >10 *Salinispora* strains. When placed on the *Salinispora* molecular clock, a clear model for *S. arenicola* and *S. tropica* emerges, however this is not as apparent with *S. pacifica* likely because it represents multiple species (Millán-Aguíñaga et al., 2017). *S. arenicola* began its independent evolutionary history 24 mya. The striking lack of diversity within this genus relative to *S. pacifica* could potentially be the result of a selective sweep that eliminated diversity across the species (Fraser et al., 2009). *S. tropica*, meanwhile, may have undergone a purge of genetic diversity more recently through a selective sweep that confined it to an ecological niche found in the Bahamas and the Yucatan.

Congruent with previous BGC studies of *Salinispora*, the *lym* BGC appears to be replaced by *sal* in clade 5 (Letzel et al., 2017). Those analyses predicted that *lym* was present in the common ancestor of the genus while *sal* was acquired more recently. For strains containing either *sal* or *lym*, the BGCs are located in the same inter-island region between GI16 and GI17 ((Letzel et al., 2017) Fig. 3).

The acquisition of *sta* by some *S. pacifica* strains is notable because this species has never been shown to produce staurosporine. Further investigation into this gene

cluster and its composition is warranted as there could be subtle differences in the BGCs that render the *S. pacifica* version silent. Many BGCs are identified as clade-specific under the newly proposed *S. pacifica* species delineations. Numerous acquisition and loss events are of BGCs with products that have yet to be identified. Yet, it is still quite apparent that there have been extensive acquisition and loss events involving biosynthetic gene clusters in *Salinispora* (Ziemert et al., 2014). As more genomes are sequenced, it is likely that the many BGC gain/loss events detected thus far represent only a small part of the evolutionary history of this genus.

The data derived from the DarkHorse analysis provides an interesting new perspective on potential evolutionary drivers of the genus. Prior to this analysis, only two *Salinispora* genomes had been investigated for this kind of genetic exchange (Penn et al., 2009). Other LGT investigations have determined that more genes are exchanged within phyla than between phyla (Beiko et al., 2005; Andam and Gogarten, 2011). This pattern appears to be consistent between *Salinispora* and the rest of the Actinobacteria. Almost three-quarters of the genes linked to LGT originated from four Actinobacterial genera. While three of these are relatively close to *Salinispora* on the 16S phylogenetic tree (*Micromonospora*, *Actinoplanes*, and *Verrucosispora*), *Streptomyces* represents a large proportion of gene donors despite being more phylogenetically distant. While DarkHorse accounts for database bias, sampling and study biases might account for these high numbers rather than a prowess for genetic exchange by *Streptomyces* since other Actinobacterial taxa are less well studied (*Streptomyces* alone has over 550 species).

It is a common occurrence when working with whole genomes that upwards of half of the genes have been uncharacterized and annotate as hypothetical proteins. This is likewise the case here; however, many genes are still informative. Unsurprisingly, transposable elements are in abundance. Understanding the mechanisms by which these transposons function could render them useful as genetic tools for biotechnological ap-

plications (Muñoz-López and García-Pérez, 2010). Secondary metabolism genes including NRPS and PKS genes show evidence of widespread swapping between taxa. Whether these include full biosynthetic clusters should be further investigated. The presence of multidrug transporters highlights the importance of protecting the cell from a secondary metabolite that it or its neighbor may produce (Goodsell, 1999).

Lateral gene transfer events from non-Actinobacteria appear to be dominated by Gram-negative Proteobacteria and Gram-positive Firmicutes. Genes from a Firmicute in the genus *Bacillus* encoding spore germination proteins were heavily represented. It is unclear if this affects *Salinispora* spore formation in any way. The presence of genes annotated as cell wall anchor proteins is an interesting find due to their utilization by pathogenic bacteria to adhere to surfaces and produce biofilms (Tettelin, 2005; Geoghegan and Foster, 2017). While *Salinispora* has not been known to be pathogenic or produce biofilms, these proteins could perhaps contribute to the differences in growth phenotypes seen in the lab even on a strain by strain basis.

The average number of LGT events detected shows that *S. pacifica* acquires twice as many genes from other Actinobacteria as *S. arenicola* and *S. tropica*. We know that *S. pacifica* is a relatively diverse species (Millán-Aguíñaga et al., 2017) and LGT may play a larger evolutionary role in what may turn out to be an amalgam of species. It is interesting to note that for transfer events from non-Actinobacterial strains, *S. tropica* has the highest average number of LGT events. This is surprising given the clonal nature of the species, however rates across all species are still relatively low.

The aims in this chapter were to understand how lateral gene transfer shaped *Salinispora* evolution as well as contextualize these dynamics with relation to a species molecular clock. I have found that genetic exchange predominantly takes place with members of the same phylum, bolstering the idea that non-homologous recombination occurs more frequently between more closely related individuals (Beiko et al., 2005).

I have also created the first molecular clock for *Salinispora* evolution. Dating of this genus has contextualized speciation on a geologic timescale, drastically contrasting the age of this rather young lineage to the commonly associated 'ancient' lineage, *Streptomyces*. Using divergence times rather than sequence similarity, we are able to gain a new perspective on bacterial evolution over small and large evolutionary timescales.

Chapter 4 is coauthored with Podell S and PR Jensen. The dissertation author was the primary investigator and author of this chapter.

Chapter 5

Final Remarks

The trajectory for comparative genomics was set when the first bacterial genomes were sequenced over two decades ago (Fleischmann et al., 1995). Genome sequences provided new opportunities to classify bacteria and understand how groups of related strains differ in gene content. Prior to this, phylogenetic approaches using the small subunit of the 16S rRNA were often used for species designations (Woese and Fox, 1977). However, this conserved gene did not provide the necessary resolution to study closely related taxa, so instead of a single gene, many genes were used in a method called MultiLocus Sequence Analysis (MLSA) (Maiden et al., 1998). Although this increased the representative portion of the genome that was analyzed, it was the availability of whole genome sequencing that allowed researchers to see the full complement of genes, known as the pangenome (Vernikos et al., 2015), and apply phylogenomic approaches and measures such as ANI to species designations (Goris et al., 2007). The field of comparative genomics has come a long way in a short amount of time. With a wealth of genome sequence data comes an unprecedented opportunity to study the levels of genetic similarity among bacteria that maintain the properties of species and the evolutionary processes that drive speciation events.

The goal of this dissertation was to explore the genus *Salinispora* through a comparative genomics lens. I was in a unique position to work in a laboratory with a large collection of strains that covered three different species of a closely related genus. Some of these species were isolated from the same location and it gave me the opportunity to ask what kinds of competitive strategies these co-occurring species use to occupy the same environments. Addressing this question requires delving into the world of bacterial species concepts and understanding ecotype models. To this end, a large-scale sequencing project was undertaken and 119 *Salinispora* strains were sequenced consisting of: 62 *S. arenicola*, 12 *S. tropica* and 45 *S. pacifica* strains. With these genomes in hand, I was able to identify the *Salinispora* pangenome, how many more genomes should be sequenced to obtain saturation and which types of genes potentially define two of the species (Chapter 2). I then analyzed global transcriptome data to identify shared genes that were differentially expressed, such that expression levels can also be incorporated into how bacterial species are determined (Chapter 3). In Chapter 4, I address an important component that also drives bacterial evolution, lateral gene transfer (LGT) and place these events in a temporal context onto the species phylogeny by generating a molecular clock for the genus.

The findings of Chapter 2 show that the pangenome of *Salinispora* has not yet reached saturation. Statistical analyses approximate that 20% of potential genes have yet to be sequenced. The core genome appears to reach an asymptote, therefore most yet to be discovered diversity likely falls into the flexible genome, representing genes typically associated with species-specific or adaptive traits. A large proportion of the genome falls into a function unknown or general function prediction (~40%). This proportion of ambiguity is not uncommon with bioinformatic analyses and represents a considerable amount of genetic potential that cannot be readily interpreted. Notwithstanding, I was able to define species-specific genetic cores for *S. arenicola* and *S. tropica*. *S. arenicola*

has more genes related to secondary metabolite biosynthesis, transport and catabolism, as well as amino acid transport and metabolism. *S. tropica*, on the other hand, had a core enriched in genes relating to cellular processes and signaling, cell division, cell wall biogenesis, and coenzyme transport and metabolism. It has been observed that *S. arenicola*, which has a larger genome and more secondary metabolism biosynthetic gene clusters, devotes more energy towards secondary metabolism production whereas *S. tropica*'s tradeoff is faster growth and a smaller genome. This provides a genetic basis for the competitive strategy that has been previously proposed between the two species (Patin et al., 2016).

Chapter 3 harnessed another powerful -omics tool, transcriptomics, to look at expression levels of genes from the pangenome. Previous studies suggest that species delineations should not be limited by gene content but also by differential expression of those genes. There was evidence in *Salinispora* for differences in gene expression related to degrading chitin, the most abundant biopolymer in the ocean. Analysis of transcriptomics data for two species found that despite having the same chitinase genes, expression is statistically higher in *S. tropica*. We have learned that *S. tropica* grows faster than *S. arenicola* (Patin et al., 2016). In general *S. tropica* has more numbers and types of chitinase genes than *S. arenicola*. Could they utilize chitin more effectively, contributing to their faster growth and utilization of multiple chitin sources? Shared genes between the two species were also identified and analyzed for differential expression to look for traits that would differentiate them as species. Bioinformatic predictions suggest *S. tropica* has the ability to better cope with osmotic stress and may be more affected by nutrient or oxidative stress while *S. arenicola* has more energetic needs during stationary phase growth potentially due to costly secondary metabolism. Thus, in addition to differences in gene content, the differential expression of shared genes appears to play a key role in what differentiates *Salinispora* species.

The goal of Chapter 4 was to take a broader view of the evolutionary history of the genus *Salinispora*. Because this marine genus does not exist in isolation, but rather, in microbial communities, opportunities for genetic exchange exist. There is clear evidence that many genes have been acquired via lateral gene transfer (LGT) throughout the evolutionary history of this taxon. In addition to identifying genes with evidence of LGT, a molecular clock is presented. This provides a new perspective on the time frames within which the three named species evolved. Layering LGT events onto the calibrated phylogeny provides a temporal framework with which to understand how genetic information moves across species and strains at both the gene and biosynthetic gene cluster level. The emergence of *Salinispora* 68 mya corresponds with the Cretaceous-Tertiary (K-T) mass extinction event 66 mya. The rise of the genus also coincided with the loss of the *mscL* gene, which may have prevented them from returning to the land. This study also supports previous findings that LGT events occur more frequently within a phylum than between phyla (Beiko et al., 2005).

Comparative genomics is the application of bioinformatics methods to the analysis of whole genome sequences with the objective of identifying biological principles, i.e. biology, *in silico*. In many ways, this statement greatly underplays the real value of comparative genomics: an extremely powerful technique that provides biological insights that could not have been achieved in any other way. In *Salinispora*, we have a model genus comprised of closely related species that can be used to ask questions regarding species concepts and understand bacterial evolution at the population level.

References

- Ahlgren, N. A., Rocap, G., and Chisholm, S. W. (2006). Measurement of *Prochlorococcus* ecotypes using realtime polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. *Environmental Microbiology*, 8(3):441–454.
- Ahmed, L., Jensen, P. R., Freel, K. C., Brown, R., Jones, A. L., Kim, B.-Y., and Goodfellow, M. (2013). *Salinispora pacifica* sp. nov., an actinomycete from marine sediments. *Antonie van Leeuwenhoek*, 103(5):1069–1078.
- Aldhebiani, A. Y. (2018). Species concept and speciation. *Saudi Journal of Biological Sciences*, 25(3):437–440.
- Aluwihare, L. I., Repeta, D. J., Pantoja, S., and Johnson, C. G. (2005). Two chemically distinct pools of organic nitrogen accumulate in the ocean. *Science (New York, NY)*, 308(5724):1007–1010.
- Ammendola, S., Pasquali, P., Pistoia, C., Petrucci, P., Petrarca, P., Rotilio, G., and Battistoni, A. (2007). High-affinity Zn²⁺ uptake system ZnuABC is required for bacterial zinc homeostasis in intracellular environments and contributes to the virulence of *Salmonella enterica*. *Infection and immunity*, 75(12):5867–5876.
- Amos, G. C. A., Awakawa, T., Tuttle, R. N., Letzel, A.-C., Kim, M. C., Kudo, Y., Fenical, W., S Moore, B., and Jensen, P. R. (2017). Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. *PNAS*, 114(52):E11121–E11130.
- Andam, C. P. and Gogarten, J. P. (2011). Biased gene transfer and its implications for the concept of lineage. *Biology Direct*, 6:47.
- Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10):–R106.
- Avery, O. T. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* type III. *Journal of Experimental Medicine*, 79(2):137–158.

- Azam, F., Fenchel, T., Field, J. G., Gray, J. S., Meyerreil, L. A., and Thingstad, F. (1983). The Ecological Role of Water-Column Microbes in the Sea. *Marine Ecology-Progress Series*, 10(3):257–263.
- Bambach, R. K., Knoll, A. H., and Wang, S. C. (2004). Origination, extinction, and mass depletions of marine diversity. *Paleobiology*, 30(4):522–542.
- Battistuzzi, F. U., Feijao, A., and Hedges, S. B. (2004). A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evolutionary Biology*, 4:44.
- Beddow, H. M., Liebrand, D., Sluijs, A., Wade, B. S., and Lourens, L. J. (2016). Global change across the Oligocene-Miocene transition: High-resolution stable isotope records from IODP Site U1334 (equatorial Pacific Ocean). *Paleoceanography*, 31(1):81–97.
- Beiko, R. G., Harlow, T. J., and Ragan, M. A. (2005). Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences*, 102(40):14332–14337.
- Bennett, S. (2004). Solexa Ltd. *Pharmacogenomics*, 5(4):433–438.
- Bentley, S. (2009). Sequencing the species pan-genome. *Nature Reviews Microbiology*, 7(4):258–259.
- Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A.-M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H., and Harper, D. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3 (2). *Nature*, 417(6885):141–147.
- Bérdy, J. (2005). Bioactive microbial metabolites. *The Journal of antibiotics*, 58(1):1–26.
- Bobay, L.-M., Traverse, C. C., and Ochman, H. (2015). Impermanence of bacterial clones. *PNAS*, 112(29):8893–8900.
- Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. A., Boudreau, M. E. R., Nesbø, C. L., Case, R. J., and Doolittle, W. F. (2003). Lateral Gene Transfer and the Origins of Prokaryotic Groups. *Annual Review of Genetics*, 37(1):283–328.
- Bown, P. (2005). Selective calcareous nannoplankton survivorship at the Cretaceous-Tertiary boundary. *Geology*, 33(8):653–656.
- Brocks, J. J., Logan, G. A., Buick, R., and Summons, R. E. (1999). Archean molecular fossils and the early rise of eukaryotes. *Science (New York, NY)*, 285(5430):1033–1036.

- Bucarey, S. A., Penn, K., Paul, L., Fenical, W., and Jensen, P. R. (2012). Genetic Complementation of the Obligate Marine Actinobacterium *Salinispora tropica* with the Large Mechanosensitive Channel Gene *mscL* Rescues Cells from Osmotic Down-shock. *Applied and Environmental Microbiology*, 78(12):4175–4182.
- Buchanan, G. O., Williams, P. G., Feling, R. H., Kauffman, C. A., Jensen, P. R., and Fenical, W. (2005). Sporolides A and B: Structurally Unprecedented Halogenated Macrolides from the Marine Actinomycete *Salinisporatropica*. *Organic Letters*, 7(13):2731–2734.
- Bull, A. T. and Stach, J. E. M. (2007). Marine actinobacteria: new opportunities for natural product search and discovery. *Trends in Microbiology*, 15(11):491–499.
- Bull, A. T., Stach, J. E. M., Ward, A. C., and Goodfellow, M. (2005). Marine actinobacteria: perspectives, challenges, future directions. *Antonie van Leeuwenhoek*, 87(1):65–79.
- Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008). ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome research*, 18(5):810–820.
- Cadillo-Quiroz, H., Didelot, X., Held, N. L., Herrera, A., Darling, A., Reno, M. L., Krause, D. J., and Whitaker, R. J. (2012). Patterns of Gene Flow Define Species of Thermophilic Archaea. *PLoS biology*, 10(2):e1001265.
- Campoy, S., Jara, M., Busquets, N., Pérez De Rozas, A. M., Badiola, I., and Barbé, J. (2002). Role of the high-affinity zinc uptake *znuABC* system in *Salmonella enterica* serovar typhimurium virulence. *Infection and immunity*, 70(8):4721–4725.
- Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G., and Parkhill, J. (2005). ACT: the Artemis comparison tool. *Bioinformatics*, 21(16):3422–3423.
- Challis, G. L. (2008). Genome mining for novel natural product discovery. *Journal of medicinal chemistry*, 51(9):2618–2628.
- Chang, D.-E., Smalley, D. J., and Conway, T. (2002). Gene expression profiling of *Escherichia coli* growth transitions: an expanded stringent response model. *Molecular Microbiology*, 45(2):289–306.
- Choudoir, M. J. and Buckley, D. H. (2018). Phylogenetic conservatism of thermal traits explains dispersal limitation and genomic differentiation of *Streptomyces* sister-taxa. *The ISME Journal*, pages 1–11.
- Cohan, F. M. (2002). What are Bacterial Species? *Annual Review of Microbiology*, 56(1):457–487.

- Cohan, F. M. and Koeppel, A. F. (2008). The Origins of Ecological Diversity in Prokaryotes. *Current Biology*, 18(21):1024–1034.
- Cottrell, M. T. and Kirchman, D. L. (2000). Natural assemblages of marine proteobacteria and members of the Cytophaga-Flavobacter cluster consuming low-and high-molecular-weight dissolved organic matter. *Applied and Environmental Microbiology*.
- Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164–1165.
- Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature methods*, 9:772 EP –.
- Darwin, C. (1859). On the Origin of Species by Means of Natural Selection. New York: D. Appleton and Company.
- Daubin, V. and Szöllősi, G. J. (2016). Horizontal Gene Transfer and the History of Life. *Cold Spring Harbor perspectives in biology*, 8(4):a018036.
- Davies, J. (1994). Inactivation of antibiotics and the dissemination of resistance genes. *Science (New York, NY)*, 264(5157):375–382.
- Davies, J. (1996). Origins and evolution of antibiotic resistance. *Microbiologia (Madrid, Spain)*, 12(1):9–16.
- de Jong, A., van der Meulen, S., Kuipers, O. P., and Kok, J. (2015). T-REx: Transcriptome analysis webserver for RNA-seq Expression data. *BMC Genomics*, 16:663.
- de Queiroz, K. (2005). Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences*, 102 Suppl 1:6600–6607.
- Denef, V. J., Kalnejais, L. H., Mueller, R. S., Wilmes, P., Baker, B. J., Thomas, B. C., VerBerkmoes, N. C., Hettich, R. L., and Banfield, J. F. (2010). Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *PNAS*, 107(6):2383–2390.
- Doolittle, R. F., Feng, D. F., Tsang, S., Cho, G., and Little, E. (1996). Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science (New York, NY)*, 271(5248):470–477.
- Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science (New York, NY)*, 284(5423):2124–2129.
- Doolittle, W. F. and Papke, R. T. (2006). Genomics and the bacterial species problem. *Genome biology*, 7(9):116.

- Doolittle, W. F. and Zhaxybayeva, O. (2009a). On the origin of prokaryotic species. *Genome research*, 19(5):744–756.
- Doolittle, W. F. and Zhaxybayeva, O. (2009b). On the origin of prokaryotic species. *Genome research*, 19(5):744–756.
- Doroghazi, J. R. and Buckley, D. H. (2010). Widespread homologous recombination within and between *Streptomyces* species. *The ISME Journal*, 4(9):1136–1143.
- Doroghazi, J. R. and Metcalf, W. W. (2013). Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics*, 14:611.
- Du, Y., Shi, W.-W., He, Y.-X., Yang, Y.-H., Zhou, C.-Z., and Chen, Y. (2011). Structures of the substrate-binding protein provide insights into the multiple compatible solute binding specificities of the *Bacillus subtilis* ABC transporter OpuC. *The Biochemical journal*, 436(2):283–289.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.
- Elnitski, L., Burhans, R., Riemer, C., Hardison, R., and Miller, W. (2010). MultiPip-Maker: a comparative alignment server for multiple DNA sequences. *Current protocols in bioinformatics / editorial board, Andreas D. Baxevanis ... [et al.]*, Chapter 10:Unit10.4.
- Eustáquio, A. S., Nam, S.-J., Penn, K., Lechner, A., Wilson, M. C., Fenical, W., Jensen, P. R., and Moore, B. S. (2010). The Discovery of Salinosporamide K from the Marine Bacterium “*Salinispora pacifica*” by Genome Mining Gives Insight into Pathway Evolution. *ChemBioChem*, 12(1):61–64.
- Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research*, 8(3):186–194.
- Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research*, 8(3):175–185.
- Fang, G., Bhardwaj, N., Robilotto, R., and Gerstein, M. B. (2010). Getting started in gene orthology and functional analysis. *PLoS computational biology*, 6(3):e1000703.
- Feling, R. H., Buchanan, G. O., Mincer, T. J., Kauffman, C. A., Jensen, P. R., and Fenical, W. (2003). Salinosporamide A: a highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus *Salinispora*. *Angewandte Chemie (International ed. in English)*, 42(3):355–357.
- Fenical, W. and Jensen, P. R. (2014). Developing a new resource for drug discovery:

- marine actinomycete bacteria. *Nature Chemical Biology*, 111(12):E1130–E1139.
- Fenical, W., Jensen, P. R., Palladino, M. A., Lam, K. S., Lloyd, G. K., and Potts, B. C. (2009). Discovery and development of the anticancer agent salinosporamide A (NPI-0052). *Bioorganic & Medicinal Chemistry*, 17(6):2175–2180.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., and Merrick, J. M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science (New York, NY)*, 269(5223):496–512.
- Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G., and Hanage, W. P. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. *Science*, 323(5915):741–746.
- Frederiksen, R. F., Paspaliari, D. K., Larsen, T., Storgaard, B. G., Larsen, M. H., Ingmer, H., Palcic, M. M., and Leisner, J. J. (2013). Bacterial chitinases and chitin-binding proteins as virulence factors. *Microbiology*, 159(Pt 5):833–847.
- Fredrickson, J. K., Romine, M. F., Beliaev, A. S., Auchtung, J. M., Driscoll, M. E., Gardner, T. S., Nealson, K. H., Osterman, A. L., Pinchuk, G., Reed, J. L., Rodionov, D. A., Rodrigues, J. L. M., Saffarini, D. A., Serres, M. H., Spormann, A. M., Zhulin, I. B., and Tiedje, J. M. (2008). Towards environmental systems biology of *Shewanella*. *Nature Reviews Microbiology*, 6(8):592–603.
- Freel, K. C., Edlund, A., and Jensen, P. R. (2011). Microdiversity and evidence for high dispersal rates in the marine actinomycete ‘*Salinispora pacifica*’. 14(2):480–493.
- Friedman, N., Cai, L., and Xie, X. S. (2006). Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Physical review letters*, 97(16):168302.
- Funkhouser, J. D. and Aronson, N. N. (2007). Chitinase family GH18: evolutionary insights from the genomic history of a diverse protein family. *BMC Evolutionary Biology*, 7(1):96.
- Gan, H. M., Hudson, A. O., Rahman, A. Y. A., Chan, K. G., and Savka, M. A. (2013). Comparative genomic analysis of six bacteria belonging to the genus *Novosphingobium*: insights into marine adaptation, cell-cell signaling and bioremediation. *BMC Genomics*, 14(1):1–1.
- Garvin, J., Buick, R., Anbar, A. D., Arnold, G. L., and Kaufman, A. J. (2009). Isotopic evidence for an aerobic nitrogen cycle in the latest Archean. *Science*, 323(5917):1045–1048.
- Gaupp, R., Schlag, S., Liebeke, M., Lalk, M., and Götz, F. (2010). Advantage of up-

- regulation of succinate dehydrogenase in *Staphylococcus aureus* biofilms. *Journal of Bacteriology*, 192(9):2385–2394.
- Ge, F., Wang, L.-S., and Kim, J. (2005). The Cobweb of Life Revealed by Genome-Scale Estimates of Horizontal Gene Transfer. *PLoS biology*, 3(10):e316.
- Geoghegan, J. A. and Foster, T. J. (2017). Cell Wall-Anchored Surface Proteins of *Staphylococcus aureus*: Many Proteins, Multiple Functions. *Current topics in microbiology and immunology*, 409:95–120.
- Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F. L., and Swings, J. (2005). Opinion: Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, 3(9):733–739.
- Gibbons, S. M. and Gilbert, J. A. (2015). Microbial diversity—exploration of natural ecosystems and microbiomes. *Current opinion in genetics & development*, 35:66–72.
- Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *PNAS*, 108(4):1513–1518.
- Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Molecular biology and evolution*, 19(12):2226–2238.
- Goodsell, D. S. (1999). The molecular perspective: the multidrug transporter. *The oncologist*, 4(5):428–429.
- Gordon, D., Abajian, C., and Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome research*, 8(3):195–202.
- Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, 57(1):81–91.
- Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, 52(5):696–704.
- Hahne, H., Mäder, U., Otto, A., Bonn, F., Steil, L., Bremer, E., Hecker, M., and Becher, D. (2010). A comprehensive proteomics and transcriptomics analysis of *Bacillus subtilis* salt stress adaptation. *Journal of Bacteriology*, 192(3):870–882.
- Hallin, P. F., Binnewies, T. T., and Ussery, D. W. (2008). The genome BLASTatlas—a

- GeneWiz extension for visualization of whole-genome homology. *Molecular bioSystems*, 4(5):363–371.
- Hanage, W. P., Fraser, C., and Spratt, B. G. (2006). The impact of homologous recombination on the generation of diversity in bacteria. *Journal of Theoretical Biology*, 239(2):210–219.
- Haubold, B. and Wiehe, T. (2004). Comparative genomics: methods and applications. *Naturwissenschaften*, 91(9).
- Hederstedt, L. and Rutberg, L. (1980). Biosynthesis and membrane binding of succinate dehydrogenase in *Bacillus subtilis*. *Journal of Bacteriology*, 144(3):941–951.
- Hederstedt, L. and Rutberg, L. (1981). Succinate dehydrogenase—a comparative review. *Microbiological reviews*, 45(4):542–555.
- Hoehler, T. M. and Jørgensen, B. B. (2013). Microbial life under extreme energy limitation. *Nature Reviews Microbiology*, 11(2):83–94.
- Holt, K. E., Thomson, N. R., Wain, J., Langridge, G. C., Hasan, R., Bhutta, Z. A., Quail, M. A., Norbertczak, H., Walker, D., Simmonds, M., White, B., Bason, N., Mungall, K., Dougan, G., and Parkhill, J. (2009). Pseudogene accumulation in the evolutionary histories of *Salmonella enterica* serovars Paratyphi A and Typhi. *BMC Genomics*, 10:36.
- Hutchison, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Gill, J., Kannan, K., Karas, B. J., Ma, L., Pelletier, J. F., Qi, Z.-Q., Richter, R. A., Strychalski, E. A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K. S., Smith, H. O., Glass, J. I., Merryman, C., Gibson, D. G., and Venter, J. C. (2016). Design and synthesis of a minimal bacterial genome. *Science*, 351(6280):aad6253.
- Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119.
- Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M., and Ōmura, S. (2003). Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nature Biotechnology*, 21(5):526–531.
- Ishihama, A. (1997). Adaptation of gene expression in stationary phase bacteria. *Current opinion in genetics & development*, 7(5):582–588.
- Itoh, Y., Kawase, T., Nikaidou, N., Fukada, H., Mitsutomi, M., Watanabe, T., and Itoh, Y. (2002). Functional analysis of the chitin-binding domain of a family 19 chitinase from *Streptomyces griseus* HUT6037: substrate-binding affinity and cis-

- dominant increase of antifungal function. *Bioscience, biotechnology, and biochemistry*, 66(5):1084–1092.
- Jensen, P. R. (2016). Natural Products and the Gene Cluster Revolution. *Trends in Microbiology*, 24(12):968–977.
- Jensen, P. R., Chavarria, K. L., Fenical, W., Moore, B. S., and Ziemert, N. (2013). Challenges and triumphs to genomics-based natural product discovery. *Journal of Industrial Microbiology & Biotechnology*, 41(2):203–209.
- Jensen, P. R., Dwight, R., and Fenical, W. (1991). Distribution of actinomycetes in near-shore tropical marine sediments. *Applied and Environmental Microbiology*, 57(4):1102–1108.
- Jensen, P. R., Gontang, E., Mafnas, C., Mincer, T. J., and Fenical, W. (2005). Culturable marine actinomycete diversity from tropical Pacific Ocean sediments. *Environmental Microbiology*, 7(7):1039–1048.
- Jensen, P. R. and Mafnas, C. (2006). Biogeography of the marine actinomycete *Salinispora*. *Environmental Microbiology*, 8(11):1881–1888.
- Jensen, P. R., Moore, B. S., and Fenical, W. (2015a). The marine actinomycete genus *Salinispora*: a model organism for secondary metabolite discovery. *Natural Product Reports*, 32(5):738–751.
- Jensen, P. R., Moore, B. S., and Fenical, W. (2015b). The marine actinomycete genus *Salinispora*: a model organism for secondary metabolite discovery. *Natural Product Reports*, 32(5):738–751.
- Jensen, P. R., Williams, P. G., Oh, D. C., Zeigler, L., and Fenical, W. (2007). Species-Specific Secondary Metabolite Production in Marine Actinomycetes of the Genus *Salinispora*. *Applied and Environmental Microbiology*, 73(4):1146–1152.
- Jeuniaux, C. and Voss-Foucart, M. F. (1991). Chitin biomass and production in the marine environment. *Biochemical Systematics and Ecology*, 19(5):347–356.
- Johnson, Z. I., Zinser, E. R., Coe, A., McNulty, N. P., Woodward, E. M. S., and Chisholm, S. W. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science (New York, NY)*, 311(5768):1737–1740.
- Knapp, S., Polaszek, A., and Watson, M. (2007). Spreading the word. *Nature*, 446:261–262.
- Koga, T., Hirota, N., and Takumi, K. (1999). Bactericidal activities of essential oils of basil and sage against a range of bacteria and the effect of these essential oils on *Vibrio parahaemolyticus*. *Microbiological Research*, 154(3):267–273.

- Koga, T. and Takumi, K. (1995). Nutrient starvation induces cross protection against heat, osmotic, or H₂O₂ challenge in *Vibrio parahaemolyticus*. *Microbiology and immunology*, 39(3):213–215.
- Kogoma, T., Cadwell, G. W., Barnard, K. G., and Asai, T. (1996). The DNA replication priming protein, PriA, is required for homologous recombination and double-strand break repair. *Journal of Bacteriology*, 178(5):1258–1264.
- Kolter, R., Siegele, D. A., and Tormo, A. (1993). The stationary phase of the bacterial life cycle. *Annual Reviews in Microbiology*, 47:855–874.
- Konstantinidis, K. T., Serres, M. H., Romine, M. F., Rodrigues, J. L. M., Auchtung, J., McCue, L.-A., Lipton, M. S., Obraztsova, A., Giometti, C. S., Nealson, K. H., Fredrickson, J. K., and Tiedje, J. M. (2009). Comparative systems biology across an evolutionary gradient within the *Shewanella* genus. *PNAS*, 106(37):15909–15914.
- Koonin, E. V., Makarova, K. S., and Aravind, L. (2001). Horizontal Gene Transfer in Prokaryotes: Quantification and Classification 1. *Annual Review of Microbiology*, 55(1):709–742.
- Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular biology and evolution*, 33(7):1870–1874.
- Laird, C. D., McConaughy, B. L., and McCarthy, B. J. (1969). Rate of Fixation of Nucleotide Substitutions in Evolution. *Nature*, 224:149 EP –.
- Lechevalier, H. A. and Lechevalier, M. P. (1967). Biology of actinomycetes. *Annual Reviews in Microbiology*, 21:71–100.
- Lederberg, J. and Tatum, E. L. (1946). Gene recombination in *Escherichia coli*. *Nature*, 158(4016):558.
- Lerat, E., Daubin, V., Ochman, H., and Moran, N. A. (2005). Evolutionary Origins of Genomic Repertoires in Bacteria. *PLoS biology*, 3(5):e130.
- Letzel, A.-C., Li, J., Amos, G. C. A., Millán-Aguíñaga, N., Ginigini, J., Abdelmohsen, U. R., Gaudêncio, S. P., Ziemert, N., Moore, B. S., and Jensen, P. R. (2017). Genomic insights into specialized metabolism in the marine actinomycete *Salinispora*. *Environmental Microbiology*, 19(9):3660–3673.
- Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595.
- Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9):2178–2189.

- Liao, C.-H., Xu, Y., Rigali, S., and Ye, B.-C. (2015). DasR is a pleiotropic regulator required for antibiotic production, pigment biosynthesis, and morphological development in *Saccharopolyspora erythraea*. *Applied Microbiology and Biotechnology*, 99(23):10215–10224.
- Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550.
- Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5):0955–0964.
- Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., and Spratt, B. G. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95(6):3140–3145.
- Maldonado, L. A., Fenical, W., Jensen, P. R., Kauffman, C. A., Mincer, T. J., Ward, A. C., Bull, A. T., and 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(Pt 5):1759–1766.
- Maldonado, L. A., Fragoso-Yáñez, D., Pérez-García, A., Rosellón-Druker, J., and Quintana, E. T. (2009). Actinobacterial diversity from marine sediments collected in Mexico. *Antonie van Leeuwenhoek*, 95(2):111–120.
- Manivasagan, P., Venkatesan, J., Sivakumar, K., and Kim, S.-K. (2013). Marine actinobacterial metabolites: Current status and future perspectives. *Microbiological Research*, 168(6):311–332.
- Mao, D. and Grogan, D. (2012). Genomic evidence of rapid, global-scale gene flow in a *Sulfolobus* species. *The ISME Journal*, 6(8):1613–1616.
- Markowitz, V. M., Mavromatis, K., Ivanova, N. N., Chen, I.-M. A., Chu, K., and Kyrpides, N. C. (2009). IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics*, 25(17):2271–2278.
- Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P. J., Naeem, S., Øvreås, L., Reysenbach, A.-L., Smith, V. H., and Staley, J. T. (2006). Microbial biogeography: putting microorganisms on the map. *Nature Reviews Mi-*

- crobiology*, 4(2):102–112.
- Maughan, H. (2007). Rates of molecular evolution in bacteria are relatively constant despite spore dormancy. *Evolution*, 61(2):280–288.
- Mayr, E. (1942). *Systematics and the Origin of Species, from the Viewpoint of a Zoologist*. Harvard University Press.
- McDonald, B. R. and Currie, C. R. (2017). Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus *Streptomyces*. *mBio*, 8(3).
- Millán-Aguinaga, N., Chavarria, K. L., Ugalde, J. A., Letzel, A.-C., Rouse, G. W., and Jensen, P. R. (2017). Phylogenomic Insight into *Salinispora* (Bacteria, Actinobacteria) Species Designations. *Scientific Reports*, 7(1):1917.
- Mincer, T. J., Fenical, W., and Jensen, P. R. (2005). Culture-Dependent and Culture-Independent Diversity within the Obligate Marine Actinomycete Genus *Salinispora*. *Applied and Environmental Microbiology*, 71(11):7019–7028.
- Mincer, T. J., Jensen, P. R., Kauffman, C. A., and 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.
- Mojzsis, S. J., Arrhenius, G., McKeegan, K. D., Harrison, T. M., Nutman, A. P., and Friend, C. (1996). Evidence for life on Earth before 3,800 million years ago. *Nature*, 384(6604):55–59.
- Moore, L. R., Goericke, R., and Chisholm, S. W. (1995). Comparative Physiology of *Synechococcus* and *Prochlorococcus* - Influence of Light and Temperature on Growth, Pigments, Fluorescence and Absorptive Properties. *Marine Ecology-Progress Series*, 116(1-3):259–275.
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628.
- Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., Zhang, J., Weinstock, G. M., Isaacs, F., Rozowsky, J., and Gerstein, M. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology*, 17:53.
- Muñoz-López, M. and García-Pérez, J. L. (2010). DNA transposons: nature and applications in genomics. *Current genomics*, 11(2):115–128.
- Navarro Llorens, J. M., Tormo, A., and Martínez-García, E. (2010). Stationary phase in gram-negative bacteria. *FEMS Microbiology Reviews*, 34(4):476–495.
- Nieselt, K., Battke, F., Herbig, A., Bruheim, P., Wentzel, A., Jakobsen, Ø. M., Sletta,

- H., Alam, M. T., Merlo, M. E., Moore, J., Omara, W. A. M., Morrissey, E. R., Juarez-Hermosillo, M. A., Rodríguez-García, A., Nentwich, M., Thomas, L., Iqbal, M., Legaie, R., Gaze, W. H., Challis, G. L., Jansen, R. C., Dijkhuizen, L., Rand, D. A., Wild, D. L., Bonin, M., Reuther, J., Wohlleben, W., Smith, M. C. M., Burroughs, N. J., Martín, J. F., Hodgson, D. A., Takano, E., Breitling, R., Ellingsen, T. E., and Wellington, E. M. H. (2010). The dynamic architecture of the metabolic switch in *Streptomyces coelicolor*. *BMC Genomics*, 11:10.
- Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304.
- Ochman, H. and Wilson, A. C. (1987). Evolution in Bacteria - Evidence for a Universal Substitution Rate in Cellular Genomes. *Journal of Molecular Evolution*, 26(1-2):74–86.
- Oh, D.-C., Williams, P. G., Kauffman, C. A., Jensen, P. R., and Fenical, W. (2006). Cyanosporasides A and B, Chloro- and Cyano-cyclopenta[a]indene Glycosides from the Marine Actinomycete “*Salinispora pacifica*”. *Organic Letters*, 8(6):1021–1024.
- Ortega, A. P., Villagra, N. A., Urrutia, I. M., Valenzuela, L. M., Talamilla-Espinoza, A., Hidalgo, A. A., Rodas, P. I., Gil, F., Calderón, I. L., Paredes-Sabja, D., Mora, G. C., and Fuentes, J. A. (2016). Infection, Genetics and Evolution. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 45(C):111–121.
- Osamura, T., Kawakami, T., Kido, R., Ishii, M., and Arai, H. (2017). Specific expression and function of the A-type cytochrome c oxidase under starvation conditions in *Pseudomonas aeruginosa*. *PLoS ONE*, 12(5):e0177957.
- Park, S. J. and Gunsalus, R. P. (1995). Oxygen, iron, carbon, and superoxide control of the fumarase *fumA* and *fumC* genes of *Escherichia coli*: role of the *arcA*, *fnr*, and *soxR* gene products. *Journal of Bacteriology*, 177(21):6255–6262.
- Pati, A., Ivanova, N. N., Mikhailova, N., Ovchinnikova, G., Hooper, S. D., Lykidis, A., and Kyrpides, N. C. (2010). GenePrimP: a gene prediction improvement pipeline for prokaryotic genomes. *Nature Publishing Group*, 7(6):455–457.
- Patin, N. V., Duncan, K. R., Dorrestein, P. C., and Jensen, P. R. (2016). Competitive strategies differentiate closely related species of marine actinobacteria. *ISME Journal*, 10(2):478–490.
- Penn, K., Jenkins, C., Nett, M., Udworthy, D. W., Gontang, E. A., McGlinchey, R. P., Foster, B., Lapidus, A., Podell, S., Allen, E. E., Moore, B. S., and Jensen, P. R. (2009). Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. *The ISME Journal*, 3(10):1193–1203.

- Penn, K. and Jensen, P. R. (2012). Comparative genomics reveals evidence of marine adaptation in *Salinispora* species. *BMC Genomics*, 13(1):1–12.
- Petrarca, P., Ammendola, S., Pasquali, P., and Battistoni, A. (2010). The Zur-regulated ZinT protein is an auxiliary component of the high-affinity ZnuABC zinc transporter that facilitates metal recruitment during severe zinc shortage. *Journal of Bacteriology*, 192(6):1553–1564.
- Podell, S. and Gaasterland, T. (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome biology*, 8(2).
- Pospichal, J. J. (1996). Calcareous nannofossils and clastic sediments at the Cretaceous-Tertiary boundary, northeastern Mexico. *Geology*, 24(3):255.
- Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., and Glockner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35(21):7188–7196.
- Qin, Q.-L., Xie, B.-B., Yu, Y., Shu, Y.-L., Rong, J.-C., Zhang, Y.-J., Zhao, D.-L., Chen, X.-L., Zhang, X.-Y., Chen, B., Zhou, B.-C., and Zhang, Y.-Z. (2013). Comparative genomics of the marine bacterial genus *Glaciecola* reveals the high degree of genomic diversity and genomic characteristic for cold adaptation. *Environmental Microbiology*, 16(6):1642–1653.
- Redfield, R. J. (2001). Do bacteria have sex? *Nature reviews. Genetics*, 2(8):634–639.
- Reeve, C. A., Amy, P. S., and Matin, A. (1984). Role of protein synthesis in the survival of carbon-starved *Escherichia coli* K-12. *Journal of Bacteriology*, 160(3):1041–1046.
- Remenant, B., Coupat-Goutaland, B., Guidot, A., Cellier, G., Wicker, E., Allen, C., Fegan, M., Pruvost, O., Elbaz, M., Calteau, A., Salvignol, G., Mornico, D., Mangenot, S., Barbe, V., Médigue, C., and Prior, P. (2010). Genomes of three tomato pathogens within the *Ralstonia solanacearum* species complex reveal significant evolutionary divergence. *BMC Genomics*, 11:379.
- Renne, P. R., Deino, A. L., Hilgen, F. J., Kuiper, K. F., Mark, D. F., Mitchell, W. S., Morgan, L. E., Mundil, R., and Smit, J. (2013). Time scales of critical events around the Cretaceous-Paleogene boundary. *Science*, 339(6120):684–687.
- Reno, M. L., Held, N. L., Fields, C. J., Burke, P. V., and Whitaker, R. J. (2009). Biogeography of the *Sulfolobus islandicus* pan-genome. *Proceedings of the National Academy of Sciences*, 106(21):8605–8610.
- Retchless, A. C. and Lawrence, J. G. (2010). Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. *PNAS*, 107(25):11453–11458.

- Rigali, S., Titgemeyer, F., Barends, S., Mulder, S., Thomae, A. W., Hopwood, D. A., and van Wezel, G. P. (2008). Feast or famine: the global regulator DasR links nutrient stress to antibiotic production by *Streptomyces*. *EMBO reports*, 9(7):670–675.
- Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., Rubin, E. M., Hugenholtz, P., and Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437.
- Roberts, M. F. (2005). Organic compatible solutes of halotolerant and halophilic microorganisms. *Saline systems*, 1:5.
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.
- Rosing, M. (1999). ^{13}C -Depleted carbon microparticles in ~ 3700 -Ma sea-floor sedimentary rocks from west greenland. *Science*, 283(5402):674–676.
- Saito, A., Shinya, T., Miyamoto, K., Yokoyama, T., Kaku, H., Minami, E., Shibuya, N., Tsujibo, H., Nagata, Y., Ando, A., Fujii, T., and Miyashita, K. (2007). The dasABC Gene Cluster, Adjacent to dasR, Encodes a Novel ABC Transporter for the Uptake of N,N'-Diacetylchitobiose in *Streptomyces coelicolor* A3(2). *Applied and Environmental Microbiology*, 73(9):3000–3008.
- Scanlan, D. J. and West, N. J. (2002). Molecular ecology of the marine cyanobacterial genera *Prochlorococcus* and *Synechococcus*. *FEMS microbiology ecology*, 40(1):1–12.
- Schrempf, H. (2001). Recognition and degradation of chitin by streptomycetes. *Antonie van Leeuwenhoek*, 79(3-4):285–289.
- Schuster, C. F., Bellows, L. E., Tosi, T., Campeotto, I., Corrigan, R. M., Freemont, P., and Gründling, A. (2016). The second messenger c-di-AMP inhibits the osmolyte uptake system OpuC in *Staphylococcus aureus*. *Science Signaling*, 9(441):ra81–ra81.
- Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabo, G., Polz, M. F., and Alm, E. J. (2012). Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science*, 336(6077):48–51.
- Shapiro, B. J. and Polz, M. F. (2015). Microbial Speciation. *Cold Spring Harbor perspectives in biology*, 7(10):a018143.
- Simpson, G. G. (1951). The Species Concept. *Evolution*, 5(4):285–298.

- Soucy, S. M., Huang, J., and Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Publishing Group*, 16(8):472–482.
- Spero, M. A., Aylward, F. O., Currie, C. R., and Donohue, T. J. (2015). Phylogenomic analysis and predicted physiological role of the proton-translocating NADH:quinone oxidoreductase (complex I) across bacteria. *mBio*, 6(2).
- Stach, E. M. and Bull, A. T. (2005). Estimating and comparing the diversity of marine actinobacteria. *Antonie van Leeuwenhoek*, 87(1):3–9.
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.
- Stojiljkovic, I., Bäumler, A. J., and Hantke, K. (1994). Fur regulon in gram-negative bacteria. Identification and characterization of new iron-regulated *Escherichia coli* genes by a fur titration assay. *Journal of molecular biology*, 236(2):531–545.
- Sukharev, S. I., Blount, P., Martinac, B., and Kung, C. (1997). Mechanosensitive channels of *Escherichia coli*: the MscL gene, protein, and activities. *Annual review of physiology*, 59:633–657.
- Tamura, K., Battistuzzi, F. U., Billings-Ross, P., Murillo, O., Filipinski, A., and Kumar, S. (2012). Estimating divergence times in large molecular phylogenies. *PNAS*, 109(47):19333–19338.
- Tettelin, H. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *PNAS*, 102(39):13950–13955.
- Thomas, C. M. and Nielsen, K. M. (2005). Mechanisms of and Barriers to, Horizontal Gene Transfer between Bacteria. *Nature Reviews Microbiology*, 3(9):711–721.
- Touati, D. (1988). Transcriptional and posttranscriptional regulation of manganese superoxide dismutase biosynthesis in *Escherichia coli*, studied with operon and protein fusions. *Journal of Bacteriology*, 170(6):2511–2520.
- Tutar, Y. (2012). Pseudogenes. *Comparative and functional genomics*, 2012:424526.
- Udwary, D. W., Zeigler, L., Asolkar, R. N., Singan, V., Lapidus, A., Fenical, W., Jensen, P. R., and Moore, B. S. (2007). Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *Proceedings of the National Academy of Sciences*, 104(25):10376–10381.
- Valenzuela, L. M., Hidalgo, A. A., Rodríguez, L., Urrutia, I. M., Ortega, A. P., Villagra, N. A., Paredes-Sabja, D., Calderón, I. L., Gil, F., Saavedra, C. P., Mora, G. C., and Fuentes, J. A. (2015). Infection, Genetics and Evolution. *Infection, genetics and*

- evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 33(C):131–142.
- van Munster, J. M., Nitsche, B. M., Krijgheld, P., van Wijk, A., Dijkhuizen, L., Wosten, H. A., Ram, A. F., and van der Maarel, M. J. E. C. (2013). Chitinases CtcB and CfcI modify the cell wall in sporulating aerial mycelium of *Aspergillus niger*. *Microbiology*, 159(Pt9) : 1853 – –1867.
- Van Valen, L. (1976). Ecological Species, Multispecies, and Oaks. *Taxon*, 25(2/3):233–239.
- Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G. F., Chater, K. F., and van Sinderen, D. (2007). Genomics of Actinobacteria: Tracing the Evolutionary History of an Ancient Phylum. *Microbiology and molecular biology reviews : MMBR*, 71(3):495–548.
- Vernikos, G., Medini, D., Riley, D. R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology*, 23:148–154.
- Vital, M., Chai, B., stman, B. o. r. O., Cole, J., Konstantinidis, K. T., and Tiedje, J. M. (2014). Gene expression analysis of *E. coli* strains provides insights into the role of gene regulation in diversification. 9(5):1130–1140.
- Vos, M. (2009). Why do bacteria engage in homologous recombination? *Trends in Microbiology*, 17(6):226–232.
- Whitaker, R. J., Grogan, D. W., and Taylor, J. W. (2003). Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science (New York, NY)*, 301(5635):976–978.
- Williams, P. G., Buchanan, G. O., Feling, R. H., Kauffman, C. A., Jensen, P. R., and Fenical, W. (2005). New cytotoxic salinosporamides from the marine Actinomycete *Salinispora tropica*. *The Journal of organic chemistry*, 70(16):6196–6203.
- Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090.
- Yawata, Y., Cordero, O. X., Menolascina, F., Hehemann, J.-H., Polz, M. F., and Stocker, R. (2014). Competition-dispersal tradeoff ecologically differentiates recently speciated marine bacterioplankton populations. *Proceedings of the National Academy of Sciences*, 111(15):5622–5627.
- Ye, J.-Z., Lin, X.-M., Cheng, Z.-X., Su, Y.-B., Li, W.-X., Ali, F.-M., Zheng, J., and Peng, B. (2018). Identification and efficacy of glycine, serine and threonine metabolism in potentiating kanamycin-mediated killing of *Edwardsiella piscicida*. *Journal of proteomics*,

183:34–44.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–829.

Ziemert, N., Lechner, A., Wietz, M., Millan-Aguinaga, N., Chavarria, K. L., and Jensen, P. R. (2014). Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *PNAS*, 111(12):E1130–E1139.

Zinder, N. D. and Lederberg, J. (1952). Genetic exchange in *Salmonella*. *Journal of Bacteriology*, 64(5):679–699.

Appendix A

***Salinispora* strain metadata**

IMG Genome ID	NCBI Taxon ID	Species	Strain	Sample	Sequence type	Location	Date collected	Latitude	Longitude	Depth (m)	Isolated from
2515154183	1137251	<i>S. arenicola</i>	CNH-905	BA00-16	ST	Bahamas	August 2000	24° 01.133 N	74° 32.669 W	30	Marine sediment
2517572153	1169162	<i>S. arenicola</i>	CNY-011	BA10-262-A	ST	Bahamas	July 2010	26° 33.838 N	77° 53.436 W	26	Marine sediment
2517572210	1136425	<i>S. arenicola</i>	CNB-458	BA89-42	ST	Bahamas	June 1989	26° 37.57 N	77° 55.19 W	NA	Marine sediment
2515154093	1137250	<i>S. arenicola</i>	CNB-527	BA89-48	ST	Bahamas	1989	22° 21 N	74° 1 W	NA	Marine sediment
2515154181	1136427	<i>S. arenicola</i>	CNH-646	BA93-3-30-C	ST	Bahamas	1999	26° 36 N	77° 53 W	15	Marine sediment
2519103192	1169176	<i>S. arenicola</i>	CNH-877	BA00-18	ST	Bahamas	August 2000	24° 01.133 N	74° 32.669 W	12	Marine sediment
2561511037	1408326	<i>S. arenicola</i>	CNH-643	BA99-1-0-A	ST	Bahamas	July 1999	NA	NA	NA	Marine sediment
2571042016	1408334	<i>S. arenicola</i>	CNS-342	036F	ST	Fiji	August 2004	NA	NA	NA	Marine sediment
2521172655	1169164	<i>S. arenicola</i>	CNY-282	USP-016	ST	Fiji	June 2011	NA	NA	NA	Sponge
2561511110	1408337	<i>S. arenicola</i>	CNY-244	FJ09-FS068	ST	Fiji	October 2009	16° 40.31 S	179° 52.25 W	NA	Marine sediment
2561511115	1408336	<i>S. arenicola</i>	CNY-230	USP-017	ST	Fiji	June 2011	NA	NA	NA	Sponge
2516143022	999546	<i>S. arenicola</i>	CNS-991	FJ06-126-4	ST	Fiji	July 2006	18° 45.667 S	178° 33.829 E	20-24	Marine sediment
2518285558	1169179	<i>S. arenicola</i>	CNY-231	FJ07-FS-057	ST	Fiji	September 2007	17° 39.462 S	178° 50.608 E	NA	Marine sediment
2519103195	1169165	<i>S. arenicola</i>	CNY-234	USP-26	ST	Fiji	June 2011	NA	NA	NA	Sponge
2517572163	1169185	<i>S. arenicola</i>	CNY-237	FJ08-SS-079	ST	Fiji	February 2008	17° 16.402 S	177° 06.052 E	NA	Marine sediment
2518285559	1169166	<i>S. arenicola</i>	CNY-256	USP-24	ST	Fiji	June 2011	NA	NA	NA	Sponge
2518285560	1169177	<i>S. arenicola</i>	CNY-260	FJ07-FS-023	ST	Fiji	August 2007	17° 59.778 S	179° 11.284 E	NA	Marine sediment
2519103185	1144930	<i>S. arenicola</i>	CNS-673	FJ06-126-3	ST	Fiji	July 2006	18° 45.667 S	178° 33.829 E	20-24	Marine sediment
2518285554	1169168	<i>S. arenicola</i>	CNS-744	FJ06-125-7	ST	Fiji	July 2006	18° 45.342 S	178° 31.425 E	4-6	Marine sediment
2517572137	1137255	<i>S. arenicola</i>	CNT-005	FJ06-153-3	ST	Fiji	July 2006	18° 46.813 S	178° 33.019 E	45	Marine sediment
2517572154	1169173	<i>S. arenicola</i>	CNY-280	FJ07-FS-025	ST	Fiji	August 2007	18° 01.169 S	179° 14.138 E	NA	Marine sediment
2565956528	1408335	<i>S. arenicola</i>	CNS-820	FJ06-71	ST	Fiji	July 2006	18° 42.806 S	178° 29.438 E	18-25	Marine sediment
2571042345	1408351	<i>S. arenicola</i>	CNS-848	FJ06-31	ST	Fiji	July 2006	18° 24.374 S	178° 09.855 E	30	Marine sediment
2561511103	1408339	<i>S. arenicola</i>	CNY-281	FJ09-FS079	G	Fiji	September 2009	17° 46.18 S	179° 23.44 W	NA	Marine sediment
2561511039	1408331	<i>S. arenicola</i>	CNQ-884	GU02-222	ST	Guam	January 2002	13° 17.545 N	144° 38.793 E	86	Marine sediment
2515154180	1144929	<i>S. arenicola</i>	CNQ-748	GU02-241-2	ST	Guam	January 2002	13° 21.9 N	144° 38.8 E	NA	Marine sediment
2519103194	1169167	<i>S. arenicola</i>	CNR-107	GU02-313-7	ST	Guam	January 2002	13° 18.362 N	144° 39.026 E	36	Marine sediment
2515154186	1137256	<i>S. arenicola</i>	CNT-798	HA08-2-1C	ST	Hawaii	December 2008	20° 38.086° N	156° 29.626° W	30-33	Marine sediment
2526164509	1169172	<i>S. arenicola</i>	CNT-799	HA08-11	ST	Hawaii	December 2008	20° 38.086° N	156° 29.626° W	30-33	Marine sediment
2515154088	1137253	<i>S. arenicola</i>	CNT-800	HA08-8-1B	ST	Hawaii	December 2008	20° 38.086° N	156° 29.626° W	30-33	Marine sediment
2518285550	1137264	<i>S. arenicola</i>	CNT-849	HA08-5	ST	Hawaii	December 2008	20° 38.086° N	156° 29.626° W	30-33	Marine sediment
2515154135	1136428	<i>S. arenicola</i>	CNT-850	HA08-11-1A	ST	Hawaii	December 2008	20° 38.086° N	156° 29.626° W	30-33	Marine sediment
2515154127	1137254	<i>S. arenicola</i>	CNT-857	HA08-3-1-II	ST	Hawaii	December 2008	20° 38' 22.69° N	156° 27' 01.47° W	15	Marine sediment
2517572233	1169163	<i>S. arenicola</i>	CNT-859	HA08-28-1C	ST	Hawaii	December 2008	20° 38' 22.69° N	156° 27' 01.47° W	15	Marine sediment
2515154203	1136429	<i>S. arenicola</i>	CNR-921	PL04-094-2R	ST	Palau	March 2004	07° 09.635 N	134° 22.127 E	12	Marine sediment
2518285553	1169178	<i>S. arenicola</i>	CNS-051	PL04-020-3B	ST	Palau	March 2004	07° 09.09 N	134° 21.34 E	5-13	Marine sediment
2524614515	1288087	<i>S. arenicola</i>	CNS-243	PL04-010	ST	Palau	March 2004	07° 17.704 N	134° 30.950 E	365	Marine sediment
2565956527	1408332	<i>S. arenicola</i>	CNS-296	PL04-197	ST	Palau	March 2004	07° 10.215 N	134° 21.165 E	36	Marine sediment
2524614529	1288086	<i>S. arenicola</i>	CNS-299	PL04-100	ST	Palau	March 2004	07° 14.431 N	134° 22.855 E	9	Marine sediment
2571042009	1408333	<i>S. arenicola</i>	CNS-325	PL04-276	ST	Palau	March 2004	07° 19.590 N	134° 26.215 E	30	Marine sediment

IMG Genome ID	NCBI Taxon ID	Species	Strain	Sample	Sequence type	Location	Date collected	Latitude	Longitude	Depth (m)	Isolated from
641228504	391037	<i>S. arenicola</i>	CNS-205	PL04-172-7D	ST	Palau	March 2004	07° 06.500 N	134° 15.200 E	15-20	Marine sediment
251828555	1169169	<i>S. arenicola</i>	CNX-481	PL09-108-BII	ST	Palmyra	September 2009	05° 52.178 N	162°04.647 E	15-25	Marine sediment
2515154137	1136426	<i>S. arenicola</i>	CNX-482	PL09-270-B	ST	Palmyra	September 2009	05° 53.545 N	162°04.609 E	3	Marine sediment
2515154188	1137257	<i>S. arenicola</i>	CNX-508	PL09-21-B	ST	Palmyra	September 2009	05° 52.233 N	162° 07.491 E	7	Marine sediment
2517572152	1169171	<i>S. arenicola</i>	CNX-814	PL09-122-G	ST	Palmyra	September 2009	05° 52.178 N	162°04.647 E	15-25	Marine sediment
2515154187	1137258	<i>S. arenicola</i>	CNX-891	PL09-106-D	ST	Palmyra	September 2009	05° 52.178 N	162°04.647 E	15-25	Marine sediment
2561511104	168697	<i>S. arenicola</i>	CNH-996	SC01-75	A	Sea of Cortez	February 2001	24° 49.49 N	110° 35.16 W	36	Marine sediment
2571042014	1408330	<i>S. arenicola</i>	CNH-996B	SC01-75	A	Sea of Cortez	February 2001	24° 49.49 N	110° 35.16 W	36	Marine sediment
2524023246	1288085	<i>S. arenicola</i>	CNH-963	SC00-90-BIII	A	Sea of Cortez	November 2000	NA	NA	NA	Marine sediment
2519103193	1169175	<i>S. arenicola</i>	CNH-962	SC00-60-BIII	A	Sea of Cortez	November 2000	NA	NA	NA	Marine sediment
2515154193	1137252	<i>S. arenicola</i>	CNH-941	SC00	B	Sea of Cortez	November 2000	NA	NA	NA	Marine sediment
2515154125	1136430	<i>S. arenicola</i>	CNH-964	SC00-90-BIII	B	Sea of Cortez	November 2000	NA	NA	NA	Marine sediment
2518285551	1169174	<i>S. arenicola</i>	CNP-105	SC01-69	B	Sea of Cortez	February 2001	24° 51.37 N	110° 35.32 W	21	Marine sediment
2518285552	1169170	<i>S. arenicola</i>	CNP-193	SC01-108	B	Sea of Cortez	February 2001	24° 25.71 N	110° 25.52 W	48	Marine sediment
2561511107	1408340	<i>S. arenicola</i>	CNY-486	PV11-10	A	Puerto Vallarta	December 2011	20° 32.687 N	105° 17.380 W	1-7	Marine sediment
2561511113	1408342	<i>S. arenicola</i>	CNY-679	MX12-56	ST	Yucatan	July 2012	24° 40.165 N	82° 54.509 W	20	Marine sediment
2563366734	1408343	<i>S. arenicola</i>	CNY-685	MX12-266	ST	Yucatan	July 2012	18° 24.373 N	87° 24.751 W	10	Marine sediment
2561511111	1408344	<i>S. arenicola</i>	CNY-690	MX12-252	ST	Yucatan	July 2012	18° 33.860 N	87° 25.275 W	12	Marine sediment
2561511114	1408345	<i>S. arenicola</i>	CNY-694	MX12-217	ST	Yucatan	July 2012	20° 19.479 N	87° 01.627 W	15	Marine sediment
2571042007	1408327	<i>S. arenicola</i>	CNH-713	RS00-154	ST	Red Sea	2000	NA	NA	1	Marine sediment
2561511105	1408328	<i>S. arenicola</i>	CNH-718	RS00-47	ST	Red Sea	2000	27° 34 N	33° 55.8 E	15	Marine sediment
2528311033	1298919	<i>S. arenicola</i>	CNR-425	GU02-184	ST	Guam	January 2002	13° 16.774 N	144° 39.225 E	106	Marine sediment
2515154178	1137260	<i>S. pacifica</i>	CNR-114	GU02-228-2	ST	Guam	January 2002	13° 17.209 N	114° 38.943 E	12	Marine sediment
2517572155	1169193	<i>S. pacifica</i>	CNQ-768	GU02-266-11	ST	Guam	January 2002	13° 31.099 N	114° 47.44 E	180	Marine sediment
2571042008	1408348	<i>S. pacifica</i>	CNR-510	GU02-230	ST	Guam	January 2002	13.253171° N	144.658865° E	NA	Marine sediment
2571042006	1408347	<i>S. pacifica</i>	CNH-732	RS00-470	ST	Red Sea	2000	24° 22.55 N	35° 23.03 E	18	Marine sediment
2563366531	1408356	<i>S. pacifica</i>	CNY-646	EG-55	O	Red Sea	August 2006	NA	NA	NA	Sponge
2561511035	1408352	<i>S. pacifica</i>	CNT-133A	FJ06-80	D	Fiji	July 2006	18° 42.806 S	178° 29.438 E	18-25	Marine sediment
2561511036	1408350	<i>S. pacifica</i>	CNS-801	FJ06-84	A	Fiji	July 2006	18° 42.806 S	178° 29.438 E	18-25	Marine sediment
2518285563	1169186	<i>S. pacifica</i>	CNS-860	FJ06-80-3	C	Fiji	July 2006	18° 42.806 S	178° 29.438 E	18-25	Marine sediment
2517572194	999542	<i>S. pacifica</i>	DSM45543	FJ06-54-2	C	Fiji	July 2006	18° 23.946 S	178° 02.086 E	10-15	Marine sediment
2517287019	999543	<i>S. pacifica</i>	DSM45544	FJ06-87-1	ST	Fiji	July 2006	18° 43.414 S	178° 29.423 E	7	Marine sediment
2517572157	1169189	<i>S. pacifica</i>	CNS-996	FJ06-137-3	C	Fiji	July 2006	18° 45.667 S	178° 33.829 E	20-24	Marine sediment
2515154184	1136416	<i>S. pacifica</i>	CNT-001	FJ06-147-3	ST	Fiji	July 2006	18° 44.602 S	178° 32.502 E	33	Marine sediment
2515154126	1136417	<i>S. pacifica</i>	CNT-003	FJ06-152-1	ST	Fiji	July 2006	18° 46.427 S	178° 32.784 E	43	Marine sediment
2515154177	1136418	<i>S. pacifica</i>	CNT-029	FJ06-30-5	C	Fiji	July 2006	18° 24.626 S	178° 09.494 E	37	Marine sediment
2517572158	1169190	<i>S. pacifica</i>	CNT-045	FJ06-135-1	C	Fiji	July 2006	18° 45.667 S	178° 33.829 E	20-24	Marine sediment
2515154202	1136419	<i>S. pacifica</i>	CNT-084	FJ06-58-1	D	Fiji	July 2006	18° 24.360 S	178° 01.089 E	10-15	Marine sediment
2517572159	1169188	<i>S. pacifica</i>	CNT-124	FJ06-153-5	C	Fiji	July 2006	18° 46.813 S	178° 33.019 E	45	Marine sediment
2515154200	1136420	<i>S. pacifica</i>	CNT-131	FJ06-76-4	ST	Fiji	July 2006	18° 42.806 S	178° 29.438 E	18-25	Marine sediment

IMG Genome ID	NCBI Taxon ID	Species	Strain	Sample	Sequence type	Location	Date collected	Latitude	Longitude	Depth (m)	Isolated from
2516493032	1050199	<i>S. pacifica</i>	DSM45547	FJ06-138-7	C	Fiji	July 2006	18° 45.667' S	178° 33.829' E	20-24	Marine sediment
2517287023	999544	<i>S. pacifica</i>	DSM45548	FJ06-154-9	A	Fiji	July 2006	18° 47.151' S	178° 33.155' E	43	Marine sediment
2517434008	999545	<i>S. pacifica</i>	DSM45549	FJ06-32-1	B	Fiji	July 2006	18° 25.301' S	178° 08.453' E	35	Marine sediment
2515154124	1137263	<i>S. pacifica</i>	CNT-569	FJ08-173-2	E	Fiji	February 2008	18° 15.26' S	178° 05.10' E	NA	Marine sediment
2517572160	1169191	<i>S. pacifica</i>	CNT-584	FJ08-333-1b	C	Fiji	January 2008	18° 15.26' S	178° 05.10' E	NA	Marine sediment
2515154185	1136424	<i>S. pacifica</i>	CNT-603	FJ08-178-53	ST	Fiji	January 2008	18° 15.26' S	178° 05.10' E	NA	Marine sediment
2517572161	1169184	<i>S. pacifica</i>	CNT-609	FJ08-273-44	D	Fiji	January 2008	17° 15.591' S	177° 06.478' E	NA	Marine sediment
2561511034	1408353	<i>S. pacifica</i>	CNT-403	FJ08-41	C	Fiji	January 2008	16.56.827' S	177 24.020' E	20	Marine sediment
2524614561	1288090	<i>S. pacifica</i>	CNT-239	FJ08-FS-080	ST	Fiji	September 2008	17° 44.31' S	179° 20.53' W	NA	Marine sediment
2515154170	1137265	<i>S. pacifica</i>	CNT-854	HA08-36-1A	C	Hawaii	December 2008	20° 38' 22.69' N	156° 27' 01.47' W	15	Marine sediment
2515154128	1137266	<i>S. pacifica</i>	CNT-855	HA08-11-1E	A	Hawaii	December 2008	20° 38.086' N	156° 29.626' W	30-33	Marine sediment
2515154182	1206102	<i>S. pacifica</i>	CNT-796	HA08-4-1C	D	Hawaii	December 2008	20° 38.086' N	156° 29.626' W	30-33	Marine sediment
2517572162	1169180	<i>S. pacifica</i>	CNT-851	HA08-11-1D	D	Hawaii	December 2008	20° 38.086' N	156° 29.626' W	30-33	Marine sediment
2524614807	1288089	<i>S. pacifica</i>	CNS-237	PL04-118	B	Palau	March 2004	07° 21.312' N	134° 26.409' E	22	Marine sediment
2518285562	1169182	<i>S. pacifica</i>	CNS-055	PL04-123-3C	A	Palau	March 2004	07° 17.092' N	134° 13.825' E	457	Marine sediment
2515154194	1137261	<i>S. pacifica</i>	CNR-894	PL04-008-2F	ST	Palau	March 2004	07° 17.894' N	134° 30.252' E	304	Marine sediment
2518285561	1169187	<i>S. pacifica</i>	CNR-942	PL04-003-1A	E	Palau	March 2004	07° 16' N	134° 28' E	45	Marine sediment
2515154129	1137262	<i>S. pacifica</i>	CNS-103	PL04-124-1C	ST	Palau	March 2004	07° 18.079' N	134° 13.449' E	457	Marine sediment
2561511038	1408349	<i>S. pacifica</i>	CNR-909	PL04-08	ST	Palau	March 2004	07° 17.894' N	134° 30.252' E	304	Marine sediment
2518645626	1169192	<i>S. pacifica</i>	CNY-330	SC-08-27	ST	Palau	March 2004	28.98995' N	113.3987167' W	20	Marine sediment
2518645627	1169181	<i>S. pacifica</i>	CNY-331	SC-08-28	ST	Sea of Cortez	July 2008	28.95318333' N	113.4308667' W	20	Marine sediment
2563366534	1408354	<i>S. pacifica</i>	CNY-363	AMS-72	ST	Sea of Cortez	July 2008	28.95318333' N	113.4308667' W	20	Marine sediment
2528311034	1305843	<i>S. pacifica</i>	CNY-202	AMS-301	K2	Sea of Cortez	July 2008	25.9503217' N	111.306283' W	330	Marine sediment
2563366539	1408355	<i>S. pacifica</i>	CNY-498	PV11-27	ST	Puerto Vallarta	December 2011	20° 42.099' N	105° 33.888' W	15	Marine sediment
2563366532	1408341	<i>S. pacifica</i>	CNY-666	MD12-107A	A	Madeira Islands	June 2012	32° 38.901' N	16° 49.365' W	15	Marine sediment
2563366533	1408357	<i>S. pacifica</i>	CNY-673	MD12-278	Q	Madeira Islands	June 2012	33° 03.155' N	16° 16.700' W	16	Marine sediment
2563366517	1408346	<i>S. pacifica</i>	CNY-703	MD12-562-A	P	Madeira Islands	June 2012	32° 32.193' N	16° 31.971' W	12	Marine sediment
640427140	369723	<i>S. tropica</i>	CNB-440	BA89-40E	ST	Bahamas	August 1989	25.405212° N	77.881818° W	6	Marine sediment
2517572211	1137247	<i>S. tropica</i>	CNB-476	BA89-43	ST	Bahamas	August 1989	26.4667° N	77.0833° W	18	Marine sediment
2517572212	1136431	<i>S. tropica</i>	CNB-536	BA89-49C	ST	Bahamas	August 1989	22.3500° N	74.0167° W	6	Marine sediment
2515154094	1137248	<i>S. tropica</i>	CNH-898	BA00-11	ST	Bahamas	July 2000	24° 32.693' N	75° 55.724' W	30	Marine sediment
2515154163	1137249	<i>S. tropica</i>	CNS-197	BA04-14-4B	ST	Bahamas	June 2004	26° 34.288' N	77° 53.207' W	5	Marine sediment
2517572164	1169195	<i>S. tropica</i>	CNS-416	BA04-26-6B	ST	Bahamas	June 2004	25° 49.658' N	77° 52.850' W	24	Marine sediment
2518645624	1169198	<i>S. tropica</i>	CNR-699	BA03-04	ST	Bahamas	July 2003	26° 33.527' N	77° 53.096' W	15	Marine sediment
2540341193	1169196	<i>S. tropica</i>	CNT-250	BA07-63	ST	Bahamas	June 2007	NA	NA	12	Marine sediment
2524614530	1288084	<i>S. tropica</i>	CNT-261	BA07-67	ST	Bahamas	June 2007	NA	NA	700	Marine sediment
2540341192	1169197	<i>S. tropica</i>	CNY-012	BA10-263	ST	Bahamas	July 2010	26° 33.838' N	77° 53.436' W	8	Marine sediment
2561511109	1408358	<i>S. tropica</i>	CNY-678	MX12-141	ST	Yucatan	July 2012	21° 10.262' N	86° 43.812' W	12	Marine sediment
2561511108	1408359	<i>S. tropica</i>	CNY-681	MX12-215	ST	Yucatan	July 2012	20° 19.479' N	87° 01.627' W	15	Marine sediment

Appendix B

Species-specific Core Genes for *S. arenicola* and *S. tropica*

Salinispora arenicola-specific Core

Salin Group	COG Function	Annotation of Salin Group	CatCode	Category
Salin4380	COG1151	6Fe-6S prismane cluster-containing protein	C	Energy production and conversion
Salin4303	COG1141	Ferredoxin	C	Energy production and conversion
Salin4370	COG1032	Fe-S oxidoreductase	C	Energy production and conversion
Salin4012	COG1018	Flavodoxin reductases (ferredoxin-NADPH reductases) family 1	C	Energy production and conversion
Salin4130	COG0667	Predicted oxidoreductases (related to aryl-alcohol dehydrogenases)	C	Energy production and conversion
Salin4354	COG0644	Dehydrogenases (flavoproteins)	C	Energy production and conversion
Salin4162	COG0604	NADPH:quinone reductase and related Zn-dependent oxidoreductases	C	Energy production and conversion
Salin4394	COG4176	ABC-type proline/glycine betaine transport system, permease component	E	Amino acid transport and metabolism
Salin4301	COG4175	ABC-type proline/glycine betaine transport system, ATPase component	E	Amino acid transport and metabolism
Salin4230	COG3200	3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase	E	Amino acid transport and metabolism
Salin4310	COG2113	ABC-type proline/glycine betaine transport systems, periplasmic components	E	Amino acid transport and metabolism
Salin4099	COG0757	3-dehydroquininate dehydratase II	E	Amino acid transport and metabolism
Salin4161	COG0747	ABC-type dipeptide transport system, periplasmic component	E	Amino acid transport and metabolism
Salin4384	COG0747	ABC-type dipeptide transport system, periplasmic component	E	Amino acid transport and metabolism
Salin4255	COG0665	Glycine/D-amino acid oxidases (deaminating)	E	Amino acid transport and metabolism
Salin4361	COG0346	Lactoylglutathione lyase and related lyases	E	Amino acid transport and metabolism
Salin4248	COG0346	Lactoylglutathione lyase and related lyases	E	Amino acid transport and metabolism
Salin4283	COG0169	Shikimate 5-dehydrogenase	E	Amino acid transport and metabolism
Salin4381	COG0028	Thiamine pyrophosphate-requiring enzymes [acetolactate synthase, pyruvate dehydrogenase (cytochrome), glyoxylate carboligase, phosphonopyruvate decarboxylase]	E	Amino acid transport and metabolism
Salin4025	COG0028	Thiamine pyrophosphate-requiring enzymes [acetolactate synthase, pyruvate dehydrogenase (cytochrome), glyoxylate carboligase, phosphonopyruvate decarboxylase]	E	Amino acid transport and metabolism
Salin4010	COG0006	Xaa-Pro aminopeptidase	E	Amino acid transport and metabolism
Salin4172	COG3959	Transketolase, N-terminal subunit	G	Carbohydrate transport and metabolism
Salin4288	COG3958	Transketolase, C-terminal subunit	G	Carbohydrate transport and metabolism
Salin3992	COG3387	Glucoamylase and related glycosyl hydrolases	G	Carbohydrate transport and metabolism
Salin4260	COG2140	Thermophilic glucose-6-phosphate isomerase and related metalloenzymes	G	Carbohydrate transport and metabolism
Salin4024	COG1819	Glycosyl transferases, related to UDP-glucuronosyltransferase	G	Carbohydrate transport and metabolism
Salin4015	COG1653	ABC-type sugar transport system, periplasmic component	G	Carbohydrate transport and metabolism
Salin4113	COG1175	ABC-type sugar transport systems, permease components	G	Carbohydrate transport and metabolism
Salin3996	COG0395	ABC-type sugar transport system, permease component	G	Carbohydrate transport and metabolism

Salin Group	COG Function	Annotation of Salin Group	CatCode	Category
Salin4396	COG2227	2-polyprenyl-3-methyl-5-hydroxy-6-methoxy-1,4-benzoquinol methylase	H	coenzyme transport and metabolism
Salin4011	COG2154	Pterin-4a-carbinolamine dehydratase	H	coenzyme transport and metabolism
Salin4228	COG0543	2-polyprenylphenol hydroxylase and related flavodoxin oxidoreductases	H	coenzyme transport and metabolism
Salin4275	COG0161	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	H	coenzyme transport and metabolism
Salin4300	COG3255	Putative sterol carrier protein	I	Lipid transport and metabolism
Salin4229	COG2084	3-hydroxyisobutyrate dehydrogenase and related beta-hydroxyacid dehydrogenases	I	Lipid transport and metabolism
Salin4183	COG1960	Acyl-CoA dehydrogenases	I	Lipid transport and metabolism
Salin4306	COG1597	Sphingosine kinase and enzymes related to eukaryotic diacylglycerol kinase	I	Lipid transport and metabolism
Salin4259	COG0332	3-oxoacyl-[acyl-carrier-protein] synthase III	I	Lipid transport and metabolism
Salin4395	COG0332	3-oxoacyl-[acyl-carrier-protein] synthase III	I	Lipid transport and metabolism
Salin4385	COG0236	Acyl carrier protein	I	Lipid transport and metabolism
Salin4302	COG1670	Acetyltransferases, including N-acetylases of ribosomal proteins	J	Translation, ribosomal structure and biogenesis
Salin4342	COG0143	Methionyl-tRNA synthetase	J	Translation, ribosomal structure and biogenesis
Salin4359	COG5662	Predicted transmembrane transcriptional regulator (anti-sigma factor)	K	Transcription
Salin4114	COG5662	Predicted transmembrane transcriptional regulator (anti-sigma factor)	K	Transcription
Salin4013	COG4977	Transcriptional regulator containing an amidase domain and an AraC-type DNA-binding HTH domain	K	Transcription
Salin4362	COG4977	Transcriptional regulator containing an amidase domain and an AraC-type DNA-binding HTH domain	K	Transcription
Salin4081	COG4977	Transcriptional regulator containing an amidase domain and an AraC-type DNA-binding HTH domain	K	Transcription
Salin4067	COG2909	ATP-dependent transcriptional regulator	K	Transcription
Salin4281	COG2771	DNA-binding HTH domain-containing proteins	K	Transcription
Salin4272	COG2188	Transcriptional regulators	K	Transcription
Salin4176	COG1940	Transcriptional regulator/sugar kinase	K	Transcription
Salin4069	COG1737	Transcriptional regulators	K	Transcription
Salin4178	COG1733	Predicted transcriptional regulators	K	Transcription
Salin3991	COG1609	Transcriptional regulators	K	Transcription
Salin4106	COG1609	Transcriptional regulators	K	Transcription
Salin4076	COG1595	DNA-directed RNA polymerase specialized sigma subunit, sigma24 homolog	K	Transcription
Salin4372	COG1595	DNA-directed RNA polymerase specialized sigma subunit, sigma24 homolog	K	Transcription
Salin4058	COG1309	Transcriptional regulator	K	Transcription
Salin4322	COG1309	Transcriptional regulator	K	Transcription
Salin4174	COG0640	Predicted transcriptional regulators	K	Transcription
Salin4133	COG0583	Transcriptional regulator	K	Transcription
Salin4104	COG1793	ATP-dependent DNA ligase	L	Replication, recombination and repair
Salin4182	COG0702	Predicted nucleoside-diphosphate-sugar epimerases	M	Cell wall/membrane/envelope biogenesis
Salin4356	COG0677	UDP-N-acetyl-D-mannosaminuronate dehydrogenase	M	Cell wall/membrane/envelope biogenesis

Salin Group	COG Function	Annotation of Salin Group	CatCode	Category
Salin4166	COG0463	Glycosyltransferases involved in cell wall biogenesis	M	Cell wall/membrane/envelope biogenesis
Salin4065	COG0463	Glycosyltransferases involved in cell wall biogenesis	M	Cell wall/membrane/envelope biogenesis
Salin4355	COG0399	Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis	M	Cell wall/membrane/envelope biogenesis
Salin4392	COG0399	Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis	M	Cell wall/membrane/envelope biogenesis
Salin4261	COG0451	Nucleoside-diphosphate-sugar epimerases	MG	Cell wall/membrane/envelope biogenesis
Salin4318	COG1404	Subtilisin-like serine proteases	O	Posttranslational modification, protein turnover, chaperons
Salin4073	COG0466	ATP-dependent Lon protease, bacterial type	O	Posttranslational modification, protein turnover, chaperons
Salin4258	COG3158	K ⁺ transporter	P	Inorganic ion transport and metabolism
Salin4233	COG1055	Na ⁺ /H ⁺ antiporter NhaD and related arsenite permeases	P	Inorganic ion transport and metabolism
Salin4131	COG0753	Catalase	P	Inorganic ion transport and metabolism
Salin4128	COG0607	Rhodanese-related sulfurtransferase	P	Inorganic ion transport and metabolism
Salin4210	No COG number	Terpene synthase family, metal binding domain.	Q	Secondary metabolites biosynthesis, transport and catabolism
Salin4304	COG3882	Predicted enzyme involved in methoxymalonyl-ACP biosynthesis	Q	Secondary metabolites biosynthesis, transport and catabolism
Salin4285	COG2132	Putative multicopper oxidases	Q	Secondary metabolites biosynthesis, transport and catabolism
Salin4369	COG2124	Cytochrome P450	Q	Secondary metabolites biosynthesis, transport and catabolism
Salin4349	COG2124	Cytochrome P450	Q	Secondary metabolites biosynthesis, transport and catabolism
Salin4289	COG2124	Cytochrome P450	Q	Secondary metabolites biosynthesis, transport and catabolism
Salin4195	COG2124	Cytochrome P450	Q	Secondary metabolites biosynthesis, transport and catabolism
Salin4312	COG2124	Cytochrome P450	Q	Secondary metabolites biosynthesis, transport and catabolism
Salin4373	COG2124	Cytochrome P450	Q	Secondary metabolites biosynthesis, transport and catabolism
Salin4353	COG2124	Cytochrome P450	Q	Secondary metabolites biosynthesis, transport and catabolism
Salin4360	COG1228	Imidazolonepropionase and related amidohydrolases	Q	Secondary metabolites biosynthesis, transport and catabolism
Salin4371	COG4533	ABC-type uncharacterized transport system, periplasmic component	R	General function prediction only
Salin4388	COG4122	Predicted O-methyltransferase	R	General function prediction only
Salin4293	COG4106	Trans-aconitate methyltransferase	R	General function prediction only
Salin4277	COG4106	Trans-aconitate methyltransferase	R	General function prediction only
Salin4127	COG3568	Metal-dependent hydrolase	R	General function prediction only
Salin4059	COG3467	Predicted flavin-nucleotide-binding protein	R	General function prediction only
Salin4122	COG2823	Predicted periplasmic or secreted lipoprotein	R	General function prediction only
Salin4190	COG2823	Predicted periplasmic or secreted lipoprotein	R	General function prediction only
Salin4203	COG2229	Predicted GTPase	R	General function prediction only
Salin4009	COG2041	Sulfite oxidase and related enzymes	R	General function prediction only
Salin4242	COG2041	Sulfite oxidase and related enzymes	R	General function prediction only
Salin4188	COG2018	Uncharacterized distant relative of homeotic protein bithoraxoid	R	General function prediction only

Salin Group	COG Function	Annotation of Salin Group	CatCode	Category
Salin4340	COG0824	Predicted thioesterase	R	General function prediction only
Salin4126	COG0730	Predicted permeases	R	General function prediction only
Salin4269	COG0730	Predicted permeases	R	General function prediction only
Salin4287	COG0673	Predicted dehydrogenases and related proteins	R	General function prediction only
Salin4231	COG0673	Predicted dehydrogenases and related proteins	R	General function prediction only
Salin4132	COG0627	Predicted esterase	R	General function prediction only
Salin4180	COG0596	Predicted hydrolases or acyltransferases (alpha/beta hydrolase superfamily)	R	General function prediction only
Salin4267	COG0596	Predicted hydrolases or acyltransferases (alpha/beta hydrolase superfamily)	R	General function prediction only
Salin4026	COG0300	Short-chain dehydrogenases of various substrate specificities	R	General function prediction only
Salin4247	Hypothetical protein	Hypothetical protein	S	Function unknown
Salin4357	Hypothetical protein	Hypothetical protein	S	Function unknown
Salin4072	Protein of unknown function	Function unknown	S	Function unknown
Salin4337	Protein of unknown function	Function unknown	S	Function unknown
Salin4379	Protein of unknown function	Function unknown	S	Function unknown
Salin4345	Protein of unknown function	Function unknown	S	Function unknown
Salin4365	hypothetical protein	hypothetical protein	S	Function unknown
Salin4212	hypothetical protein	hypothetical protein	S	Function unknown
Salin4298	hypothetical protein	hypothetical protein	S	Function unknown
Salin4184	hypothetical protein	hypothetical protein	S	Function unknown
Salin4189	hypothetical protein	hypothetical protein	S	Function unknown
Salin4305	hypothetical protein	hypothetical protein	S	Function unknown
Salin4367	hypothetical protein	hypothetical protein	S	Function unknown
Salin4363	hypothetical protein	hypothetical protein	S	Function unknown
Salin4164	hypothetical protein	hypothetical protein	S	Function unknown
Salin4348	hypothetical protein	hypothetical protein	S	Function unknown
Salin4341	hypothetical protein	hypothetical protein	S	Function unknown
Salin4347	hypothetical protein	hypothetical protein	S	Function unknown
Salin4346	hypothetical protein	hypothetical protein	S	Function unknown
Salin4124	hypothetical protein	hypothetical protein	S	Function unknown
Salin4387	hypothetical protein	hypothetical protein	S	Function unknown
Salin4386	hypothetical protein	hypothetical protein	S	Function unknown
Salin4382	hypothetical protein	hypothetical protein	S	Function unknown
Salin4389	hypothetical protein	hypothetical protein	S	Function unknown
Salin4249	hypothetical protein	hypothetical protein	S	Function unknown
Salin4243	hypothetical protein	hypothetical protein	S	Function unknown
Salin4204	hypothetical protein	hypothetical protein	S	Function unknown
Salin4201	hypothetical protein	hypothetical protein	S	Function unknown
Salin4208	hypothetical protein	hypothetical protein	S	Function unknown
Salin4209	hypothetical protein	hypothetical protein	S	Function unknown
Salin4192	hypothetical protein	hypothetical protein	S	Function unknown
Salin4193	hypothetical protein	hypothetical protein	S	Function unknown
Salin4191	hypothetical protein	hypothetical protein	S	Function unknown
Salin4198	hypothetical protein	hypothetical protein	S	Function unknown
Salin4338	hypothetical protein	hypothetical protein	S	Function unknown
Salin4374	hypothetical protein	hypothetical protein	S	Function unknown
Salin4375	hypothetical protein	hypothetical protein	S	Function unknown
Salin4358	hypothetical protein	hypothetical protein	S	Function unknown
Salin4352	hypothetical protein	hypothetical protein	S	Function unknown
Salin4117	hypothetical protein	hypothetical protein	S	Function unknown
Salin4253	hypothetical protein	hypothetical protein	S	Function unknown
Salin4390	hypothetical protein	hypothetical protein	S	Function unknown
Salin4279	hypothetical protein	hypothetical protein	S	Function unknown
Salin4111	COG5617	Predicted integral membrane protein	S	Function unknown
Salin4256	COG4292	Predicted membrane protein	S	Function unknown

Salin Group	COG Function	Annotation of Salin Group	CatCode	Category
Salin4377	COG4292	Predicted membrane protein	S	Function unknown
Salin4181	COG3832	Uncharacterized conserved protein	S	Function unknown
Salin4351	COG3832	Uncharacterized conserved protein	S	Function unknown
Salin4236	COG3801	Uncharacterized protein conserved in bacteria	S	Function unknown
Salin4239	COG3222	Uncharacterized protein conserved in bacteria	S	Function unknown
Salin4063	COG2268	Uncharacterized protein conserved in bacteria	S	Function unknown
Salin4002	COG1376	Uncharacterized protein conserved in bacteria	S	Function unknown
Salin4368	COG1262	Uncharacterized conserved protein	S	Function unknown
Salin4376	No COG number	ThiS family.	S	Function unknown
Salin4200	No COG number	Pectate lyase superfamily protein	S	Function unknown
Salin4391	No COG number	NIPSNAP.	S	Function unknown
Salin4206	No COG number	Helix-turn-helix domain	S	Function unknown
Salin4343	No COG number	Carboxymuconolactone decarboxylase family.	S	Function unknown
Salin4313	No COG number	Animal haem peroxidase.	S	Function unknown
Salin4100	COG5001	Predicted signal transduction protein containing a membrane domain, an EAL and a GGDEF domain	T	Signal transduction mechanisms
Salin4185	COG4585	Signal transduction histidine kinase	T	Signal transduction mechanisms
Salin4383	COG2197	Response regulator containing a CheY-like receiver domain and an HTH DNA-binding domain	T	Signal transduction mechanisms
Salin4199	COG2197	Response regulator containing a CheY-like receiver domain and an HTH DNA-binding domain	T	Signal transduction mechanisms
Salin4205	COG1366	Anti-anti-sigma regulatory factor (antagonist of anti-sigma factor)	T	Signal transduction mechanisms
Salin4197	COG1366	Anti-anti-sigma regulatory factor (antagonist of anti-sigma factor)	T	Signal transduction mechanisms
Salin4366	COG0664	cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases	T	Signal transduction mechanisms
Salin4074	COG0642	Signal transduction histidine kinase	T	Signal transduction mechanisms
Salin4186	COG1566	Multidrug resistance efflux pump	V	Defense mechanisms
Salin4393	COG1131	ABC-type multidrug transport system, ATPase component	V	Defense mechanisms
Salin4364	COG0842	ABC-type multidrug transport system, permease component	V	Defense mechanisms
Salin4187	COG0577	ABC-type antimicrobial peptide transport system, permease component	V	Defense mechanisms

Salinispora tropica-specific Core

Salin Group	COG Function	Annotation of Salin Group	CatCode	Category
Salin6590	COG2141	Coenzyme F420-dependent N5,N10-methylene tetrahydromethanopterin reductase and related flavin-dependent oxidoreductases	C	Energy production and conversion
Salin7039	COG0584	Glycerophosphoryl diester phosphodiesterase	C	Energy production and conversion
Salin6687	COG3640	CO dehydrogenase maturation factor	D	Cell cycle control, cell division, chromosome partitioning
Salin6800	COG1063	Threonine dehydrogenase and related Zn-dependent dehydrogenases	E	Amino acid transport and metabolism
Salin7224	COG0028	Thiamine pyrophosphate-requiring enzymes	E	Amino acid transport and metabolism
Salin6797	COG2226	Methylase involved in ubiquinone/menaquinone biosynthesis	H	coenzyme transport and metabolism
Salin7225	COG0654	2-polyprenyl-6-methoxyphenol hydroxylase and related FAD-dependent oxidoreductases	H	coenzyme transport and metabolism
Salin7237	COG4799	Acetyl-CoA carboxylase, carboxyl-transferase component (subunits alpha and beta)	I	Lipid transport and metabolism
Salin7230	No COG number	transcriptional regulator, TetR family	K	Transcription
Salin7231	COG2378	Predicted transcriptional regulator	K	Transcription
Salin7234	COG1396	Predicted transcriptional regulators	K	Transcription
Salin7233	COG0789	Predicted transcriptional regulators	K	Transcription
Salin7228	COG1961	Site-specific recombinases, DNA invertase Pin homologs	L	Replication, recombination and repair
Salin7238	COG0463	Glycosyltransferases involved in cell wall biogenesis	M	Cell wall/membrane/envelope biogenesis
Salin7065	COG0463	Glycosyltransferases involved in cell wall biogenesis	M	Cell wall/membrane/envelope biogenesis
Salin7221	COG0438	Glycosyltransferase	M	Cell wall/membrane/envelope biogenesis
Salin6819	COG0438	Glycosyltransferase	M	Cell wall/membrane/envelope biogenesis
Salin7218	COG0492	Thioredoxin reductase	O	Posttranslational modification, protein turnover, chaperons
Salin7229	COG3545	Predicted esterase of the alpha/beta hydrolase fold	R	General function prediction only
Salin7235	COG0596	Predicted hydrolases or acyltransferases (alpha/beta hydrolase superfamily)	R	General function prediction only
Salin6792	COG0491	Zn-dependent hydrolases, including glyoxylases	R	General function prediction only
Salin7220	hypothetical protein	hypothetical protein	S	Function unknown
Salin7226	hypothetical protein	hypothetical protein	S	Function unknown
Salin7064	hypothetical protein	hypothetical protein	S	Function unknown
Salin7232	hypothetical protein	hypothetical protein	S	Function unknown
Salin7236	hypothetical protein	hypothetical protein	S	Function unknown
Salin7063	hypothetical protein	hypothetical protein	S	Function unknown
Salin7219	hypothetical protein	hypothetical protein	S	Function unknown
Salin6923	COG2128	Uncharacterized conserved protein	S	Function unknown
Salin7223	No COG number	BsuBI/PstI restriction endonuclease C-terminus.	S	Function unknown
Salin7227	COG4753	Response regulator containing CheY-like receiver domain and AraC-type DNA-binding domain	T	Signal transduction mechanisms
Salin7222	COG2197	Response regulator containing a CheY-like receiver domain and an HTH DNA-binding domain	T	Signal transduction mechanisms
Salin6939	COG1131	ABC-type multidrug transport system, ATPase component	V	Defense mechanisms

Appendix C

ANOVA and Tukey's Post Hoc Test for Chitinase Gene Expression

Table C.1: One-way ANOVA of chitinase gene expression between strains for both exponential and stationary phase growth. Significance values: (***) $p < 0.001$, (**) $p < 0.01$, (*) $p < 0.05$, (.) $p < .1$

ANOVA		Df	Sum Sq	Mean Sq	F value	Pr(>F)	Sig
Salin3372 Exponential	Strain	2	127.95	63.98	101.4	8.99E-05	***
	Residuals	5	3.16	0.63			
Salin3372 Stationary	Strain	2	184.8	92.4	5.072	0.0626	.
	Residuals	5	91.09	18.22			
Salin984 Exponential	Strain	2	1159.7	579.9	6.824	0.0372	*
	Residuals	5	424.9	85			
Salin984 Stationary	Strain	2	13981	6990	52.06	0.000449	***
	Residuals	5	671	134			
Salin4659 Exponential	Strain	1	0.1688	0.1688	0.437	0.556	
	Residuals	3	1.1582	0.3861			
Salin4659 Stationary	Strain	1	8.342	8.342	2.913	0.186	
	Residuals	3	8.593	2.864			

Table C.2: Tukey's HSD post hoc test to determine pairwise significance of differential expression between strains.

Tukey's Post Hoc Test	Comparison	diff	lwr	upr	p
Salin3372 Exponential	CNS-205-CNB-440	8.3766667	6.26609	10.487243	0.0001164
	CNS-991-CNB-440	0.2983333	-2.061363	2.65803	0.9125025
	CNS-991-CNS-205	-8.0783333	-10.43803	-5.718637	0.0002396
Salin3372 Stationary	CNS-205-CNB-440	9.23	-2.109707	20.56971	0.0977287
	CNS-991-CNB-440	10.78	-1.898178	23.45818	0.0854668
	CNS-991-CNS-205	1.55	-11.128178	14.22818	0.9178464
Salin984 Exponential	CNS-205-CNB-440	-21.353333	-45.84477	3.138101	0.0789741
	CNS-991-CNB-440	-28.486667	-55.86892	-1.104411	0.0434966
	CNS-991-CNS-205	-7.133333	-34.51559	20.248922	0.6927287
Salin984 Stationary	CNS-205-CNB-440	-86.223333	-117.0104	-55.43627	0.0006269
	CNS-991-CNB-440	-86.541667	-120.96265	-52.12068	0.0010418
	CNS-991-CNS-205	-0.318333	-34.73932	34.10265	0.9995009
Salin4659 Exponential	CNS-991-CNB-440	-0.375	-2.18014	1.43014	0.5557731
Salin4659 Stationary	CNS-991-CNB-440	-2.636667	-7.553368	2.280035	0.1864288

Appendix D

Differential Expression of *Salinispora*: Exponential Phase

GeneID	CNB440_JGI.ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj.pvalue	Fold	minFDR
Salin3411	640474482	COG1434	Uncharacterized conserved protein	S	Function unknown	-6.39	6.96	366.87	8.94E-77	4.30E-75	-85.00	262.62
Salin3407	640475914	NA	hypothetical protein	S	Function unknown	-5.67	5.55	177.27	2.29E-36	3.90E-35	-52.03	127.46
Salin3290	640475096	COG5178	U5 snRNP spliceosome subunit	A	RNA processing and modification	-3.78	7.05	190.40	3.52E-34	6.59E-33	-14.01	136.97
Salin0213	640475965	COG2814	Arabinose efflux permease	G	Carbohydrate transport and metabolism	-3.77	6.17	135.91	1.99E-26	2.18E-25	-14.08	97.98
Salin2629	640473917	COG3315	O-Methyltransferase involved in polyketide biosynthesis	Q	Secondary metabolites biosynthesis, transport and catabolism	-3.37	7.19	180.23	1.72E-33	3.15E-32	-10.43	129.68
Salin3284	640475390	COG2207	AraC-type DNA-binding domain-containing proteins	K	Transcription	-3.34	6.06	137.91	2.45E-29	3.16E-28	-10.22	99.40
Salin2340	640475684	COG2329	Uncharacterized enzyme involved in biosynthesis of extracellular polysaccharides	R	General function prediction only	-3.21	7.01	168.04	9.38E-26	1.22E-24	-9.66	121.01
Salin1335	640473640	COG1550	Uncharacterized protein conserved in bacteria	S	Function unknown	-3.13	5.22	87.43	5.48E-14	3.44E-13	-8.98	63.47
Salin1618	640473115	COG1846	Transcriptional regulators	K	Transcription	-2.97	4.51	52.30	1.38E-10	4.82E-10	-8.07	38.56
Salin0169	640476349	COG0513	Superfamily II DNA and RNA helicases	L	Replication, recombination and repair	-2.97	6.55	154.64	2.72E-20	2.58E-19	-8.25	111.50
Salin3204	640472923	COG2814	Arabinose efflux permease	G	Carbohydrate transport and metabolism	-2.95	4.92	69.59	1.85E-12	7.45E-12	-8.10	50.87
Salin1952	640472475	COG5459	Predicted rRNA methylase	J	Translation, ribosomal structure and biogenesis	-2.89	6.48	159.81	6.22E-29	7.87E-28	-7.64	114.99
Salin2366	640476773	COG1853	Conserved protein/domain typically associated with flavoprotein oxygenases, DIM6/NTAB family	R	General function prediction only	-2.88	5.48	81.72	1.80E-15	8.93E-15	-7.64	59.50
Salin1155	640473645	COG0534	Na ⁺ -driven multidrug efflux pump	V	Defense mechanisms	-2.88	5.34	78.25	3.89E-16	2.06E-15	-7.53	57.02
Salin0186	640472602	COG4221	Short-chain alcohol dehydrogenase of unknown specificity	R	General function prediction only	-2.82	6.05	124.23	2.50E-24	2.34E-23	-7.19	89.70
Salin1681	640474314	COG0346	Lactoylglutathione lyase and related lyases	E	Amino acid transport and metabolism	-2.80	7.72	142.85	1.67E-28	2.03E-27	-7.11	102.92
Salin2043	640475712	Unknown	Hypothetical protein	S	Function unknown	-2.74	4.84	66.47	4.71E-12	1.83E-11	-6.96	48.66
Salin1430	640472601	NA	hypothetical protein	S	Function unknown	-2.71	4.95	67.60	1.19E-13	5.19E-13	-6.70	49.47
Salin0492	640473565	COG2318	Uncharacterized protein conserved in bacteria	S	Function unknown	-2.69	7.83	132.44	8.26E-25	8.11E-24	-6.68	95.52

GeneID	CNB440_JGI_ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj.pvalue	Fold	minFDR
Salin1091	640473801	COG1232	Protoporphyrinogen oxidase	H	coenzyme transport and metabolism	-2.64	8.04	125.69	1.82E-22	1.55E-21	-6.50	90.74
Salin2450	640473173	NA	hypothetical protein	S	Function unknown	-2.62	5.44	78.28	2.45E-13	1.45E-12	-6.24	56.99
Salin2631	640472441	COG4243	Predicted membrane protein	S	Function unknown	-2.51	6.25	96.55	1.06E-20	7.99E-20	-5.78	69.97
Salin0312	640473368	COG1977	Molybdopterine converting factor, small subunit	H	coenzyme transport and metabolism	-2.46	10.18	89.54	4.94E-16	2.58E-15	-5.77	65.04
Salin0764	640476016	COG0436	Aspartate/tyrosine/aromatic aminotransferase	E	Amino acid transport and metabolism	-2.46	8.73	101.90	8.55E-20	6.01E-19	-5.64	73.79
Salin2817	640474776	COG2221	Dissimilatory sulfite reductase (desulfoviridin), alpha and beta subunits	C	Energy production and conversion	-2.45	7.63	93.97	2.07E-13	1.23E-12	-5.79	68.11
Salin1105	640475404	COG1959	Predicted transcriptional regulator	K	Transcription	-2.40	5.44	77.36	1.54E-11	7.81E-11	-5.42	56.33
Salin3247	640475452	COG1872	Uncharacterized conserved protein	S	Function unknown	-2.40	7.15	130.19	1.66E-22	1.42E-21	-5.47	93.93
Salin1232	640476489	COG5516	Conserved protein containing a Zn-ribbon-like motif, possibly RNA-binding	R	General function prediction only	-2.32	5.38	72.19	3.27E-09	1.28E-08	-5.21	52.70
Salin0461	640472556	NUDIX domain, COG0125	No COG number	S	Function unknown	-2.32	6.97	122.03	7.50E-21	5.72E-20	-5.18	88.15
Salin2518	640475836		Thymidylate kinase	F	Nucleotide transport and metabolism	-2.32	6.39	90.71	2.88E-18	1.82E-17	-5.07	65.85
Salin3007	640475178	COG3022	Uncharacterized protein conserved in bacteria	S	Function unknown	-2.27	6.30	106.40	6.72E-15	4.54E-14	-4.97	76.99
Salin3244	640472515	PAP2 superfamily, COG3832	No COG number	S	Function unknown	-2.25	5.25	62.60	8.16E-11	3.87E-10	-4.80	45.88
Salin1904	640474313		Uncharacterized conserved protein	S	Function unknown	-2.23	5.60	59.81	1.27E-07	4.14E-07	-4.94	43.93
Salin1895	640476503	COG1309	Transcriptional regulator	K	Transcription	-2.22	4.85	43.24	1.79E-09	5.61E-09	-4.73	32.15
Salin2305	640476014	COG2968	Uncharacterized conserved protein	S	Function unknown	-2.21	7.69	104.14	4.45E-17	2.51E-16	-4.87	75.46
Salin1873	640474600	COG0204	1-acyl-sn-glycerol-3-phosphate acyltransferase	I	Lipid transport and metabolism	-2.20	4.86	44.70	1.46E-06	4.26E-06	-4.74	33.24
Salin3138	640472773	COG0667	Predicted oxidoreductases (related to aryl-alcohol dehydrogenases)	C	Energy production and conversion	-2.19	8.91	117.92	1.24E-23	1.13E-22	-4.63	85.17
Salin0288	640476015	COG0287	Prephenate dehydrogenase	E	Amino acid transport and metabolism	-2.18	6.85	114.25	4.10E-23	3.64E-22	-4.60	82.57
Salin2903	640473342	COG5373	Predicted membrane protein	S	Function unknown	-2.17	5.46	52.17	4.72E-09	1.83E-08	-4.59	38.53
Salin2689	640474788	COG3573	Predicted oxidoreductase	R	General function prediction only	-2.17	5.10	52.74	6.86E-08	2.31E-07	-4.64	38.92

GeneID	CNB440_JGI_ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj_pvalue	Fold	minFDR
Salin2085	640474845	COG2847	Uncharacterized protein conserved in bacteria	S	Function unknown	-2.16	10.49	77.74	1.01E-14	4.75E-14	-4.59	56.68
Salin0431	640473800	COG0407	Uroporphyrinogen-III decarboxylase	H	coenzyme transport and metabolism	-2.15	8.97	107.00	4.69E-22	3.90E-21	-4.49	77.39
Salin0175	640475594	COG1309	Transcriptional regulator	K	Transcription	-2.15	4.36	31.03	1.48E-06	3.51E-06	-4.55	23.47
Salin2432	640473169	Unknown	Function unknown	S	Function unknown	-2.14	6.03	82.10	2.23E-09	8.89E-09	-4.66	59.73
Salin0774	640475696	COG1381	Recombinational DNA repair protein (RecF pathway)	L	Replication, recombination and repair	-2.12	5.43	64.04	4.26E-11	1.57E-10	-4.52	46.92
Salin0533	640475899	COG0216	Protein chain release factor A	J	Translation, ribosomal structure and biogenesis	-2.12	6.93	97.12	9.21E-16	6.73E-15	-4.43	70.33
Salin2846	640474329	Unknown	Hypothetical protein	S	Function unknown	-2.11	6.71	82.88	2.40E-17	1.39E-16	-4.36	60.31
Salin3137	640474962	COG3154	Putative lipid carrier protein	I	Lipid transport and metabolism	-2.10	4.81	30.51	6.87E-06	1.51E-05	-4.47	23.08
Salin1928	640475383	COG2329	Uncharacterized enzyme involved in biosynthesis of extracellular polysaccharides	R	General function prediction only	-2.09	6.61	84.54	3.92E-17	2.22E-16	-4.33	61.48
Salin2478	640474352	COG2199	FOG: GGDEF domain	T	Signal transduction mechanisms	-2.09	8.24	90.48	1.08E-15	5.44E-15	-4.42	65.72
Salin2624	640475765	NA	hypothetical protein	S	Function unknown	-2.08	7.46	107.16	8.54E-17	4.73E-16	-4.44	77.62
Salin0343	640474887	COG1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)	I	Lipid transport and metabolism	-2.08	4.65	40.29	1.84E-08	5.25E-08	-4.31	30.06
Salin0925	640476551	COG1174	ABC-type proline/glycine betaine transport systems, permease component	E	Amino acid transport and metabolism	-2.07	6.01	75.04	1.19E-15	5.99E-15	-4.26	54.76
Salin1616	640474135	COG1198	Primosomal protein N' (replication factor Y) - superfamily II helicase	L	Replication, recombination and repair	-2.05	5.66	61.30	6.02E-12	2.33E-11	-4.23	44.98
Salin0859	640474062	COG0311	Predicted glutamine amidotransferase involved in pyridoxine biosynthesis	H	coenzyme transport and metabolism	-2.04	8.56	99.46	6.26E-19	4.18E-18	-4.22	72.06
Salin1002	640475929	COG1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)	I	Lipid transport and metabolism	-2.04	7.63	86.62	8.92E-11	4.20E-10	-4.34	62.92
Salin1396	640473369	COG0031	Cysteine synthase	E	Amino acid transport and metabolism	-2.03	9.25	71.79	2.06E-12	8.22E-12	-4.25	52.44

GeneID	CNB440_JGI.ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj.pvalue	Fold	minFDR
Salin2268	640475529	COG5001	Predicted signal transduction protein containing a membrane domain, an EAL and a GGDEF domain	T	Signal transduction mechanisms	-2.01	6.76	89.89	1.80E-18	1.16E-17	-4.08	65.26
Salin2010	640472499	COG0136	Aspartate-semialdehyde dehydrogenase	E	Amino acid transport and metabolism	-2.00	6.83	75.80	1.24E-09	5.12E-09	-4.21	55.25
Salin1254	640476158	COG0024	Methionine aminopeptidase	J	Translation, ribosomal structure and biogenesis	-2.00	7.06	105.07	6.66E-14	4.12E-13	-4.13	76.05
Salin1859	640474078	COG1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)	I	Lipid transport and metabolism	-1.99	3.66	17.67	0.005981399	0.010038054	-4.20	13.96
Salin2362	640473239	COG1040	Predicted amidophosphoribosyltransferases	R	General function prediction only	-1.97	3.35	14.95	0.009354687	0.015207357	-4.12	12.00
Salin3118	640475632	NA	hypothetical protein	S	Function unknown	-1.97	6.07	68.96	1.70E-14	7.80E-14	-3.97	50.44
Salin1634	640475551	COG0524	Sugar kinases, ribokinase family	G	Carbohydrate transport and metabolism	-1.92	10.21	88.16	3.08E-17	1.76E-16	-3.85	64.06
Salin1375	640475687	COG1121	ABC-type Mn/Zn transport systems, ATPase component	P	Inorganic ion transport and metabolism	-1.91	3.61	16.39	0.006251345	0.010460628	-3.92	13.03
Salin1074	640476855	COG0357	Predicted S-adenosylmethionine-dependent methyltransferase involved in bacterial cell division	M	Cell wall/membrane/envelope biogenesis	-1.89	6.37	78.66	7.94E-15	3.73E-14	-3.79	57.33
Salin2384	640473143	COG3238	Uncharacterized protein conserved in bacteria	S	Function unknown	-1.88	5.08	30.94	6.81E-06	1.50E-05	-3.83	23.39
Salin1233	640475703	COG1385	Uncharacterized protein conserved in bacteria	S	Function unknown	-1.87	6.59	81.42	3.92E-17	2.22E-16	-3.68	59.27
Salin0876	640474155	COG0566	rRNA methylases	J	Translation, ribosomal structure and biogenesis	-1.87	5.32	47.79	1.20E-10	4.21E-10	-3.68	35.37
Salin1652	640474129	COG0284	Orotidine-5'-phosphate decarboxylase	F	Nucleotide transport and metabolism	-1.87	5.89	64.03	5.77E-12	2.23E-11	-3.73	46.92
Salin1815	640474407	COG0653	Preprotein translocase subunit SecA (ATPase, RNA helicase)	U	Intracellular trafficking, secretion and vesicular transport	-1.85	8.43	77.90	4.33E-11	2.10E-10	-3.74	56.72
Salin1121	640472472	COG0590	Cytosine/adenosine deaminases	F	Nucleotide transport and metabolism	-1.84	4.64	33.62	1.75E-05	4.43E-05	-3.65	25.37

GeneID	CNB440_JGI.ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj_pvalue	Fold	minFDR
Salin2549	640472507	COG1917	Uncharacterized conserved protein, contains double-stranded beta-helix domain	S	Function unknown	-1.83	8.36	76.71	7.46E-16	3.84E-15	-3.60	55.94
Salin2470	640473575	COG0442	Prolyl-tRNA synthetase	J	Translation, ribosomal structure and biogenesis	-1.82	7.32	74.70	1.87E-11	9.45E-11	-3.60	54.45
Salin3172	640474736	COG1304	L-lactate dehydrogenase (FMN-dependent) and related alpha-hydroxy acid dehydrogenases	C	Energy production and conversion	-1.80	6.09	27.05	1.15E-05	2.46E-05	-3.61	20.65
Salin2054	640473379	COG3217	Uncharacterized Fe-S protein	R	General function prediction only	-1.79	4.47	26.88	0.000111024	0.000245328	-3.53	20.57
Salin3451	640473625	Unknown	Function unknown	S	Function unknown	-1.79	6.46	67.21	3.83E-10	1.68E-09	-3.52	49.15
Salin2411	640473171	COG0714	MoxR-like ATPases	R	General function prediction only	-1.76	6.37	74.81	1.29E-09	5.31E-09	-3.50	54.55
Salin0512	640473749	COG0756	dUTPase	F	Nucleotide transport and metabolism	-1.75	5.88	38.61	4.00E-08	1.10E-07	-3.43	28.85
Salin0654	640474133	COG0452	Phosphopantothencysteine synthetase/decarboxylase	H	coenzyme transport and metabolism	-1.73	8.27	68.55	1.23E-13	5.35E-13	-3.37	50.14
Salin2088	640475686	COG0803	ABC-type metal ion transport system, periplasmic component/surface adhesin	P	Inorganic ion transport and metabolism	-1.70	5.68	46.26	1.60E-05	4.07E-05	-3.41	34.32
Salin1660	640473576	COG2306	Uncharacterized conserved protein	S	Function unknown	-1.69	7.78	67.86	2.85E-11	1.41E-10	-3.28	49.60
Salin1338	640473657	COG1351	Predicted alternative thymidylate synthase	F	Nucleotide transport and metabolism	-1.68	7.14	81.96	1.62E-13	9.76E-13	-3.23	59.59
Salin0702	640475524	COG5637	Predicted integral membrane protein	S	Function unknown	-1.67	7.77	66.71	3.02E-08	1.06E-07	-3.32	48.82
Salin0546	640473193	COG0782	Transcription elongation factor	K	Transcription	-1.66	9.78	68.79	1.48E-10	5.13E-10	-3.31	50.30
Salin0556	640474500	COG2519	tRNA(1-methyladenosine) methyltransferase and related methyltransferases	J	Translation, ribosomal structure and biogenesis	-1.65	6.48	68.34	4.37E-08	1.50E-07	-3.25	49.99
Salin2497	640474510	Unknown	Function unknown	S	Function unknown	-1.64	4.33	22.05	0.001593469	0.002964083	-3.23	17.10
Salin0687	640474498	COG2887	RecB family exonuclease	L	Replication, recombination and repair	-1.64	5.60	43.72	9.85E-07	2.92E-06	-3.17	32.55
Salin0936	640473780	COG3324	Predicted enzyme related to lactoylglutathione lyase	R	General function prediction only	-1.63	7.85	45.66	2.56E-07	8.10E-07	-3.17	33.93
Salin1567	640476360	COG0406	Fructose-2,6-bisphosphatase	G	Carbohydrate transport and metabolism	-1.62	7.11	64.95	1.34E-13	5.80E-13	-3.10	47.58

GeneID	CNB440_JGI.ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj.pvalue	Fold	minFDR
Salin2532	640474734	NA	hypothetical protein	S	Function unknown	-1.60	7.91	63.46	1.68E-12	6.78E-12	-3.08	46.50
Salin1418	640473875	COG1637	Predicted nuclease of the RecB family	L	Replication, recombination and repair	-1.60	8.03	64.18	7.71E-11	3.67E-10	-3.07	47.00
Salin2718	640475964	COG0457	FOG: TPR repeat	R	General function prediction only	-1.60	3.77	15.04	0.001369138	0.002265442	-3.12	12.04
Salin2107	640475785	COG3483	Tryptophan 2,3-dioxygenase (vermillion)	E	Amino acid transport and metabolism	-1.56	4.73	24.18	1.40E-05	2.98E-05	-3.00	18.62
Salin1535	640473838	COG0156	7-keto-8-aminopelargonate synthetase and related enzymes	H	coenzyme transport and metabolism	-1.55	6.03	42.51	3.58E-07	1.12E-06	-2.97	31.70
Salin0497	640475491	COG2606	Uncharacterized conserved protein	S	Function unknown	-1.55	4.70	25.61	6.72E-05	0.000131406	-3.02	19.60
Salin0581	640473059	COG1825	Ribosomal protein L25 (general stress protein Ctc)	J	Translation, ribosomal structure and biogenesis	-1.55	8.62	61.59	1.69E-09	6.83E-09	-2.98	45.19
Salin0506	640475485	COG1524	Uncharacterized proteins of the AP superfamily	R	General function prediction only	-1.54	5.84	46.53	1.13E-07	3.68E-07	-2.94	34.55
Salin2814	640474779	COG1903	Cobalamin biosynthesis protein CbiD	H	coenzyme transport and metabolism	-1.54	7.58	52.86	1.42E-08	5.19E-08	-2.95	39.01
Salin1916	640476748	COG0705	Uncharacterized membrane protein (homolog of Drosophila rhomboid)	R	General function prediction only	-1.53	8.88	59.79	2.16E-12	8.61E-12	-2.92	43.89
Salin1103	640472699	COG2017	Galactose mutarotase and related enzymes	G	Carbohydrate transport and metabolism	-1.52	6.53	64.31	2.44E-06	6.88E-06	-3.02	47.12
Salin0791	640473504	COG0059	Ketol-acid reductoisomerase	E	Amino acid transport and metabolism	-1.51	8.60	55.81	1.74E-06	5.02E-06	-2.99	41.08
Salin0428	640474125	COG0505	Carbamoylphosphate synthase small subunit	E	Amino acid transport and metabolism	-1.48	5.92	44.21	6.94E-09	2.06E-08	-2.82	32.83
Salin0821	640474058	COG1560	Lauroyl/myristoyl acyltransferase	M	Cell wall/membrane/envelope biogenesis	-1.47	6.03	47.98	4.47E-06	1.22E-05	-2.85	35.54
Salin2276	640476020	NA	hypothetical protein	S	Function unknown	-1.46	6.44	54.91	1.05E-11	4.01E-11	-2.78	40.41
Salin3380	640472533	COG1309	Transcriptional regulator	K	Transcription	-1.46	3.90	13.90	0.01097853	0.017584228	-2.84	11.23
Salin2946	640473646	COG1108	ABC-type Mn2+/Zn2+ transport systems, permease components	P	Inorganic ion transport and metabolism	-1.46	7.58	66.80	4.89E-08	1.68E-07	-2.84	48.89
Salin1399	640473095	COG1012	NAD-dependent aldehyde dehydrogenases	C	Energy production and conversion	-1.45	8.00	45.42	3.36E-08	9.33E-08	-2.80	33.67
Salin0476	640475308	COG0006	Xaa-Pro aminopeptidase	E	Amino acid transport and metabolism	-1.44	7.29	55.66	3.93E-10	1.30E-09	-2.77	40.96

GeneID	CNB440_JGI_ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj_pvalue	Fold	minFDR
Salin2550	640474432	COG2188	Transcriptional regulators	K	Transcription	-1.44	5.88	36.95	1.75E-07	4.53E-07	-2.75	27.67
Salin2223	640475493	COG0456	Acetyltransferases	R	General function prediction only	-1.43	5.37	30.69	7.78E-07	1.88E-06	-2.73	23.25
Salin1911	640475381	COG2365	Protein tyrosine/serine phosphatase	T	Signal transduction mechanisms	-1.43	5.93	35.53	2.59E-07	6.57E-07	-2.73	26.67
Salin2798	640472440	COG1651	Protein-disulfide isomerase	O	Posttranslational modification, protein turnover, chaperons	-1.41	7.86	32.71	3.59E-06	8.14E-06	-2.73	24.65
Salin2426	640475635	NA	hypothetical protein	S	Function unknown	-1.40	11.27	63.55	6.86E-09	2.59E-08	-2.71	46.57
Salin2424	640472276	NA	hypothetical protein	S	Function unknown	-1.38	10.91	58.18	2.95E-09	1.16E-08	-2.65	42.78
Salin2295	640472586	COG3246	Uncharacterized conserved protein	S	Function unknown	-1.38	6.84	59.04	2.02E-08	7.28E-08	-2.64	43.38
Salin2811	640472510	COG0707	UDP-N-acetylglucosamine:LPS transferase	M	Cell wall/membrane/envelope biogenesis	-1.37	5.82	39.84	3.19E-06	8.86E-06	-2.63	29.81
Salin1854	640473642	COG0618	Exopolyphosphatase-related proteins	R	General function prediction only	-1.36	7.35	48.32	1.44E-09	4.56E-09	-2.59	35.74
Salin1761	640476818	COG3809	Uncharacterized protein conserved in bacteria	S	Function unknown	-1.35	10.59	48.04	5.17E-10	1.70E-09	-2.58	35.55
Salin0679	640474124	COG0044	Dihydroorotase and related cyclic amidohydrolases	F	Nucleotide transport and metabolism	-1.34	4.90	23.66	1.97E-05	4.12E-05	-2.57	18.24
Salin2513	640474375	COG0515	Serine/threonine protein kinase	R	General function prediction only	-1.33	7.66	51.11	5.31E-07	1.62E-06	-2.58	37.76
Salin3698	640474048	COG0667	Predicted oxidoreductases (related to aryl-alcohol dehydrogenases)	C	Energy production and conversion	-1.31	6.27	34.17	2.34E-07	5.95E-07	-2.51	25.71
Salin2230	640475517	COG0664	cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases	T	Signal transduction mechanisms	-1.29	8.09	45.40	8.54E-08	2.83E-07	-2.47	33.76
Salin2428	640476341	COG0609	ABC-type Fe3+-siderophore transport system, permease component	P	Inorganic ion transport and metabolism	-1.25	5.84	28.83	3.25E-06	7.42E-06	-2.42	21.91
Salin0992	640474531	COG0815	Apolipoprotein N-acyltransferase	M	Cell wall/membrane/envelope biogenesis	-1.20	6.09	33.76	1.95E-07	5.02E-07	-2.32	25.42
Salin1394	640472614	COG0506	Proline dehydrogenase	E	Amino acid transport and metabolism	1.26	6.85	46.98	7.81E-11	2.81E-10	2.42	34.80

GeneID	CNB440_JGI.ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj.pvalue	Fold	minFDR
Salin0161	640474707	Unknown	Function unknown	S	Function unknown	1.35	8.99	28.60	3.75E-06	8.46E-06	2.60	21.75
Salin2376	640476606	COG0515	Serine/threonine protein kinase	R	General function prediction only	1.36	6.94	69.07	2.27E-09	9.04E-09	2.62	50.50
Salin0928	640474617	COG0242	N-formylmethionyl-tRNA deformylase	J	Translation, ribosomal structure and biogenesis	1.36	5.86	43.47	4.89E-07	1.50E-06	2.61	32.38
Salin0739	640475994	COG0661	Predicted unusual protein kinase	R	General function prediction only	1.37	6.04	43.89	4.87E-07	1.50E-06	2.63	32.68
Salin1179	640475484	COG0389	Nucleotidyltransferase/DNA polymerase involved in DNA repair	L	Replication, recombination and repair	1.38	5.63	34.01	4.60E-08	1.26E-07	2.65	25.62
Salin0915	640475949	COG0282	Acetate kinase	C	Energy production and conversion	1.42	9.35	32.47	5.36E-07	1.32E-06	2.75	24.50
Salin1589	640475556	Unknown	Function unknown	S	Function unknown	1.45	8.92	44.78	1.60E-06	4.62E-06	2.80	33.30
Salin2320	640476278	COG4262	Predicted spermidine synthase with an N-terminal membrane domain	R	General function prediction only	1.47	5.05	35.62	2.82E-08	7.88E-08	2.84	26.75
Salin3484	640475636	COG0596	Predicted hydrolases (alpha/beta hydrolase superfamily)	R	General function prediction only	1.47	6.61	52.13	1.50E-10	5.19E-10	2.86	38.44
Salin2084	640473492	COG1408	Predicted phosphohydrolases	R	General function prediction only	1.48	5.30	45.18	9.95E-06	2.59E-05	2.93	33.57
Salin1986	640472753	COG1404	Subtilisin-like serine proteases	O	Posttranslational modification, protein turnover, chaperons	1.51	5.40	51.22	2.84E-06	7.92E-06	2.99	37.84
Salin1493	640474198	COG1192	ATPases involved in chromosome partitioning	D	Cell cycle control, cell division, chromosome partitioning	1.52	8.88	55.67	2.46E-10	8.32E-10	2.96	40.96
Salin1276	640476514	COG0079	Histidinol-phosphate/aromatic aminotransferase and cobyric acid decarboxylase	E	Amino acid transport and metabolism	1.53	7.35	74.50	7.48E-12	3.92E-11	2.92	54.32
Salin0360	640472429	COG3576	Predicted flavin-nucleotide-binding protein structurally related to pyridoxine 5'-phosphate oxidase	R	General function prediction only	1.54	7.29	42.65	4.09E-07	1.27E-06	2.96	31.81
Salin3014	640474883	NA	hypothetical protein	S	Function unknown	1.55	6.23	49.00	7.89E-07	2.36E-06	3.03	36.27

GeneID	CNB440_JGI.ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj.pvalue	Fold	minFDR
Salin0309	640475940	COG0265	Trypsin-like serine proteases, typically periplasmic, contain C-terminal PDZ domain	O	Posttranslational modification, protein turnover, chaperons	1.57	8.94	64.57	9.58E-09	3.56E-08	3.06	47.30
Salin0294	640472678	COG0501	Zn-dependent protease with chaperone function	O	Posttranslational modification, protein turnover, chaperons	1.59	7.06	67.16	1.05E-08	3.90E-08	3.12	49.15
Salin3013	640473368	COG0277	FAD/FMN-containing dehydrogenases	C	Energy production and conversion	1.59	6.48	72.54	3.97E-10	1.74E-09	3.10	52.94
Salin1604	640473064	COG1087	UDP-glucose 4-epimerase	M	Cell wall/membrane/envelope biogenesis	1.60	5.65	57.14	1.11E-08	4.11E-08	3.10	42.05
Salin1996	640472519	COG2222	Predicted phosphosugar isomerases	M	Cell wall/membrane/envelope biogenesis	1.61	5.62	36.93	4.60E-06	1.26E-05	3.12	27.73
Salin1526	640473980	COG0479	Succinate dehydrogenase/fumarate reductase, Fe-S protein subunit	C	Energy production and conversion	1.62	6.15	49.64	5.42E-08	1.84E-07	3.12	36.75
Salin1424	640476512	COG3185	4-hydroxyphenylpyruvate dioxygenase and related hemolysins	ER	Amino acid transport and metabolism	1.64	9.67	68.78	6.28E-11	2.28E-10	3.29	50.28
Salin0850	640475589	COG0413	Ketopantoate hydroxymethyltransferase	H	coenzyme transport and metabolism	1.64	8.27	57.83	5.92E-08	2.00E-07	3.22	42.53
Salin1988	640475948	COG0280	Phosphotransacetylase	C	Energy production and conversion	1.64	10.27	51.70	4.04E-10	1.34E-09	3.22	38.14
Salin1170	640475354	COG0514	Superfamily II DNA helicase	L	Replication, recombination and repair	1.65	4.77	37.23	2.07E-08	5.88E-08	3.26	27.88
Salin0677	640473573	COG0541	Signal recognition particle GTPase	U	Intracellular trafficking, secretion and vesicular transport	1.65	9.27	66.37	1.08E-13	4.76E-13	3.21	48.58
Salin4621	640472909	Unknown	Function unknown	S	Function unknown	1.66	4.76	30.42	2.81E-07	7.11E-07	3.27	23.07
Salin3217	640472370	Unknown	Function unknown	S	Function unknown	1.66	8.49	55.62	1.76E-07	5.65E-07	3.27	40.97
Salin0491	640474503	COG1222	ATP-dependent 26S proteasome regulatory subunit	O	Posttranslational modification, protein turnover, chaperons	1.67	11.91	78.73	5.72E-11	2.74E-10	3.26	57.31
Salin1683	640475648	COG1994	Zn-dependent proteases	R	General function prediction only	1.68	9.45	60.54	1.90E-09	7.66E-09	3.25	44.44
Salin1799	640473675	COG1201	Lhr-like helicases	R	General function prediction only	1.68	5.01	51.33	6.95E-07	2.10E-06	3.37	37.92
Salin2344	640476533	COG5373	Predicted membrane protein	S	Function unknown	1.69	7.22	71.75	3.77E-14	1.69E-13	3.35	52.43

GeneID	CNB440_JGI_ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj_pvalue	Fold	minFDR
Salin2288	640474053	COG1397	ADP-ribosylglycohydrolase	O	Posttranslational modification, protein turnover, chaperons	1.71	5.92	58.75	5.47E-09	2.10E-08	3.35	43.18
Salin2422	640473215	COG5178	U5 snRNP spliceosome subunit	A	RNA processing and modification	1.72	7.80	72.36	7.98E-10	3.36E-09	3.40	52.82
Salin2464	640475811	COG1544	Ribosome-associated protein Y (PSrp-1)	J	Translation, ribosomal structure and biogenesis	1.72	14.39	61.96	7.89E-12	3.03E-11	3.42	45.43
Salin1866	640475792	COG2346	Truncated hemoglobins	R	General function prediction only	1.80	9.11	77.65	1.77E-17	1.04E-16	3.52	56.59
Salin2153	640476468	COG1268	Uncharacterized conserved protein	S	Function unknown	1.82	7.58	80.52	4.92E-13	2.83E-12	3.57	58.57
Salin0254	640474319	COG0376	Catalase (peroxidase I)	P	Inorganic ion transport and metabolism	1.82	10.94	77.19	8.25E-12	4.30E-11	3.61	56.22
Salin0919	640472539	COG0456	Acetyltransferases	R	General function prediction only	1.82	4.63	42.47	1.47E-06	4.29E-06	3.69	31.67
Salin2705	640472999	NA	hypothetical protein	S	Function unknown	1.85	5.59	47.50	1.25E-10	4.38E-10	3.74	35.17
Salin1903	640476560	NA	hypothetical protein	S	Function unknown	1.86	7.81	71.47	7.05E-13	2.91E-12	3.84	52.22
Salin3009	640473687	COG0745	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	T	Signal transduction mechanisms	1.87	5.10	37.67	5.10E-06	1.38E-05	3.79	28.27
Salin0318	640475576	COG0320	Lipoate synthase	H	coenzyme transport and metabolism	1.87	10.18	91.04	5.45E-15	3.72E-14	3.70	66.01
Salin1469	640473105	COG2105	Uncharacterized conserved protein	S	Function unknown	1.88	6.88	86.24	3.46E-20	2.56E-19	3.73	62.67
Salin2519	640475226	NA	hypothetical protein	S	Function unknown	1.90	8.99	70.03	1.32E-08	4.85E-08	3.93	51.17
Salin1804	640472752	COG4842	Uncharacterized protein conserved in bacteria	S	Function unknown	1.91	8.92	78.87	1.02E-11	5.24E-11	3.86	57.40
Salin1359	640472690	COG1846	Transcriptional regulators	K	Transcription	1.94	10.15	97.65	1.80E-12	1.00E-11	4.00	70.75
Salin2100	640472750	Unknown	No COG number	S	Function unknown	1.95	4.73	54.54	4.18E-08	1.44E-07	4.05	40.21
Salin1605	640473981	COG1053	Succinate dehydrogenase/fumarate reductase, flavoprotein subunit	C	Energy production and conversion	1.98	6.59	86.87	1.76E-12	9.82E-12	4.08	63.06
Salin2455	640473134	NA	hypothetical protein	S	Function unknown	2.00	7.54	88.50	1.84E-13	1.10E-12	4.07	64.22
Salin0982	640474309	COG4552	Predicted acetyltransferase involved in intracellular survival and related acetyltransferases	R	General function prediction only	2.00	7.27	85.79	3.36E-18	2.10E-17	4.12	62.36
Salin2964	640476686	COG1131	ABC-type multidrug transport system, ATPase component	V	Defense mechanisms	2.06	5.07	52.76	9.45E-08	3.13E-07	4.37	38.94

GeneID	CNB440_JGI.ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj.pvalue	Fold	minFDR
Salin2435	640475776	COG0662	Mannose-6-phosphate isomerase	G	Carbohydrate transport and metabolism	2.06	7.03	101.94	4.67E-14	2.96E-13	4.34	73.80
Salin0785	640473195	COG2120	Uncharacterized proteins, LmbE homologs	S	Function unknown	2.07	7.90	97.64	8.31E-19	5.48E-18	4.40	70.76
Salin2780	640476248	COG1670	Acetyltransferases, including N-acetylases of ribosomal proteins	J	Translation, ribosomal structure and biogenesis	2.08	4.63	46.52	5.15E-11	1.89E-10	4.43	34.47
Salin3846	640473785	Unknown	Function unknown	S	Function unknown	2.10	9.37	103.30	7.54E-16	5.55E-15	4.37	74.73
Salin1509	640474056	COG0480	Translation elongation factors (GTPases)	J	Translation, ribosomal structure and biogenesis	2.12	6.06	85.87	4.38E-19	2.96E-18	4.51	62.39
Salin3150	640476188	COG2514	Predicted ring-cleavage extradiol dioxygenase	R	General function prediction only	2.13	7.66	86.76	7.17E-20	5.09E-19	4.46	63.03
Salin0345	640472438	COG1230	Co/Zn/Cd efflux system component	P	Inorganic ion transport and metabolism	2.14	8.40	72.79	1.17E-16	6.45E-16	4.50	53.14
Salin1671	640472578	COG0745	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	T	Signal transduction mechanisms	2.16	5.89	68.47	2.14E-09	8.55E-09	4.71	50.07
Salin1364	640472661	COG0076	Glutamate decarboxylase and related PLP-dependent proteins	E	Amino acid transport and metabolism	2.19	4.99	53.02	5.35E-12	2.07E-11	4.84	39.09
Salin3162	640473164	COG1695	Predicted transcriptional regulators	K	Transcription	2.21	5.29	79.87	8.21E-12	4.28E-11	4.80	58.10
Salin3870	640473883	COG0446	Uncharacterized NAD(FAD)-dependent dehydrogenases	R	General function prediction only	2.22	10.69	91.95	4.58E-18	2.80E-17	4.89	66.75
Salin2145	640475761	COG0604	NADPH:quinone reductase and related Zn-dependent oxidoreductases	C	Energy production and conversion	2.26	8.77	80.68	1.07E-15	5.41E-15	5.07	58.76
Salin3008	640473523	NA	hypothetical protein	S	Function unknown	2.30	6.46	133.28	8.42E-31	1.32E-29	5.02	96.07
Salin1336	640473672	COG1396	Predicted transcriptional regulators	K	Transcription	2.37	6.21	116.61	6.94E-16	5.12E-15	5.44	84.26
Salin2675	640475981	COG0462	Phosphoribosylpyrophosphate synthetase	F	Nucleotide transport and metabolism	2.43	7.64	150.01	1.21E-33	1.88E-32	5.51	108.02
Salin3581	640472894	COG4447	Uncharacterized protein related to plant photosystem II stability/assembly factor	R	General function prediction only	2.48	5.09	63.67	5.53E-10	2.37E-09	5.85	46.65
Salin2392	640476001	Unknown	Function unknown	S	Function unknown	2.56	8.00	75.21	7.23E-12	3.80E-11	6.11	54.81

GeneID	CNB440_JGI.ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj.pvalue	Fold	minFDR
Salin3015	640475762	NA	hypothetical protein	S	Function unknown	2.70	8.74	124.78	7.29E-28	8.62E-27	6.71	90.04
Salin2703	640476750	COG1426	Uncharacterized protein conserved in bacteria	S	Function unknown	2.72	6.20	147.89	6.53E-21	6.40E-20	6.95	106.63
Salin1050	640476233	COG0318	AcyI-CoA synthetases (AMP-forming)/AMP-acid ligases II	I	Lipid transport and metabolism	2.73	6.88	152.46	6.44E-35	1.21E-33	6.82	109.79
Salin3111	640476470	NA	hypothetical protein	S	Function unknown	2.84	6.68	203.01	9.98E-31	1.62E-29	7.39	146.06
Salin2989	640476452	NA	hypothetical protein	S	Function unknown	2.90	5.22	115.61	3.15E-26	4.19E-25	7.77	83.50
Salin2318	640476000	Unknown	No COG number	S	Function unknown	2.92	7.54	203.23	5.20E-28	7.74E-27	8.05	146.22
Salin1302	640473982	NA	succinate dehydrogenase (or fumarate reductase)	S	Function unknown	3.07	7.01	231.82	2.00E-33	3.64E-32	8.85	166.61
Salin3033	640474936	COG3832	cytochrome b subunit, b558 family	S	Function unknown	3.15	7.75	202.84	2.32E-45	5.07E-44	9.27	145.78
Salin0369	640476628	NA	Uncharacterized conserved protein	S	Function unknown	3.18	10.92	158.13	5.36E-23	5.99E-22	9.54	113.95
Salin1674	640475825	COG3211	Function unknown Predicted phosphatase	R	General function prediction only	3.21	6.45	209.49	2.83E-33	5.12E-32	9.49	150.68
Salin2588	640475640	COG1366	Anti-anti-sigma regulatory factor (antagonist of anti-sigma factor)	T	Signal transduction mechanisms	3.26	11.17	157.90	8.40E-26	1.10E-24	9.77	113.73
Salin2608	640473614	COG0388	Predicted amidohydrolase	R	General function prediction only	3.34	7.65	256.19	5.29E-39	1.24E-37	10.50	184.02
Salin1022	640472373	COG5373	Predicted membrane protein	S	Function unknown	3.39	9.08	218.55	7.68E-33	1.37E-31	10.90	157.15
Salin2481	640475669	COG1429	Cobalamin biosynthesis protein CobN and related Mg-chelataes	H	coenzyme transport and metabolism	3.46	8.43	228.59	1.25E-34	2.36E-33	11.45	164.31
Salin2161	640473001	COG3402	Uncharacterized conserved protein	S	Function unknown	3.67	9.94	279.42	1.53E-59	7.34E-58	12.96	200.58
Salin2078	640472287	COG0517	FOG: CBS domain	R	General function prediction only	3.79	8.33	234.62	8.47E-39	1.97E-37	14.06	168.60
Salin2373	640476719	COG4961	Flp pilus assembly protein TadG	U	Intracellular trafficking, secretion and vesicular transport	4.22	7.70	252.29	4.84E-39	1.14E-37	19.77	181.23
Salin1898	640474374	COG3544	Uncharacterized protein conserved in bacteria	S	Function unknown	4.56	9.22	332.96	1.60E-53	6.48E-52	24.49	238.93
Salin3713	640473783	NA	hypothetical protein	S	Function unknown	5.02	8.76	309.65	4.45E-49	1.54E-47	34.33	222.26
Salin3053	640474676	Unknown	thiazolypeptide-type bacteriocin precursor	S	Function unknown	8.02	12.80	484.46	1.50E-82	2.10E-80	265.80	346.90

Appendix E

Differential Expression of *Salinispora*: Stationary Phase

GeneID	CNB440_JGI.ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj_pvalue	Fold	minFDR
Salin2977	640474741	COG2814	Arabinose efflux permease	G	Carbohydrate and metabolism	-3.86	5.70	70.14	1.01384E-12	1.34313E-11	-15.22	49.76
Salin3156	640472508	COG2345	Predicted transcriptional regulator	K	Transcription	-3.41	5.21	46.43	2.58713E-09	1.48309E-08	-11.28	33.24
Salin2759	640474735	COG1167	(MocR family) Transcriptional regulators containing a DNA-binding HTH domain	K	Transcription	-3.35	8.09	94.79	5.84485E-18	1.05741E-16	-10.79	66.86
Salin1687	640473957	COG1134	ABC-type polysaccharide/polyol phosphate transport system, ATPase component	G	Carbohydrate and metabolism	-3.18	4.53	43.26	1.21079E-07	6.70261E-07	-9.48	31.12
Salin3381	640476654	COG0726	Predicted xylanase/chitin deacetylase	G	Carbohydrate and metabolism	-3.04	4.70	41.09	4.92261E-08	2.30065E-07	-8.66	29.56
Salin1520	640473102	COG1473	Metal-dependent amidase-aminoacylase-carboxypeptidase	R	General function prediction only	-2.68	7.66	53.50	8.82E-11	6.38267E-10	-6.68	38.11
Salin0925	640476551	COG1174	ABC-type proline/glycine betaine transport systems, permease component	E	Amino acid transport and metabolism	-2.50	6.90	53.52	2.09014E-10	1.43538E-09	-5.90	38.10
Salin1849	640475878	COG0842	ABC-type multidrug transport system, permease component	V	Defense mechanisms	-2.40	5.10	28.26	8.35222E-06	2.73417E-05	-5.56	20.70
Salin2291	640476612	COG5373	Predicted membrane protein	S	Function unknown	-2.37	9.18	48.84	1.01827E-09	6.41668E-09	-5.38	34.91
Salin0379	640476502	COG0251	Putative translation initiation inhibitor, yjgF family	J	Translation, ribosomal structure and biogenesis	-2.30	6.37	35.46	2.01789E-06	8.62797E-06	-5.12	25.75
Salin0527	640475548	COG1622	Heme/copper-type cytochrome/quinol oxidases, subunit 2	C	Energy production and conversion	-2.27	9.87	47.17	1.17534E-09	7.25699E-09	-4.98	33.75
Salin3075	640473455	COG1609	Transcriptional regulators	K	Transcription	-2.26	4.98	21.14	0.000532334	0.001322781	-5.09	15.90
Salin0537	640476108	COG0533	Metal-dependent proteases with possible chaperone activity	O	Posttranslational modification, protein turnover, chaperons	-2.26	6.19	42.52	9.8429E-08	4.36957E-07	-5.04	30.57
Salin3169	640475001	COG4430	Uncharacterized protein conserved in bacteria	S	Function unknown	-2.25	5.96	35.70	6.62434E-06	2.54749E-05	-5.01	25.93
Salin2235	640476002	NA	hypothetical protein	S	Function unknown	-2.23	6.95	38.21	1.03326E-07	4.55699E-07	-4.88	27.59
Salin1097	640476146	COG0352	Thiamine monophosphate synthase	H	coenzyme transport and metabolism	-2.21	9.32	39.74	1.14968E-07	5.01775E-07	-4.85	28.64

GeneID	CNB440_JGI.ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj_pvalue	Fold	minFDR
Salin2288	640474053	COG1397	ADP-ribosylglycohydrolase	O	Posttranslational modification, protein turnover, chaperons	-2.19	7.42	36.47	1.48898E-07	6.39119E-07	-4.72	26.39
Salin2729	640474597	COG3096	Uncharacterized protein involved in chromosome partitioning	D	Cell cycle control, cell division, chromosome partitioning	-2.17	5.56	28.43	8.27835E-06	2.71525E-05	-4.70	20.81
Salin2451	640475428	COG1132	ABC-type multidrug transport system, ATPase and permease components	V	Defense mechanisms	-2.09	5.50	23.77	4.99466E-05	0.000142458	-4.47	17.61
Salin0946	640475528	COG4106	Trans-aconitate methyltransferase	R	General function prediction only	-2.09	7.53	46.96	3.32567E-08	2.10355E-07	-4.38	33.68
Salin0699	640476140	COG0422	Thiamine biosynthesis	H	coenzyme transport and metabolism	-2.08	11.05	42.83	1.11991E-08	5.87001E-08	-4.38	30.77
Salin1979	640472420	COG4934	protein ThiC Predicted protease	O	Posttranslational modification, protein turnover, chaperons	-2.08	8.22	41.87	2.38669E-08	1.18982E-07	-4.38	30.09
Salin1288	640473471	COG0524	Sugar kinases, ribokinase family	G	Carbohydrate transport and metabolism	-2.05	9.40	42.07	4.69605E-07	2.31704E-06	-4.30	30.31
Salin0694	640473549	COG0227	Ribosomal protein L28	J	Translation, ribosomal structure and biogenesis	-2.01	10.33	30.59	1.84865E-05	6.42408E-05	-4.21	22.42
Salin2585	640474143	COG3706	Response regulator containing a CheY-like receiver domain and a GGDEF domain	T	Signal transduction mechanisms	-1.95	5.50	24.60	0.000105441	0.000280933	-4.09	18.15
Salin2839	640474866	COG1960	Acyl-CoA dehydrogenases	I	Lipid transport and metabolism	-1.94	4.21	18.87	0.000884586	0.001925612	-4.04	14.19
Salin1241	640476011	COG2171	Tetrahydridipicolinate N-succinyltransferase	E	Amino acid transport and metabolism	-1.91	6.64	28.81	5.31881E-05	0.000167426	-3.91	21.19
Salin1264	640476028	COG1051	ADP-ribose pyrophosphatase	F	Nucleotide transport and metabolism	-1.91	6.39	28.09	1.61314E-05	4.98929E-05	-3.92	20.57
Salin1246	640476087	COG1183	Phosphatidylserine synthase	I	Lipid transport and metabolism	-1.90	8.55	34.40	8.42838E-07	3.21929E-06	-3.87	24.94
Salin2855	640474815	COG0807	GTP cyclohydrolase II	H	coenzyme transport and metabolism	-1.87	7.33	31.84	7.63152E-06	2.91158E-05	-3.76	23.27
Salin2919	640473266	NA	hypothetical protein	S	Function unknown	-1.79	10.16	28.49	4.01222E-05	0.000129445	-3.58	20.98
Salin2180	640474857	COG1309	Transcriptional regulator	K	Transcription	-1.77	10.88	36.56	6.56241E-06	2.5266E-05	-3.55	26.50
Salin1812	640472461	NA	hypothetical protein	S	Function unknown	-1.76	9.52	35.27	1.08581E-05	4.01509E-05	-3.52	25.63
Salin0793	640474636	COG1003	Glycine cleavage system protein P (pyridoxal-binding), C-terminal domain	E	Amino acid transport and metabolism	-1.75	9.61	27.64	0.000138976	0.00039398	-3.52	20.38

GeneID	CNB440_JGIID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj_pvalue	Fold	minFDR
Salin0783	640476326	COG0838	NADH:ubiquinone oxidoreductase subunit 3 (chain A)	C	Energy production and conversion	-1.73	9.51	27.48	0.000171549	0.000475505	-3.49	20.27
Salin3136	640474959	COG2128	Uncharacterized conserved protein	S	Function unknown	-1.73	5.93	24.09	0.000462738	0.001167253	-3.47	17.92
Salin1255	640474505	No COG number	No COG number	S	Function unknown	-1.71	6.42	27.47	0.000122171	0.000352166	-3.39	20.27
Salin3162	640473164	COG1695	Predicted transcriptional regulators	K	Transcription	-1.70	5.86	19.16	0.000306871	0.00073327	-3.36	14.43
Salin3008	640473523	NA	hypothetical protein	S	Function unknown	-1.67	5.77	23.13	0.000636509	0.001561121	-3.31	17.25
Salin1934	640474043	Unknown	Function unknown	S	Function unknown	-1.65	6.21	25.75	2.41391E-05	7.2598E-05	-3.24	18.97
Salin1796	640474033	COG0438	Glycosyltransferase	M	Cell wall/membrane/envelope biogenesis	-1.59	7.96	27.75	5.73406E-05	0.000178612	-3.09	20.47
Salin0256	640475896	COG0009	Putative translation factor (SUA5)	J	Translation, ribosomal structure and biogenesis	-1.59	7.87	24.54	7.72145E-05	0.000211767	-3.13	18.12
Salin0567	640476542	COG0029	Aspartate oxidase	H	coenzyme transport and metabolism	-1.58	5.90	21.26	0.000505065	0.001160138	-3.13	15.86
Salin3023	640474672	COG1226	Kef-type K+ transport systems, predicted NAD-binding component	P	Inorganic ion transport and metabolism	-1.57	8.36	27.88	0.000100202	0.00029448	-3.08	20.55
Salin1342	640475402	COG1522	Transcriptional regulators	K	Transcription	-1.57	4.89	11.69	0.01000775	0.01724756	-3.14	9.21
Salin2884	640474931	COG1053	Succinate dehydrogenase/fumarate reductase, flavoprotein subunit	C	Energy production and conversion	-1.56	5.02	18.33	0.001545231	0.003452806	-3.03	13.95
Salin1321	640473578	COG1837	Predicted RNA-binding protein (contains KH domain)	R	General function prediction only	-1.55	5.53	17.47	0.002522039	0.005314357	-3.03	13.34
Salin0575	640474115	COG0082	Chorismate synthase	E	Amino acid transport and metabolism	-1.52	7.96	21.46	0.000208633	0.000518569	-2.98	16.01
Salin1509	640474056	COG0480	Translation elongation factors (GTPases)	J	Translation, ribosomal structure and biogenesis	-1.51	4.39	14.47	0.006519498	0.012347331	-2.95	11.24
Salin1536	640472316	NA	hypothetical protein	S	Function unknown	-1.50	6.08	14.21	0.005153148	0.009541953	-3.00	10.97
Salin2559	640475675	COG1292	Choline-glycine betaine transporter	M	Cell wall/membrane/envelope biogenesis	-1.47	7.88	17.66	0.003277269	0.006729727	-2.88	13.47
Salin0767	640475603	COG0406	Fructose-2,6-bisphosphatase	G	Carbohydrate transport and metabolism	-1.45	6.31	17.05	0.000720325	0.001603368	-2.80	12.97
Salin2669	640473937	COG2199	FOG; GGDEF domain	T	Signal transduction mechanisms	-1.44	4.55	12.46	0.008483229	0.014833857	-2.83	9.76
Salin1377	640475721	COG0406	Fructose-2,6-bisphosphatase	G	Carbohydrate transport and metabolism	-1.37	8.08	19.83	0.000283055	0.000679127	-2.67	14.88

GeneID	CNB440_JGI.ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj_pvalue	Fold	minFDR
Salin0331	640475415	COG0069	Glutamate synthase domain 2	E	Amino acid transport and metabolism	1.47	9.61	19.87	0.00147786	0.003319798	2.87	15.01
Salin1292	640476357	COG1333	ResB protein required for cytochrome c biosynthesis	O	Posttranslational modification, protein turnover, chaperons	1.49	4.54	15.91	0.001046742	0.002251983	2.96	12.18
Salin1436	640475461	COG0707	UDP-N-acetylglucosamine:LPS N-acetylglucosamine transferase	M	Cell wall/membrane/envelope biogenesis	1.63	6.91	22.69	0.000612319	0.001506571	3.22	16.95
Salin1383	640475331	COG2146	Ferredoxin subunits of nitrite reductase and ring-hydroxylating dioxygenases	P	Inorganic ion transport and metabolism	1.65	7.38	29.77	4.97316E-05	0.000157559	3.22	21.86
Salin1517	640473750	Unknown	Function unknown	S	Function unknown	1.71	6.92	18.55	0.002541931	0.005352778	3.44	14.09
Salin2446	640474093	COG2271	Sugar phosphate permease	G	Carbohydrate transport and metabolism	1.77	5.96	26.06	0.000326897	0.0008514	3.61	19.30
Salin2797	640474658	COG1690	Uncharacterized conserved protein	S	Function unknown	1.79	7.33	23.03	0.000764781	0.001847446	3.66	17.17
Salin2695	640474877	COG0183	Acetyl-CoA acetyltransferase	I	Lipid transport and metabolism	1.91	7.56	33.09	2.87949E-05	9.56102E-05	3.92	24.12
Salin3487	640473788	NA	hypothetical protein	S	Function unknown	2.18	7.07	38.52	2.19623E-06	9.36669E-06	4.72	27.85
Salin2354	640474107	No COG number	No COG number	S	Function unknown	2.24	5.79	36.73	2.53901E-06	1.06531E-05	4.94	26.62
Salin1302	640473982	No COG number	succinate dehydrogenase (or fumarate reductase) cytochrome b subunit, b558 family	S	Function unknown	2.29	5.97	37.27	1.70733E-06	7.42267E-06	5.08	26.99
Salin0392	640472545	COG0277	FAD/FMN-containing dehydrogenases	C	Energy production and conversion	2.43	6.22	43.86	2.48765E-07	1.29983E-06	5.65	31.56
Salin2343	640476755	COG0371	Glycerol dehydrogenase and related enzymes	C	Energy production and conversion	2.56	7.47	58.09	1.82743E-09	1.42345E-08	6.12	41.46
Salin1820	640473463	COG1801	Uncharacterized conserved protein	S	Function unknown	2.59	7.45	58.95	3.36824E-13	3.60052E-12	6.30	41.84
Salin0352	640474799	COG0735	Fe2+/Zn2+ uptake regulation proteins	P	Inorganic ion transport and metabolism	2.72	9.10	56.50	1.0046E-12	1.00101E-11	6.89	40.16
Salin1134	640476803	COG2896	Molybdenum cofactor biosynthesis enzyme	H	coenzyme transport and metabolism	3.23	10.32	90.91	7.20917E-15	1.30424E-13	9.69	64.36
Salin0254	640474319	COG0376	Catalase (peroxidase I)	P	Inorganic ion transport and metabolism	3.51	12.04	101.71	5.18687E-16	1.11171E-14	12.01	71.89
Salin1805	640472435	COG1819	Glycosyl transferases, related to UDP-glucuronosyltransferase	G	Carbohydrate transport and metabolism	3.83	9.14	87.77	1.25226E-14	2.14993E-13	14.89	62.17

GeneID	CNB440_JGI.ID	COG Function	Annotation of SalinGroup	CatCode	Category	logFC	logCPM	LR	pvalue	adj_pvalue	Fold	minFDR
Salin0955	640473613	COG0542	ATPases with chaperone activity, ATP-binding subunit	O	Posttranslational modification, protein turnover, chaperons	4.52	8.50	129.19	7.68558E-22	3.1159E-20	23.90	91.12
Salin2609	640473869	COG3115	Cell division protein	D	Cell cycle control, cell division, chromosome partitioning	4.65	8.93	118.13	2.91865E-20	1.0125E-18	26.15	83.39
Salin2709	640476825	COG1807	4-amino-4-deoxy-L-arabinose transferase and related glycosyltransferases of PMT family	M	Cell wall/membrane/envelope biogenesis	6.74	12.27	243.36	1.11912E-49	5.37976E-47	111.08	171.26

Appendix F

KEGG Maps





-  Upregulated Gene in *S. tropica* CNB-440
-  Downregulated Gene in *S. tropica* CNB-440
-  Positional Cluster Gene
-  Other genes in *S. tropica* CNB-440

Figure F.1: KEGG Map key of genes found in biosynthetic pathways of a genome. Green boxes indicate genes that are upregulated in *S. tropica* CNB-440. Red boxes indicate genes that are downregulated in *S. tropica* CNB-440. Yellow boxes indicate positional cluster genes that are co-located on the chromosome. Purple boxes indicate genes found in *S. tropica* CNB-440 that are also found in the pathway.

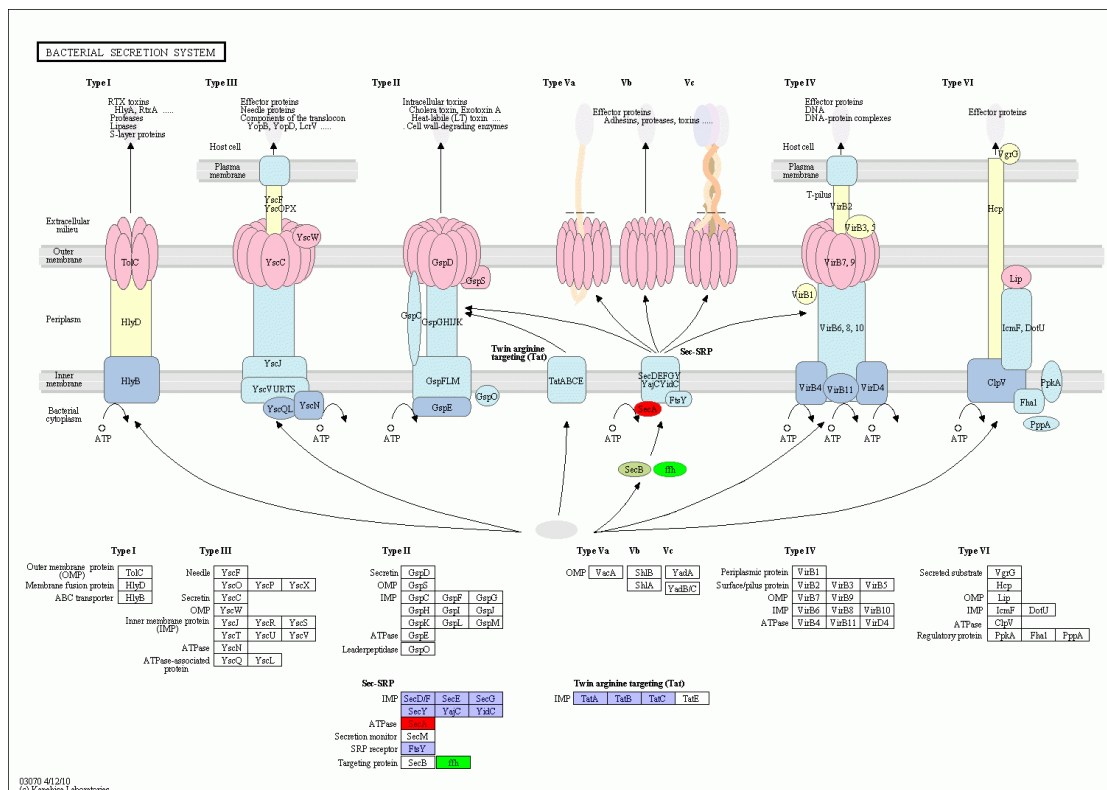


Figure F.3: Differential expression of genes during exponential phase growth found in pathway: Bacterial Secretion System in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

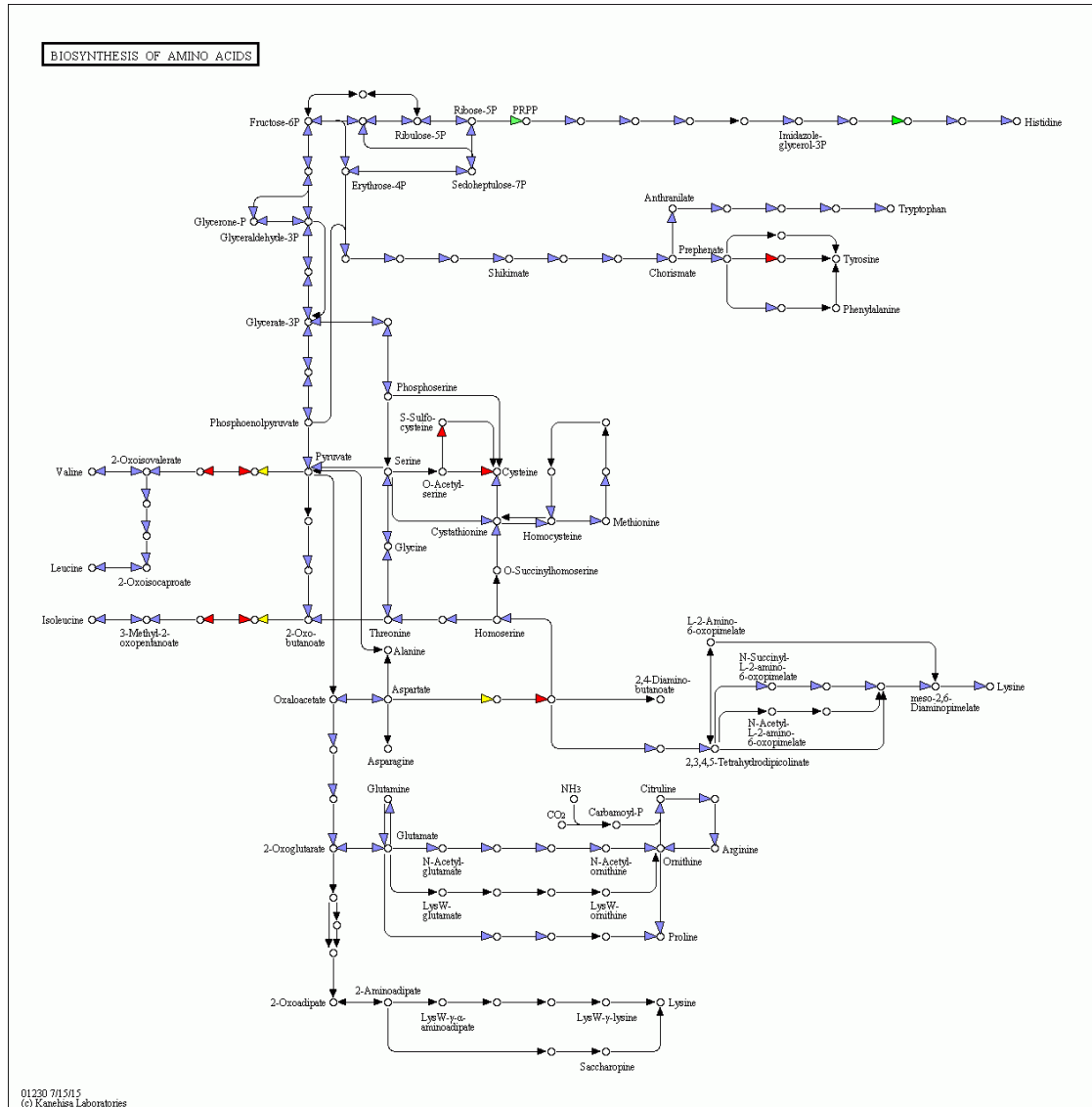


Figure F4: Differential expression of genes during exponential phase growth found in pathway: Biosynthesis of Amino Acids in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

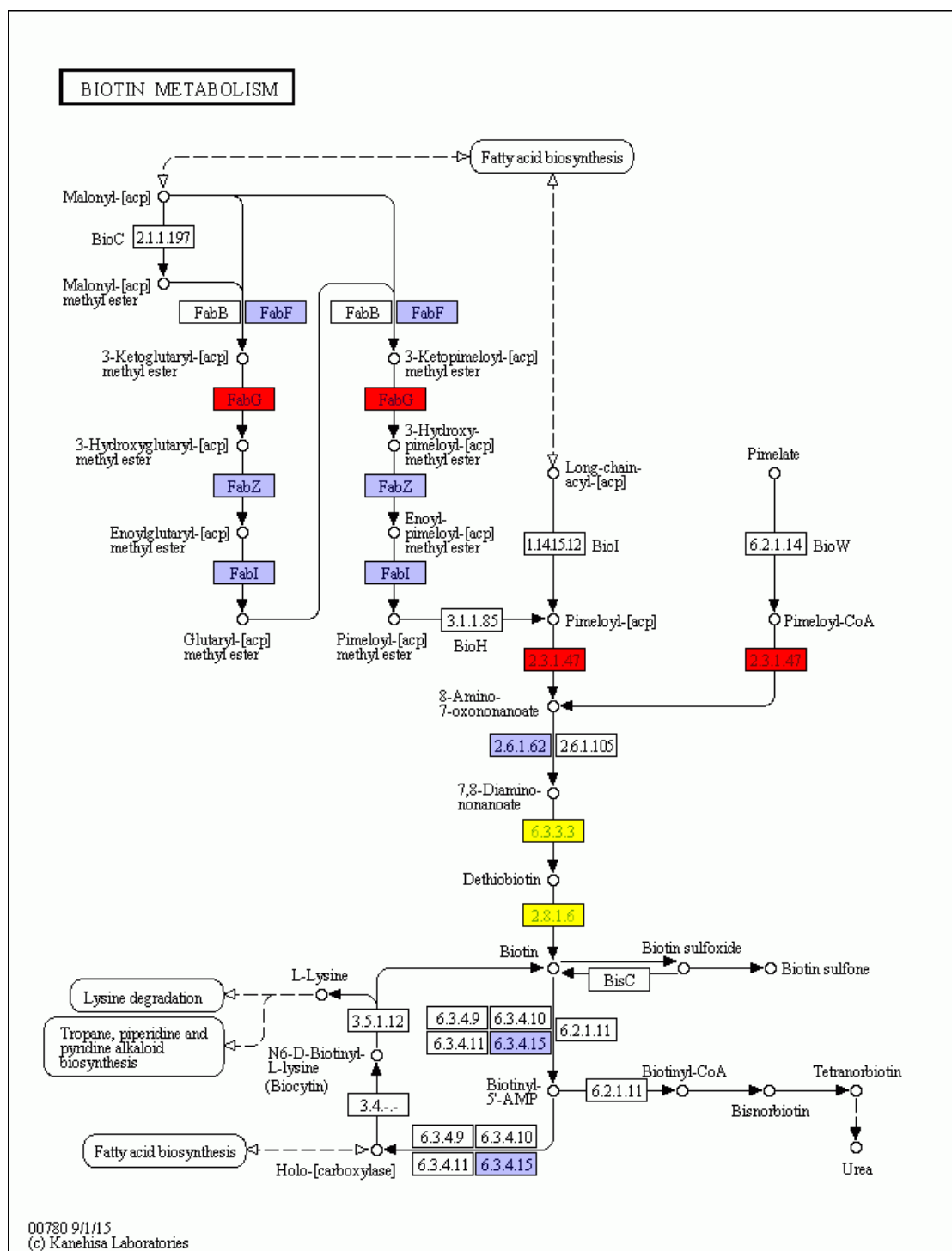


Figure F.5: Differential expression of genes during exponential phase growth found in pathway: Biotin Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

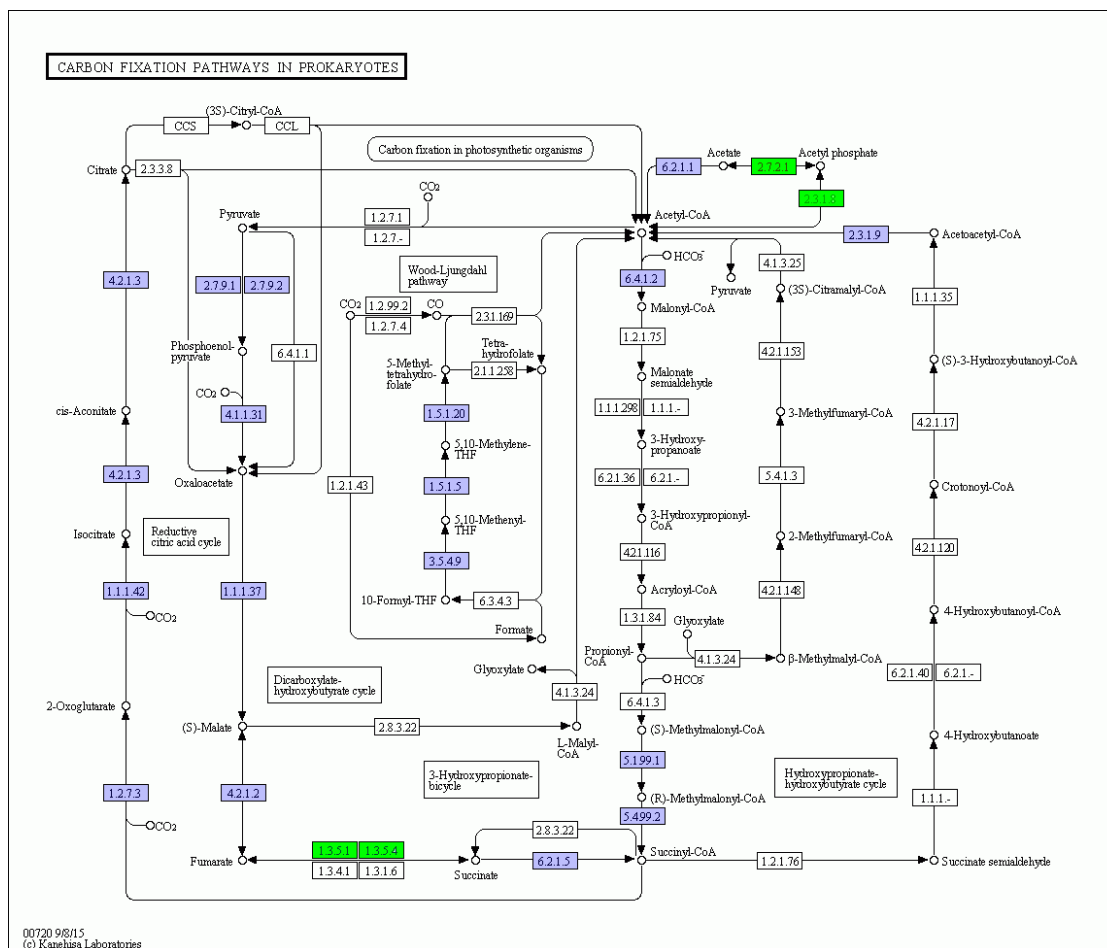


Figure F.7: Differential expression of genes during exponential phase growth found in pathway: Carbon Fixation Pathways Prokaryotes in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

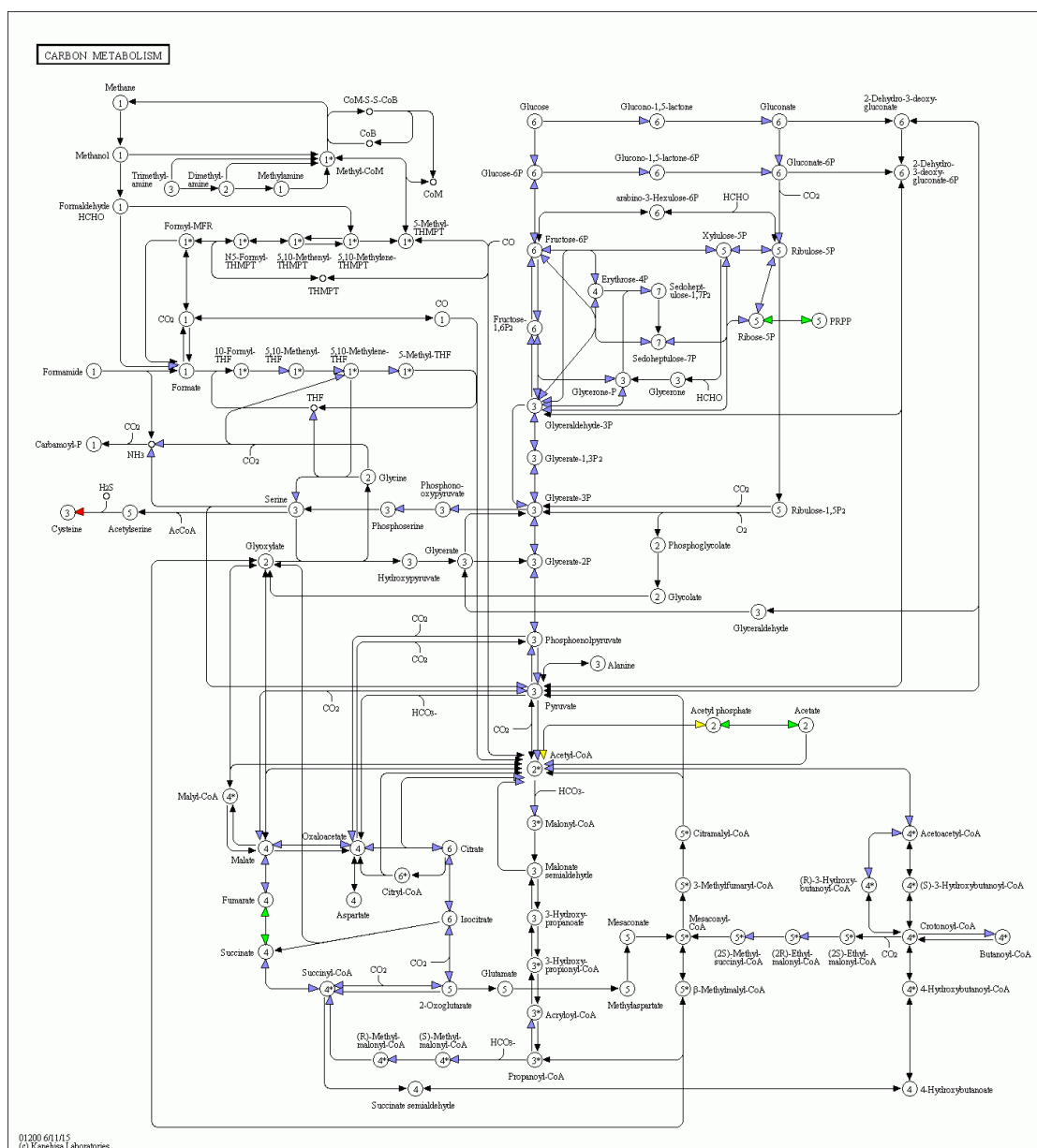


Figure E.8: Differential expression of genes during exponential phase growth found in pathway: Carbon Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

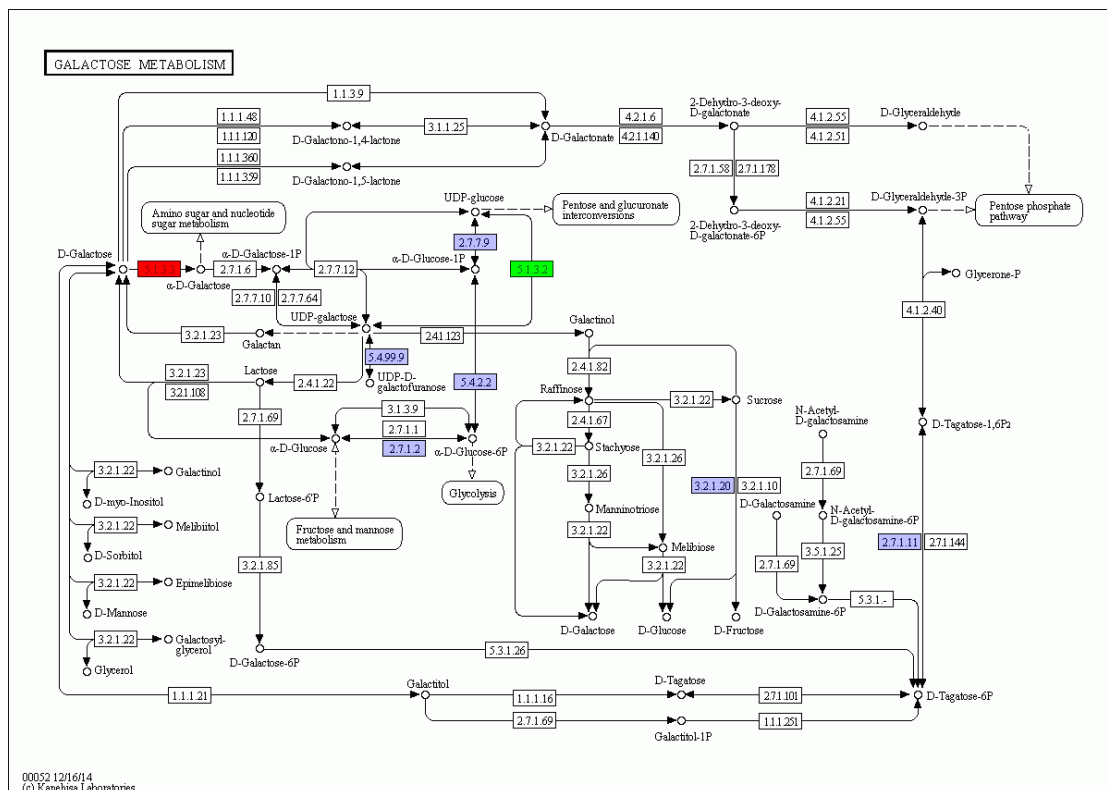


Figure F.10: Differential expression of genes during exponential phase growth found in pathway: Galactose Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

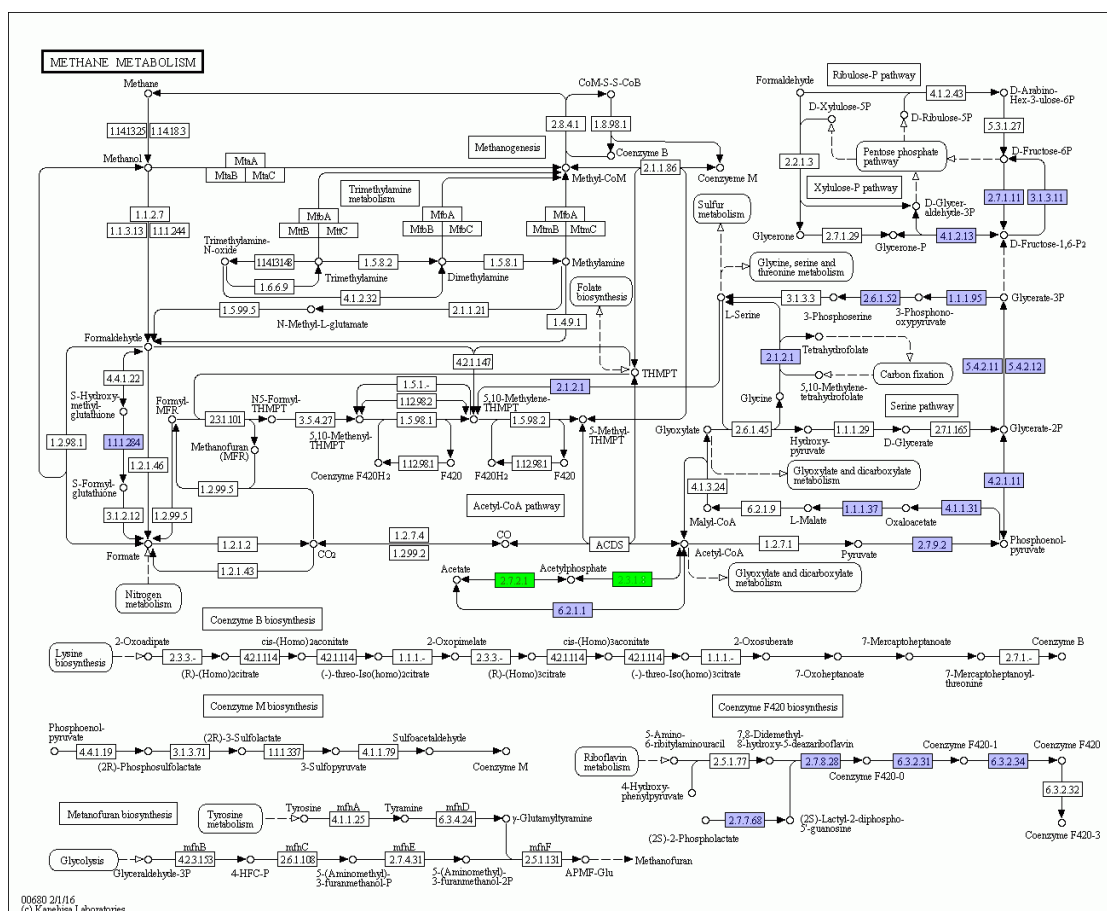


Figure F.11: Differential expression of genes during exponential phase growth found in pathway: Methane Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

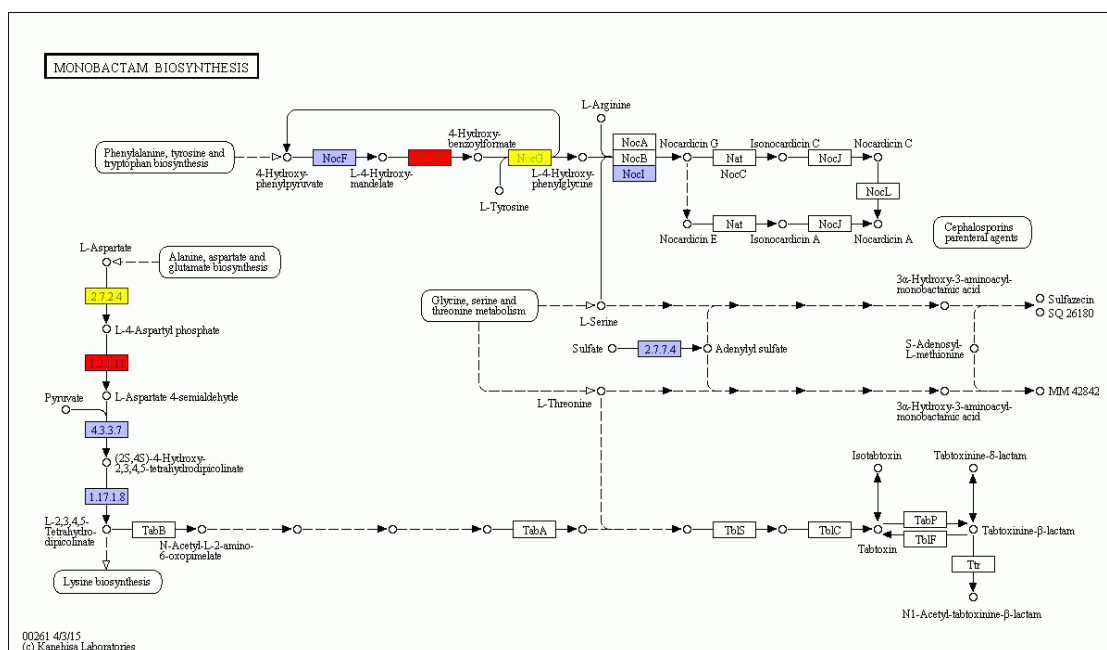


Figure F.12: Differential expression of genes during exponential phase growth found in pathway: Monobactam Biosynthesis in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

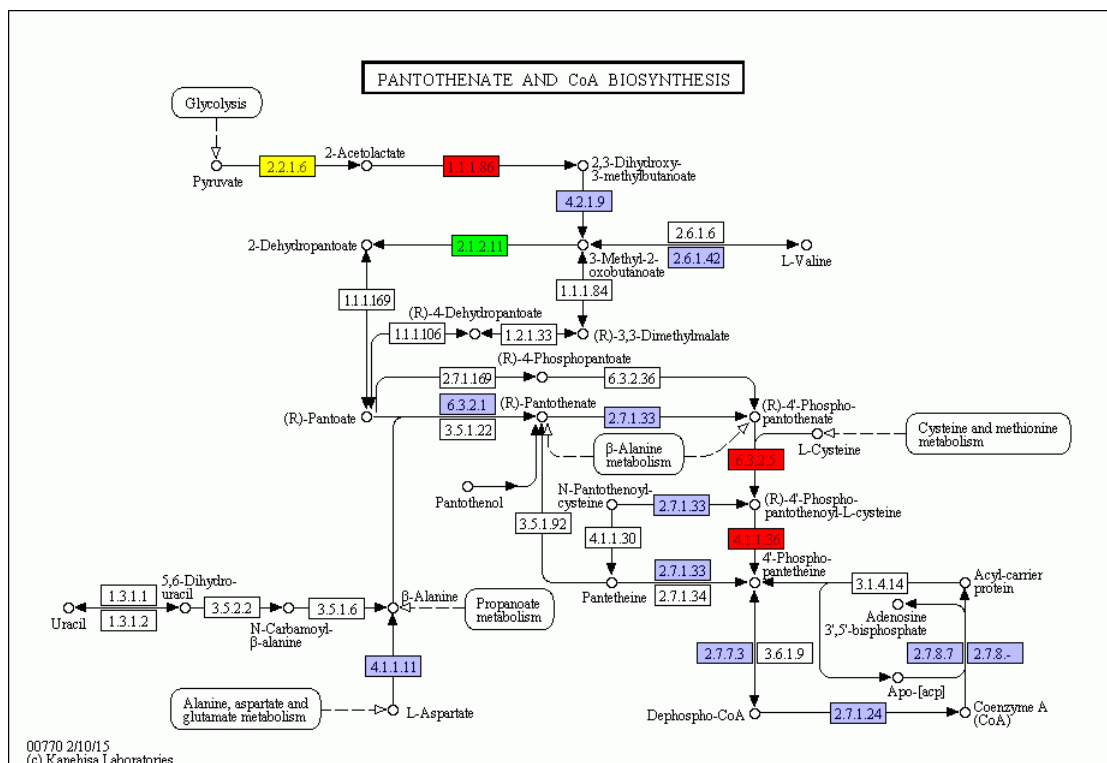


Figure F.14: Differential expression of genes during exponential phase growth found in pathway: Pantothenate CoA Biosynthesis in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

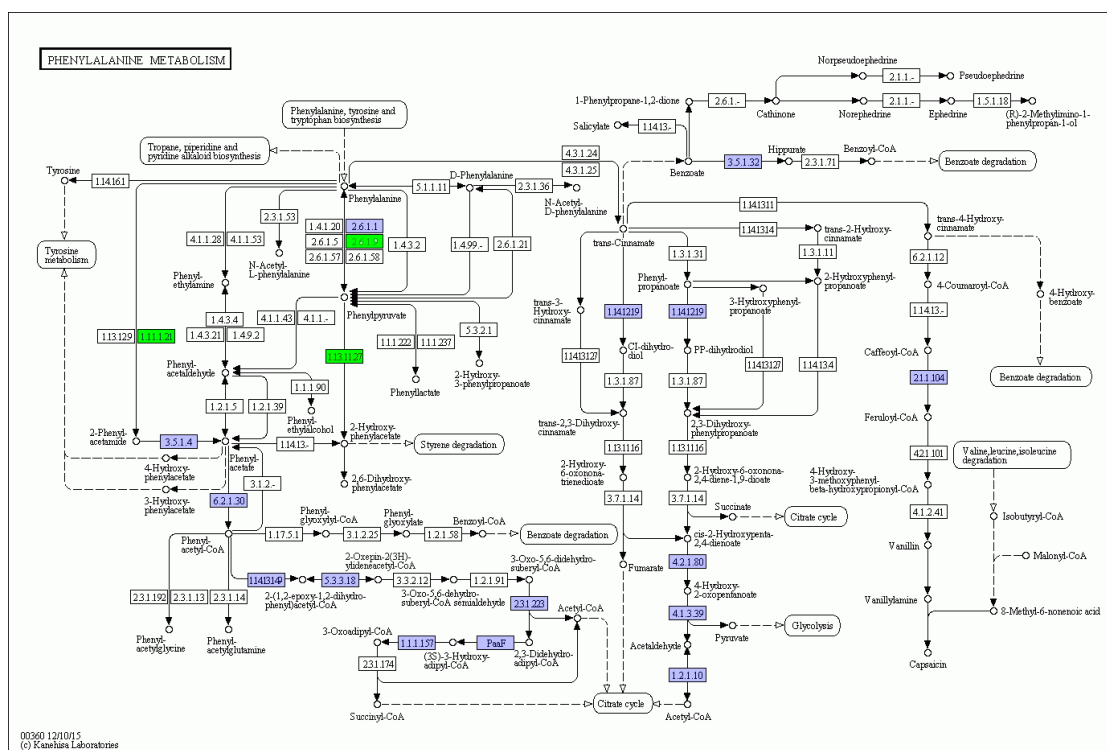


Figure F.15: Differential expression of genes during exponential phase growth found in pathway: Phenylalanine Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

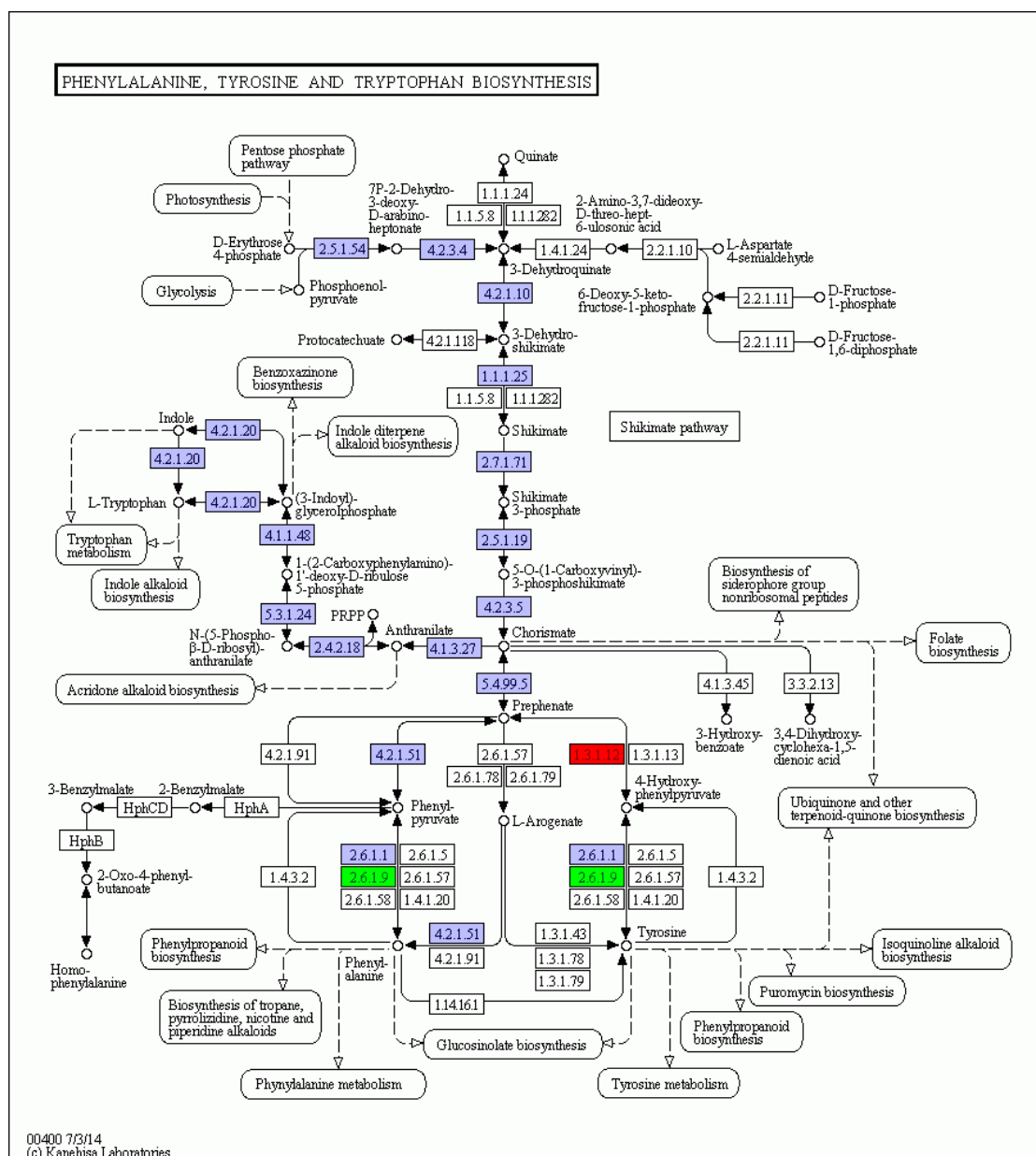


Figure F.16: Differential expression of genes during exponential phase growth found in pathway: Phenylalanine Tyrosine Tryptophan Biosynthesis in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

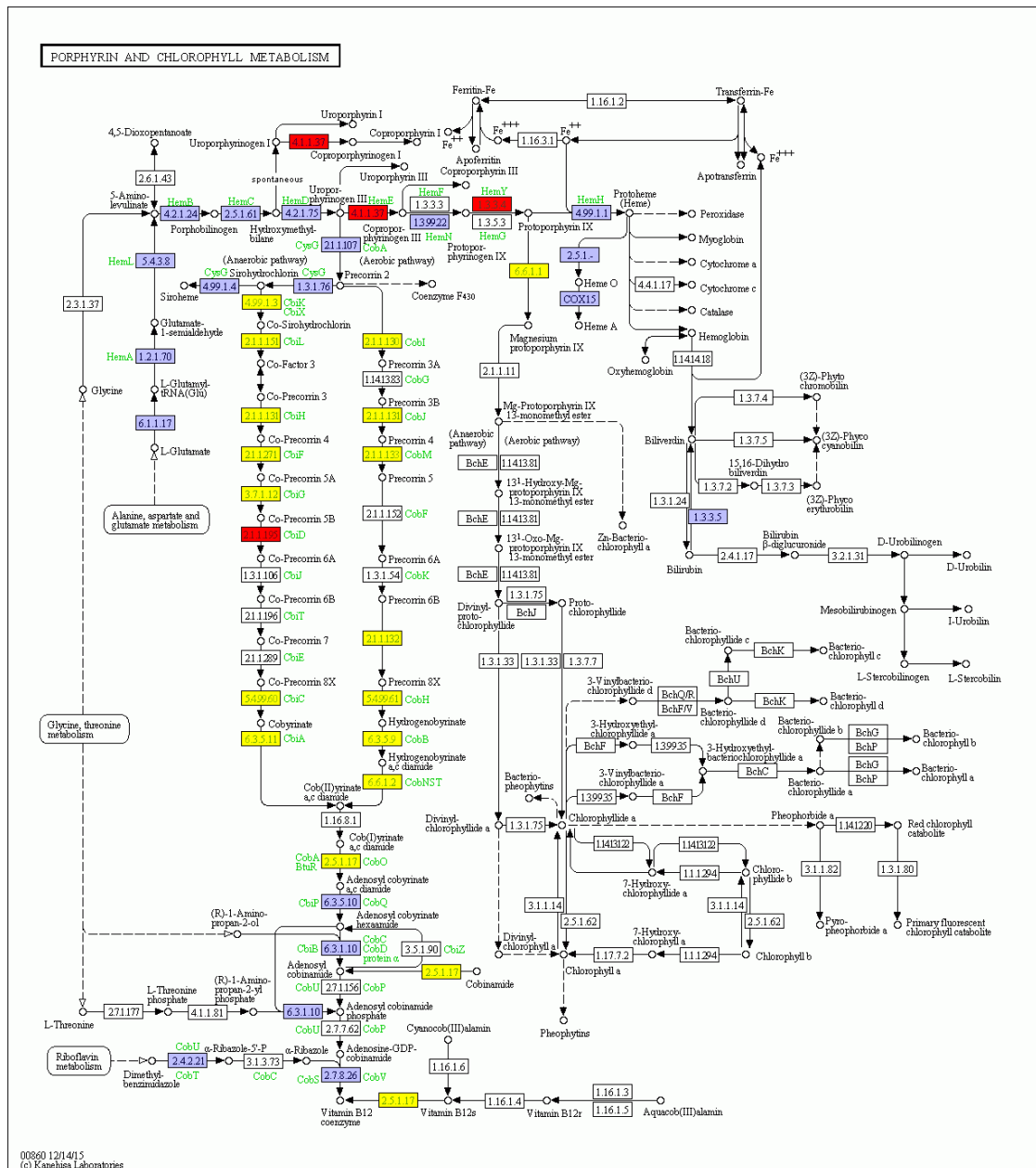


Figure F.17: Differential expression of genes during exponential phase growth found in pathway: Porphyrin Chlorophyll Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

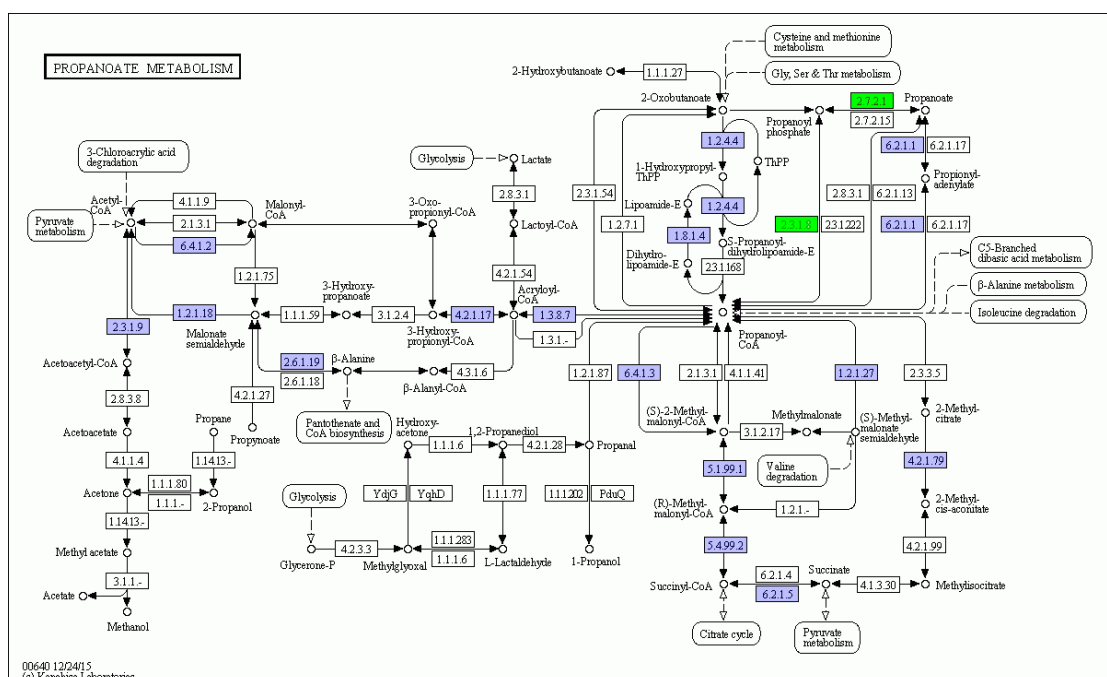


Figure F.18: Differential expression of genes during exponential phase growth found in pathway: Propanoate Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

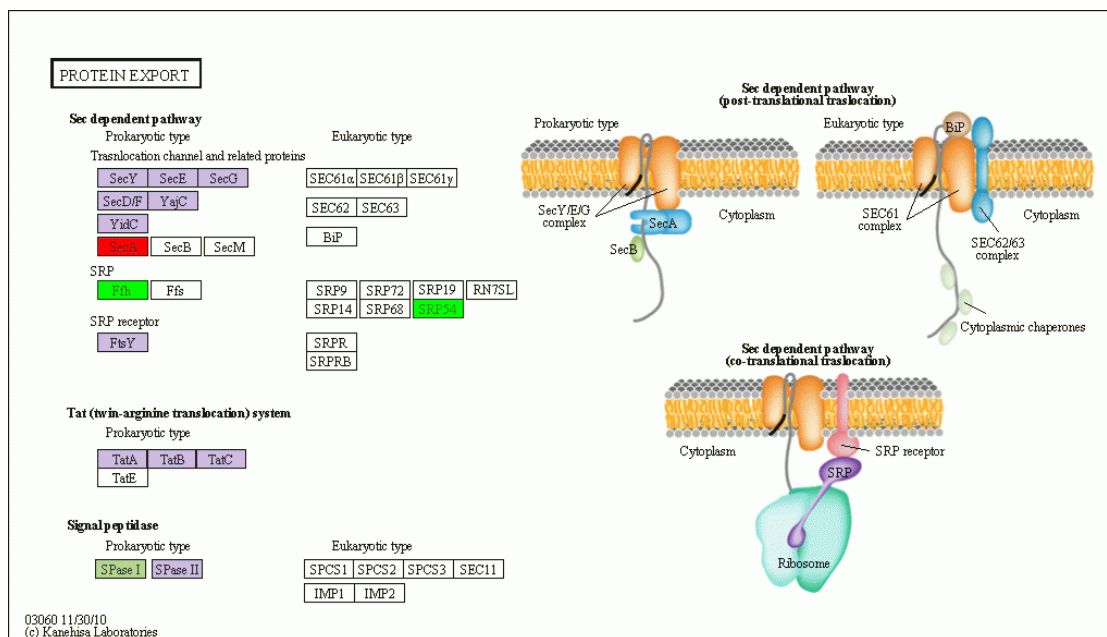


Figure F.19: Differential expression of genes during exponential phase growth found in pathway: Protein Export in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

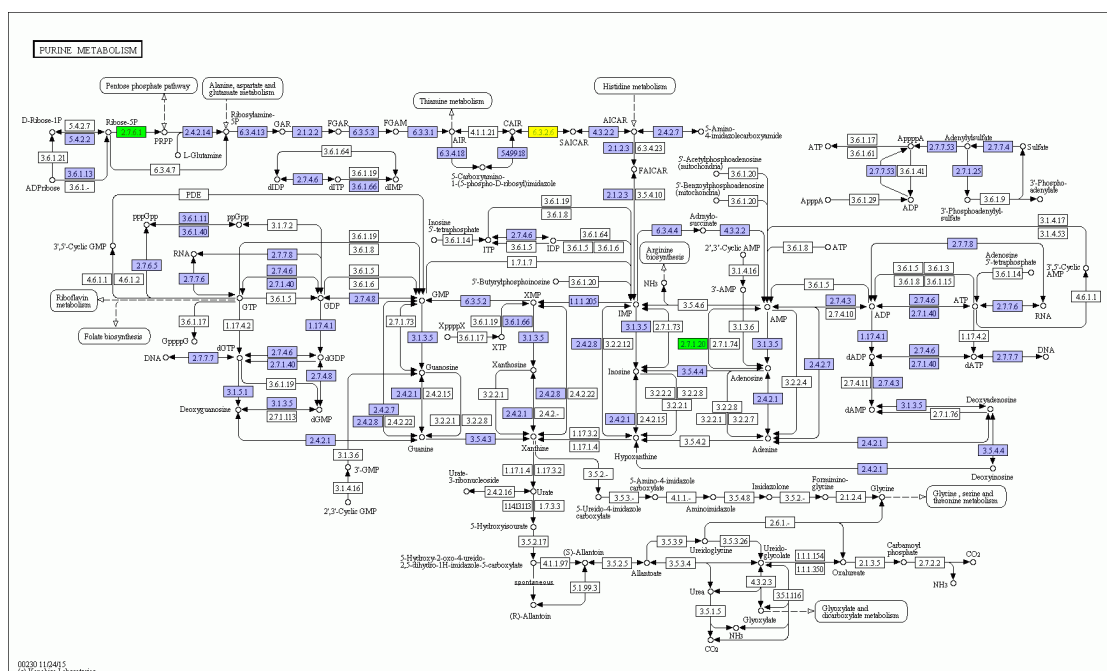


Figure F.20: Differential expression of genes during exponential phase growth found in pathway: Purine Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

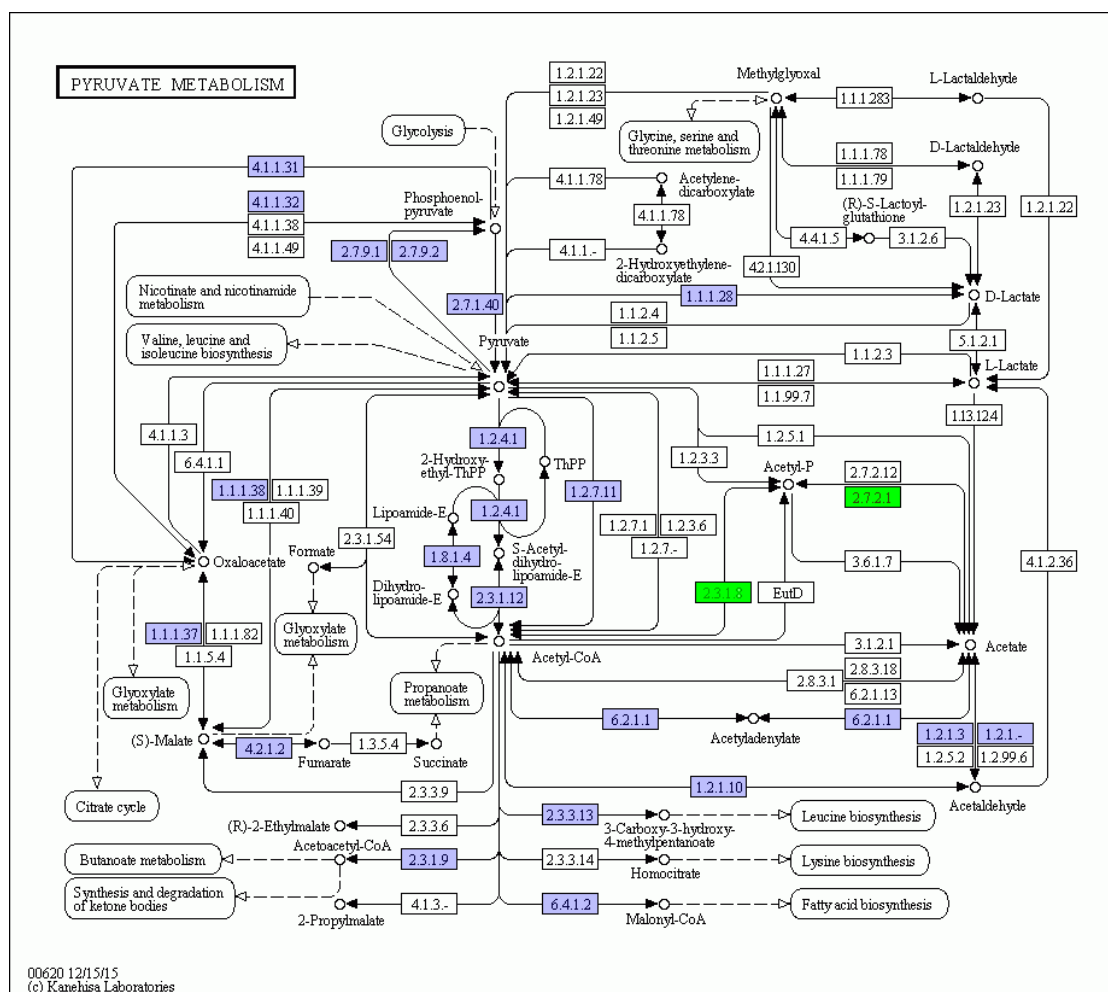


Figure F.22: Differential expression of genes during exponential phase growth found in pathway: Pyruvate Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

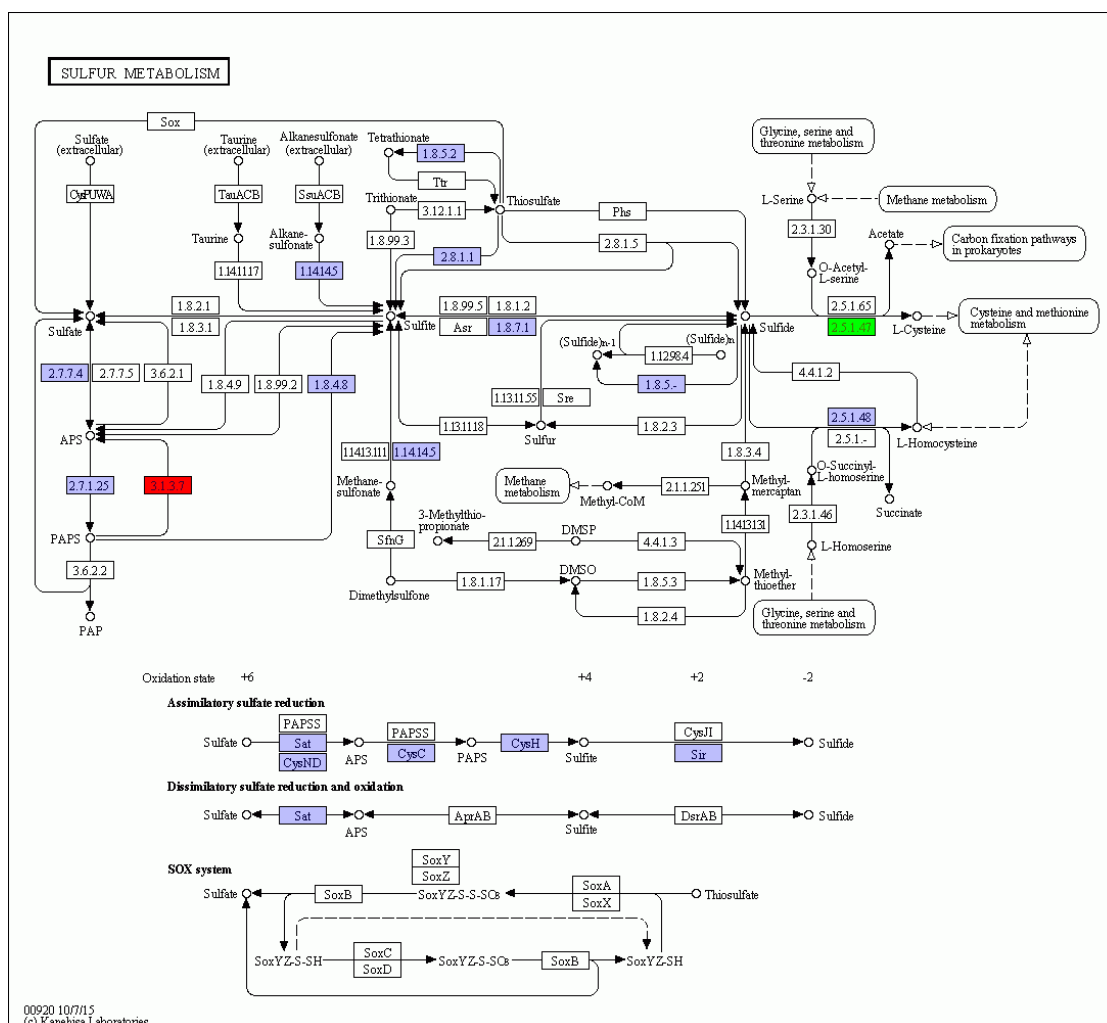


Figure F.23: Differential expression of genes during exponential phase growth found in pathway: Sulfur Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

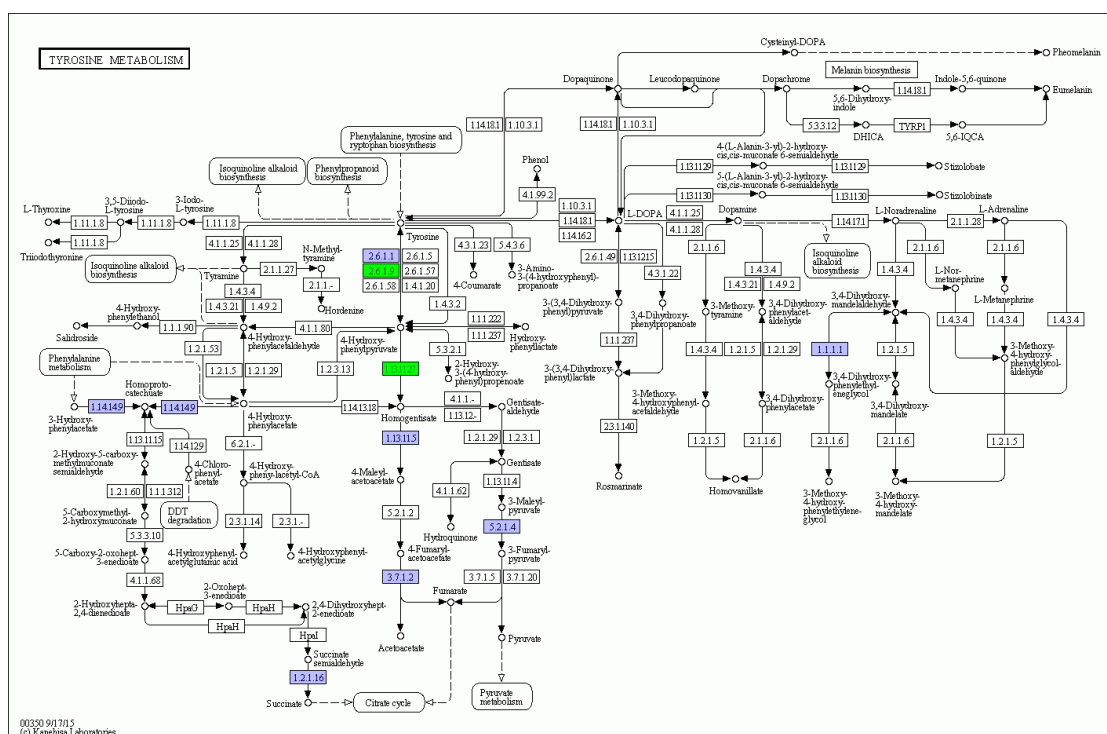


Figure F.26: Differential expression of genes during exponential phase growth found in pathway: Tyrosine Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

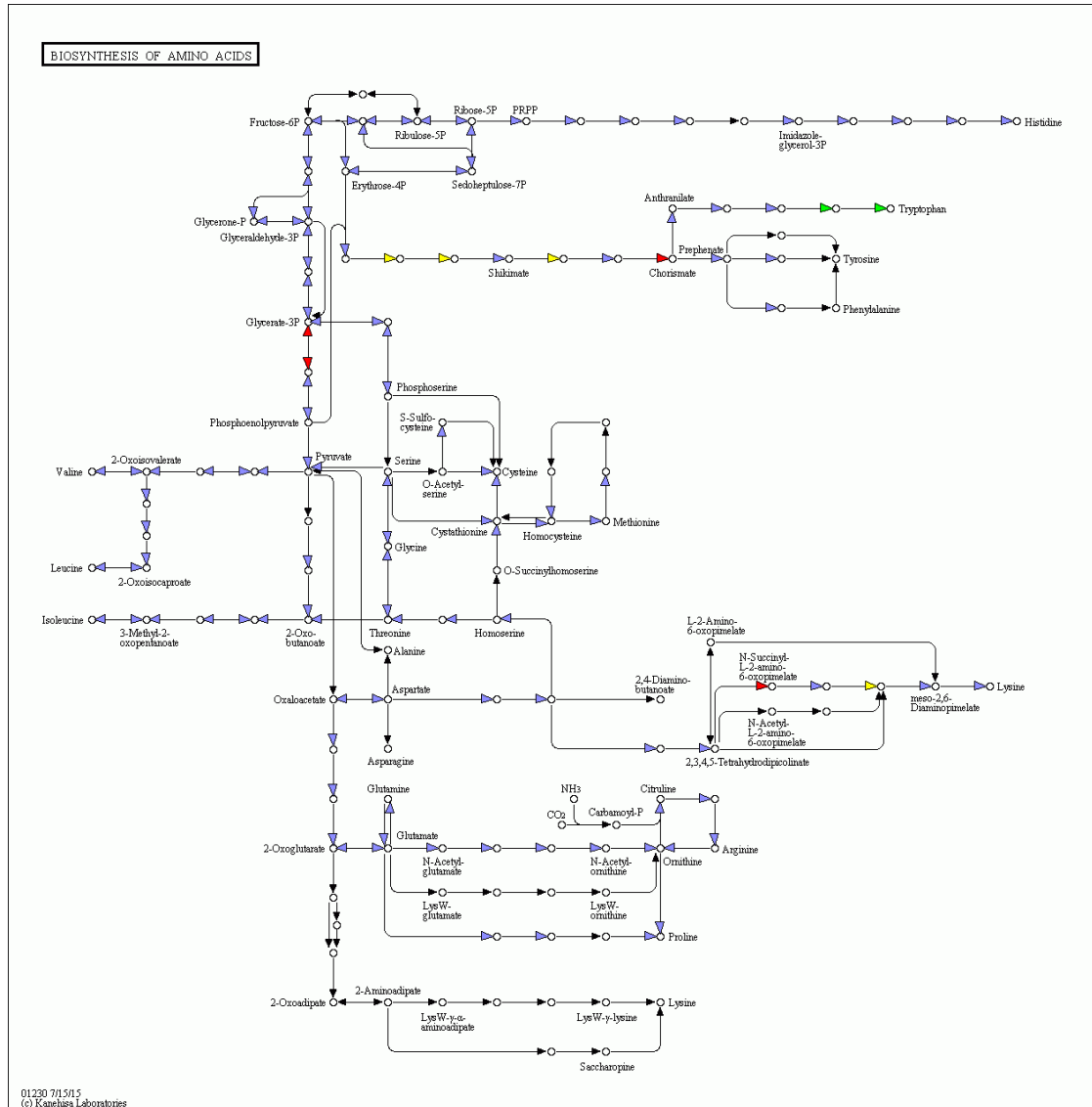


Figure F.28: Differential expression of genes during stationary phase growth found in pathway: Biosynthesis of Amino Acids in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

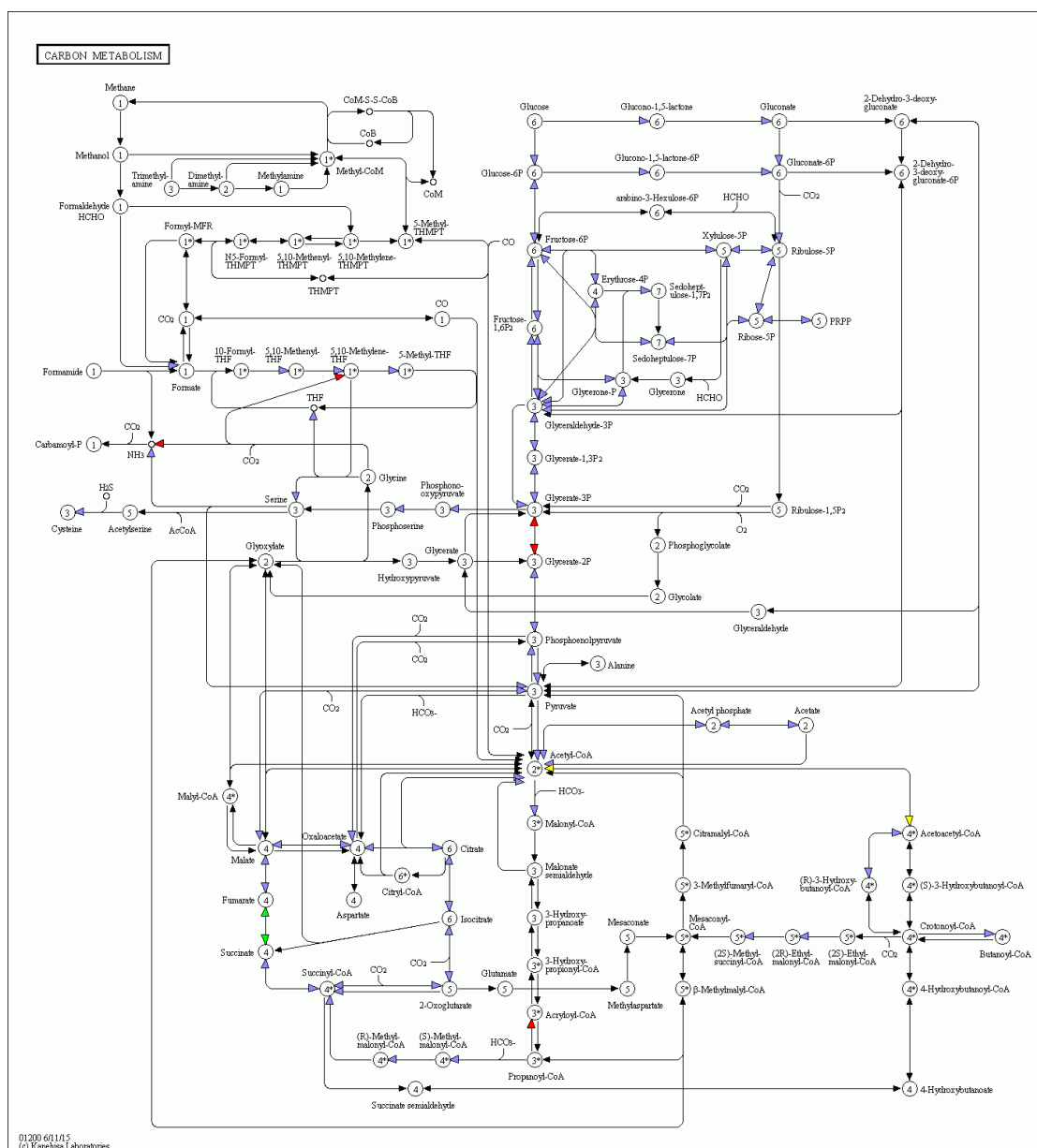


Figure F.29: Differential expression of genes during stationary phase growth found in pathway: Carbon Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

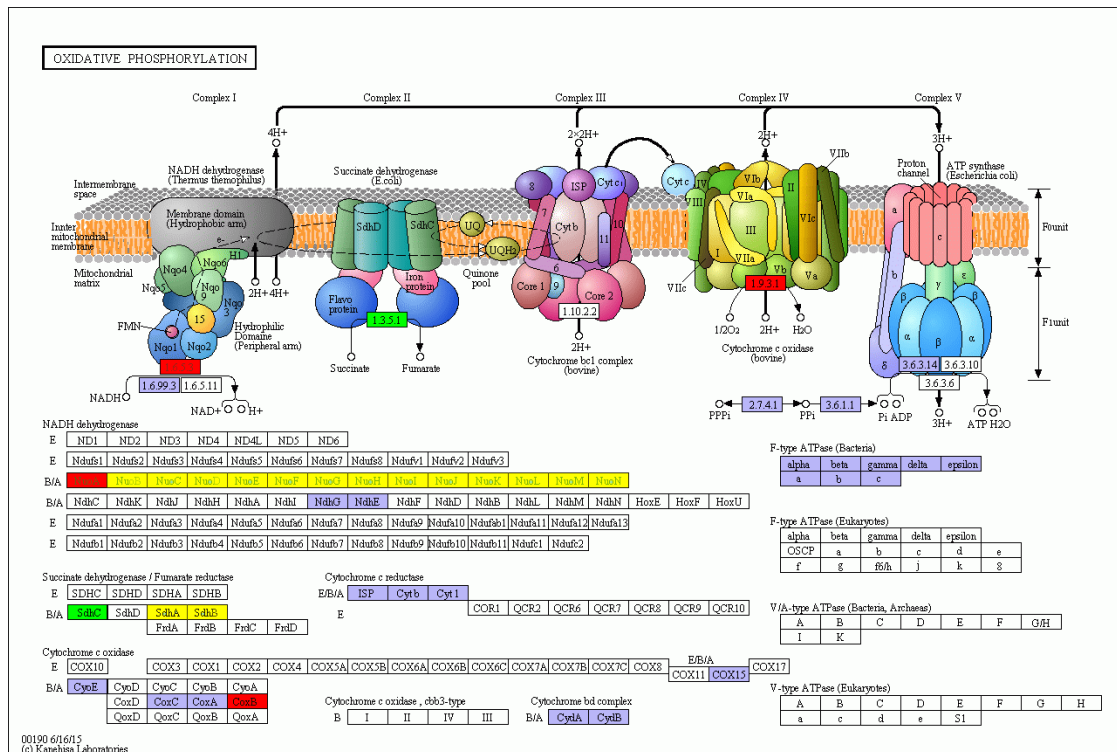


Figure F32: Differential expression of genes during stationary phase growth found in pathway: Oxidative Phosphorylation in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

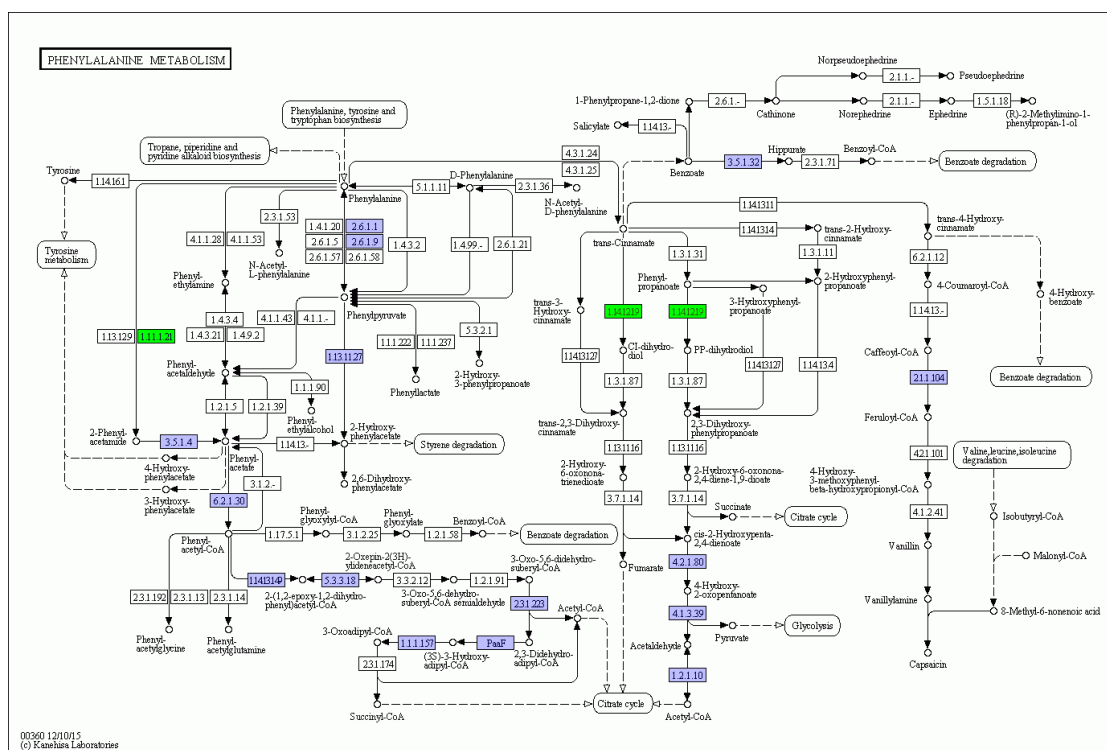


Figure F.33: Differential expression of genes during stationary phase growth found in pathway: Phenylalanine Metabolism in strain *S. tropica* CNB-440 compared to *S. arenicola* strains CNS-205 and DSM45545.

Appendix G

GenProp0799

Gene Product Name	Gene Symbol	Accession ID	CNB-440 Gene ID	Salin Group
Alanyl-tRNA synthetase	alaS	TIGR00344: alanine-tRNA ligase	640474098	Salin530
Arginyl-tRNA synthetase	argS	pfam00750: arginine-tRNA ligase	640475908	Salin688
Aspartyl-tRNA synthetase	aspS	TIGR00459: aspartate-tRNA ligase	640474091	Salin224
Obg family GTPase CgtA	cgtA	TIGR02729: Obg family GTPase CgtA	640475728	Salin682
dephospho-CoA kinase	coaE	TIGR00152: dephospho-CoA kinase	640475386	Salin745
cysteinyl-tRNA synthetase	cysS	TIGR00435: cysteine-tRNA ligase	640474431	Salin651
chromosomal replication initiator protein DnaA	dnaA	TIGR00362: chromosomal replication initiator protein DnaA	640472264	Salin1028
DNA primase	dnaG	TIGR01391: DNA primase	640475674	Salin390
chaperone protein DnaK	dnaK	TIGR02350: chaperone protein DnaK	640472380	Salin143
DNA polymerase III, beta subunit	dnaN	TIGR00663: DNA polymerase III, beta subunit	640472266	Salin1736
DNA polymerase III, subunits gamma and tau	dnaX	TIGR02397: DNA polymerase III, subunit gamma and tau	640472485	Salin652
ribosome-associated GTPase EngA	engA	TIGR03594: ribosome-associated GTPase EngA	640474203	Salin659
GTP-binding protein Era	era	TIGR00436: GTP-binding protein Era	640475698	Salin847
signal recognition particle protein	ffh	TIGR00959: signal recognition particle protein	640473573	Salin677
methionyl-tRNA formyltransferase	fmt	TIGR00460: methionyl-tRNA formyltransferase	640474139	Salin826
ribosome recycling factor	frr	TIGR00496: ribosome recycling factor	640473597	Salin894
signal recognition particle-docking protein FtsY	ftsY	TIGR00064: signal recognition particle-docking protein FtsY	640473567	Salin551
glycyl-tRNA synthetase	glyS	TIGR00389: glycine-tRNA ligase	640475682	Salin514
guanylate kinase	gmK	TIGR03263: guanylate kinase	640474131	Salin782
co-chaperone GrpE	grpE	pfam01025: co-chaperone GrpE	640472381	Salin301
DNA gyrase, A subunit	gyrA	TIGR01063: DNA gyrase, A subunit	640472272	Salin696
DNA gyrase, B subunit	gyrB	TIGR01059: DNA gyrase, B subunit	640472271	Salin1045
histidyl-tRNA synthetase	hisS	TIGR00442: histidine-tRNA ligase	640474089	Salin1722
isoleucyl-tRNA synthetase	ileS	TIGR00392: isoleucine-tRNA ligase	640475736	Salin111
translation initiation factor IF-2	infB	TIGR00487: translation initiation factor IF-2	640473639	Salin427
translation initiation factor IF-3	infC	TIGR00168: translation initiation factor IF-3	640474152	Salin531
dimethyladenosine transferase	ksgA	TIGR00755: ribosomal RNA small subunit methyltransferase A	640473050	Salin382
GTP-binding protein LepA	lepA	TIGR01393: elongation factor 4	640475711	Salin526
leucyl-tRNA synthetase	leuS	TIGR00396: leucine-tRNA ligase	640475473	Salin134
DNA ligase, NAD-dependent	ligA	TIGR00575: DNA ligase, NAD-dependent	640473486	Salin498
tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase	mnmA	TIGR00420: tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase	640473483	Salin761
MraW methylase family	mraW	pfam01795: MraW methylase family	640475468	Salin868
transcription termination factor NusA	nusA	TIGR01953: transcription termination factor NusA	640473637	Salin596
transcription termination/antitermination factor NusG	nusG	TIGR00922: transcription termination/antitermination factor NusG	640476195	Salin576
phosphoglycerate kinase	pgk	pfam00162: phosphoglycerate kinase	640475347	Salin472
phenylalanyl-tRNA synthetase, alpha subunit	pheS	TIGR00468: phenylalanine-tRNA ligase, alpha subunit	640474156	Salin622
phenylalanyl-tRNA synthetase, beta subunit	pheT	TIGR00471: phenylalanine-tRNA ligase, beta subunit	640474157	Salin743
peptide chain release factor 1	prfA	TIGR00019: peptide chain release factor 1	640475899	Salin533
prolyl-tRNA synthetase	proS	TIGR00408: proline-tRNA ligase	640473575	Salin2470
CTP synthase	pyrG	TIGR00337: CTP synthase	640474193	Salin449
recA protein	recA	TIGR02012: protein RecA	640473693	Salin334
ribosome-binding factor A	rfaA	TIGR00082: ribosome-binding factor A	640473641	Salin1427
ribonuclease III	rnc	TIGR02191: ribonuclease III	640473559	Salin410
ribosomal protein L1	rplA	TIGR01169: ribosomal protein uL1	640476193	Salin503
ribosomal protein L2	rplB	TIGR01171: ribosomal protein uL2	640476177	Salin673

Gene Product Name	Gene Symbol	Accession ID	CNB-440 Gene ID	Salin Group
ribosomal protein L3	rplC	pfam00297: ribosomal protein uL3	640476180	Salin585
ribosomal protein L4	rplD	pfam00573: ribosomal protein uL4	640476179	Salin843
ribosomal protein L5	rplE	pfam00281: ribosomal protein uL5	640476168	Salin529
ribosomal protein L6	rplF	pfam00347: ribosomal protein uL6	640476165	Salin720
ribosomal protein L9	rplI	TIGR00158: ribosomal protein bL9	640476820	Salin1090
ribosomal protein L10	rplJ	pfam00466: ribosomal protein uL10	640476192	Salin765
ribosomal protein L11	rplK	TIGR01632: ribosomal protein uL11	640476194	Salin835
ribosomal protein L7/L12	rplL	TIGR00855: ribosomal protein bL12	640476191	Salin640
ribosomal protein L13	rplM	TIGR01066: ribosomal protein uL13	640476121	Salin711
ribosomal protein L14	rplN	TIGR01067: ribosomal protein uL14	640476170	Salin405
ribosomal protein L15	rplO	TIGR01071: ribosomal protein uL15	640476161	Salin685
ribosomal protein L16	rplP	TIGR01164: ribosomal protein uL16	640476173	Salin740
ribosomal protein L17	rplQ	TIGR00059: ribosomal protein bL17	640476151	Salin779
ribosomal protein L18	rplR	TIGR00060: ribosomal protein uL18	640476164	Salin603
ribosomal protein L19	rplS	TIGR01024: ribosomal protein bL19	640473581	Salin457
ribosomal protein L20	rplT	TIGR01032: ribosomal protein bL20	640474154	Salin407
ribosomal protein L21	rplU	TIGR00061: ribosomal protein bL21	640475730	Salin856
ribosomal protein L22	rplV	TIGR01044: ribosomal protein uL22	640476175	Salin804
ribosomal protein L23	rplW	pfam00276: ribosomal protein uL23	640476178	Salin416
ribosomal protein L24	rplX	TIGR01079: ribosomal protein uL24	640476169	Salin661
ribosomal protein L27	rpmA	TIGR00062: ribosomal protein bL27	640475729	Salin430
ribosomal protein L28	rpmB	TIGR00009: ribosomal protein bL28	640473549	Salin694
ribosomal protein L29	rpmC	TIGR00012: ribosomal protein uL29	640476172	Salin619
ribosomal protein L32	rpmF	TIGR01031: ribosomal protein bL32	640473557	Salin1455
ribosomal protein L34	rpmH	TIGR01030: ribosomal protein bL34	640476860	Salin1115
ribosomal protein L35	rpmI	TIGR00001: ribosomal protein bL35	640474153	Salin1522
DNA-directed RNA polymerase, alpha subunit	rpoA	TIGR02027: DNA-directed RNA polymerase, alpha subunit	640476152	Salin279
DNA-directed RNA polymerase, beta subunit	rpoB	TIGR02013: DNA-directed RNA polymerase, beta subunit	640476190	Salin902
DNA-directed RNA polymerase, beta' or beta'' subunit	rpoC	TIGR02386: DNA-directed RNA polymerase, beta' subunit	640476189	Salin386
ribosomal protein S2	rpsB	TIGR01011: ribosomal protein uS2	640473594	Salin495
ribosomal protein S3	rpsC	TIGR01009: ribosomal protein uS3	640476174	Salin480
ribosomal protein S4	rpsD	TIGR01017: ribosomal protein uS4	640476153	Salin1086
ribosomal protein S4	rpsD	TIGR01017: ribosomal protein uS4	640476632	Salin2125
ribosomal protein S5	rpsE	TIGR01021: ribosomal protein uS5	640476163	Salin860
ribosomal protein S6	rpsF	TIGR00166: ribosomal protein bS6	640476823	Salin1051
ribosomal protein S7	rpsG	TIGR01029: ribosomal protein uS7	640476184	Salin507
ribosomal protein S8	rpsH	pfam00410: ribosomal protein uS8	640476166	Salin467
ribosomal protein S9	rpsI	pfam00380: ribosomal protein uS9	640476120	Salin594
ribosomal protein S10	rpsJ	TIGR01049: ribosomal protein uS10	640476181	Salin701
ribosomal protein S11	rpsK	pfam00411: ribosomal protein uS11	640476154	Salin396
ribosomal protein S12	rpsL	TIGR00981: ribosomal protein uS12	640476185	Salin644
ribosomal protein S13	rpsM	pfam00416: ribosomal protein uS13	640476155	Salin893
ribosomal protein S15	rpsO	TIGR00952: ribosomal protein uS15	640473649	Salin489
ribosomal protein S16	rpsP	TIGR00002: ribosomal protein bS16	640473577	Salin425
ribosomal protein S17	rpsQ	pfam00366: ribosomal protein uS17	640476171	Salin874
ribosomal protein S18	rpsR	TIGR00165: ribosomal protein bS18	640476821	Salin1065
ribosomal protein S19	rpsS	TIGR01050: ribosomal protein uS19	640476176	Salin544
ribosomal protein S20	rpsT	TIGR00029: ribosomal protein bS20	640475714	Salin655
preprotein translocase, SecA subunit	secA	TIGR00963: preprotein translocase, SecA subunit	640473242	Salin710
preprotein translocase, SecE subunit	secE	TIGR00964: preprotein translocase, SecE subunit	640476196	Salin1141
preprotein translocase, SecG subunit	secG	TIGR00810: preprotein translocase, SecG subunit	640475345	Salin1645
preprotein translocase, SecY subunit	secY	TIGR00967: preprotein translocase, SecY subunit	640476160	Salin566
seryl-tRNA synthetase	serS	TIGR00414: serine-tRNA ligase	640475645	Salin333
SmpB protein	smpB	TIGR00086: SsrA-binding protein	640473252	Salin632
threonyl-tRNA synthetase	thrS	TIGR00418: threonine-tRNA ligase	640474054	Salin394
trigger factor	tig	TIGR00115: trigger factor	640475752	Salin760

Gene Product Name	Gene Symbol	Accession ID	CNB-440 Gene ID	Salin Group
tRNA(Ile)-lysine synthetase	tilS	TIGR02432: tRNA(Ile)-lysine synthetase	640476572	Salin795
tRNA threonylcarbamoyladenosine modification protein TsaD	tsaD	TIGR03723: tRNA threonylcarbamoyladenosine modification protein TsaD	640476108	Salin537
translation elongation factor Ts	tsf	TIGR00116: translation elongation factor Ts	640473595	Salin756
tyrosyl-tRNA synthetase	tyrS	TIGR00234: tyrosine-tRNA ligase	640474175	Salin752
excinuclease ABC, B subunit	uvrB	TIGR00631: excinuclease ABC subunit B	640475384	Salin880
valyl-tRNA synthetase	valS	TIGR00422: valine-tRNA ligase	640475742	Salin1449
16S rRNA maturation RNase YgeY	ybeY	TIGR00043: rRNA maturation RNase YbeY	640475701	Salin675
GTP-binding protein YchF	ychF	TIGR00092: GTP-binding protein YchF	640476039	Salin697

Appendix H

Biosynthetic Gene Cluster Presence/Absence Matrix

Presence/Absence matrix of biosynthetic gene clusters found in greater than 10 strains (48 in total). This matrix was used for the likelihood analysis conducted in Chapter 4. The last row indicates the lowest number of evolutionary steps for the most parsimonious gain/loss events.

Strain	PKS4	sio	Bac2	aminoacylcl	NRPS4	salinipositi	lym	des	sta	sid1/2	NRPS1	PKS2	Rif	PKS1A	PKS3A	PKS3B	Terp1	NRPS2	Bac4	PKS16	PKS5	PKS1C	cya	slm	sal	NRPS19	NRPS20	sid5	salinicheilin	Terp6	PKS15	PKS19	NRPS14	PKS7	Lan1	PKSNRPS2	Lan9	NRPS3	sid3	sid4	PKS25	Betalactame	NRPS27	
SisCNB336 BA	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SisCNY012 BA	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SisCNH898 BA	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SisCNT261 BA	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SisCNB476 BA	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SisCNCN416 BA	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SisCNT250 BA	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SisCNS197 BA	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SisCNR699 BA	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SisCNY681 YU	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SisCNY678 YU	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SisCNB440 BA	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNT569 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNR942 PL	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNY202 SC	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNY646 RS	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNS237 PL	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpDSM45549 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNT854 HI	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNT584 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNT124 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpDSM45547 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNT029 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNY703 MD	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNY673 MD	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNT045 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNS996 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNT403 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNS860 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpDSM45543 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNY666 MD	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNS055 PL	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpDSM45548 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNS801 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNT609 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNT084 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SpCNT133 FJ	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

[illegible]

Strain	PKS4	sio	Bac2	aminocyclitol	NRP54	saliniposin	lym	des	sta	sid1/2	NRP51	PKS2	Lan2	Rif	PKS1A	PKS3A	PKS1B	Terp1	NRP52	Bac4	PKS16	PKS5	lom	PKS1C	cya	slm	sal	NRP519	NRP520	sid5	salinichelin	Terp6	PKS15	PKS19	NRP514	PKS7	Lan1	PKS2NRP52	Lan9	NRP53	sid3	sid4	spo	PKS25	Betalactame	PKS12	NRP527		
SaCNT859 HI	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNT857 HI	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNT850 HI	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNT799 HI	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNT849 HI	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNT798 HI	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNT800 HI	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNY011 BA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNH877 BA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNY679 YU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNH643 BA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNH646 BA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNH905 BA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNX482 PM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNX814 PM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNX508 PM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNX481 PM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNX891 PM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNH718 RS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNR425 GU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNR107 GU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNQ748 GU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SaCNQ884 GU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1</																																

Strain	PKS4	sio	Bac2	aminocyclitol	NRPS4	salinipostin	lym	des	sta	sid1/2	NRPS1	PKS2	Lan2	Rif	PKS1A	PKS3A	PKS3B	PKS1B	Terp1	NRPS2	Bac4	PKS16	PKS5	lom	PKS1C	cya	slm	sal	NRPS19	NRPS20	sid5	salinichelin	Terp6	PKS15	PKS19	NRPS14	PKS7	Lan1	PKS5NRPS2	Lan9	NRPS3	sid3	sid4	spo	PKS25	Betalactame	PKS12	NRPS27		
SaCNY244 FJ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
SaCNS342 FJ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
SaCNY237 FJ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
SaCNY282 FJ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
SaCNY234 FJ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
Micromonospora aurantiaca	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
ATCC27029																																																		
Number of Steps	1	1	1	3	4	4	4	4	3	9	3	4	3	1	1	1	1	1	1	1	1	8	4	8	3	3	5	3	8	5	5	2	2	2	1	2	1	2	1	1	2	1	1	1	1	1	1	1	1	2

Appendix I

Contributed Publications

Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*

Nadine Ziemert, Anna Lechner, Matthias Wietz, Natalie Millán-Aguinaga, Krystle L. Chavarria, and Paul Robert Jensen¹

Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093

Edited* by Christopher T. Walsh, Harvard Medical School, Boston, MA, and approved February 6, 2014 (received for review December 30, 2013)

Access to genome sequence data has challenged traditional natural product discovery paradigms by revealing that the products of most bacterial biosynthetic pathways have yet to be discovered. Despite the insight afforded by this technology, little is known about the diversity and distributions of natural product biosynthetic pathways among bacteria and how they evolve to generate structural diversity. Here we analyze genome sequence data derived from 75 strains of the marine actinomycete genus *Salinispora* for pathways associated with polyketide and nonribosomal peptide biosynthesis, the products of which account for some of today's most important medicines. The results reveal high levels of diversity, with a total of 124 pathways identified and 229 predicted with continued sequencing. Recent horizontal gene transfer accounts for the majority of pathways, which occur in only one or two strains. Acquired pathways are incorporated into genomic islands and are commonly exchanged within and between species. Acquisition and transfer events largely involve complete pathways, which subsequently evolve by gene gain, loss, and duplication followed by divergence. The exchange of similar pathway types at the precise chromosomal locations in different strains suggests that the mechanisms of integration include pathway-level homologous recombination. Despite extensive horizontal gene transfer there is clear evidence of species-level vertical inheritance, supporting the concept that secondary metabolites represent functional traits that help define *Salinispora* species. The plasticity of the *Salinispora* secondary metabolome provides an effective mechanism to maximize population-level secondary metabolite diversity while limiting the number of pathways maintained within any individual genome.

genome sequencing | comparative genomics

Microbial secondary metabolites have long benefited human health and industry. They include important pharmaceutical agents such as the antibiotic penicillin, the anticancer agent vancomycin, and the immunosuppressant rapamycin among the more than 20 thousand biologically active microbial natural products reported as of 2002 (1). Secondary metabolites also have important ecological roles for the organisms that produce them, particularly in terms of nutrient acquisition, chemical communication, and defense (2). Many of these compounds are the products of polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) pathways or hybrids thereof. These pathways are generally organized into gene clusters that can exceed 100 kb and include regulatory, resistance, and transport elements (3), thus making them well-suited for horizontal gene transfer (HGT) (4, 5). The architectures and functional attributes of PKS and NRPS genes have been reviewed in detail (3, 6, 7) and account for much of the structural diversity that is the hallmark of microbial natural products. Remarkably, PKS and NRPS enzymes build these complex secondary metabolites via the controlled assembly of simple biosynthetic building blocks such as acetate, propionate, and amino acids. These building blocks are incorporated in a combinatorial fashion via a series of sequential chemical condensation reactions encoded by ketosynthase (KS) and condensation (C) domains within PKS and NRPS genes, respectively (3).

The pathways responsible for secondary metabolite biosynthesis are among the most rapidly evolving genetic elements known (5). It has been shown that gene duplication, loss, and HGT have all played important roles in the distribution of PKSs among microbes (8, 9). Changes within PKS and NRPS genes also include mutation, domain rearrangement, and module duplication (5), all of which can account for the generation of new small-molecule diversity. The evolutionary histories of specific PKS and NRPS domains have proven particularly informative, with KS and C domains providing insight into enzyme architecture and function (10, 11). These studies have helped establish the extensive nature of HGT among biosynthetic genes (4, 12), which is reflected in the incongruence between PKS and NRPS gene phylogenies and those of the organisms in which they reside (13). Although resolving the evolutionary histories of entire pathways remains more challenging than individual genes or domains, comparative analyses of biosynthetic gene clusters have proven useful for the identification of pathway boundaries (14).

The exchange of PKS and NRPS pathways by HGT confounds the relationships between taxonomy and secondary metabolite

Significance

Microbial natural products are a major source of new drug leads, yet discovery efforts are constrained by the lack of information describing the diversity and distributions of the associated biosynthetic pathways among bacteria. Using the marine actinomycete genus *Salinispora* as a model, we analyzed genome sequence data from 75 closely related strains. The results provide evidence for high levels of pathway diversity, with most being acquired relatively recently in the evolution of the genus. The distributions and evolutionary histories of these pathways provide insight into the mechanisms that generate new chemical diversity and the strategies used by bacteria to maximize their population-level capacity to produce diverse secondary metabolites.

Author contributions: N.Z. and P.R.J. designed research; N.Z., A.L., M.W., N.M.-A., and K.L.C. performed research; N.Z., A.L., M.W., N.M.-A., K.L.C., and P.R.J. analyzed data; and N.Z. and P.R.J. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: Genome sequences reported in this paper have been deposited in the Joint Genome Institute's Integrated Microbial Genomes (IMG) database, <http://img.jgi.doe.gov/cgi-bin/w/main.cgi> (accession nos. 2517572210, 2515154093, 2515154181, 2519103192, 2515154183, 2515154193, 2519103193, 2515154125, 2518285551, 2518285552, 2515154180, 2519103194, 2515154203, 641228504, 2516143022, 2518285553, 2519103185, 2518285554, 2517572137, 2515154186, 2515154088, 2515154135, 2515154127, 2517572233, 2518285555, 2515154137, 2515154188, 2517572152, 2515154187, 2517572153, 2518285558, 2519103195, 2518285559, 2518285560, 2517572154, 2517572155, 2515154178, 2515154194, 2518285561, 2518285562, 2515154129, 2518285563, 2517572157, 2515154184, 2515154126, 2515154177, 2517572158, 2515154202, 2517572159, 2515154200, 2515154124, 2517572160, 2515154185, 2517572161, 2515154182, 2518285550, 2517572162, 2515154170, 2515154128, 2517572163, 2518645626, 2518645627, 2517572194, 2517287019, 2516653042, 2516493032, 2517287023, 2517434008, 640427140, 2517572211, 2517572212, 2515154094, 2518645624, 2515154163, and 2517572164).

¹To whom correspondence should be addressed. E-mail: pjensen@ucsd.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1324161111/-DCSupplemental.

production. This may in part explain the historical reliance on chance for the discovery of natural product drug leads from chemically prolific but taxonomically complex taxa such as the genus *Streptomyces*. Genome sequencing has changed the playing field by providing bioinformatic opportunities to “mine” the biosynthetic potential of strains before chemical analysis and to target the products of specific pathways that are predicted to yield compounds of interest (15). Sequence-based methodologies not only hold great promise for natural product discovery; they are providing a wealth of information that will ultimately improve our understanding of pathway diversity and distributions and the evolutionary events that generate new chemical diversity.

Here we report the analysis of PKS and NRPS biosynthetic gene clusters in 75 *Salinispora* genome sequences. This obligate marine actinomycete is composed of three closely related species (16, 17) that are clearly delineated using phylogenetic approaches (18). *Salinispora* spp. share 99% 16S rRNA gene sequence identity (19), thus making them more narrowly defined than many taxa, which can include up to 3% sequence divergence (20). *Salinispora* spp. are a rich source of secondary metabolites, including salinosporamide A (21), which has undergone a series of phase I clinical trials for the treatment of cancer (22). They devote ca. 10% of their genomic content to secondary metabolism (23, 24) and represent a tractable model with which to address correlations between fine-scale molecular systematics and secondary metabolite production (25). The results presented here describe the diversity and distributions of biosynthetic pathways among a closely related group of bacteria and reveal high levels of pathway acquisition via horizontal gene transfer, with more than half of the pathways occurring in only one or two strains. The data provide evidence of the evolutionary mechanisms that generate new pathway diversity and a striking example of the plasticity of the bacterial secondary metabolome.

Results

Pathway Identification. Draft and complete genome sequences from 75 *Salinispora* strains were analyzed (Table S1). These

strains encompass the major biogeographic regions from which the three currently described species have been reported (Fig. S1) and include representatives of 11 previously identified 16S rRNA gene sequence variants that differ by as little as a single nucleotide change (26). KS and C domains were extracted from the sequence data and used for the initial identification of PKS and NRPS pathways, respectively. In total, 2,079 KS and 1,693 C domains were detected. Of the KS domains, BLAST, antiSMASH (27), and manual analyses that included the gene environments in which these domains occurred linked 75 to fatty acid biosynthesis (one per strain), whereas 80 were identified as false positives (*N*-acetyltransferases) and the remaining 1,924 (92.5%) were associated with secondary metabolism. All of the C domains were linked to secondary metabolism.

The next step was to assemble pathways that appeared to be split among different contigs, which was generally the case for highly repetitive modular type I PKSs and some NRPSs. This was accomplished using reference pathways from prior studies (23, 24) and better-assembled *Salinispora* genomes that included seven strains that were assembled into single contigs that exceeded 5 Mb. In the absence of a reference pathway, contigs were assembled when the KS- or C-domain phylogenies indicated close evolutionary relationships. Pathways that contained similar gene content and organization were grouped into “operational biosynthetic units” (OBUs) based on predictions they produced related secondary metabolites. These groups were defined based on sequence identity (SI) values of 90% and 85%, respectively, among homologous KS and C domains (10). The stringency of these cutoff values is supported by the *cya* and *spo* enediyne KSs, which share ca. 88% SI yet yield compounds that possess fundamentally different carbon skeletons (Fig. 1) (28, 29). Likewise, homologous C domains associated with NRPS4 and 19 share ca. 80% SI, yet they occur in pathways that differ not only in gene content but also in the composition of the NRPS genes (Fig. S2). In all cases where the secondary metabolic products of the pathways were known, fundamentally different

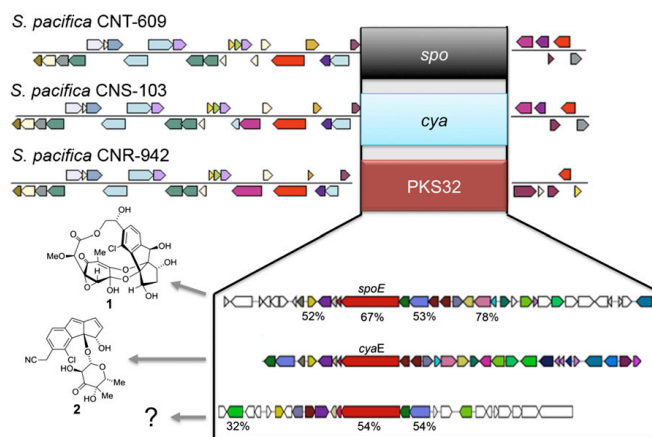


Fig. 1. Enediyne pathway exchange. Three different OBUs were detected in the same chromosomal position in three different *Salinispora* strains. These OBUs were classified as enediyne PKSs based on a NaPDoS analysis of the KS domains derived from the type I PKS genes (in red) in each pathway. The different OBU assignments are supported by the products of the sporolide (*spo*) and cyanosporaside (*cya*) gene clusters, which include sporolide A (1) and cyanosporaside A (2), respectively. These compounds, which are shown to the left of the pathways responsible for their production, possess fundamentally different carbon skeletons and are predicted to originate from enediyne precursors (28, 65, 66). Amino acid sequence identities relative to orthologs in the *cya* pathway are shown for representative genes. Products have yet to be identified from PKS32, which appears at the bottom.

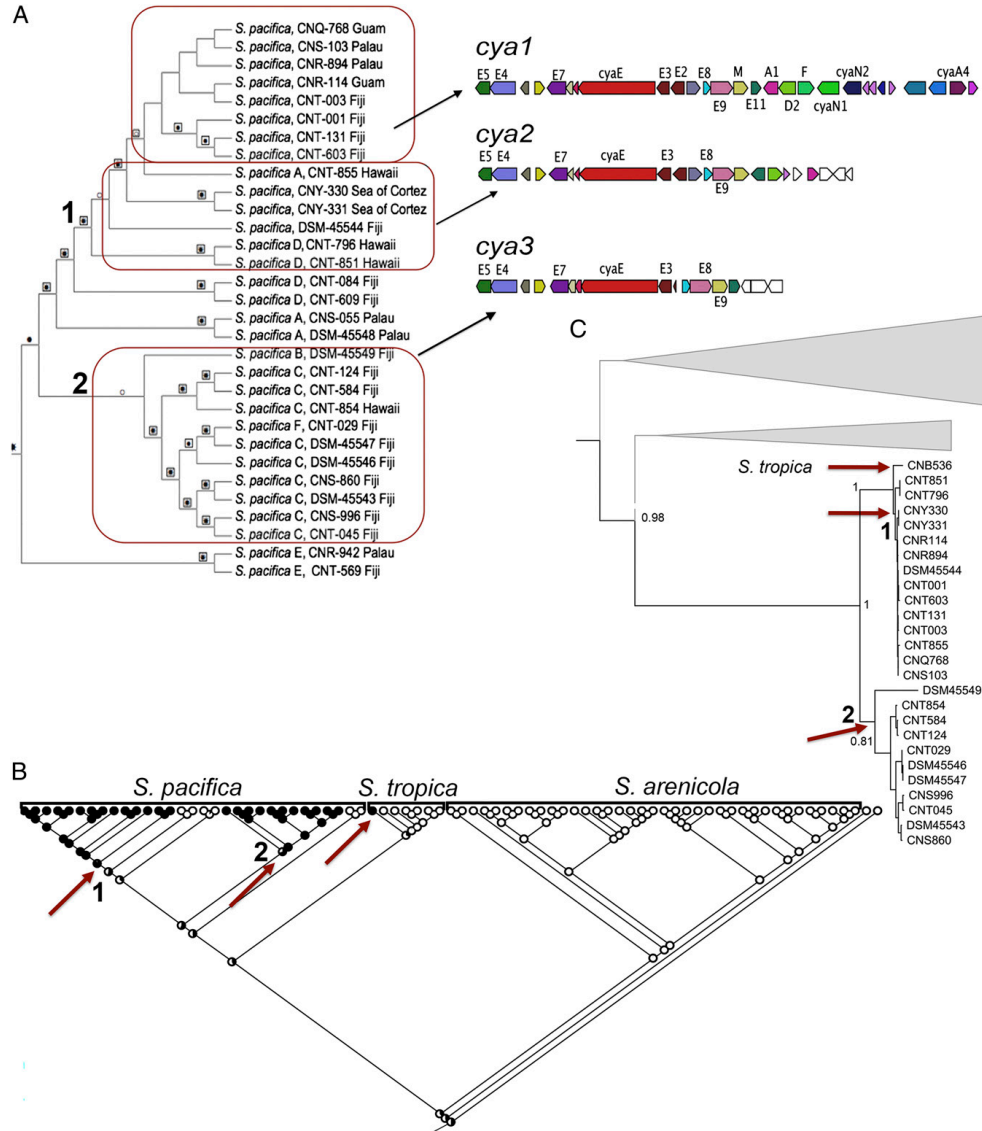


Fig. 2. Evolution of the *cya* pathway in *S. pacifica*. (A) The *cya* OBU contains three different versions of the pathway. The *cya1* version is responsible for the biosynthesis of cyanosporaside A, which is derived from an enediynes precursor (28). It was observed in all strains in the uppermost clade of the *Salinispora* species tree (boxed in red). Two truncated versions of the pathway were observed, with *cya2* appearing ancestral to *cya1* in the species tree. Genes missing in *cya2* and 3 include *cyaA4* (O-acyltransferase), *cyaN1* (epoxide hydrolase), and *cyaN2* (oxidoreductase), which are predicted to encode tailoring enzymes. Together, the *cya1*- and 2-containing strains form a single clade (clade 1) in this region of the *S. pacifica* species tree. The *cya3* pathway occurs in a separate *S. pacifica* lineage (clade 2). Products have yet to be identified from *cya2* and 3. (B) A likelihood analysis predicts three independent acquisition events for the *cya* pathway, one in *S. tropica* and two in *S. pacifica* (red arrows). The *S. pacifica* acquisition events correspond to clades 1 and 2 in the species tree. (C) Maximum-likelihood phylogeny of the *cyaE* enediynes PKS gene including the top 10 BLASTp matches (bootstrap values for 100 replicates are shown at major nodes) reveals two major lineages (red arrows, numbers 1 and 2) that correspond to the strains in clades 1 and 2 of the species tree. This supports the vertical inheritance of this gene subsequent to acquisition. The position of the *S. tropica* (CNB-536) *cyaE* homolog within *S. pacifica* clade 1 suggests that the acquisition event in *S. tropica* is the result of horizontal gene transfer with *S. pacifica* clade 1. Gene gain and loss are assumed to account for the variations in the *cya1*–3 pathways.

Table 1. *Salinispora* genome composition and pathway (OBU) statistics

Species	No. genomes analyzed	Avg. genome size, Mb	Avg. no. contigs*	Avg. no. OBUs per genome	OBU richness†	Avg. no. singletons‡ per genome
<i>S. arenicola</i>	37	5.7 ± 0.14	78 ± 19	17.5 ± 1.9	47	0.49
<i>S. pacifica</i>	31	5.4 ± 0.19	93 ± 33	14.1 ± 2.7	88	1.00
<i>S. tropica</i>	7	5.4 ± 0.19	90 ± 19	13.6 ± 1.8	19	0.57

Averages reported are ±1 SD.

*Does not include the closed genomes of CNB-440 and CNS-205.

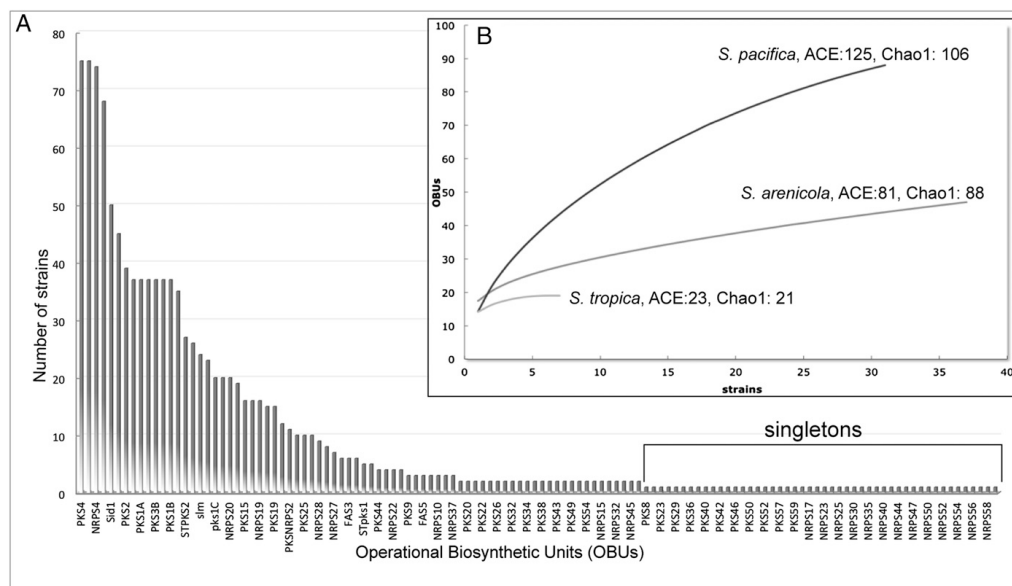
†Number of different OBUs observed.

‡OBUs observed in only one strain.

carbon skeletons were observed when the KS- and C-domain SIs fell below these cutoff values. MultiGeneBlast analyses were performed on pathways that occurred in at least five strains to better assess the OBU assignments. These analyses revealed high levels of synteny and SI among the shared genes within each OBU and sharply lower cumulative BLAST bit scores for strains that lacked the pathway. Intra-OBU differences occurred largely among genes predicted to encode tailoring enzymes, as observed in the *cya* pathway (Fig. 2). Nonetheless, given that minor structural differences can have a major impact on secondary metabolite biological activity, the products of different versions of a pathway that have been grouped into a single OBU may have different ecological functions. Although the KS- and C-domain clustering values used here appear appropriate for *Salinispora* species, it remains to be seen how well they will apply to other taxonomic groups.

Pathway Diversity. Comparable to prior studies (23, 24), the *Salinispora* genomes were enriched in PKS and NRPS biosynthetic pathways. On average, *S. arenicola* genomes were 300 kb larger

and contained four more OBUs per genome than *S. pacifica* or *S. tropica* (Table 1). Although more OBUs were detected per *S. arenicola* genome, considerably more OBU diversity was observed in *S. pacifica*, which contained a total of 88 different OBUs compared with 47 and 19 for *S. arenicola* and *S. tropica*, respectively. In total, 124 distinct OBUs were identified, including representatives of diverse PKS types (Fig. S3). Only nine of these OBUs have been formally linked to the production of specific secondary metabolites. These are *sal* (salinosporamides) (30), *slm* (salinilactam) (23), *cyl* (cyclomarins) (31), *cya* (cyanosporasides) (28), *spo* (sporolides) (29), *arn* (arenimycin) (32), *rif* (rifamycins) (33), *lym* (lymphostin) (34), and *lom* (lomaiviticin) (35), whereas two others are predicted to yield enterocin (36) (PKS31) and arenicolide (37) (PKS28) based on bioinformatic analyses. Although there is no evidence that all pathways are functional, the 113 remaining OBUs far exceed the four *Salinispora* secondary metabolites (arenamides, pacificanones, salinipyrones, and salini-quinones) that have yet to be linked to specific pathways, suggesting

**Fig. 3.** Distribution and diversity of PKS and NRPS OBUs. (A) Rank-abundance curve showing the abundance of each OBU among the 75 strains analyzed (representative OBU names are shown). (B) Rarefaction curves with diversity estimators for each species.

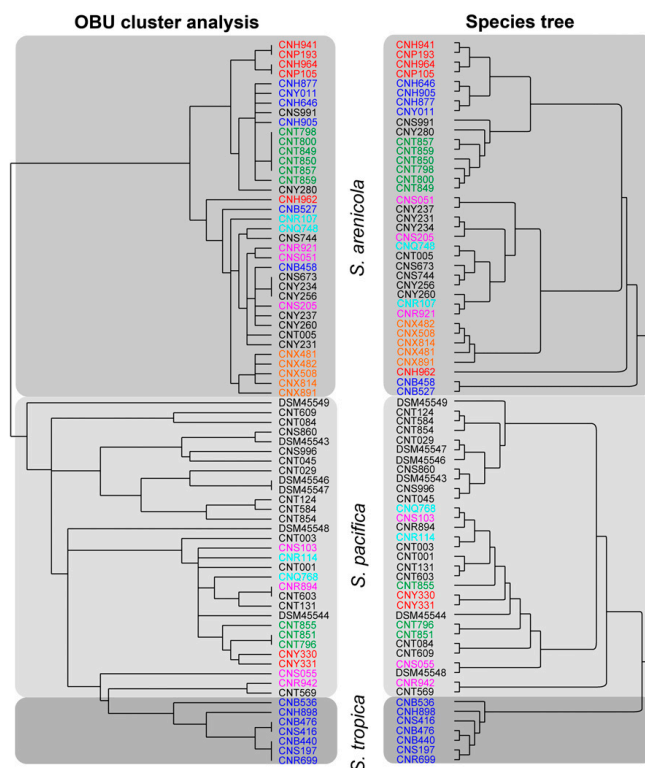


Fig. 4. OBU hierarchical cluster analysis and *Salinispora* species tree. Hierarchical cluster analysis based on OBU presence or absence. Maximum-likelihood species phylogeny generated from 10 housekeeping genes. Colors indicate the collection site: blue, Bahamas; pink, Palau; black, Fiji; red, Sea of Cortez; orange, Palmyra; turquoise, Guam; green, Hawaii.

that considerable chemical diversity remains to be discovered from this genus.

A rank-abundance curve describing the distribution of the OBUs among the 75 strains reveals a long right-hand tail, as is characteristic of a highly diverse community (Fig. 3). Remarkably, 48 of the OBUs were only observed in one strain (singletons), with an additional 24 occurring in two strains. These 72 OBUs account for 58% of the total number observed in the 75 genomes and illustrate extensive acquisition via horizontal gene transfer. In the case of *S. pacifica*, the most phylogenetically diverse of the three species (26), an average of one singleton was detected per genome sequenced (Table 1). Rarefaction curves, used primarily in community ecology to assess species richness (38), provide an assessment of OBU richness for the given sequencing effort and reveal that considerable diversity has yet to be detected (Fig. 3). This is particularly evident for *S. pacifica*, which shows little evidence of saturation, and is further supported by ACE and Chao1 diversity estimators, which predict as many as 229 distinct OBUs with continued sequencing of the three species (Fig. 3). This represents an extraordinary level of biosynthetic diversity for three bacterial species that share 99% 16S rRNA sequence identity (19).

Pathway Distributions. We next generated a well-supported *Salinispora* species phylogeny (Fig. S4) and a hierarchical cluster analysis based on pathway presence or absence. Despite the large

number of OBUs that occur in only one or two strains, these two dendrograms are highly congruent, with the exception that *S. pacifica* is paraphyletic with respect to *S. tropica* in the OBU cluster analysis (Fig. 4). Contributing to this congruence are species-specific OBUs (i.e., pathways commonly observed in one species but generally not in others). In *S. arenicola*, these include *rif*, *PKS1A/B*, *PKS2*, *PKS3A/B*, *PKS5*, *NRPS1*, and *NRPS2* (Table S2). In *S. tropica*, these include *spo*, *slm*, *sal*, *Sid3*, *Sid4*, *NRPS3*, and *STPKS1* (Table S3). Interestingly, only one OBU (*NRPS20*) appears commonly in *S. pacifica* and not in others (Table S4). These results support previous culture-based studies and KS fingerprinting analyses that revealed species-specific patterns of secondary metabolite production and gene distributions in *S. arenicola* and *S. tropica* (25, 39). There is some evidence of OBU clustering based on the location from which the strains originate (Fig. 4); however, a permutational multivariate analysis of variance (PERMANOVA) revealed a significant correlation between OBU and species ($R^2 = 0.54$, $P = 0.001$) and not location ($P = 0.075$), indicating the importance of taxonomy over biogeographic origin in terms of OBU distributions.

Pathway Evolution. To explore the evolutionary history of the pathways in relation to the strains in which they reside, likelihood analyses were performed on the KS- and C-domain sequences to assign the ancestral node(s) for each OBU in the

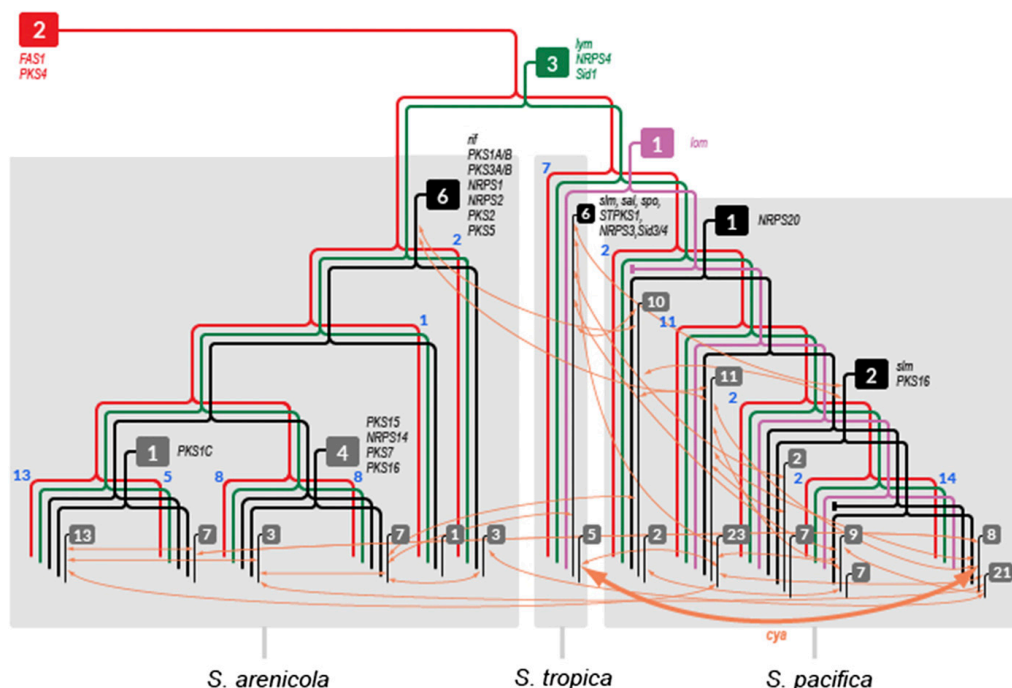


Fig. 5. *Salinispora* phylogeny depicting OBU inferred ancestry. A simplified species tree generated from 10 housekeeping genes (Fig. S4) shows 12 major *Salinispora* lineages with the number of strains in each indicated in blue adjacent to the node branch points. Boxes indicate the number of OBUs originating at various points in the species tree. Red, shared with a common ancestor of the genus; green, genus-specific; purple, shared with *S. tropica* and *S. pacifica*; black, species-specific; gray, clade-specific. Representative OBU names are indicated next to the point of acquisition. Orange arrows describe inter- and intraspecies OBU exchange events (*cya* exchange between *S. tropica* and *S. pacifica* is indicated in bold).

species tree. The results for the 124 OBUs were overlaid onto a simplified *Salinispora* phylogeny (Fig. 5) generated by collapsing the species tree (Fig. S4) into 12 lineages. The analysis reveals that only five OBUs were present in the common ancestor of the genus and only two of these (FAS1 and PKS4) were shared with the closely related genus *Micromonospora*. It can thus be inferred that the remaining pathways (96% of the total) were acquired by HGT at various points during the evolution of the genus. Phylogenetic analyses of key biosynthetic genes from each OBU confirm these evolutionary histories and indicate, based on congruence with the species tree, vertical inheritance for 65 of the OBUs subsequent to acquisition. Seven OBUs appear to have been acquired early in the evolution of *S. arenicola*. These include *rif*, which supports the consistent production of rifamycins by *S. arenicola* (25, 40). Likewise, six OBUs appear early in the evolutionary history of *S. tropica* and one in *S. pacifica* (Fig. 5). Most of the OBUs, however, were acquired relatively recently in the evolution of the genus, appearing toward the branch terminals in the tree. Based on BLAST analyses of the singleton PKS and NRPS genes, it appears that most of these pathways were acquired from other high-G+C bacteria such as *Streptomyces* spp. (Fig. S5), which also occur in marine sediments (41). The results for PKS17 suggest the independent acquisition of this pathway by four *S. arenicola* strains from Fiji and one *S. pacifica* strain from the Sea of Cortez (Fig. S6). Although these results may reflect sampling

effort, they suggest that location-dependent pathway acquisition warrants future study.

Phylogenetic analyses of key biosynthetic genes were also used to infer that 36 of the 124 OBUs identified (29%) were exchanged within or between species. One example is the *cya* pathway, which was exchanged between *S. pacifica* and *S. tropica* (Fig. 2). These transfer events were added to the simplified species tree to depict the complexity of pathway movement within the genus (Fig. 5). In total, it could be inferred that 23 OBUs moved once, 9 moved twice, and 4 (PKS17, *sal*, Sid1, and PKS-NRPS2) moved three times. There was no evidence for KS- or C-domain exchange among OBUs or the formation of chimeric pathways, although events of these types may have been missed with the assembly methods used. Instead, OBUs evolved largely by gene gain, gene loss, and duplication followed by divergence. In the last case, NRPS4 is a genus-specific pathway observed in 72 of the 75 strains. A subset of *S. arenicola* (clade 6) and *S. pacifica* (clade 12) contains a second copy of this pathway (NRPS19) that is sufficiently diverged (i.e., shares <85% C-domain SI) to be considered a new OBU (Fig. S2). Thus, pathway duplication followed by divergence appears to be another mechanism by which OBU diversity is created in *Salinispora* spp.

Genomic Islands as Hot Spots for Secondary Metabolism. Pseudochromosomes were generated by mapping sequence contigs onto the closed genomes of *S. tropica* (CNB-440), *S. arenicola* (CNS-205), and a number of high-quality *S. pacifica* draft genomes that

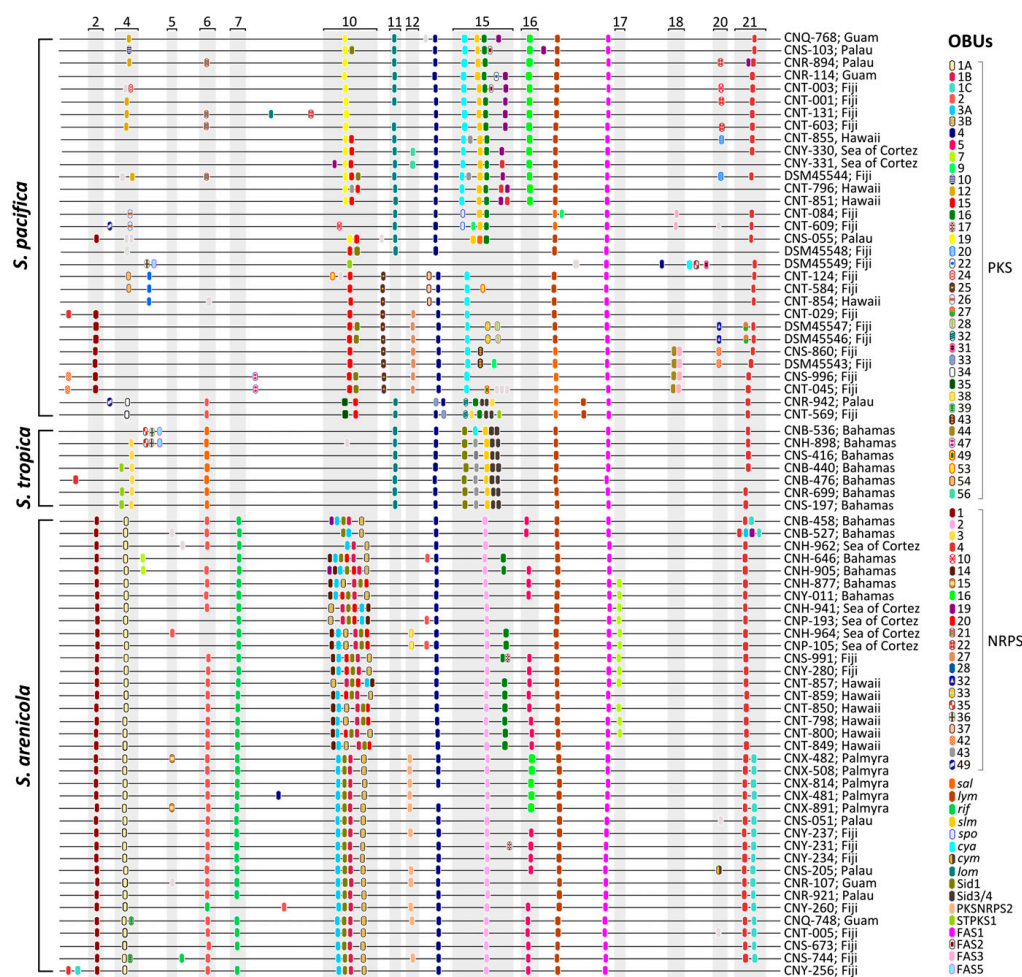


Fig. 6. Linear pseudochromosomes reveal the positioning of *Salinispora* OBUs within genomic islands (numbered and shaded) as identified based on previously defined boundaries (24). Pathway positions were mapped using PKS and NRPS genes as reference. Only the pathways that could be mapped onto the pseudochromosomes are depicted. Pathways are color-coded and listed on the right next to the strain name and geographic origin. Singletons are depicted in light gray.

were generated as part of this study. The results show that the OBUs are clustered in genomic islands (GIs) (Fig. 6), regions of bacterial chromosomes known to encode acquired, adaptive traits (42). *Salinispora* GIs were also enriched in mobile genetic elements, which may play a role in OBU acquisition and transfer, relative to other regions of the genome (Wilcoxon rank-sum test, $P < 0.05$). Remarkably, the flanking regions of 21 previously identified *Salinispora* GIs (24) are conserved across all 75 genome sequences, suggesting that island boundaries can be used as queries to identify similar regions in other strains. In some cases, OBUs that encode the biosynthesis of similar classes of compounds were exchanged at precisely the same island location. This type of pathway “swapping” was observed with three enediyne OBUs in *S. pacifica* (Fig. 1), and may represent

an example of pathway-level homologous recombination that is yet to be described.

Discussion

Major advances in our understanding of the molecular genetics of natural product biosynthesis have created unprecedented opportunities for pathway engineering (43) and the generation of new chemical diversity in high-priority scaffolds (44). Coupled with increased access to genome sequence data and the revelation that even well-studied taxa can harbor a wealth of biosynthetic pathways for which the products have yet to be discovered (45, 46), natural product research is undergoing a renaissance driven by the development of new discovery methods (47). Despite these advances, we have yet to gain perspective on the diversity and

distributions of the pathways responsible for secondary metabolism among groups of related bacteria and how these pathways evolve to generate new chemical diversity.

The 75 genome sequences analyzed here provide insight into the remarkable levels of pathway diversity that can be maintained among a group of bacteria that share 99% 16S rRNA gene sequence identity. This diversity can largely be attributed to the many pathways that were observed in only one or two strains and that are inferred to be the result of HGT events that occurred relatively recently in the evolutionary history of the genus. Although the effects of geographic origin on the OBUs maintained by individual strains warrant further study, the potential for location-specific acquisition suggests that differences in the local gene pool may account for some of the diversity reported here. Although the total number of OBUs maintained by these three closely related species remains unknown, it is extraordinarily high relative to the numbers observed in the individual strains (Table 1), with a total of 229 distinct PKS and NRPS OBUs predicted with continued sequencing.

Mapping the inferred ancestral nodes of the individual OBUs onto the *Salinispora* species phylogeny made it possible to trace pathway evolutionary histories relative to the strains in which they reside (Fig. 5). These analyses reveal that 105 of the 124 OBUs (85%) were acquired subsequent to the speciation events within the genus, which suggests that the ecological functions of secondary metabolites act largely at the subspecies level. However, the congruence observed between the species tree and the OBU cluster analysis (Fig. 4) suggests that secondary metabolites nonetheless represent functional traits that help define *Salinispora* spp. (25). The fixation of certain pathways within *S. arenicola* and *S. tropica* could be the result of periodic selection (48), which if driven by the products of these OBUs would indicate that they provide a strong selective advantage. Species-specific OBUs include *rif* and *sal*, which encode the production of the potent antibiotic rifamycin and the proteasome inhibitor salinosporamide A in *S. arenicola* and *S. tropica*, respectively. In *S. pacifica*, the most diverse of the three species (26), similar levels of fixation are not observed, yet many OBUs appear fixed among major clades within the species. Based on this, it could be speculated that *S. pacifica* is undergoing a series of nascent speciation events, with ecological divergence preventing periodic selection from fixing pathways at the currently defined species level.

The OBUs were concentrated in GIs whose boundaries were highly conserved among all strains. These GIs were enriched in mobile genetic elements, suggesting they are hot spots for pathway acquisition and evolution. The observed swapping of enediynes OBUs at the precise chromosomal locations in different strains (Fig. 1) suggests that recombination may function at the pathway level in a manner comparable to the domain-level homologous recombination observed in PKS and NRPS analyses (49). The absence of KS- or C-domain exchange among OBUs, a process that is generally considered important in PKS and NRPS evolution (5, 50), suggests that pathway HGT followed by gene gain or loss events is the major force driving the creation of OBU diversity in *Salinispora* spp. Although it is unclear how these results apply to other bacteria, the continued sequencing of large numbers of closely related strains will provide additional insight into the evolutionary processes by which bacteria generate new secondary metabolite diversity.

A better understanding of the taxonomic distributions and evolutionary histories of the pathways responsible for secondary metabolite biosynthesis will provide opportunities for the development of theory-based sampling strategies that capitalize on the genetic potential of individual strains to produce new chemical scaffolds or compounds within a privileged chemical class. Recognition that some pathways diverge in lineage-specific patterns indicates that related strains within the same species can be the source of related compounds within the same chemical

class (39), thus providing an alternative to synthetic chemistry as an approach to generating structural diversity. The plasticity of secondary metabolism in *Salinispora* spp. provides a glimpse into the evolutionary strategies by which bacteria capitalize on the benefits afforded by these compounds. Despite not knowing the ecological functions of most *Salinispora* secondary metabolites, extensive pathway sampling provides a mechanism to maximize the population-level secondary metabolome while limiting the number of pathways maintained within any individual genome. The potentially vast array of molecules produced at the population level would increase the likelihood of an effective response to new selective pressures and thus provide an ecological rationale for the extensive pathway diversity observed in this study.

Materials and Methods

Genome Sequencing and Assembly. *Salinispora* strains were obtained in culture as previously described (41, 51). DNA was extracted following US Department of Energy Joint Genome Institute (JGI) protocols (<http://my.jgi.doe.gov/general/protocols.html>) and submitted to the JGI for sequencing, assembly, and annotation. The sequencing and annotation of *S. arenicola* CNS-205 and *S. tropica* CNB-440 were as previously described (23, 24). For the remaining 73 strains, short- and long-insert paired-end libraries were constructed and sequenced by the JGI using the Illumina HiSeq 2000 system. Filtered reads were assembled using Velvet (52) and ALLPATHS-LG (53), and possible misassemblies were corrected with manual editing in Consed (54). Gap closure was accomplished using repeat resolution software and sequencing of bridging PCR fragments with Sanger and/or PacBio technologies. Genes were identified using Prodigal (55), followed by a round of manual curation using GenePRIMP (56). Predicted coding DNA sequences were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant (nr) database, UniProt, TrEMBL, Pfam, Kyoto Encyclopedia of Genes and Genomes, Clusters of Orthologous Groups, and InterPro databases. Strains and accession numbers are provided in Table S1.

Pathway and OBU Identification. Genome sequences in FASTA format were screened for PKS and NRPS genes by searching for KS and C domains, respectively, using NaPDos (<http://napdos.ucsd.edu>) with default settings (57). The associated genes and gene environments were then analyzed using BLAST (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>), IMG/ER (<https://img.jgi.doe.gov/cgi-bin/er/main.cgi>), and antiSMASH (27) to confirm association with secondary metabolism, assess the similarities among pathways, and create links to known secondary metabolites based on homology to experimentally characterized pathways. Pathways split onto different contigs were assembled with the aid of complete pathways or when KS- and C-domain phylogenies revealed that the sequences claded together. Pathways were grouped into OBUs when BLAST analyses revealed that homologous KS and C domains shared $\geq 90\%$ and $\geq 85\%$ amino acid sequence identity, respectively. OBUs were assigned a unique identifier (e.g., PKS1, NRPS1) or a formal name if linked to an experimentally characterized pathway. A database of all genomes was created and OBU assignments were verified for pathways that occurred in five or more strains using MultiGeneBlast (58) based on the synteny and SI of conserved genes in each pathway and cumulative BLAST bit scores, which dropped precipitously in strains that did not possess the pathway (10).

***Salinispora* Species Phylogeny.** Nucleotide sequences for 10 unlinked, single-copy genes (*dnaA*, *gyrB*, *pyrH*, *recA*, *pgi*, *trpB*, *atpD*, *sucC*, *rpoB*, *topA*) were extracted, aligned using Muscle in Geneious Pro v5.5 (Biomatters; www.geneious.com), and concatenated using Mesquite v2.75 (59). MODELTEST (60) was run and the best model [generalized time reversible (GTR)+G] was used to create a maximum-likelihood (ML) tree using PhyML 3.0 (61) and a neighbor-joining tree using MEGA5 (62). Nodal support values were obtained using 1,000 bootstrap replicates. Concatenated Bayesian tree and posterior probabilities were created using MRBAYES (63) with 1 million generations.

OBU Phylogeny. Nucleotide sequences from at least two conserved genes from each OBU observed in two or more of the 12 major *Salinispora* clades presented in Fig. 5 were aligned in Muscle and manually curated. ML phylogenies were created using PhyML 3.0 under the GTR model of nucleotide substitution with 100 bootstrap replicates or a fast approximate likelihood-ratio test performed as a measure of branch support. Other common models

of nucleotide substitution were used with no significant changes in the results. If the phylogenies for the genes within an OBU were congruent, this phylogeny was assumed for the whole pathway.

Statistics. Hierarchical cluster analyses were performed using Cluster 3.0 (<http://bonsai.hgc.jp/~mdehoon/software/cluster/software>) with presence/absence OBU matrices as the input files (Tables S2–S4) using a correlation-centered similarity metric with the complete linkage clustering method. A PERMANOVA was implemented with the vegan package in R (www.r-project.org). EstimateS (<http://viceroy.eeb.uconn.edu/estimates>) was used to generate rarefaction curves and diversity estimates. The Wilcoxon rank-sum test implemented in R was used to compare the fraction of mobile genetic elements inside and outside of GIs.

Ancestral State Reconstruction. The ancestral node for each OBU was inferred in the species tree using the trace character history function implemented in Mesquite v2.75 (59). A categorical character matrix was created for all OBUs, and likelihood calculations were performed using the Mk1 model. Likelihood scores >50% were used to infer the points of OBU acquisition (ancestral nodes) in the species tree. OBU ML phylogenies were used to corroborate points of acquisition based on congruence with the species tree and to infer inter- and intraspecies exchange events as shown in Fig. 2.

Pseudochromosome Assembly, OBU Localization, and Genomic Island Analysis. Draft genomes were assembled into linear “pseudochromosomes” using the CONTIGuator 2 web application (64) and oriented with *dnaA* as the first gene. The closed genomes *S. arenicola* CNS-205 (24) and *S. tropica* CNB-440 (23) were used as templates for the assembly of these species. High-quality draft *S. pacifica* genomes (one 5-Mb scaffold and one to three contigs of 10–100 kb) from strains DSM-45544, DSM-45548, and DSM-45543 were used as reference templates for the assembly of *S. pacifica* phylotypes ST, A, and C. For other phylotypes, the template that gave the best assembly was used. The chromosomal position of the OBUs present in ≥3 strains was determined using the Assembly function in Geneious Pro v5.5 and a PKS or NRPS gene

from the predicted OBU as reference. All remaining OBUs were mapped by searching for KS- and C-domain amino acid sequences using Custom-BLAST in Geneious Pro v5.5. In a previous study of *S. arenicola* CNS-205 and *S. tropica* CNB-440, 21 GIs were identified based on regions of conservation flanking regions >20 kb that shared <40% gene orthology (24). Conserved regions 5 kb up- and downstream of genomic islands were extracted from CNS-205 and located in the pseudochromosomes by BLAST in Geneious Pro v5.5. Mobile genetic elements were quantified in closed and high-quality draft genomes (*S. arenicola* CNS-205 and CNS-991, *S. tropica* CNB-440, and *S. pacifica* DSM-45543, DSM-45544, DSM-45546, DSM-45547, DSM-45548, and DSM-45549) by counting annotated recombinase, transposase, phage, integrase, and tRNA genes inside and outside of GIs.

Source of Singleton OBUs. All KS and C domains that occurred in one *Salinispora* strain (singletons) were subjected to BLAST analyses using the NCBI/nr protein database to assess the taxonomic distribution of homologous domains in other microorganisms. A total of 330 KS domains (from 16 pathways) and 1,100 C domains (from 26 pathways) was analyzed. The top 10 BLAST hits of every query were sorted by taxonomy in Geneious Pro v5.5 to calculate the distribution per taxonomic group.

ACKNOWLEDGMENTS. We acknowledge the Joint Genome Institute/Community Sequencing Program for providing sequence data, assembly, and annotation and for helpful advice on sample preparation. Greg Rouse and Nastassia Patin are acknowledged for assistance with the phylogenetic and statistical analyses, respectively. Brad Moore is acknowledged for helpful discussions about the data. Kelley Gallagher, Anindita Sarkar, Eun Ju Choi, Kevin Penn, and Nastassia Patin assisted with DNA extractions. Financial support was provided by the National Institutes of Health under Grants U01-TW0007401, GM085770, and GM086261 (to P.R.J.); the National Science Foundation Graduate Research Fellowship under Grant DGE-1144086 (to K.L.C.); and the German Academic Exchange Service [Deutscher Akademischer Austauschdienst (DAAD)] (to M.W.). The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract no. DE-AC02-05CH11231.

- Bérdy J (2005) Bioactive microbial metabolites. *J Antibiot (Tokyo)* 58(1):1–26.
- Wietz M, Duncan K, Patin NV, Jensen PR (2013) Antagonistic interactions mediated by marine bacteria: The role of small molecules. *J Chem Ecol* 39(7):879–891.
- Fischbach MA, Walsh CT (2006) Assembly-line enzymology for polyketide and non-ribosomal peptide antibiotics: Logic, machinery, and mechanisms. *Chem Rev* 106(8):3468–3496.
- Jenke-Kodama H, Dittmann E (2009) Evolution of metabolic diversity: Insights from microbial polyketide synthases. *Phytochemistry* 70(15–16):1858–1866.
- Fischbach MA, Walsh CT, Clardy J (2008) The evolution of gene collectives: How natural selection drives chemical innovation. *Proc Natl Acad Sci USA* 105(12):4601–4608.
- Hertweck C (2009) The biosynthetic logic of polyketide diversity. *Angew Chem Int Ed Engl* 48(26):4688–4716.
- Mootz HD, Schwarzer D, Marahiel MA (2002) Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *ChemBioChem* 3(6):490–504.
- Jenke-Kodama H, Sandmann A, Müller R, Dittmann E (2005) Evolutionary implications of bacterial polyketide synthases. *Mol Biol Evol* 22(10):2027–2039.
- Kroken S, Glass NL, Taylor JW, Yoder OC, Turgeon BG (2003) Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proc Natl Acad Sci USA* 100(26):15670–15675.
- Ziemert N, Jensen PR (2012) Phylogenetic approaches to natural product structure prediction. *Methods Enzymol* 517:161–182.
- Rausch C, Hoof I, Weber T, Wohlleben W, Huson DH (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol Biol* 7:78.
- Ginolhac A, et al. (2005) Type I polyketide synthases may have evolved through horizontal gene transfer. *J Mol Evol* 60(6):716–725.
- Metsä-Ketelä M, et al. (2002) Molecular evolution of aromatic polyketides and comparative sequence analysis of polyketide ketosynthase and 16S ribosomal DNA genes from various *Streptomyces* species. *Appl Environ Microbiol* 68(9):4472–4479.
- Doroghazi JR, Metcalf WW (2013) Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics* 14:611.
- Corre C, Challis GL (2009) New natural product biosynthetic chemistry discovered by genome mining. *Nat Prod Rep* 26(8):977–986.
- Ahmed L, et al. (2013) *Salinispora pacifica* sp. nov., an actinomycete from marine sediments. *Antonie van Leeuwenhoek* 103(5):1069–1078.
- Maldonado LA, et al. (2005) *Salinispora arenicola* gen. nov., sp. nov. and *Salinispora tropica* sp. nov., obligate marine actinomycetes belonging to the family Microsporaceae. *Int J Syst Evol Microbiol* 55(Pt 5):1759–1766.
- Freel KC, Millán-Aguilera N, Jensen PR (2013) Multilocus sequence typing reveals evidence of homologous recombination linked to antibiotic resistance in the genus *Salinispora*. *Appl Environ Microbiol* 79(19):5997–6005.
- Jensen PR, Mafnas C (2006) Biogeography of the marine actinomycete *Salinispora*. *Environ Microbiol* 8(11):1881–1888.
- Gevers D, et al. (2005) Re-evaluating prokaryotic species. *Nat Rev Microbiol* 3(9):733–739.
- Feling RH, et al. (2003) Salinosporamide A: A highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus *Salinispora*. *Angew Chem Int Ed Engl* 42(3):355–357.
- Finical W, et al. (2009) Discovery and development of the anticancer agent salinosporamide A (NPI-0052). *Bioorg Med Chem* 17(6):2175–2180.
- Udwary DW, et al. (2007) Genome sequencing reveals complex secondary metabolism in the marine actinomycete *Salinispora tropica*. *Proc Natl Acad Sci USA* 104(25):10376–10381.
- Penn K, et al. (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. *ISME J* 3(10):1193–1203.
- Jensen PR, Williams PG, Oh DC, Zeigler L, Finical W (2007) Species-specific secondary metabolite production in marine actinomycetes of the genus *Salinispora*. *Appl Environ Microbiol* 73(4):1146–1152.
- Freel KC, Edlund A, Jensen PR (2012) Microdiversity and evidence for high dispersal rates in the marine actinomycete ‘*Salinispora pacifica*’. *Environ Microbiol Rep* 14(2):480–493.
- Medema MH, et al. (2011) antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 39(Web Server issue, Suppl 2):W339–W346.
- Lane AL, et al. (2013) Structures and comparative characterization of biosynthetic gene clusters for cyanosporasides, enediene-derived natural products from marine actinomycetes. *J Am Chem Soc* 135(11):4171–4174.
- McGlinchey RP, Nett M, Moore BS (2008) Unraveling the biosynthesis of the sporolide cyclohexenone building block. *J Am Chem Soc* 130(8):2406–2407.
- Eustáquio AS, et al. (2009) Biosynthesis of the salinosporamide A polyketide synthase substrate chloroethylmalonyl-coenzyme A from S-adenosyl-L-methionine. *Proc Natl Acad Sci USA* 106(30):12295–12300.
- Schultz AW, et al. (2008) Biosynthesis and structures of cyclomarins and cyclomarinins, prenylated cyclic peptides of marine actinobacterial origin. *J Am Chem Soc* 130(13):4507–4516.
- Kersten RD, et al. (2013) Glycomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. *Proc Natl Acad Sci USA* 110(47):E4407–E4416.
- Wilson MC, Gulder TAM, Mahmud T, Moore BS (2010) Shared biosynthesis of the saliniketals and rifamycins in *Salinispora arenicola* is controlled by the sare1259-encoded cytochrome P450. *J Am Chem Soc* 132(36):12757–12765.
- Myanaga A, et al. (2011) Discovery and assembly-line biosynthesis of the lymphostin pyrroloquinoline alkaloid family of mTOR inhibitors in *Salinispora* bacteria. *J Am Chem Soc* 133(34):13311–13313.
- Kersten RD, et al. (2013) Bioactivity-guided genome mining reveals the lomaiviticin biosynthetic gene cluster in *Salinispora tropica*. *ChemBioChem* 14(8):955–962.

36. Piel J, et al. (2000) Cloning, sequencing and analysis of the enterocin biosynthesis gene cluster from the marine isolate '*Streptomyces maritimus*': Evidence for the derailment of an aromatic polyketide synthase. *Chem Biol* 7(12):943–955.
37. Williams PG, Miller ED, Asolkar RN, Jensen PR, Fenical W (2007) Arenicolides A–C, 26-membered ring macrolides from the marine actinomycete *Salinispora arenicola*. *J Org Chem* 72(14):5025–5034.
38. Gotelli NJ, Colwell RK (2001) Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecol Lett* 4(4):379–391.
39. Freel KC, Nam S-J, Fenical W, Jensen PR (2011) Evolution of secondary metabolite genes in three closely related marine actinomycete species. *Appl Environ Microbiol* 77(20):7261–7270.
40. Kim TK, Hewavitharana AK, Shaw PN, Fuerst JA (2006) Discovery of a new source of rifamycin antibiotics in marine sponge actinobacteria by phylogenetic prediction. *Appl Environ Microbiol* 72(3):2118–2125.
41. Jensen PR, Gontang E, Mafnas C, Mincer TJ, Fenical W (2005) Culturable marine actinomycete diversity from tropical Pacific Ocean sediments. *Environ Microbiol* 7(7):1039–1048.
42. Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* 2(5):414–424.
43. Walsh CT (2002) Combinatorial biosynthesis of antibiotics: Challenges and opportunities. *ChemBioChem* 3(2–3):125–134.
44. Weissman KJ, Leadlay PF (2005) Combinatorial biosynthesis of reduced polyketides. *Nat Rev Microbiol* 3(12):925–936.
45. Bentley SD, et al. (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 417(6885):141–147.
46. Nett M, Ikeda H, Moore BS (2009) Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat Prod Rep* 26(11):1362–1384.
47. Koehn FE, Carter GT (2005) The evolving role of natural products in drug discovery. *Nat Rev Drug Discov* 4(3):206–220.
48. Cohan FM (2002) What are bacterial species? *Annu Rev Microbiol* 56:457–487.
49. Jenke-Kodama H, Dittmann E (2009) Bioinformatic perspectives on NRPS/PKS megasynthases: Advances and challenges. *Nat Prod Rep* 26(7):874–883.
50. Hopwood DA (1997) Genetic contributions to understanding polyketide synthases. *Chem Rev* 97(7):2465–2498.
51. Gontang EA, Fenical W, Jensen PR (2007) Phylogenetic diversity of Gram-positive bacteria cultured from marine sediments. *Appl Environ Microbiol* 73(10):3272–3282.
52. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821–829.
53. Gnerre S, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* 108(4):1513–1518.
54. Gordon D, Abajian C, Green P (1998) Consed: A graphical tool for sequence finishing. *Genome Res* 8(3):195–202.
55. Hyatt D, et al. (2010) Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.
56. Pati A, et al. (2010) GenePRIMP: A gene prediction improvement pipeline for prokaryotic genomes. *Nat Methods* 7(6):455–457.
57. Ziemert N, et al. (2012) The natural product domain seeker NaPDoS: A phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS ONE* 7(3):e34064.
58. Medema MH, Takano E, Breitling R (2013) Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol Biol Evol* 30(5):1218–1223.
59. Maddison WP, Maddison DR (2011) Mesquite: A Modular System for Evolutionary Analysis, Version 2.75. Available at <http://mesquiteproject.org>. Accessed February 17, 2014.
60. Posada D, Crandall KA (1998) MODELTEST: Testing the model of DNA substitution. *Bioinformatics* 14(9):817–818.
61. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307–321.
62. Tamura K, et al. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28(10):2731–2739.
63. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8):754–755.
64. Galardini M, Biondi EG, Bazzicalupo M, Mengoni A (2011) CONTIGuator: A bacterial genomes finishing tool for structural insights on draft genomes. *Source Code Biol Med* 6:11.
65. Oh D-C, Williams PG, Kauffman CA, Jensen PR, Fenical W (2006) Cyanosporasides A and B, chloro- and cyano-cyclopenta[a]indene glycosides from the marine actinomycete '*Salinispora pacifica*'. *Org Lett* 8(6):1021–1024.
66. Fenical W, Jensen PR (2006) Developing a new resource for drug discovery: Marine actinomycete bacteria. *Nat Chem Biol* 2(12):666–673.

Challenges and triumphs to genomics-based natural product discovery

Paul R. Jensen · Krystle L. Chavarria ·
William Fenical · Bradley S. Moore · Nadine Ziemert

Received: 7 August 2013 / Accepted: 18 September 2013 / Published online: 9 October 2013
© Society for Industrial Microbiology and Biotechnology 2013

Abstract Genome sequencing is rapidly changing the field of natural products research by providing opportunities to assess the biosynthetic potential of strains prior to chemical analysis or biological testing. Ready access to sequence data is driving the development of new bioinformatic tools and methods to identify the products of silent or cryptic pathways. While genome mining has fast become a useful approach to natural product discovery, it has also become clear that identifying pathways of interest is much easier than finding the associated products. This has led to bottlenecks in the discovery process that must be overcome for the potential of genomics-based natural product discovery to be fully realized. In this perspective, we address some of these challenges in the context of our work with the marine actinomycete genus *Salinispora*, which is proving to be a useful model with which to apply genome mining as an approach to natural product discovery.

Keywords Genomics · Natural product biosynthesis · Genome mining · *Salinispora*

Introduction

Bacterial natural product discovery once relied heavily upon luck. Thousands of strains were typically cultured in a limited number of fermentation conditions in the hope that a minimum number would yield compounds of interest. The odds could be improved by creative fermentation

techniques and targeting poorly studied taxa, however the process remained deeply invested in serendipity. While this strategy initially proved successful, it became less tenable by the close of the 20th century as the rates of new compound discovery dropped to levels that could not be supported by the pharmaceutical industry. These diminishing returns, coupled with advances in combinatorial chemistry and high-throughput screening, led major pharmaceutical companies worldwide to move en masse away from natural products as a resource for drug discovery [16], leaving this area of research largely in the realm of academia and small biotechnology companies.

Ready access to microbial genome sequencing has now changed the playing field for natural product discovery. Genome sequencing provides a highly informed approach by which strains can be prioritized based on a bioinformatic assessment of their biosynthetic potential. This potential can be used to infer the production of known compounds (de-replication), to make generalized predictions about the types of compounds that can be expected (e.g., polyketides, terpenes, etc.), and in some cases to make precise structural predictions. These capabilities provide opportunities to identify strains with the potential to produce compounds of interest, which once identified can be subjected to detailed fermentation studies, biological screening, and chemical analysis. Key genes in a targeted pathway can also be monitored for expression to help ensure that the appropriate fermentation conditions have been selected. Alternatively, entire pathways can be targeted for heterologous expression, thereby bypassing regulatory hurdles in the native host.

Genome sequencing has fundamentally changed the way we think about natural product discovery. Nonetheless, it is far from a panacea, as many technical challenges remain. These challenges include the bioinformatic expertise

P. R. Jensen (✉) · K. L. Chavarria · W. Fenical · B. S. Moore · N. Ziemert
Scripps Institution of Oceanography, University of California,
San Diego, La Jolla, CA 92093, USA
e-mail: pjensen@ucsd.edu

required to handle large data sets and the lack of comprehensive pathway databases that can be used for rapid comparative analysis. One major challenge arises from the fact that genome sequences are rarely closed, with the number of contigs dependent upon the depth of sequencing and the efficiency of the assembly process. In the case of secondary metabolism, where biosynthetic gene clusters can exceed 100 kb, it is uncommon to capture large pathways on a single contig. In addition, the highly repetitive sequence motifs associated with many biosynthetic genes create assembly challenges that are not readily surmountable regardless of sequencing depth. Despite these challenges, it has become widely recognized that bacterial genomes harbor many more biosynthetic pathways than the number of compounds discovered from them would predict [23]. While pathways can be readily identified from sequence data using tools such as antiSMASH [22] and SBSPKS [2], it has become increasingly clear that the identification and structure elucidation of the compounds they produce, and the establishment of formal links between pathways and products, remain major bottlenecks.

The recognition that even well-studied species such as *S. coelicolor* can harbor a large number of pathways for which the products remain unknown came as something of a surprise [4]. This observation implies that the associated compounds are either not being produced or are not being detected using the techniques employed. Both of these issues can be addressed but not without significant effort. An alternative is heterologous expression, which may ultimately provide the most effective approach, but currently remains limited in application. Here we provide perspectives on these various topics derived from our experience with the marine actinomycete genus *Salinispora*. While it remains unclear how broadly applicable the results obtained for this model organism will be to other bacteria, the challenges are similar to those faced with better known secondary metabolite producing taxa such as the genus *Streptomyces*.

Salinispora genomics

The marine actinomycete genus *Salinispora* is comprised of only three species [1, 20], yet has yielded an impressive array of structurally diverse secondary metabolites [10]. Most significant among these is salinosporamide A [9], which has advanced to clinical trials for the treatment of cancer [11]. The first *Salinispora* genome to be sequenced revealed a surprisingly large number of biosynthetic pathways relative to the compounds that had been discovered [28]. The second genome sequence provided clear evidence that these pathways were clustered in genomic islands [24] and additional support for the observation that secondary

metabolites were produced in species-specific patterns [14]. The analysis of additional genome sequences is providing new insight into the biosynthetic diversity within this taxon and information about the processes driving secondary metabolite gene evolution. These efforts are being made possible through the acquisition of more than 100 *Salinispora* genome sequences through the Joint Genome Institute Community Sequencing Program (<http://www.jgi.doe.gov/CSP/overview>). This program provides high-quality, annotated draft genomes and is linked to a variety of tools that can be used to assist in genome analyses (<http://img.jgi.doe.gov/>).

Pathway assemblies

The poor assemblies observed for many secondary metabolite biosynthetic pathways creates challenges for bioinformatic-based structure predictions. However, the quality of the assembly can vary greatly depending not only upon the depth of sequencing but also on the type of biosynthetic pathway encountered. For example, of the 11 different type I modular PKS pathways (containing more than three modules) that have been detected to date in *Salinispora* genomes, none were assembled. This was readily apparent from the detection of highly similar KS domains on different contigs and by the use of well-defined pathways, such as that for rifamycin biosynthesis [29], as templates for manual contig assembly. Type I modular PKSs are highly repetitive and thus it is not surprising that they create challenges for assembly algorithms. In some cases, modules are collapsed within the assemblies while in others they simply fail to assemble. Another interesting observation is that the same PKS pathway can be truncated in the identical location in different genome sequences. This is exemplified by the *cya* gene cluster, which is responsible for the biosynthesis of the cyanosporasides in *S. pacifica* strain CNS-143 [17]. In multiple strains that possess this pathway, the contigs truncate within *orf7*, which is annotated as a dihydrofolate reductase (Fig. 1). Further examination reveals that each contig ends at the same nucleotide, which suggests that the termination may be linked to the sequencing technology itself. Certainly new sequencing technologies that acquire longer read lengths, such as that marketed by Pacific Biosciences (<http://www.pacificbiosciences.com/>), will help solve this problem. Conversely, the majority of type II PKS pathways, which lack the highly repetitive structure of modular PKSs, are fully assembled in the *Salinispora* genomes.

Defining pathway boundaries

Identifying the boundaries of a biosynthetic gene cluster is a subjective process. Outside of the core biosynthetic genes and those associated with regulation and transport, there

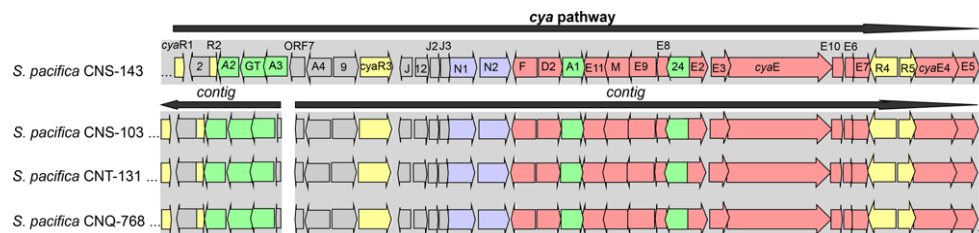
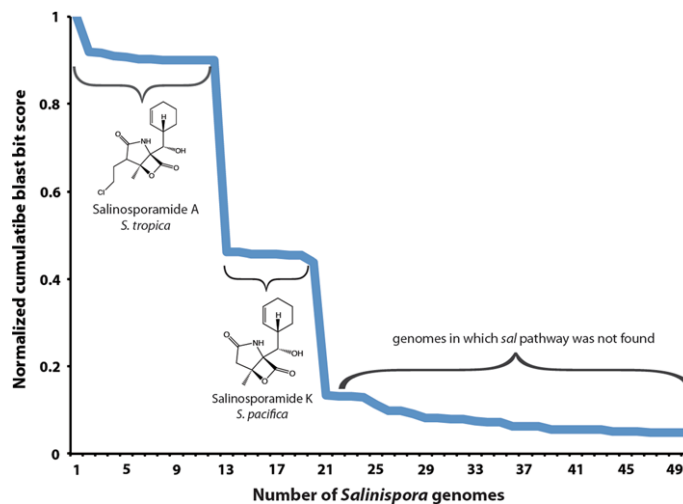


Fig. 1 The cyanosporaside pathway *cya*. This pathway was originally characterized in *S. pacifica* strain CNS-143 [17] using a combination of fosmid sequencing and primer walking. The three *S. pacifica* draft genome sequences (strains CNS-103, CNT-131, and CNQ-768) all possess the pathway, which occurs on two contigs. These contigs all terminate at the same nucleotide positions in ORF7

Fig. 2 Distribution of the *sal* pathway among *Salinispora* genomes. MultiGeneBlast [21] was used to BLAST the *sal* pathway from *S. tropica* (which is responsible for salinosporamide A production) against a database of *Salinispora* genomes. The resulting cumulative BLAST bit scores were normalized to the genome from which the query pathway was derived. A drop in scores to ca. 0.5 is observed for *S. pacifica* genomes that possess the version of the pathway responsible for salinosporamide K production. These strains lack the genes encoding the 26-kb chloroethylmalonyl-CoA portion of the pathway [7]. Scores then drop further to <0.2 in strains that do not possess the pathway



are often uncertainties about other genes in the cluster, especially in the flanking regions and for those with hypothetical annotations. Having access to multiple genomes from strains that produce the same compound provides a useful method to predict the minimum pathway required for compound production. MultiGeneBlast [21] provides a useful tool for this type of analysis. The search output includes cumulative blast bit scores, which represent the sum of the BlastP bit scores for all genes in a genome that match the query sequence. This score provides a quantitative method to estimate the presence/absence of pathways in genome sequences as scores generally drop precipitously when a pathway is not present. Furthermore, this tool can be used to identify strains that contain variations of related pathways, which can be predictive of structural variations within a compound class. This is exemplified by

the production of salinosporamides A and K by *S. tropica* and *S. pacifica*, respectively [7, 8]. In this case, plotting the normalized blast bit scores shows a clear stepwise decrease that corresponds to the distribution of the salinosporamide A and K biosynthetic pathways in the two *Salinispora* species (Fig. 2). A subsequent decrease in scores to <0.2 is then observed for genomes that do not possess the pathway. It is particularly interesting to consider that the variations in the *sal* pathway correspond to a speciation event and to speculate on the potential ecological significance of the structural changes [12].

Sequence tags

Given that many biosynthetic pathways are not fully assembled in most draft genome sequences, an alternative

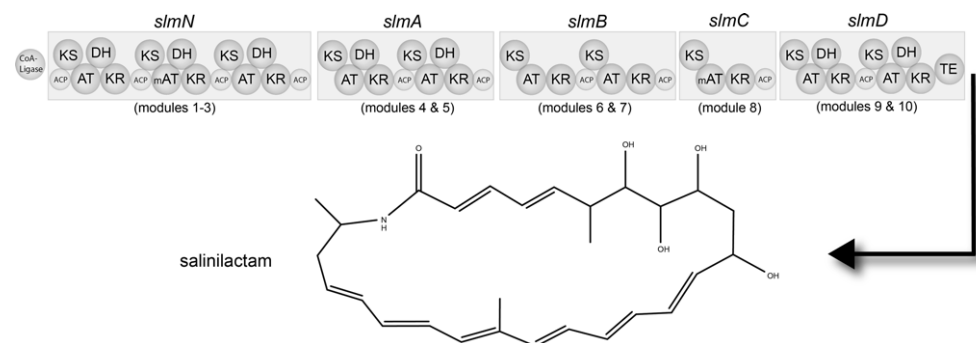


Fig. 3 Modular organization of the *slm* pathway as observed in *S. tropica* strain CNB-440 and its associated product salinilactam. *ACP* acyl-carrier protein, *KS* ketosynthase, *AT* acyl-transferase, *mAT*

methyl-malonyl-CoA-specific AT domain, *DH* dehydratase, *KR* ketoreductase, *TE* thioesterase

approach to predict the class of compounds that will be produced is through the use of sequence tags. NaPDOS was specifically developed using this concept for the classification of PKS and NRPS genes based on sequence tags corresponding to KS and C domains, respectively [31, 32]. Despite sequence lengths of only 200–300 amino acids, these tags can be used to make predictions about the pathway type (e.g., type I or type II PKS), the structural class of the product (e.g., an enediyne or a PKS-NRPS hybrid), and in the case of close matches (i.e., >90 % nucleotide sequence identity), the structure of the product. The analysis of sequence tags can be a particularly useful strategy in the case of poorly assembled genome sequences and to make a quick assessment of the biosynthetic potential of individual strains including the de-replication of well-known compound classes. This approach can also be readily applied to environmental DNA in an effort to identify the best sample types from which to target cultivation efforts.

Iterative pathway-product analysis

While genome sequences can be viewed as the ultimate predictors of secondary metabolite biosynthesis, there are numerous examples where generalized biosynthetic logic, such as the co-linearity rule, has been violated. These examples include the siderophore coelichelin, which included one more amino acid than predicted based on the associated tri-modular NRPS [18], and have led to an improved understanding of processes such as module skipping and stuttering [3]. Exceptions from traditional paradigms have helped expand our understanding of biosynthetic processes and emphasize the importance of having structurally characterized compounds that can be compared

to the pathways responsible for their production [25]. Interestingly, structures can also be used to help resolve ambiguities in genome sequence data. Salinilactam A isolated from *S. tropica* provides one such example (Fig. 3). The initial characterization of this compound revealed a carbon skeleton that would require ten extension modules. The candidate *slm* pathway, however, was found on separate contigs and, due to the high level of sequence similarity among modules (>99 % in many regions), it was not clear how they should be assembled. An understanding of the basic structure of the compound provided the logic to assemble the pathway into ten modules, which ultimately helped to close the *S. tropica* CNB-440 genome sequence [28]. The organization of the dehydratase domains could then be used to help assign the eight conjugated olefins in the compound, which were difficult to resolve by NMR, along with the position of a methyl group at C-18 based on the methyl-malonyl-CoA specificity of the associated AT domain. This type of iterative process between structure assignment and gene cluster assembly can be especially useful for large modular PKSs.

The bottlenecks

Two major bottlenecks to genomics-based natural product research are readily apparent. One is the large number of biosynthetic pathways that appear to be silent. If in fact these pathways are not being expressed, it is clear that improved cultivation techniques, e.g., those that seek to better mimic natural conditions, must be sought. Alternatively, it is possible that many are being expressed yet the products are simply not being detected with the analytical methods employed. This could be due to low yield, the absence of a UV chromophore, or the complexity of the mixture

within which they reside. Alternatively, organic extraction methods, which frequently select for more lipophilic compounds, may not be appropriate for many of these products. A second bottleneck remains the establishment of formal links between specific pathways and structurally characterized compounds. In cases such as modular type I PKSs, bioinformatic predictions may correlate well with structures, however in others correlations are less apparent. Experimentally verified links are extremely important as once made they inform all future discovery efforts. Yet establishing these links requires considerable effort, such as knocking out key genes in the biosynthetic pathway, and thus remains a time-consuming endeavor.

Salinispora genetics

While *Salinispora* genomics has provided considerable insight into the natural product biosynthetic diversity of the genus, the development of genetic protocols to work with these bacteria has been crucial to experimentally link biosynthetic pathways to specific metabolites. The first validated *Salinispora* biosynthetic pathway was the *sal* locus in *S. tropica* CNB-440, which is responsible for the construction of the β -lactone proteasome inhibitor salinosporamide A [6]. Since then, numerous *Salinispora* biosynthetic gene clusters have been validated, including those associated with the production of the cyclic peptide cyclomarin (*cym*) in *S. arenicola* CNS-205 [26] and the enediyne polyketide cyanosporaside (*cya*) in *S. pacifica* CNS-143 [17]. The general methodology for interrogating the function of *Salinispora* genes involves PCR targeting via Red/ET recombineering, which is also useful in other actinomycetes [13]. In addition to facilitating gene deletions, λ -Red-mediated recombination has also been used to replace *Salinispora* genes with homologues in order to alter native pathway functions as in the genetic engineering of fluorosalinosporamide [5]. ϕ C31 phage-based vectors have also been employed to integrate DNA into *Salinispora* chromosomes at pseudo-*attB* sites [19]. These studies have shown that modifications to genetic methods commonly employed with terrestrial actinomycetes, such as the model organism *S. coelicolor* A3(2), are appropriate in *Salinispora* after taking into account its requirement for saline growth media. A recently constructed synthetic promoter library for actinomycetes based on -10 and -35 consensus sequences of native promoters proved equally effective in *S. tropica* CNB-440 as in several terrestrial actinomycete strains [27], thereby adding to the notion that *Salinispora* isolates are amenable to genetic protocols that are commonly employed in *Streptomyces* spp. [15]. Thus, it comes at some surprise that a *Salinispora* biosynthetic pathway has yet to be heterologously expressed in a surrogate host, which is a well-established practice with terrestrial and

more recently marine *Streptomyces* spp. [30]. The development of an effective expression system for *Salinispora* strains will be needed if we are to effectively capture a greater percentage of the secondary metabolite biosynthetic potential of this marine actinomycete genus.

Natural products chemistry in the post-genomic era

The identification of diverse and abundant secondary metabolite biosynthetic pathways in bacterial strains has created considerable excitement about opportunities for the isolation of new metabolites. Of course, it is unclear how many of these pathways are expressed when strains are grown in the laboratory and, if they are, at what levels the associated compounds are produced. While bioinformatic analyses provide the opportunity to evaluate the biosynthetic potential of individual strains and predict, to varying degrees of accuracy, the structures and even stereochemical details of metabolites, a major difficulty remains the translation of this potential into purified molecules that can be evaluated using spectroscopic techniques and tested for biological activity. One major obstacle is compound yield. While sensitive techniques such as high-resolution mass spectroscopy can detect the presence of compounds that occur in very low yields, the isolation and structure elucidation of these compounds generally requires that they be obtained in milligram quantities. Given that laboratory cultures can produce compounds at the microgram per liter level, there remain significant challenges in the purification of sufficient quantities for identification. Although large-scale cultivation technologies can address this problem, these facilities are seldom available in academic settings. These issues contribute to the gap between the pathways observed in genome sequence data and the isolation and characterization of the associated compounds. The continued development of new isolation and spectroscopic methods that accommodate smaller sample sizes will surely facilitate the discovery of a greater percentage of these minor metabolites.

Conclusions

The increasing availability of DNA sequence data has brought natural products research into the genomic era. Genome sequences provide valuable blueprints that can speed the de-replication process and direct the selection of strains for detailed chemical and genetic studies. As the utility of genome sequence data becomes apparent, so do a number of challenges that need to be overcome for the potential of genomics-based natural product research to be fully realized. These challenges include the accurate assembly of highly repetitive sequence motifs, the

isolation and structural characterization of compounds that are produced in low yields, and the creation of formal links between pathways and compounds, which requires tractable genetic approaches that are applicable to diverse organisms. While heterologous expression remains a particularly promising approach, considerable work remains before this technique will become broadly applicable to natural product discovery. Despite these challenges, genomics has taken center stage in the field of natural product discovery. The renewed interest in this field, coupled with increasingly cost-effective genome sequencing, will undoubtedly continue to drive future discovery efforts and help realize the potential of microorganisms to yield new chemical scaffolds that can be explored for applications in medicine and biotechnology.

Acknowledgments PJ and WF acknowledge financial support from the National Institutes of Health (NIH R37 CA 044848 and R01-GM086261) and the Fogarty Center International Cooperative Biodiversity Groups program (grant U01-TW007401-01). PJ, WF, and BSM acknowledge support from the NIH (grant R01-GM085770).

References

- Ahmed L, Jensen P, Freil K, Brown R, Jones A, Kim B-Y, Goodfellow M (2013) *Salinispora pacifica* sp. nov., an actinomycete from marine sediments. *Antonie Van Leeuwenhoek* 103:1069–1078
- Anand S, Prasad MVR, Yadav G, Kumar N, Shehara J, Ansari MZ, Mohanty D (2010) SBSPKS: structure-based sequence analysis of polyketide synthases. *Nucleic Acids Res* 38:487–496
- Bachmann BO, Ravel J (2009) Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol* 458:181–217
- Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD et al (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 417:141–147
- Eustáquio AS, O'Hagan D, Moore BS (2010) Engineering fluorometabolite production: fluorinase expression in *Salinispora tropica* yields fluorosalinosporamide A. *J Nat Prod* 73:378–382
- Eustáquio AS, Pojer F, Noe JP, Moore BS (2008) Discovery and characterization of a marine bacterial SAM-dependent chlorinase. *Nat Chem Biol* 4:69–74
- Eustáquio AS, Nam S-J, Penn K, Lechner A, Wilson MC, Fenical W et al (2011) The discovery of salinosporamide K from the marine bacterium "*Salinispora pacifica*" by genome mining gives insight into pathway evolution. *ChemBioChem* 12:61–64
- Eustáquio AS, McGlinchey RP, Liu Y, Hazzard C, Beer LL, Florova G et al (2009) Biosynthesis of the salinosporamide A polyketide synthase substrate chloroethylmalonyl-coenzyme A from S-adenosyl-L-methionine. *Proc Natl Acad Sci* 106:12295–12300
- 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 *Salinispora*. *Angew Chem* 115:369–371
- Fenical W, Jensen PR (2006) Developing a new resource for drug discovery: marine actinomycete bacteria. *Nat Chem Biol* 2:666–673
- Fenical W, Jensen PR, Palladino MA, Lam KS, Lloyd GK, Potts BC (2009) Discovery and development of the anticancer agent salinosporamide A (NPI-0052). *Bioorg Med Chem* 17:2175–2180
- Freel KC, Nam S-J, Fenical W, Jensen PR (2011) Evolution of secondary metabolite genes in three closely related marine actinomycete species. *Appl Environ Microbiol* 77:7261–7270
- Gust B, Challis GL, Fowler K, Kieser T, Chater KF (2003) PCR-targeted *Streptomyces* gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. *Proc Natl Acad Sci* 100:1541–1546
- Jensen PR, Williams PG, Oh DC, Zeigler L, Fenical W (2007) Species-specific secondary metabolite production in marine actinomycetes of the genus *Salinispora*. *Appl Environ Microbiol* 73:1146–1152
- Kieser T, Bibb MJ, Buttner MJ, Chater KF, Hopwood DA (2000) Practical *Streptomyces* genetics. John Innes Foundation, Norwich
- Koehn FE, Carter GT (2005) The evolving role of natural products in drug discovery. *Nat Rev Drug Discov* 4:206–220
- Lane AL, Nam S-J, Fukuda T, Yamanaka K, Kauffman CA, Jensen PR et al (2013) Structures and comparative characterization of biosynthetic gene clusters for cyanosporasides, enediyne-derived natural products from marine actinomycetes. *J Am Chem Soc* 135:4171–4174
- Lautru S, Deeth RJ, Bailey LM, Challis GL (2005) Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nat Chem Biol* 1:265–269
- Lechner A, Eustáquio A, Gulder TAM, Hafner M, Moore BS (2011) Selective overproduction of the proteasome inhibitor salinosporamide A via precursor pathway regulation. *Chem Biol* 18:1527–1536
- Maldonado LA, Fenical W, Jensen PR, Kauffman CA, Mincer TJ, Ward AC et al (2005) *Salinispora arenicola* gen. nov., sp. nov. and *Salinispora tropica* sp. nov., obligate marine actinomycetes belonging to the family Micromonosporaceae. *Int J Syst Evol Microbiol* 55:1759–1766
- Medema MH, Takano E, Breitling R (2013) Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol Biol Evol* 30:1218–1223
- Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA et al (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 39:W339–W346
- Nett M, Ikeda H, Moore BS (2009) Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Natural Product Reports* 26:1362–1384
- Penn K, Jenkins C, Nett M, Udawary DW, Gontang EA, McGlinchey RP et al (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. *ISME J* 3:1193–1203
- Ross AC, Xu Y, Lu L, Kersten RD, Shao Z, Al-Suwailem AM et al (2012) Biosynthetic multitasking facilitates thalassosporamide structural diversity in marine bacteria. *J Am Chem Soc* 135:1155–1162
- Schultz AW, Oh DC, Carney JR, Williamson RT, Udawary DW, Jensen PR et al (2008) Biosynthesis and structures of cyclomarin and cyclomarazines, prenylated cyclic peptides of marine actinobacterial origin. *J Am Chem Soc* 130:4507–4516
- Siegl T, Tokovenko B, Myronovskiy M, Luzhetskyy A (2013) Design, construction and characterisation of a synthetic promoter library for fine-tuned gene expression in actinomycetes. *Metab Eng* 19:98–106
- Udawary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W et al (2007) Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *Proc Natl Acad Sci* 104:10376–10381

29. Wilson MC, Gulder TAM, Mahmud T, Moore BS (2010) Shared biosynthesis of the saliniketals and rifamycins in *Salinispora arenicola* is controlled by the sare1259-encoded cytochrome P450. J Am Chem Soc 132:12757–12765
30. Yamanaka K, Ryan KS, Gulder TAM, Hughes CC, Moore BS (2012) Flavoenzyme-catalyzed atropo-selective N,C-bipyrrole homocoupling in marinopyrrole biosynthesis. J Am Chem Soc 134:12434–12437
31. Ziemert N, Jensen PR (2012) Phylogenetic approaches to natural product structure prediction. Methods Enzymol 517:161–182
32. Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR (2012) The natural product domain seeker NaPDoS: a phylogeny-based bioinformatic tool to classify secondary metabolite gene diversity. PLoS ONE 7:e34064

SCIENTIFIC REPORTS

OPEN Phylogenomic Insight into *Salinispora* (Bacteria, Actinobacteria) Species Designations

Received: 15 February 2017
Accepted: 18 April 2017
Published online: 15 June 2017

Natalie Millán-Aguinaga^{1,2}, Krystle L. Chavarria¹, Juan A. Ugalde^{1,3}, Anne-Catrin Letzel¹, Greg W. Rouse⁴ & Paul R. Jensen^{1,4}

Bacteria represent the most genetically diverse kingdom of life. While great progress has been made in describing this diversity, it remains difficult to identify the phylogenetic and ecological characteristics that delineate groups of bacteria that possess species-like properties. One major challenge associated with species delineations is that not all shared genes have the same evolutionary history, and thus the choice of loci can have a major impact on phylogenetic reconstruction. Sequencing the genomes of large numbers of closely related strains provides new opportunities to distinguish ancestral from acquired alleles and assess the effects of recombination on phylogenetic inference. Here we analyzed the genomes of 119 strains of the marine actinomycete genus *Salinispora*, which is currently comprised of three named species that share 99% 16S rRNA gene sequence identity. While 63% of the core genome showed evidence of recombination, this had no effect on species-level phylogenomic resolution. Recombination did however blur intra-species relationships and biogeographic resolution. The genome-wide average nucleotide identity provided a new perspective on *Salinispora* diversity, revealing as many as seven new species. Patterns of orthologous group distributions reveal a genetic basis to delineation the candidate taxa and insight into the levels of genetic cohesion associated with bacterial species.

The concept that bacteria can be grouped into phylogenetically cohesive clusters with properties that allow them to be regarded as “species” remains controversial^{1,2}. It is challenging to determine which clusters represent species level units of diversity and if ecological or evolutionary theory can be invoked to explain the circumstances that led to their formation³. As Gevers *et al.* lament⁴, “any effort to produce a robust species definition is hindered by the lack of a solid theoretical basis explaining the effects of biological processes on cohesion within and divergence between species”. Nonetheless, identifying meaningful groups of bacteria and ascribing formal Latinized names remains useful in clinical, environmental, and experimental contexts⁵. In the absence of a robust species concept for bacteria, we are left with a series of metrics used to gauge the relatedness among strains and phylogenetic frameworks within which species level units of diversity are often arbitrarily assigned.

It is widely recognized that bacterial species concepts should consider both genetic diversity and ecology^{2,6,7}. Buckley and Roberts stated that, “in moving forward with microbial taxonomy, it is critical to determine whether microorganisms cluster in groups with meaningful commonalities or determine what commonalities may be best used to cluster microorganisms into meaningful groups”⁸. The ecotype model states that bacterial species should fall into well-supported sequence clusters that evolve under cohesive processes and are ecologically distinct and irreversibly separated from each other⁶. A fundamental tenant of this model is that ecologically distinct populations can be recognized as clades in phylogenetic trees and that these clades correspond to fundamental units of diversity or species^{2,6}.

¹Center for Marine Biotechnology and Biomedicine Scripps Institution of Oceanography, University of California San Diego, San Diego, California, United States. ²Universidad Autónoma de Baja California. Facultad de Ciencias Marinas, Ensenada, Baja California, Mexico. ³Centro de Bioinformática y Biología Integrativa, Facultad de Ciencias Biológicas, Universidad Andrés Bello, Santiago, Chile. ⁴Marine Biology Research Division Scripps Institution of Oceanography, University of California San Diego, San Diego, California, United States. Correspondence and requests for materials should be addressed to P.R.J. (email: pjensen@ucsd.edu)

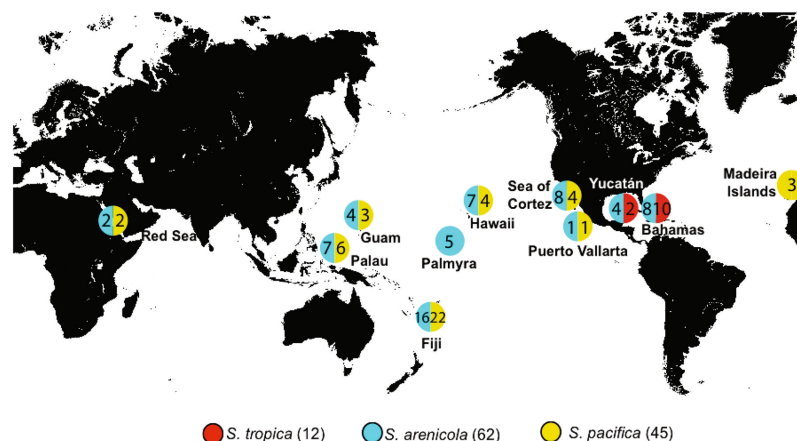


Figure 1. Strain origins. Numbers of strains sequenced at each site for each species with totals in parentheses. Modified with permission from Freel *et al.*²⁴, Environ. Microbiol. 14:480–493.

Confounding the common ancestry inferred by phylogenetic reconstruction is homologous recombination. While the efficiency of homologous recombination decreases with increasing genetic distance⁹, it nonetheless occurs between different species¹⁰. The homologous exchange of genes encoding common housekeeping functions creates challenges for species delineations based on single gene phylogenies and led to the use of techniques such as multi-locus sequence analysis¹. However, even when multiple loci are considered, an accurate model of vertical inheritance can be difficult to depict due to widespread recombination between species^{11,12} including ancestral events that have subsequently become fixed among subclades¹³. While the rates of recombination vary widely among bacteria¹⁴, it remains largely unknown how this process affects species-level phylogenetic resolution when whole genomes are considered.

Whole-genome sequencing has become an indispensable tool for studying genome evolution, genetic diversity, and bacterial species concepts. It has recently been suggested that genome sequences should be used as a source of taxonomic information¹⁵. One genome-based metric that is gaining acceptance is the Average Nucleotide Identity (ANI) of the sequences shared between strains. It has been shown that an ANI of 95% corresponds to the 70% DNA:DNA hybridization value traditionally used to delineate bacterial species¹⁶ thus establishing a link to bacterial systematics. Genome sequences also provide unique opportunities to generate highly resolved phylogenies, with automated pipelines to build genomic phylogenies from concatenated protein markers now available¹⁷. While there is no agreement regarding how many genes it takes to generate a robust phylogenomic evolutionary tree, genome sequences provide comprehensive datasets from which to address evolutionary relationships and predict lateral gene transfer events¹⁸.

The marine actinomycete genus *Salinispora* provides a valuable model to address bacterial species concepts^{19,20}. It is comprised of three closely related species (*S. arenicola*, *S. tropica*, and *S. pacifica*) within the family Micromonosporaceae^{21,22} whose relationships could not be confidently resolved based on 16S rRNA gene phylogeny^{23,24}. The genus is a rich source of structurally diverse natural products²⁵, and there is evidence that certain compounds²⁶ and their associated biosynthetic gene clusters (BGCs)²⁷ are fixed at the species level. This has been used to suggest that secondary metabolites represent ecotype-defining traits for *S. tropica* and *S. arenicola*. Similar patterns were not observed for *S. pacifica*²⁶, the most diverse of the three species²⁴. This greater diversity, coupled with the relatively low recombination to mutation rates observed within the *S. pacifica* clade, were used to suggest it represents an amalgam of ecotypes or newly diverged species¹⁹. While all three species are prolific in terms of natural product biosynthesis, it was shown that *S. arenicola* differentially invests in interference competition, while *S. tropica* invests in growth thus establishing these co-occurring lineages as distinct ecotypes²⁸. Here we present a phylogenomic analysis of the genus *Salinispora* based on the shared gene content among 119 strains. The goals were to assess species level diversity and address the effects of recombination on species level phylogenetic reconstruction.

Results

General genome characteristics. The 119 *Salinispora* genome sequences were derived from 12 *S. tropica*, 45 *S. pacifica*, and 62 *S. arenicola* strains isolated from 11 global locations (Fig. 1). All strains were obtained from marine sediment samples collected at depths from 1–700 meters with the exception of four that were derived from marine sponges (Supplementary Table S1). No heterogeneity was observed in the 2–5 copies of the 16S rRNA gene observed in any of the strains. The draft genome sequences averaged 86.3 contigs (Supplementary Table S2) with the majority of sequence data accounted for by a few large contigs in each genome. The average genome size

Taxa	Genome Size (Mbp)	Gene Count	Scaffold Count	GC Content (%)
<i>Salinispora</i>	5.57	5148	85	69.7
<i>S. arenicola</i>	5.74	5234	80	69.8
<i>S. pacifica</i>	5.42	5079	90	69.9
<i>S. tropica</i>	5.31	4959	89	69.2

Table 1. Average genome statistics for the genus *Salinispora* and each species.

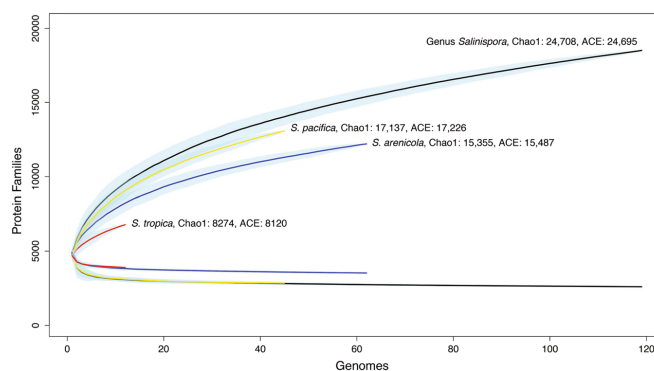


Figure 2. Rarefaction curves. Orthologous groups (protein families) plotted vs. the number of sequenced genomes. Core genomes (lower curves) and pan-genomes (upper curves) are shown for the genus and each species. Black: genus, red: *S. tropica*, blue: *S. arenicola*, yellow: *S. pacifica*. Blue shading indicates standard error. Diversity estimates using Chao1 and ACE are given.

was 5.49 Mb, with the *S. arenicola* genomes being larger and containing more genes than the other two species (Table 1).

Orthologous groups. The program FastOrtho was used to predict a total of 13,512 orthologous groups (OGs) and 4,980 single copy genes (singletons) among the 119 *Salinispora* genomes revealing a pan-genome that totaled 18,492 protein families. The core genome consists of 2603 OGs shared by all 119 strains, with 2362 of these occurring as a single-copy in each strain. The core genome represents 51% of the average gene content across the genus. Based on the annotation or putative function of the OGs, more than 50% of the pan-genome is comprised of poorly characterized genes (Supplementary Fig. S1). As observed in other genera³⁹, the core genome is enriched in functionally annotated genes with the largest group (35%) attributed to metabolism. Similar analyses performed at the species level reveal that *S. tropica* has the largest core genome representing 78.68% of the average gene content. Conversely, *S. pacifica* displays the smallest core genome at 56.10% of the average gene content while *S. arenicola* was intermediate at 67.42%. As expected, the core genomes vary inversely as a function of the diversity of the strains sequenced within each species.

Rarefaction curves were computed to estimate how effectively gene content had been sampled (Fig. 2). There is clear evidence for saturation when the genus or species-level core genomes are considered and thus the common genetic features that characterize the cultured representatives of these taxa have largely been identified. It is notable that the curves generated from the *S. tropica* and *S. arenicola* core genomes are largely identical, while the curve for *S. pacifica* resembles that describing the genus. For the pan-genomes however, it can be predicted that additional sequencing will reveal additional genetic diversity at all levels. Diversity estimators (Chao1 and ACE) predict more than 24,000 protein families at the genus-level relative to the 18,492 observed. Of the three species, *S. pacifica* shows the highest observed and predicted genetic diversity.

Effects of recombination on *Salinispora* phylogeny. The 2362 single copy genes identified in the core *Salinispora* genome (hereafter referred to as the single copy core or SCC) were used to generate a concatenated phylogeny that clearly resolved the genus into three well supported clades in accordance with prior species-level relationships (Fig. 3)¹⁹. This phylogeny supports the relatively high level of diversity reported for *S. pacifica*. We next used the program PhiPack to address the effects of recombination on phylogenetic reconstruction⁴⁰. This led to the detection of 1,486 SCC genes (62.9%) with evidence of recombination. The remaining 876 genes had no evidence of recombination and are considered the “minimum” core genome. We generated a second concatenated phylogenomic tree using the minimum core genome (Fig. 3) and manually compared this to the individual gene trees for each of the 1,486 SCC genes with evidence of recombination. We identified 635 genes (42.7% of those

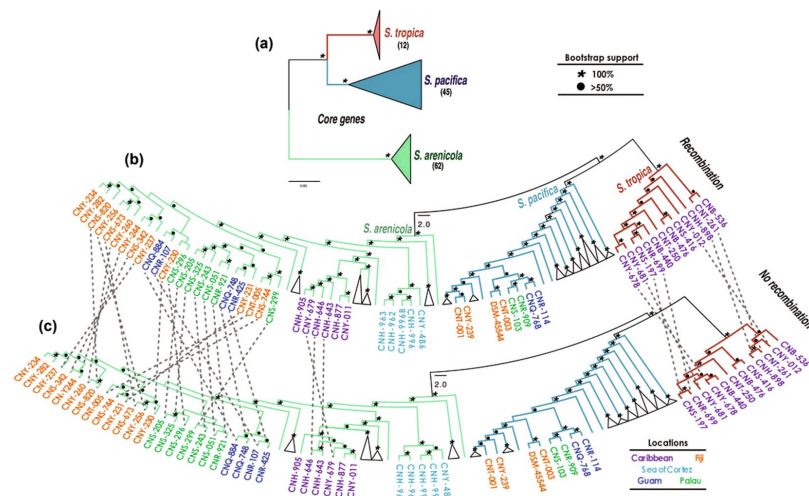


Figure 3. *Salinispora* maximum likelihood phylogeny. (a) Collapsed phylogenomic tree based on a concatenation of 2362 shared, single-copy genes. Number of strains analyzed for each species is shown in parentheses. Non-collapsed trees are presented in Fig. 5. (b) Phylogeny based on genes with evidence of recombination. (c) Phylogeny based on genes with no evidence of recombination. Strain numbers are given in cases where the tree topologies differ. When possible, branches with the same topology in both trees were collapsed. Dashed lines depict positional changes of strains in the trees. Branches are color-coded by species. Symbols on the branches represent the support from 1,000 bootstrap replicates. Strain numbers are color-coded by location.

under recombination and 26.9% of the SCC) that displayed incongruent species level phylogenies for at least one strain relative to the concatenated phylogenomic tree (Supplementary Fig. S2). To test for the aggregate effects of recombination, a third concatenated phylogeny was generated using the 1486 SCC genes with evidence of recombination (Fig. 3). Surprisingly, all three trees were both congruent and similarly well supported in terms of the three major clades associated with the named *Salinispora* species. Thus, recombination did not affect *Salinispora* species-level phylogenomic resolution. The large numbers of genes that displayed incongruent species-level phylogenies were insufficient to affect interspecies relationships when taken in the context of the larger gene pools. Notably, the tree generated from the minimum core genome reveals clear biogeographic patterns within *S. arenicola* that were obscured when genes subject to recombination were included (Fig. 3).

These phylogenies were based on the concatenation of various gene sets into a single multiple alignment and the estimation of a single tree from this super-alignment. Given that alternative phylogenetic methods can infer different relationships, the data were re-analyzed using ASTRAL (Accurate Species TRee ALgorithm), a coalescent-based method to summarize individual gene trees into a single species tree³¹. ASTRAL identifies the species tree that agrees with the largest number of individual trees and can be more accurate than maximum likelihood analyses when using a concatenated gene set³². Given this, we performed a similar set of analyses using ASTRAL, which resulted in trees that were congruent at the species level with the concatenated trees (Supplementary Figs S3 and S4), thus providing further support for these phylogenetic patterns.

ANI and ANI-AF metrics. We next asked if the species assignments inferred from the three primary clades observed in the SCC phylogenomic tree, which have been used to distinguish among the three *Salinispora* species²⁴, were in accordance with the proposal that ANI values between members of the same species should be $\geq 95\%$ ¹⁶. A distance matrix based on ANI values reveals a dendrogram with three primary bifurcations that are congruent with the phylogenomic tree (Fig. 4). However, many strains within the three primary clades fall below the 95% ANI metric, suggesting the existence of additional species-level diversity. More specifically, seven branches within the primary *S. pacifica* lineage could be considered distinct species based on this metric. The most populated branch includes the *S. pacifica* type strain (CNR-114)²² and 22 additional strains isolated from seven of the global collection sites. The second most populated branch includes 12 strains recovered largely from Fiji while the remaining five branches include one to three strains. The strains comprising these seven lineages are clearly resolved in the expanded phylogenomic tree (Fig. 5) and suggest that the primary clade sister to *S. tropica* is comprised of as many as seven distinct species of which *S. pacifica* is one. Similarly, the *S. arenicola* clade includes two branches that fall below the 95% ANI level. These consist of the single strain CNY-281 and a second branch that contains all of the remaining *S. arenicola* strains including the type strain. Conversely, the *S. tropica*

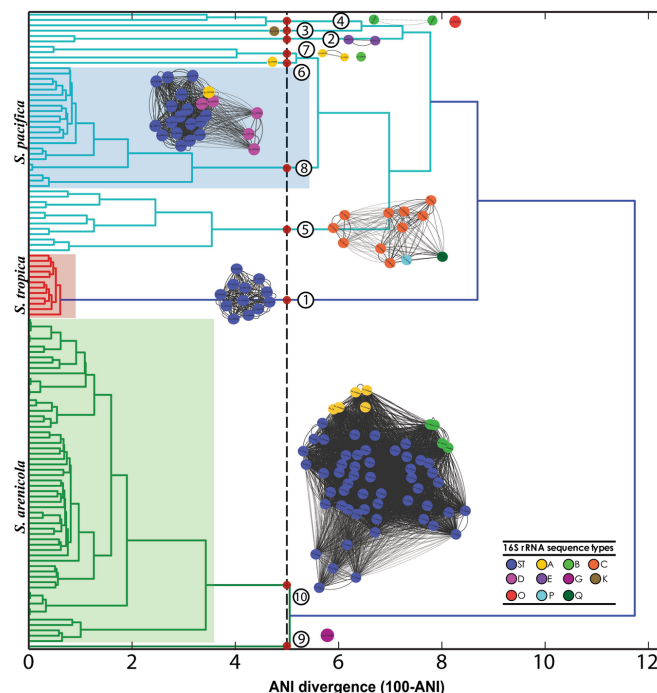


Figure 4. Average Nucleotide Identity (ANI) distance matrix. The vertical dashed line represents 95% ANI. Strains associated with the three primary clades are delineated by green (*S. arenicola*), red (*S. tropica*), and light blue (*S. pacifica*) branches. ANI clades that share >95% and are associated with type strains are shaded. Red circles and corresponding numbers represent all lineages that share <95% ANI values including seven (2–7, 9) that do not contain type strains. ANI-AF networks are shown adjacent to the corresponding regions in the dendrogram. Each node represents a strain and is color-coded based on the 16S rRNA gene sequence types (single nucleotide polymorphisms) observed for each species.

clade is represented by a single branch within which all strains share >95% ANI. Thus, according to the ANI analyses, the 119 *Salinispora* strains represent as many as 10 different species.

We analyzed the data further using the ANI-AF method³³, which considers only coding orthologous groups (CDS: From Coding DNA Sequences) and the alignment fraction (AF) between genomes as a measure of relatedness. The values suggested to delineate species are ANI >96.5 and AF >0.6. The ANI-AF results for *S. tropica* and *S. arenicola* remain the same, however within the *S. pacifica* clade, CNS-055 and CNY-646 are delineated as two additional species. Based on the ANI species designations, we re-investigated the effects of recombination on species-level phylogenomic resolution and once again found no effect (Fig. 5). The 10 candidate *Salinispora* species are all clearly resolved both from their minimum core genomes and the SCC genes with evidence of recombination. Thus, recombination does not affect the phylogenetic resolution of the major lineages associated with the three currently named *Salinispora* species or the ten candidate species into which these lineages could be delineated based on ANI.

Salinispora 16S rRNA sequence types (single nucleotide polymorphisms) correspond surprisingly well to the ANI-AF clustering (Fig. 4). To further explore these relationships, we plotted 16S rRNA sequence divergence vs. ANI (Fig. 6). Interspecies comparisons based on the three primary clades in the *Salinispora* phylogeny revealed from five (St-Sp) to 14 (Sa-Sp) changes in the 16S rRNA gene. All *S. arenicola* and *S. tropica* intra-species comparisons are above 95% ANI and reveal at most three 16S polymorphisms while many of the *S. pacifica* intraspecies comparisons fall below this line and include up to six SNPs. A linear regression of the data and best-fit line reveals that a 95% ANI value corresponds to 3.1 changes in the 16S rRNA gene (Supplementary Fig. S5). Given that many of the intra-clade comparisons for the major clade that includes *S. pacifica* fall below 95% ANI, we performed a separate analysis of these seven lineages (Fig. 6). As expected, all comparisons within these seven clades fall above and all between clade comparisons fall below 95% ANI. In this case however, the inter-clade comparisons differ from 0–6 16S rRNA SNPs.

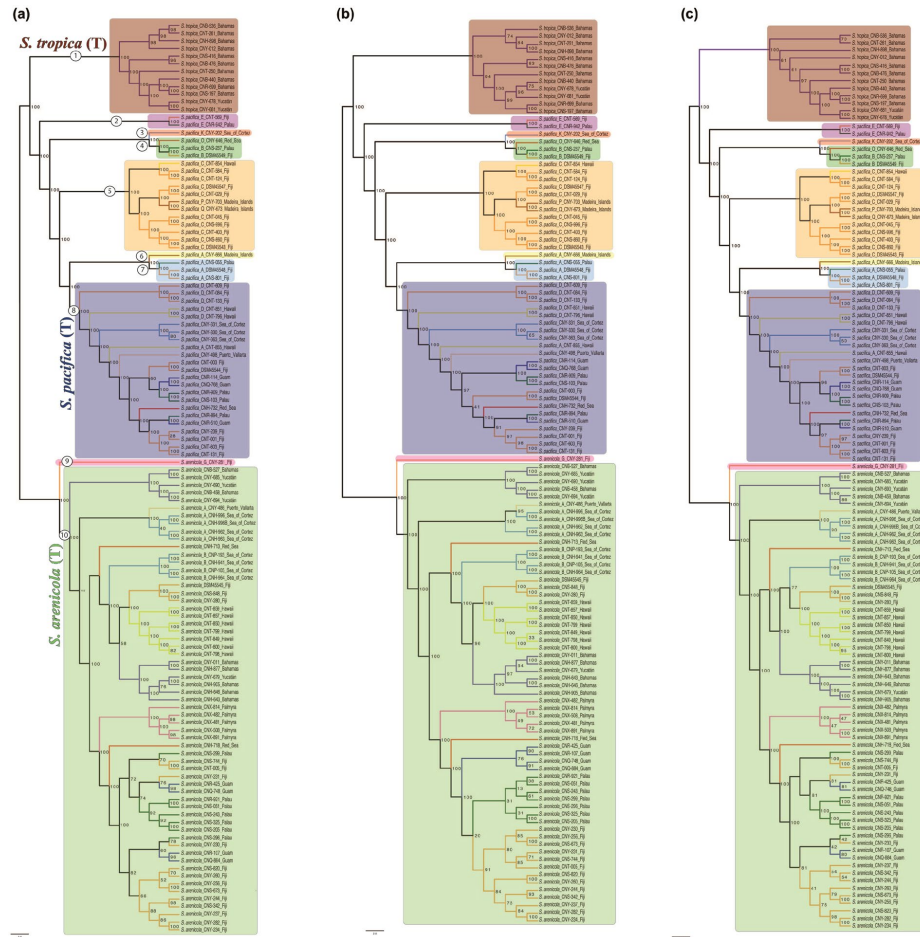


Figure 5. Effects of recombination on phylogenetic resolution using ANI species designations. **(a)** Phylogenomic tree based on a concatenation of 2362 shared, single-copy genes. Each sequence or clade that shares <95% ANI with neighboring strains is numbered 1–10 (corresponding to Fig. 4) and shaded with a different color. Species names are listed corresponding to the clades associated with the type strains (T). **(b)** Phylogeny based on genes with no evidence of recombination. **(c)** Phylogeny based on genes with evidence of recombination.

Genetic basis for species delineations. We previously reported species-specific patterns of secondary metabolite production in *S. arenicola* and *S. tropica*²⁶, however similar patterns were not observed for *S. pacifica*²⁵. To further explore this concept in *S. pacifica*, we identified biosynthetic gene clusters (BGCs) associated with secondary metabolism using antiSMASH³⁴ and manual annotations. We then prepared a similarity matrix using the presence/absence of BGCs in each strain as input (Supplementary Fig. S6). Except for the position of CNY-666, the BGC dendrogram and the phylogenomic tree are largely identical. To further test for evidence of genetic or functional traits that differentiate the candidate *Salinispora* species, we performed similar analyses based on the presence or absence of orthologous groups associated with 23 COG categories (Supplementary Table S3) and found that categories C (energy production and conversion, Supplementary Fig. S7), E (amino acid transport and metabolism), G (carbohydrate transport and metabolism), H (coenzyme transport and metabolism), I (lipid transport and metabolism), and R (general function prediction) consistently delineated the candidate species within the primary *S. pacifica* lineage in accordance with the phylogenomic tree (Fig. 5). Thus, in addition to secondary metabolism, there appear to be major genetic differences among the candidate *S. pacifica* species.

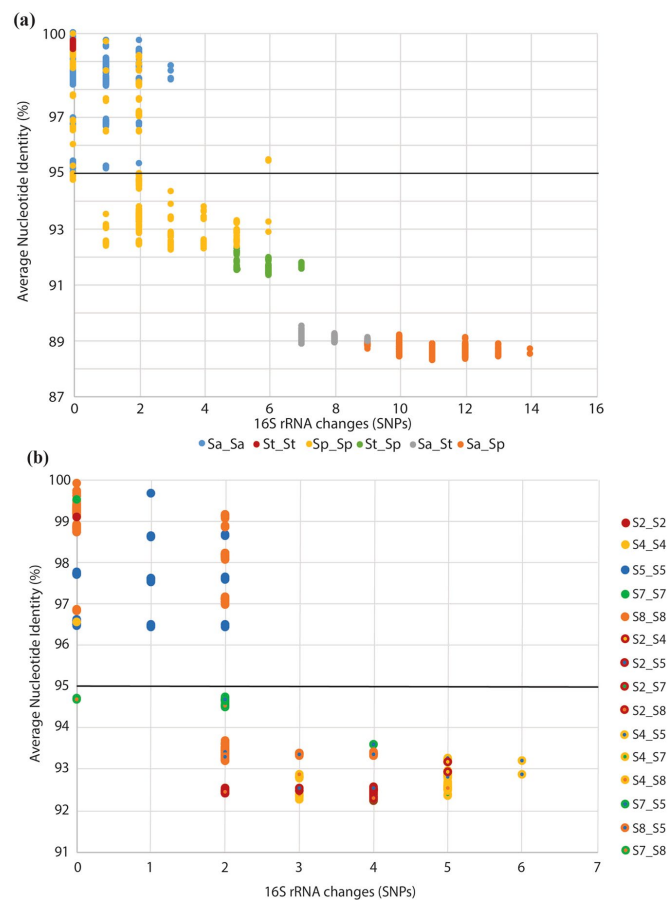


Figure 6. Relationships between 16S rRNA diversity and Average Nucleotide Identity (ANI). The black line indicates 95% ANI. (a) Inter- and intraclade comparisons among the three major lineages represented by *S. tropica* (St), *S. arenicola* (Sa), *S. pacifica* (Sp). (b) Inter- and intraclade comparisons among the *S. pacifica* clades (2–8) as identified in Figs 4 and 5.

While differences in gene content provide one mechanism to distinguish among species, it can also be expected that the same species will share a certain level of genetic homogeneity. To explore these concepts, we plotted OG distributions across various taxonomic levels (Fig. 7). All histograms clearly show that most genes are either rare or occur in all strains. When the genus is assessed, the core genome represents only 14% of the pan-genome and the relatively large spike in the left portion of the graph provides little evidence for genetic cohesion, as might be expected from a genus comprised of multiple species²⁹. Conversely, when *S. arenicola* and *S. tropica* are plotted, the core genomes represent 29% and 58% of the respective pan-genomes, and the numbers of OGs observed in all strains exceed those observed in only one strain. In the primary *S. pacifica* lineage however, the pattern is similar to that detected for the genus, with the core genome representing only 22% of the pan-genome. As was observed in the rarefaction curves, these results are more similar to those for the genus than for either *S. tropica* or *S. arenicola*. We performed similar analyses using the two most populated candidate species within the primary *S. pacifica* lineage and observed OG distributions that resemble *S. tropica* and *S. arenicola*, with core genomes between 40% and 44% of the pan-genomes. These patterns may provide added insight into the levels of genetic cohesion expected for a bacterial species.

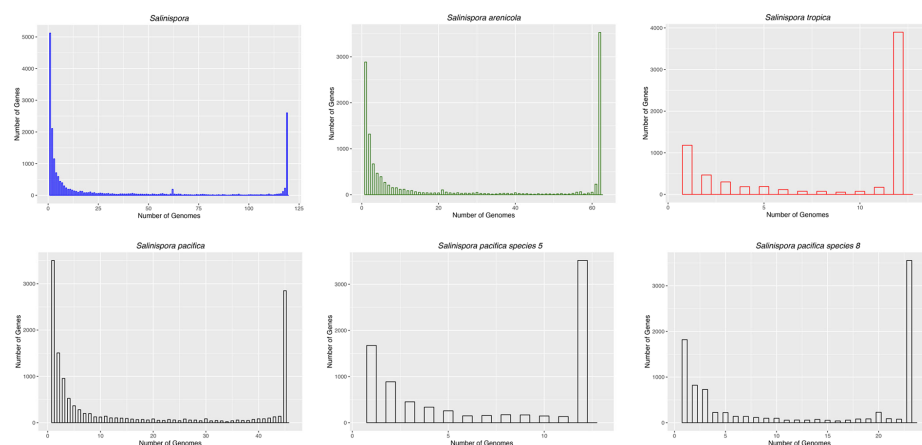


Figure 7. Numbers of orthologous groups found across all genomes (upper left), *S. arenicola* genomes (upper center), *S. tropica* genomes (upper right), *S. pacifica* genomes (bottom left), *S. pacifica* candidate species “Sp1” (bottom center), *S. pacifica* candidate species “Sp2” (bottom right). The histograms were generated from the pan-genomes excluding singletons and recent paralogs.

Discussion

The comparison of large numbers of genome sequences derived from closely related bacteria provides a unique opportunity to address bacterial species concepts and the metrics commonly employed to assess sequence-based relationships. Fundamental to this process is the identification of the core genome, which defines the common genomic features that characterize the strains under consideration. As can be expected, core genomes vary widely depending on the diversity of strains and number of genomes examined^{35–38}. Nonetheless, this shared gene pool provides unparalleled opportunities to assess levels of sequence divergence and generate comprehensive molecular phylogenies that can be used to infer evolutionary relationships and identify alleles that have been exchanged by homologous recombination.

Homologous recombination provides a mechanism to repair damaged DNA and generate genetic diversity within bacterial genomes³⁹. While molecular phylogeny is the primary tool used to assess bacterial diversity, it is well documented that homologous recombination blurs species boundaries and can prevent accurate species delineations⁴⁰. By analyzing the single copy core (SCC) genome associated with 119 closely related *Salinispora* strains, it was possible to generate a detailed and highly supported phylogeny that revealed three primary lineages in agreement with previously established relationships among the three currently named species¹⁹. Although 63% of the SCC showed evidence of recombination for at least one strain, this had no effect on the evolutionary relationships among the three primary clades. However, removing loci that showed evidence of recombination from the analyses revealed enhanced biogeographic patterning within the *S. arenicola* clade and new evidence for endemism among the structured populations. A majority of genes that displayed evidence of recombination generated phylogenies that were congruent with the established species phylogeny, indicating that most of these events occurred within the three primary lineages as opposed to between them. This is in agreement with the concept that recombination provides a cohesive force that maintains species level units of diversity⁴¹. However, the large number of core genes that generated incongruent species phylogenies (27%) reveals the importance of selecting the appropriate phylogenetic markers and the power of phylogenomics to overcome this potential source of misleading phylogenetic inference.

ANI analyses revealed that the three primary *Salinispora* clades could be further delineated into as many as 10 different species, all of which could be confidently resolved even when recombinant alleles were included. While three of these lineages are associated with named species^{21,22}, six belong to the relatively diverse clade that is sister to *S. tropica* and contains the *S. pacifica* type strain. This supports the previous suggestion that this clade represents an amalgam of ecotypes or newly diverges species based on its relative low rates of recombination to mutation¹⁹. The possibility that 10 species are represented among a group of strains that share 99% 16S sequence identity supports the concept that this conserved phylogenetic marker is not the best choice for species-level resolution⁴. Nonetheless, 95% ANI corresponded to approximately three changes in the 16S gene thus indicating that any change in this conserved marker may be meaningful from a taxonomic perspective.

The phylogenetic resolution achieved here is in stark contrast to the genus *Streptomyces*, where high recombination to mutation rates detected using MLSA approaches led to the suggestion that phylogenetic relationships within this genus were better represented by a reticulate network¹². It remains unclear why the effects of recombination on phylogenetic resolution differ between two taxa within the same bacterial order, however it may relate to the diversity of the strains examined and the number of alleles assessed in the different studies. Furthermore,

it is interesting to speculate that among *Streptomyces* spp., the acquisition of alleles resistant to the many antibiotics they produce may contribute to the high levels of homologous recombination observed, as was shown for the *rpoB* phylogeny in *Salinispora* spp.¹⁹ and exploited to identify the biological targets of secondary metabolites prior to their discovery⁴².

In support of this concept, natural product BGCs are frequently exchanged by horizontal gene transfer²⁷ and often include a resistant version of the target on which the encoded compounds act⁴³. These resistance genes often have homologs in the core genome and can appear as a second copy of a housekeeping gene⁴⁴. In other cases, the resistant housekeeping gene associated with the BGC is the only copy in the genome⁴⁵, suggesting the ancestral allele was subsequently lost. These later events are difficult to distinguish from homologous recombination and may account for some of the single copy genes identified as under recombination in this study. Thus, the ability to produce and be resistant to secondary metabolites may represent a major factor confounding phylogenetic resolution among bacteria enriched in this metabolic capacity. Nonetheless, phylogenomic approaches were sufficient to overcome these incongruences, leading to the generation of stable trees with highly supported clades that can be further evaluated for species-like properties.

Linking strains that can be delineated based on phylogeny or sequence similarity with distinct ecological traits remains a critical and challenging component of microbial ecology. In this regard, it was possible to show that the distributions of secondary metabolite BGCs and six COG categories were largely congruent with the 10 candidate *Salinispora* species delineated based on ANI and resolved in the phylogenomic tree. Thus, there appears to be considerable genetic cohesion among these lineages including within the category of secondary metabolism, which has been reported to represent an important species defining trait for this genus⁴⁵. Ultimately, resolving the genetic and ecological differences among these closely related groups of bacteria, as initially demonstrated between strains of *S. tropica* and *S. arenicola*²⁸, will be an essential component of testing the hypothesis that they maintain the properties expected of different species. While it remains to be determined if these results apply more broadly to other groups of bacteria, the expansive growth of genome sequence data will provide ample opportunities to explore species concepts in the future.

Methods

Genome sequencing. Genome sequencing was conducted by the U.S. Department of Energy Joint Genome Institute as part of the Community Science Program (<http://jgi.doe.gov/user-program-info/community-science-program/>). DNA was extracted and the sequence annotation and assembly carried out as previously described²⁷. Genomic data is available from the Integrated Microbial Genomes (IMG) database (<https://img.jgi.doe.gov>). IMG genomes ID and NCBI taxon numbers are provided in Supplementary Table S1.

Orthologous group computation. A total of 119 *Salinispora* strains (12*S. tropica*, 62*S. arenicola* and 45*S. pacifica*) from 11 different locations (Fig. 1, Supplementary Table S1) were analyzed using the program FastOrtho⁴⁶ to identify groups of orthologous protein coding genes (orthologous groups, OGs). This program is a reimplementation of OrthoMCL⁴⁷ and performs a bidirectional best blast amino-acid analysis. Clustering based on a percent match was performed using default parameters (cutoff = 70, e-value cutoff = $1e^{-05}$, and inflation index (I) = 1.5) (<https://github.com/juanu/MicroCompGenomics>). Rarefaction curves and diversity estimates were generated using the vegan package in R (<http://www.R-project.org>). The output matrix of FastOrtho was processed to identify species-specific orthologous groups using an Excel macro (<https://github.com/joseluisrc/FindSharedGenes>). Histograms were plotted from the presence-absence matrix of OGs using the qplot function and the ggplot2 package in R (<http://www.R-project.org>).

Identification of the core genome and the detection of recombination. A series of custom python scripts (<https://github.com/juanu/MicroCompGenomics>) were applied to the FastOrtho results to identify the OG members that included gene duplications (paralogs). Orthologous groups that included paralogs were removed to generate the single copy core (SCC) gene pool. The nucleotide sequences of the individual SCC genes in each strain were aligned using MUSCLE with default parameters and trimmed for quality using GBlocks. The SCC genes were screened for evidence of recombination using PhiPack⁴⁸, which included the statistical tests PHI, MaxChi, and Neighbor Similarity Score, all with default parameters. Recombination was inferred when p-values less than 0.01 were detected. Attempts to use the Recombination Detection Program⁴⁹ failed due to the large number of loci examined.

Phylogenetic analyses. A maximum likelihood (ML) tree was generated for each SCC gene using the program RAXML (command line version) with mid-point rooting and 100 bootstraps (Stamatakis, 2006). The individual gene trees were visualized using the program FigTree v1.3.1 (<http://tree.bio.ed.ac.uk/software/figtree>). Trimmed alignments of each gene were then concatenated and used to build ML phylogenies using RAXML⁵⁰ implemented on the CIPRES portal v2.2 at the San Diego Supercomputer Center⁵¹. Analyses included 1,000 bootstrap replicates using the most complex model (GTR + GAMMA) for both bootstrapping and final ML optimization using default parameter settings. The resulting tree was rooted at the mid-point and visualized using FigTree. Individual SCC gene trees that showed incongruence at the species level with the concatenated tree were scored as under recombination. Two additional concatenated SCC gene trees were then generated for the subsets of this gene pool that included only genes with evidence of recombination and only genes with no evidence of recombination using the methods described above. A similar set of SCC species trees was also generated using the program ASTRAL³¹, which uses the best RAXML trees for each gene tree.

Average nucleotide identity and alignment fraction. The average nucleotide identity (ANI) and alignment fraction (AF) were determined for all 119 *Salinispora* genomes using published methods^{16,33}. ANI values were calculated for all pairwise comparisons and used to compile a distance matrix representing ANI

divergence (100 - ANI). The custom scripts used to perform these analyses and generate the ANI dendrogram are available (https://github.com/juanu/ANI_analysis/blob/master/ANI_blastn.py and <https://ani.jgi-psf.org/html/download.php>). Cytoscape 3.3.0 was used to visualize the results⁵².

Clustering based on COG category and functional traits. The OGs were classified into five major functional categories based on the FastOrtho results and further divided into clusters of orthologous groups (COGs, Supplementary Table S3). These classifications were used to build hierarchical cluster analyses based on the presence/absence of OGs assigned to each COG category using the function `hclust` and the method “average” in the R package (<http://www.R-project.org>). A hierarchical cluster analysis was similarly generated using the presence/absence of secondary metabolite BGCs predicted for the 119 *Salinispora* genomes using antiSMASH⁵⁴ as previously described²⁷.

References

1. Hanage, W. P., Fraser, C. & Spratt, B. G. Sequences, sequence clusters and bacterial species. *Phil Trans Royal Soc B: Biol Sci* **361**, 1917–1927, doi:10.1098/rstb.2006.1917 (2006).
2. Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G. & Hanage, W. P. The bacterial species challenge: Making sense of genetic and ecological diversity. *Science* **323**, 741–746, doi:10.1126/science.1159388 (2009).
3. Doolittle, W. F. & Zhaxybayeva, O. On the origin of prokaryotic species. *Genome Res* **19**, 744–756, doi:10.1101/gr.086645.108 (2009).
4. Gevers, D. *et al.* Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* **3**, 733–739, doi:10.1038/nrmicro1236 (2005).
5. Stackebrandt, E. *et al.* Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Inter J Syst Evol Microbiol* **52**, 1043–1047, doi:10.1099/00207713-52-3-1043 (2002).
6. Cohan, F. What are bacterial species? *Annu Rev Microbiol* **56**, 457–487, doi:10.1146/annurev.micro.56.012302.160634 (2002).
7. Doolittle, W. F. & Papke, R. T. Genomics and the bacterial species problem. *Genome Biol* **7**, 116, doi:10.1186/gb-2006-7-9-116 (2006).
8. Buckley, M. & Roberts, R. Reconciling microbial systematics and genomics. *Amer Acad Microbiol Rep* 2006 (2007).
9. Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M. & Dowson, C. G. Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. *J Bacteriol* **182**, 1016–1023, doi:10.1128/JB.182.4.1016-1023.2000 (2000).
10. Fraser, C., Hanage, W. P. & Spratt, B. G. Recombination and the nature of bacterial speciation. *Science* **315**, 476–480, doi:10.1126/science.1127573 (2007).
11. Doroghazi, J. R. & Buckley, D. H. Widespread homologous recombination within and between *Streptomyces* species. *ISME J.* **4**, 1136–1143, doi:10.1038/ismej.2010.45 (2010).
12. Cheng, K., Rong, X. & Huang, Y. Widespread interspecies homologous recombination reveals reticulate evolution within the genus *Streptomyces*. *Mol Phylogenetics Evol* **102**, 246–254, doi:10.1016/j.ympev.2016.06.004 (2016).
13. Andam, C. P., Choudoir, M. J., Nguyen, A. V., Park, H. S. & Buckley, D. H. Contributions of ancestral inter-species recombination to the genetic diversity of extant *Streptomyces* lineages. *ISME J* **10**, 1731–1741, doi:10.1038/ismej.2015.230 (2016).
14. Vos, M. & Didelot, X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J* **3**, 199–208, doi:10.1038/ismej.2008.93 (2009).
15. Thompson, C. C. *et al.* Microbial taxonomy in the post-genomic era: Rebuilding from scratch? *Arch Microbiol* **197**, 359–370, doi:10.1007/s00203-014-1071-2 (2015).
16. Goris, J. *et al.* DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**, 81–91, doi:10.1099/ijso.0.64483-0 (2007).
17. Wu, M. & Eisen, J. A. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* **9**, 1 (2008).
18. Eisen, J. A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* **8**, 163–167, doi:10.1101/gr.8.3.163 (1998).
19. Freil, K. C., Millan-Aguinaga, N. & Jensen, P. R. Multilocus sequence typing reveals evidence of homologous recombination linked to antibiotic resistance in the genus *Salinispora*. *Appl Environ Microbiol* **79**, 5997–6005, doi:10.1128/AEM.00880-13 (2013).
20. Jensen, P. R. Linking species concepts to natural product discovery in the post-genomic era. *J Ind Microbiol Biotechnol* **37**, 219–224, doi:10.1007/s10295-009-0683-z (2010).
21. Maldonado, L. A. *et al.* *Salinispora arenicola* gen. nov., sp. nov. and *Salinispora tropica* sp. nov., obligate marine actinomycetes belonging to the family Micromonosporaceae. *Int. J. Syst. Evol. Microbiol.* **55**, 1759–1766, doi:10.1099/ijso.0.63625-0 (2005).
22. Ahmed, L. *et al.* *Salinispora pacifica* sp. nov., an actinomycete from marine sediments. *Antonie Van Leeuwenhoek* **103**, 1069–1078, doi:10.1007/s10482-013-9886-4 (2013).
23. Vidgen, M. E., Hooper, J. N. A. & Fuerst, J. A. Diversity and distribution of the bioactive actinobacterial genus *Salinispora* from sponges along the Great Barrier Reef. *Antonie Van Leeuwenhoek* **101**, 603–618, doi:10.1007/s10482-011-9676-9 (2012).
24. Freil, K. C., Edlund, A. & Jensen, P. R. Microdiversity and evidence for high dispersal rates in the marine actinomycete ‘*Salinispora pacifica*’. *Environ Microbiol* **14**, 480–493, doi:10.1111/j.1462-2920.2011.02641.x (2012).
25. Jensen, P. R., Moore, B. S. & Fenical, W. The marine actinomycete genus *Salinispora*: a model organism for secondary metabolite discovery. *Nat Prod Rep* **32**, 738–751, doi:10.1039/c4np00167b (2015).
26. Jensen, P. R., Williams, P. G., Oh, D. C., Zeigler, L. & Fenical, W. Species-specific secondary metabolite production in marine actinomycetes of the genus *Salinispora*. *Appl Environ Microbiol* **73**, 1146–1152, doi:10.1128/AEM.01891-06 (2007).
27. Ziemert, N. *et al.* Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc Natl Acad Sci* **111**, E1130–E1139, doi:10.1073/pnas.1324161111 (2014).
28. Patin, N. V., Duncan, K. R., Dorrestein, P. C. & Jensen, P. R. Competitive strategies differentiate closely related species of marine actinobacteria. *ISME J* **10**, 478–490, doi:10.1038/ismej.2015.128 (2015).
29. Qin, Q. L. *et al.* Comparative genomics of the marine bacterial genus *Glaciecola* reveals the high degree of genomic diversity and genomic characteristic for cold adaptation. *Environ Microbiol* **16**, 1642–1653, doi:10.1111/emi.2014.16.issue-6 (2014).
30. Bruen, T. & Bruen, T. PhiPack: PHI test and other tests of recombination. *McGill University, Montreal, Quebec* (2005).
31. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, 1541–1548, doi:10.1093/bioinformatics/btu462 (2014).
32. Kubatko, L. S. & Degnan, J. H. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol* **56**, 17–24, doi:10.1080/10635150601146041 (2007).
33. Varghese, N. J. *et al.* Microbial species delineation using whole genome sequences. *Nucl Acids Res*, doi: 10.1093/nar/gkv1657 (2015).
34. Weber, T. *et al.* antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucl Acids Res* **43**, 237–243, doi:10.1093/nar/gkv437 (2015).
35. Deng, X., Phillippy, A. M., Li, Z., Salzberg, S. L. & Zhang, W. Probing the pan-genome of *Listeria monocytogenes*: new insights into intraspecific niche expansion and genomic diversification. *BMC Genomics* **11**, 1, doi:10.1186/1471-2164-11-500 (2010).
36. Donati, C. *et al.* Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol* **11**, 1, doi:10.1186/gb-2010-11-10-r107 (2010).

37. Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 sequenced *Escherichia coli* genomes. *Microbial Ecol* **60**, 708–720, doi:10.1007/s00248-010-9717-3 (2010).
38. Biller, S. J. *et al.* Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *Scientific Data* **1**, 140034, doi:10.1038/sdata.2014.1034 (2014).
39. Rocha, E. P., Cornet, E. & Michel, B. Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet* **1**, e15, doi:10.1371/journal.pgen.0010015 (2005).
40. Hanage, W. P., Fraser, C. & Spratt, B. G. Fuzzy species among recombinogenic bacteria. *BMC Biol.* **3**, 1, doi:10.1186/1741-7007-3-6 (2005).
41. Majewski, J. & Cohan, F. M. DNA sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics* **153**, 1525–1533 (1999).
42. Tang, X. *et al.* Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem Biol* **10**, 2841–2849, doi:10.1021/acschembio.5b00658 (2015).
43. Nett, M., Ikeda, H. & Moore, B. S. Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat Prod Rep* **26**, 1362–1384, doi:10.1039/b817069j (2009).
44. Kale, A. J., McGlinchey, R. P., Lechner, A. & Moore, B. S. Bacterial self-resistance to the natural proteasome inhibitor salinosporamide A. *ACS Chem Biol* **6**, 1257–1264, doi:10.1021/cb2002544 (2011).
45. Jensen, P. Natural products and the gene cluster revolution. *Trends Microbiol.* doi: 10.1026/j.tim.2106.07.006 (2016).
46. Wattam, A. R. *et al.* PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* gkt1099 (2013).
47. Li, L., Stoekert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178–2189, doi:10.1101/gr.1224503 (2003).
48. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681, doi:10.1534/genetics.105.048975 (2006).
49. Martin, D. P. *et al.* RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462–2463, doi:10.1093/bioinformatics/btq467 (2010).
50. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690, doi:10.1093/bioinformatics/btl446 (2006).
51. Miller, M. A., Pfeiffer, W. & Schwartz, T. In *Gateway Computing Environments Workshop (GCE)* 1–8 (IEEE).
52. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504, doi:10.1101/gr.1239303 (2003).

Acknowledgements

This research was supported by the National Science Foundation (OCE-1235142) and the National Institutes of Health (2U19TW007401 and 5R01GM085770). JU was supported by a Conicyt Grant (Fondecyt Iniciación 11140666) and a research grant from Amazon Web Services. NM-A acknowledges a graduate fellowship from Consejo Nacional de Ciencia y Tecnología (CONACyT-213497). Susana Gaudêncio (REQUIMTE, LAQV) and the Portuguese funding agency FCT/MEC (grant PTDC/QUI-QUI/119116/2010 and IF/00700/2014) are acknowledged for support of sample acquisition from the Madeira Islands, PT. Genome sequencing was conducted by the U.S. Department of Energy Joint Genome Institute and supported by the Office Of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank U. Hentschel and U. Abdelomohsen for kindly providing strain CNY-646 and W. Aalbersberg and J. Ginigini for strains CNY-230, CNY-234, CNY-256, CNY-282, and CNY-342.

Author Contributions

N.M.-A. designed and performed the majority of the research, analyzed the data, and drafted the initial manuscript, K.L.C. assisted with generating the genome sequences and the bioinformatic analyses, J.A.U. wrote the scripts and assisted with the bioinformatics analyses, A.-C.L. generated the secondary metabolite gene cluster annotations, G.W.R. assisted with the phylogenetic analyses, P.R.J. helped design the experiments, analyze the data, and write the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-02845-3

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017

Appendix I is compilation of publications I have participated in writing and is in full, a reprint of the material as it appears in their respective journals:

Millán-Aguiñaga N, KL Chavarría, JA Ugalde, A-C Letzel, GW Rouse, PR Jensen. Phylogenomic Insight into *Salinispora* (Bacteria, Actinobacteria) Species Designations. *Nature Scientific Reports* 7, 3564. 2017.

Ziemert N, A Lechner, M Wietz, N Millán-Aguiñaga, KL Chavarría, PR Jensen. Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proceedings of the National Academy of Sciences* 111, E1130-E1139. 2014.

Jensen PR, KL Chavarría, W Fenical, BS Moore, N Ziemert. Challenges and triumphs to genomics-based natural product discovery. *Journal of Industrial Microbiology & Biotechnology* 41:203-209. 2014.