Decoding the Mechanisms of Disease-Suppressive Soil Microbial Communities

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Stephen Clifford Heinsch

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Prof. Michael J. Smanski

February 2024

Acknowledgments

I cannot begin to describe how grateful I am to my advisor, Mike Smanski. As a mentor, Mike showed me how to stay focused on what is important, how to fail fast and move in a new direction, and how to critically evaluate my own ideas. Mike was consistent, kind, humble, and taught me what it means to be a great leader.

I would also like to thank Linda Kinkel. Linda showed me nothing but kindness and respect when we had the chance to interact during my graduate studies. Linda has an infectious enthusiasm about research, and has true dedication to her students.

Thank you to Dr. Larry Wackett, Dr. Erin Carlson, and Dr. Mike Sadowsky, who along with Dr. Smanski and Dr. Kinkel served on my committee.

I would not have made it here without the support of many people. Thank you to Robert Barni, my first college-level biology instructor who encouraged me to study molecular biology. Thank you to Dr. Claudia Schmidt-Dannert and Dr. Jeff Gralnick, you opened your labs to me, and gave me the opportunity to start doing research. And to the past members of those labs who helped me think more critically about the science, especially James Ellinger, Mark Held, Sarah Bloch, and Aunica Kane; thank you.

Thank you to all of the members of the Smanski lab. I have loved getting to know you, and appreciate all of the support you have given me. I want to specially thank Suzie Hsu, Maciej Maselko, Chris Stach,

Dimitri Perusse, Adam Sychla, and Maxime Boneza. You have each played a major part throughout my studies in different ways, providing insight, deep scientific discussion, opportunities for mentoring, and great distractions when we needed them.

This work is dedicated to all those who have supported me.

To my partner Naomi, and our children Sebastian and Fiona, for putting up with long hours, missed weekends, and providing a beacon to always navigate back to what is most important.

To my parents, Cliff and Pat, for encouraging me to stay curious, and to never stop exploring my interests.

Abstract

As the global population is expected to reach 9.4 billion by 2050, food demand is expected to require a doubling of food production. One of the main obstacles in achieving this goal is the prevalence of diseases caused by plant pests and pathogens, which are responsible for approximately 30% of pre-harvest crop loss. This study seeks to learn the design rules for engineering disease-suppressive soils (DSS), soils in which plants thrive despite the presence of pathogens. Microbial communities and their interactions contribute to the establishment, pathogen-suppressive properties, and maintenance of DSS. This work focuses on members of the bacterial genus *Streptomyces*. *Streptomyces* is well-known as a source of numerous antibiotics, and has previously been implicated in DSS.

We examine the genomes of three distinct *Streptomyces* isolates from unique soil environments, including DSS. The isolates were selected for their exquisite inhibitory effects, and ability to interact with a wide range of *Streptomyces* species. Interestingly, comparative genomic analysis does not reflect their phenotypic distinctiveness, suggesting that their unique characteristics may arise from specific gene expression responses to environmental stimuli, including other microbial taxa.

We explore how gene expression changes in *Streptomyces* communities of varying complexities. We measure genome-wide transcriptional changes, with a focus on biosynthetic gene clusters, the source of small molecules like antibiotics. A key finding is the potential role of iron in influencing microbial community interactions, shedding new light on the complex dynamics within DSS.

This work includes a study focused on simulation modeling techniques to optimize gene expression within engineered multi-gene systems. This approach, while initially inspired by small molecule titer improvement, could be extended to the design of microbial consortia for environmental and agricultural applications.

The dissertation concludes by integrating these findings within the context of soil microbial ecology and the potential for engineering microbial consortia. Finally, we propose promising threads to follow as future research directions.

Contents

Listing of figures

List of Tables

Chapter 1

Introduction

Existential threats, such as global food crises and climate change, are poised to be the most significant challenges confronting humanity in this century. It is projected that by 2050, the global population will reach 9.4 billion people. Concurrently, the global demand for food is expected to double by 2050[1], necessitating a 30-70% increase in crop yields to meet this escalating demand[2,3]. Achieving higher crop yields is a complex issue with multiple influencing factors. A primary factor that hinders achieving higher yields is the diseases caused by plant pests and pathogens. Similar to humans, plants are vulnerable to diseases induced by various pathogens, including bacteria, fungi, and other pests[4]. Roughly 30% of crops are lost before harvest due to these pests and pathogens[4,5]. Management strategies predominantly involve the use of chemical pesticides. While these pesticides can help control pests, they come with significant downsides: they increase production costs, do not always result in yield improvements, and pose substantial risks to the environment[6,7]. Therefore, innovative strategies to prevent crop loss are essential to sustainably feed the world's growing population.

Nature has already devised ways to combat plant pathogens, one of which is through disease-suppressive soils (DSS). These are soils in which plants prosper, despite the presence of pathogens. The study of DSS dates back to the late 1800s[8], but it wasn't until the 1930s[9] that researchers began attributing some of the properties of these soils to their microbial inhabitants. Today, it is widely understood that beneficial

taxa in the soil microbiome are the primary contributors to the anti-pathogen properties of DSS.

One significant bacterial family involved in DSS is *Streptomyces*[10]. These bacteria are known for producing a variety of biologically active compounds with a range of uses including antibacterial, anti-fungal, nematicidal, as well as anti-cancer. In fact, *Streptomyces* are the source of over two-thirds of the antibiotics that are currently used in clinical settings[11]. It is therefore likely that a major part of the disease-suppressive qualities of DSS can be attributed to the antibiotics produced by these beneficial bacteria.

The aim of this thesis is to build upon build upon what is known regarding soil microbial communities, with the purpose of one day being able to leverage these findings towards engineering of disease-suppressive microbial consortia inoculants for agriculture. There are several examples of engineered microbiomes, or communities in the literature. Engineered consortia have been utilized in a distribution of labor scheme[12,13] in which separate modules of metabolic pathways reside in different members of the consortia, between which intermediates for the final product are shared. However, these are limited to pathways of limited complexity. Interactions in soil microbial communities are immensely complex, containing examples of both competition, and cooperation. Co-inoculation of microbes either on seed or root has been demonstrated to be useful in improving nitrogen uptake[14], phosphorus solubilization[15] and uptake[16], reducing disease incidence[17,18], and improving plant yield[14,15,18,19]. While these are promising examples showing desired crop outcomes as a function executed by consortia, and have been utilized in commercial agriculture[20], these are transient effects. One of the most exciting features of DSS that have developed in response to a pathogen outbreak is the capacity of the pathogen inhibiting microbial taxa to stay resident and respond to new typically disease-inducing pressure events from the same pathogen[21].

If we hope to engineer soil microbiomes in a bottom-up, approach we must develop a deeper understanding of how all these diverse modes of interaction come together to build consortia with defined output functions, coupled with robustness to the unknown variables encountered in soil. With recent ad-

vances in DNA sequencing[22], proteomics[23], metabolomics[24], and computational frameworks[25,26]; we are amassing the foundational technologies required to address the scale necessary to begin learning the design rules of microbiome engineering[27]. To this end we apply genomics, transcriptomics, and simulaton methods. The scope of this thesis work is focused on a set of microbes sourced from unique ecosystems, with a subset of these microbes having been shown to engage with each other in communication networks ranging in degree of complexity[28].

Leaning on what is known regarding *Streptomyces* genomes and microbial signaling, in Chapter 1 we investigate the genomes of three streptomycetes isolated from unique soil ecosystems including disease-suppressive soils. Each of these isolates have been featured due to their exquisite inhibitory phenotypes, or their ability to signal a high number of other *Streptomyces* isolates. We learn that although they stand out phenotypically among many other isolates, this uniqueness is not not easily seen through comparative genomics. While not explicitly stated in Chapter 1, this finding suggests that it may be changes in gene expression in response to their environment that lead to these unique phenotypes.

Chapter 2, the major component of my thesis work, aims to determine how gene expression changes in a set of streptomycetes, thought to have co-evolved and known to engage in communication, at different levels of community complexity. The work provides general insights into how transcription changes across the genome in response to community participation. We also explore how gene expression changes within biosynthetic gene clusters, groups of genes that work together to produce small molecules including antibiotics. Narrowing down on changes in expression of primary metabolic genes, we find clues to the possibility of iron acting as a modulator of community interactions. Finally, we set the stage for future iterations of transcriptomics experiments targeting microbial communities by detailing best practices to avoid some of the limitations to our experiment.

Chapter 3 explores simulation modeling strategies for optimization of multi-gene expression experiments. The focus of this study is on increasing titers of small molecules produced by engineered biosynthetic gene clusters. It is inspired by the goal of optimizing production, rather than discovery, of valu-

able small molecules. However, while not discussed in Chapter 3, this approach could also be useful in the design of microbial consortia in two ways. First, given that microbial communities are shaped, in part, by these small molecule metabolites, it follows that the tuning of their production may be a useful lever when engineering microbial consortia. Second, the simulation framework could also be used to explore the combinatorial space of naturally-occurring isolates, at different abundance ratios, and their performance on a measurable objective function (e.g. plant health, toxin remediation, etc.). We find that the search algorithms compared are sensitive to landscape ruggedness, as well as breadth and size of sampling when measuring the dependent variable.

Chapter 4, the concluding chapter, attempts to fit this body of work within the field of soil microbial communities, communication within microbial consortia, and their engineering. I provide thoughts on how this work could be expanded upon as well as an outlook on the field in general.

Chapter 2

Complete genome sequences of Streptomyces spp. isolated from disease-suppressive soils

The following is a reprint of the article Heinsch, S. C., Hsu, S. Y., Otto-Hanson, L., Kinkel, L., & Smanski, M. J. (2019). Complete genome sequences of *Streptomyces spp.* isolated from disease-suppressive soils. *BMC genomics*, 20(1), 1-13.

Article hyperlink:

https://doi.org/10.1186/s12864-019-6279-8

LK and MJS conceived the study. SCH, SH, and MJS isolated genomic DNA, performed data processing, genome assembly, and computational analyses. LH and LK performed strain isolation, cultivation, and phenotypic assays. SCH, LK, and MJS wrote the manuscript. All authors read and approved the final manuscript.

2.1 Summary

Bacteria within the genus *Streptomyces* remain a major source of new natural product discovery and as soil inoculants in agriculture where they promote plant growth and protect from disease. Recently, *Streptomyces spp.* have been implicated as important members of naturally disease-suppressive soils. To shine more light on the ecology and evolution of disease-suppressive microbial communities, we have

sequenced the genome of three *Streptomyces* strains isolated from disease-suppressive soils and compared them to previously sequenced isolates. Strains selected for sequencing had previously showed strong phenotypes in competition or signaling assays. Here we present the de novo sequencing of three strains of the genus *Streptomyces* isolated from disease-suppressive soils to produce high-quality complete genomes. *Streptomyces sp.* GS93–23, *Streptomyces sp.* 3211–3, and *Streptomyces sp.* S3–4 were found to have linear chromosomes of 8.24 Mb, 8.23 Mb, and greater than 7.5 Mb, respectively. In addition, two of the strains were found to have large, linear plasmids. Each strain harbors between 26 and 38 natural product biosynthetic gene clusters, on par with previously sequenced *Streptomyces spp.* We compared these newly sequenced genomes with those of previously sequenced organisms. We see substantial natural product biosynthetic diversity between closely related strains, with the gain/loss of episomal DNA elements being a primary driver of genome evolution. Long read sequencing data facilitates large contig assembly for high-GC *Streptomyces* genomes. While the sample number is too small for a definitive conclusion, we do not see evidence that disease suppressive soil isolates are particularly privileged in terms of numbers of biosynthetic gene clusters. The strong sequence similarity between GS93–23 and previously isolated *Streptomyces lydicus* suggests that species recruitment may contribute to the evolution of disease-suppressive microbial communities.

2.2 Introduction

Roughly one third of pre-harvest crops are lost each year worldwide due to agricultural pests and disease[5]. Ninety percent of the 2000 major diseases of the 31 principle crops in the US are caused by soil-borne pathogens[29,30], and soil microbial communities can have a protective effect[31]. Crops are particularly susceptible to disease during their establishment period and when introduced into a new geographic location[32,33]. With the predicted changes in agricultural land use that will accompany climate change or a shift towards crops that support biofuel production, it is important to develop innovative

approaches to combat crop losses to disease.

Natural and agricultural disease-suppressive soils (DSSs) have been identified that provide long-lasting and stable protection against numerous bacterial and fungal pathogens[34]. In addition to preventing crop loss, DSSs can lower the cost of production by removing the need for pesticide application. They have been reported against many major crop pathogens, including wheat take-all disease, potato scab, and wilt on melon[35–39]. Disease-suppression is correlated with increased antagonistic or competitive capacities in one or more isolates from the soil microbial community, and this behavior can emerge in a soil following long-term monoculture[34,40–43]. However, long-term monoculture is not an attractive management strategy to create DSSs, as it generally takes a decade or more for DSSs to emerge and there would be increased plant losses in the short-term. A better understanding of the composition and ecology of DSSs will facilitate engineering soil communities for crop protection.

Recent investigations into the mechanisms of disease suppression, including metagenomic analyses of DSSs[10,34] and phenotypic characterization of microbial isolates[44,45], point to the importance of natural product biosynthesis within a few privileged microbial taxa. Not only are known natural product producers, *Actinomycetes* and *Pseudomonads*, enriched in DSS samples, but interruption of natural product biosynthesis genes interferes with disease-suppression[10]. Further, ecological models that describe the emergence and maintenance of DSSs propose a link between plant biodiversity and the evolution of DSSs. In soils supporting diverse plant species, root exudates and decomposing biomass supply diverse nutrients to soil microbes, which can evolve to co-exist via niche-differentiation. However, in long-term mono-species plant plots, the abundant but non-diverse plant nutrients create a competitive soil environment that favors the evolution of antagonism through antibiosis[34].

Because the metagenomics, phenotypic, and theoretical work all point to the importance of natural products in the formation and maintenance of DSSs, we have sought to better understand natural product biosynthesis in these communities. The observation that isolates from DSSs are more likely to produce antibiotics that target sympatric isolates[46] [20] supports several alternative hypotheses surround-

ing natural product biosynthesis. Highly antagonistic microbial strains should either (i) encode more natural product biosynthetic gene clusters (BGCs) in their genomes than isolates from non-suppressive soils, (ii) encode the same number but actively express a greater percentage of their BGCs, or (iii) produce the same number of natural products, but these compounds are enriched in the biological activities that are important for the formation of DSSs. The first hypothesis is directly testable through whole genome sequencing and comparison.

Here we present the first genome sequences for *Streptomyces spp.* isolated from DSSs. Genomes were sequenced with both long-read PacBio and short-read Illumina technology to produce high-quality and nearly complete sequences for each strain. Bioinformatic analyses highlight the importance of natural product biosynthesis in these isolates, and comparative genomics provides insight to the evolution and ecology of DSSs.

2.3 Results

2.3.1 Isolation and phenotypic characterization of strains

Each of the strains sequenced for this study were selected because (i) they were isolated from soils with measurable disease-suppressive characteristics, and (ii) they displayed strong phenotypes in competition or signaling assays.

Streptomyces sp. GS93–23 was isolated from a potato scab-suppressive plot in Grand Rapids, MN using the Anderson Air Sampler isolation method[47,48]. This strain performed the best of 800 isolated strains at combating potato scab[47]. GS93–23 also shows antifungal activity against Phytophthora medicaginis and Phytophthora sojae, two fungal pathogens of alfalfa. This activity extended to soil studies, where GS93–23 protected alfalfa, reducing the percentage of dead plants from 50 to 0% when pathogens were seeded at low density[49]. Further, compared to no-treatment controls, GS93–23 increased plant growth and yield (forage weight per pot), suggesting direct or indirect plant growth promotion activ-

ity. Lastly, GS93–23 was found to be strongly antagonistic against other *Streptomyces spp.*, but did not reduce nodule production by rhizobial bacteria[49].

Streptomyces spp. S3–4 and 3211–3 were isolated from pathogen suppressive soils located in the Cedar Creek Ecosystem Science Reserve (CCESR), an NSF long-term ecological research site[50]. S3–4 was isolated from soil in a long-term big bluestem (*Andropogon gerardii*) monoculture plot and is antagonistic against sympatrically evolved soil isolates[51]. Strain 3211–3 was isolated from a native prairie control plot at CCESR. It has a strong signaling phenotype, defined as the ability to elicit antibiotic/antifungal production in strains with which it is cultured on close spatial proximity[52].

2.3.2 PacBio sequencing and assembly of genomes

Initial genome sequencing and scaffold assembly was performed on a Pacific Biosciences (PacBio) RS single molecule sequencer (October 2014). Genomic DNA was size-selected using Blue-Pippen 20kb and sequenced in three SMRTcells each. The first two SMRTcells for each genome were run using P4 chemistry, and third SMRTcell was run for each genome with P6 chemistry. Initial read assembly using the PacBio HGAP2 algorithm and sequence polishing using the PacBio Resequencing algorithm produced genome sizes of and contig numbers shown in Table 2.1. Final coverage was >100x for each genome.

Table 2.1: Comparison of general chromosome characteristics

	GS93-23	3211-3	S3-4
Assembled genome size (bp)	8,243,179	8,991,292	8,056,350
Chromosome size (bp)	8,243,179	8,232,231	>7,504,752
Chromosome topology	Linear	Linear	Linear
Chromosome G+C content	72%	71%	73%
rRNA operons	7	7	8
tRNA genes	66	77	73
Protein-coding genes G+C content	7188	8087	7071
Natural product BGCs	26	38	28

The high GC-content of *Streptomyces* genomes produces many homopolymer G and C stretches,

which can produce errors during base-calling and genome assembly. Low-coverage Illumina sequence data was collected for error correction. Illumina sequencing was performed on a Mi-seq instrument to collect 2 × 250 base paired end reads equating to 110-fold (3211–3), 118-fold (GS93–23), or 155-fold (S3–4) coverage for each genome. Final, error-corrected genome sequences were generated by mapping Illumina short reads to PacBio-generated reference genomes using the BreSeq algorithm[53], and incorporating single nucleotide polymorphisms (SNPs) and short Indels using the Pilon algorithm[54] .

2.3.3 Comparison of Illumina-corrected and PacBio-alone genome sequences

The short-read corrected genome sequences were compared to the PacBio-only assemblies, and 70, 295, and 335 SNP/Indels were present between the two assemblies for GS93–23, S3–4, and 3211–3, respectively. In each case, the vast majority were single base insertions in homopolymer stretches. We next sought to verify that the short-read corrected sequences were indeed a better representation of the actual genome sequence, as the two sequencing platforms are known to generate different types of errors. To determine which sequence variant was correct for each SNP/indel, translated protein sequences at each of the 295 SNP/indel loci in the S3–4 genome were compared against the NCBI GenBank non-redundant database, with the assumption that a frameshift resulting from an indel will result in a worse top blast hit for a stretch of DNA. Figure A.1 shows the comparison of significance score for BLASTx results of searching a fragment of DNA +/– 150 bases from the variant loci. This analysis is only expected to reveal the correct sequence variant when (i) the indel is present within a coding DNA sequence (CDS), (ii) correct protein sequences for close homologs are present in GenBank, and (iii) the 300 base-pair window that is searched is sufficiently focused such that top BLAST hits align to the translated query in the region of the variant locus (i.e. at the center of the query, not the edges). We find that the Illumina-corrected sequence returns a top BLASTx hit with lower (better) E-value twice as often as the uncorrected sequence. The average E-values for the top BLASTx hit alignment are six orders of magnitude lower (better) for the short-read corrected sequences compared to the PacBio-only sequences.

Because of this, we use the short-read corrected genome sequences for the remaining analyses.

2.3.4 General characteristics of the genome sequences

We were able to assemble the chromosome as a single large contig for strains GS93–23 (8.24 Mb) and 3211–3 (8.23 Mb), and as two large contigs for S3–4 (4.19 Mb and 3.31 Mb) (Figure 2.1 and Table 2.1). For S3–4, the two chromosome arms can be oriented relative to one other with high confidence based on GC-skew, orientation of rRNA operons, and enrichment of specialized metabolite gene clusters at chromosome arms (Figure 2.1, rings 8, 6, and 4, respectively). Manual attempts to close the gap by retrieving PacBio reads that mapped to each contig were unsuccessful. The gap is present in a locus that is especially repetitive, with 3 rRNA operons in close proximity. The overall G+C content (71–73%) and differences in G/C skew for the chromosome arms in each genome are similar to what has been reported for other genomes from this genus[55–60]. In addition to the large linear chromosomes, strains 3211–3 and S3–4 each contain two large linear plasmids (519 Kb and 240 Kb for 3211–3, 349 Kb and 203 Kb for S3–4).

Annotation of the genomes with the Prokka software tool[61] identified 7188 CDSs, 7 ribosomal RNA operons, and 66 tRNAs for GS93–23. Similar numbers of annotated genes were present in the S3–4 genome (7071 CDSs, 8 rRNA operons, 73 tRNAs), and slightly more in the 3211–3 genome (8087 CDSs, 7 rRNA operons, 77 tRNAs), accounting for its larger total size. Gene products were assigned to Clusters of Orthologous Groups (COGs) using the BASys platform[62]. Functional categorization of proteins reported in Table 2.2 in comparison to the model organism, S. coelicolor A3 (2) were performed with EggNOG-mapper[63].

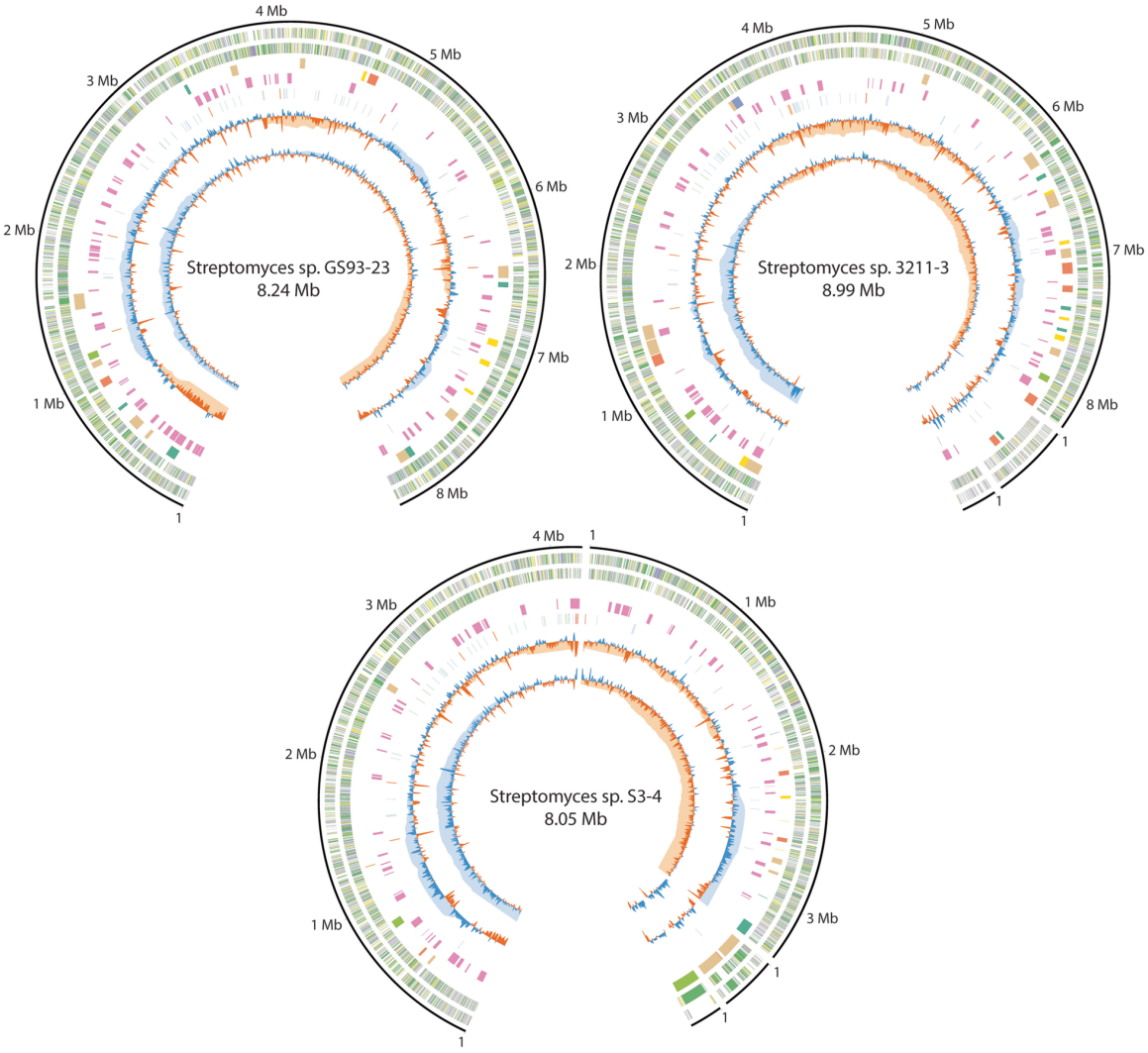


Figure 2.1: Schematic representation of genome sequences for strains GS93–23, 3211–3, and S3–4. Outer, solid black ring shows contig length in Mb. Second and third rings show annotated CDSs in the forward or reverse orientation, respectively, colored by functional classification. Genes involved in metabolism are green, information storage and processing are purple, cellular processes and signaling are yellow, and unknown functions are grey (see Table 2.2. Fourth and fifth rings show high-confidence and putative natural product BGCs, respectively. High-confidence BGCs are colored by biosynthetic class, with polyketides light green, non-ribosomal peptides orange, terpenes yellow, nucleosides purple, RIPPs dark green, and hybrid clusters tan. Sixth ring shows functional RNA elements, including rRNA (reverse orientation orange, forward orientation red) and tRNAs (reverse orientation blue, forward orientation green). Seventh and eighth rings show G+C content and G+C skew, respectively. Each is shown for two window sizes: 10 Kb (dark blue above average, dark orange below average) and 1 Mb (light blue above average, light orange below average). Only the 10 Kb resolution data is shown for plasmids.

Table 2.2: COG functional categories

COG	GS93-23		3211-3		S3-4		*S. coelicolor*	
	%	Num.	%	Num.	%	Num.	%	Num.
Cellular processes and signaling								
Cell division and cytoskeleton	0.5	38	0.5	40	0.5	38	0.5	38
Defense mechanisms	1.4	106	1.3	112	1.2	90	1.4	115
Signal transduction mechanisms	4.5	335	4.5	394	4.4	328	4.6	385
Cell wall/membrane/envelope biogenesis	2.9	217	2.6	228	2.8	209	2.8	235
Secretion	0.5	34	0.5	40	0.5	34	0.5	43
Posttranslational modification	2.0	151	2.1	186	2.1	154	1.9	158
Information storage and processing								
Translation, ribosomal structure and biogenesis	2.4	180	2.1	182	2.5	188	2.3	192
Transcription and RNA processing	9.4	708	7.6	666	8.1	597	9.4	786
Replication, recombination and repair	2.7	204	6.3	551	4.0	295	3.8	318
Metabolism								
Energy production and conversion	4.7	353	3.7	330	4.3	316	4.5	374
Carbohydrate transport and metabolism	5.2	392	3.7	325	4.2	308	6.1	509
Amino acid transport and metabolism	5.9	439	4.5	393	5.1	376	4.7	395
Nucleotide transport and metabolism	1.6	120	1.2	106	1.4	106	1.2	103
Coenzyme transport and metabolism	2.0	147	1.7	148	2.0	146	1.7	143
Lipid transport and metabolism	2.8	207	2.7	240	2.7	199	2.4	199
Inorganic ion transport and metabolism	3.5	261	3.2	280	3.2	237	4.0	335
Secondary metabolism	2.5	189	2.2	194	2.8	209	2.0	168
Poorly characterized								
Function unknown	30.9	2314	30.1	2658	32.2	2386	30.8	2564
No COG in database	14.7	1102	19.8	1743	16.2	1198	15.2	1265

2.3.5 Annotation of natural product biosynthetic gene clusters

Because natural product biosynthesis is thought to play a mechanistic role that underpins the ecology of disease suppressive soils[10,64], we have analyzed the genomes for their biosynthetic potential using the antiSMASH 3.0 toolkit[65]. We conservatively assigned specific molecules to these BGCs only when the annotated gene clusters share 100% of the biosynthetic genes from previously characterized BGCs by manual comparison (Additional file Information). For ribosomally produced and post-translationally modified peptides (RiPPs), we predict the production of minor structural variants when the sequence of precursor peptides is slightly different than in characterized BGCs. The 26 high-confidence BGCs identified in the GS93–23 genome include known pathways for RiPP cyclothiazomycin[66], the dienoyltetramic acid streptolydigin[67], and the lipoglycopeptide mannopeptimycin[68]. The 38 high-confidence BGCs in the 3211–3 genome include known pathways for the chlorinated non-ribosomal peptide tambromycin[69], the siderophore coelichelin[70], and terpenoid 2-methylisoborneol[71]. The 28 high-confidence BGCs in the S3–4 genome include known pathways for 2-methylisoborneol, and the aminoglycoside streptothricin[72]. In addition, all three genomes contain the highly conserved BGCs for the siderophore desferrioxamine b[73], terpenes geosmin[74] and hopene[75], minor structural variants of lantibiotic SapB[76], and osmoprotectant ectoine[77].

The majority of BGCs identified in these genomes remain uncharacterized. Intriguing pathways include a 178 Kb polyketide cluster on a plasmid in S3–4 that putatively encodes a 60-member macrolide, and a pyrrolopyrrole-containing metabolite in 3211–3.

2.3.6 Comparison to closest sequenced relatives

We compared the draft genome sequences to a collection of 500 publicly available actinomycete genomes using multi-locus sequence comparison to identify the closest sequenced relative of each (Figure 2.2). S3–4 groups with the small *Streptomyces katrae* clade near type strain NRRL-ISP 5550[78]. Strain

3211–3 is in the neighboring *Streptomyces virginiae* clade defined by the type strain NRRL ISP-5094[79]. GS93–23 clusters with the *Streptomyces lydicus* type strain NRRL-ISP 5461[80].

We identified closely related genomes in the available whole-genome sequence databases for each of our DSS isolates (Figure 2.3). For each of our newly sequenced strains, a previously published genome was available with high sequence similarity in several common phylogenetic markers (16S rRNA, rpoB, and multi-locus sequencing (MLS) using ribosomal proteins) (Figure 2.3a). Our closest pair of new and previously reported genomes is GS93–23 and *S. lydicus* NRRL ISP-5461, which share 100% identity of 16S rRNA and 99.92% identity using MLS comparison. Even our most divergent pair, S3–4 to *Streptomyces sp.* WM6372, shared >98% identity at the 16S rRNA level and >96% identity at the rpoB level, and 93.72% by four-gene MLS comparison (atpD, gyrB, rpoB, trpB).

Genome pairs were compared to determine the amount of shared sequence across the entire genome (Figure 2.3b). Alignments were constructed in Mauve and alignment gaps were mapped back to the new high-quality reference genomes. Alignment gaps between of GS93–23 and ISP-5461 are uniformly distributed across the chromosome. Insertions or deletion events greater than 100 bp account for only 4.5% of the genome sequence as a whole (Figure 2.3b), with a similar proportion being lost/gained in BGCs as in the rest of the genome (Figure 2.3b).

The high-level of sequence conservation between GS93–23 and ISP-5461 allowed us to examine the micro-scale evolution of these genomes. There are approximately 40,000 SNPs between the two, making the sequence identity in the aligning sequences greater than 99.5%. Interestingly, the position of SNPs relative to CDSs shows a marked de-enrichment in (i) the approximate position of the Shine-Dalgarno sequence in the 5′-UTR, and (ii) the 5′ end of the CDS (Figure 2.3c). This suggests a selection for maintaining relative translation rates of encoded genes, as both loci are important in determining translation initiation rates in bacteria[81]. Most of the 33,000 SNPs in CDSs encode silent mutations. Of the missense mutations, the majority are conservative in terms of amino acid chemistry (Figure 2.3d). The ratio of synonymous to non-synonymous mutations (dS/dN) is 1.8, which is substantially lower than seen in

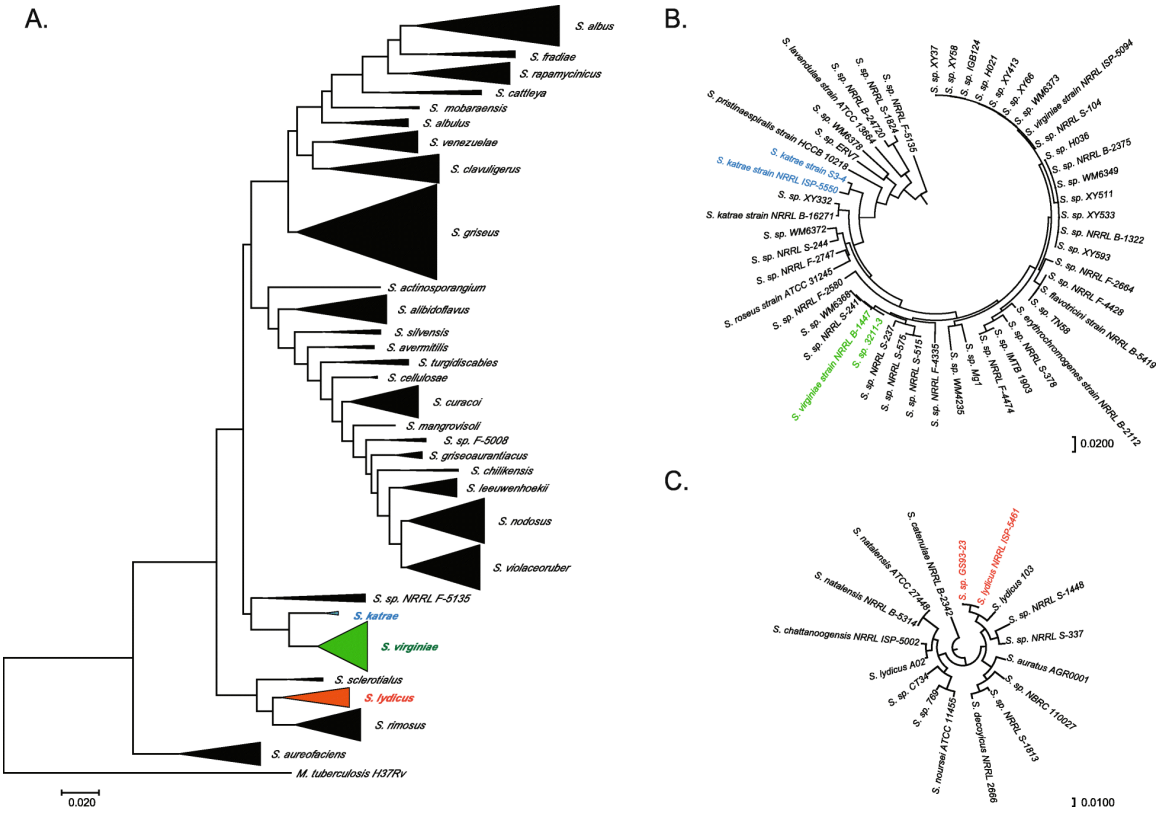


Figure 2.2: Molecular phylogeny of newly sequenced strains. (a) Phylogenetic tree of 496 publicly available *Streptomyces* genomes. Mycobacterium tuberculosis H37Rv was used as outgroup. Select regions of the atpD, gyrB, recA, rpoB, and trpB genes were concatenated and used to generate a multi-locus alignment in the MEGA7 software package. Genetic distances (average nucleotide identity) generated from the multisequence alignment were used to build a phylogenetic tree using the maximum likelihood method. Clades containing the newly sequenced genomes are S. katrae (S3–4, blue), S. virginiae (3211–3, green), and S. lydicus (GS93–23, red). Subtrees composed of S. katrae and S. virginiae (b), and S. lydicus (c) showing the newly sequenced isolates and their closest relatives.

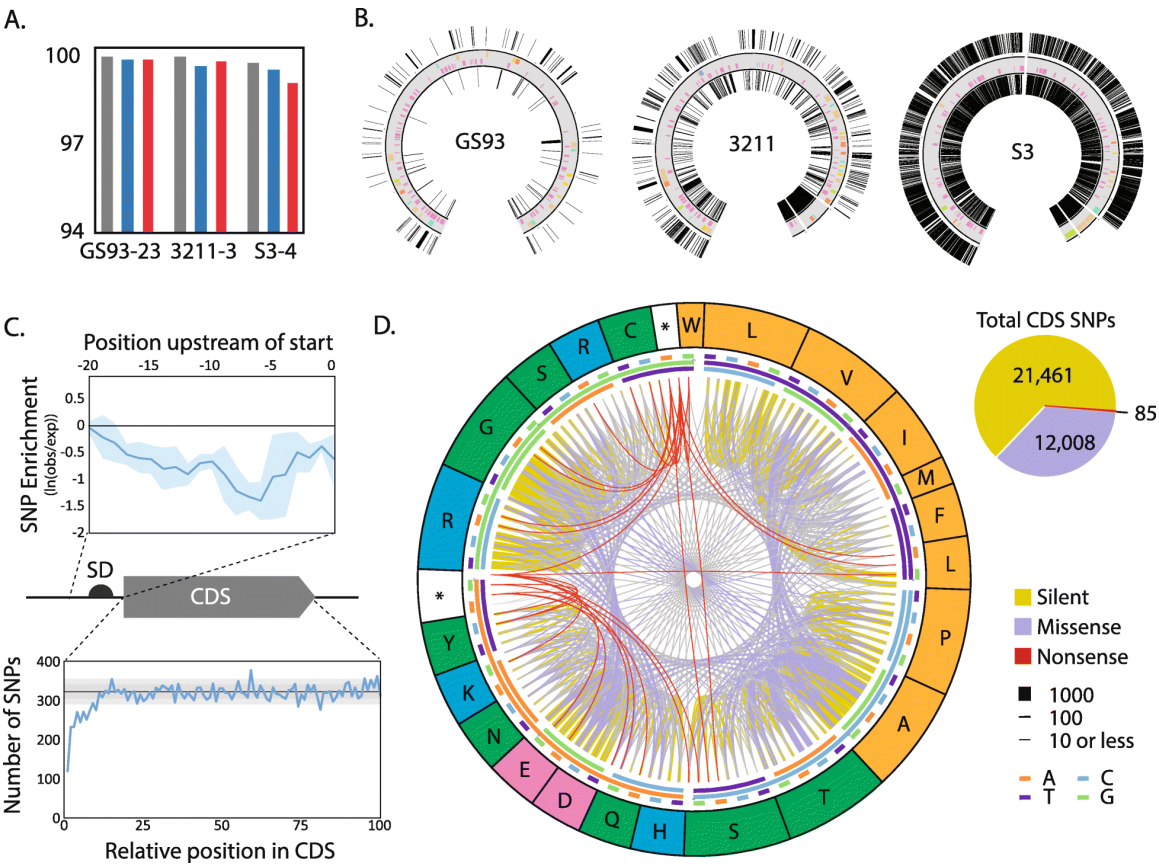


Figure 2.3: Comparative analysis of with closest sequenced relatives. (a) Sequence identity between newly sequenced strains and closest relatives, with 16S rDNA (black), rpoB gene (dark grey) and multilocus sequence comparison (light grey) shown for each pair of strains. (b) Genomic location of alignment gaps larger than 100 bp. Grey ring represents newly sequenced genomes, with high-confidence and putative BGCs labeled as in Figure 2.1 Outer ring shows location of extra sequence present in closest relative but missing in our newly sequenced strain. Inner ring shows location of extra sequence present in newly sequenced strains but missing from closest relative. (c) SNP analysis of strain GS93–23 and its closest relative ISP-5461. Pie-chart in upper right shows relative proportion of silent, missense, and nonsense mutations. Circle chart at left shows frequency of all SNPs found in CDSs, with the outer ring showing one-letter code for amino acids, colored according to chemical property (hydrophobic, orange; hydrophilic, green; basic, blue; acidic, pink). Three-base code is shown using three inner rings, with innermost representing the first codon position and outermost representing the last codon position. For each codon, there are two nodes on the graph. In the clockwise direction, the first node corresponds to a codon in GS93–23 and the second node to ISP-5461. Each CDS SNP is represented by an arc connecting a codon in GS93–23 to a codon in ISP-5461, with the width of the arc indicating number of instances of that mutation. (d) Location of SNPs relative to CDS position. Top line graph shows enrichment of SNPs upstream of start codons using absolute positions, with the solid blue line showing average value for a sliding 3-base window and the light-blue filled region showing one standard deviation in either direction. Bottom line graph shows SNP abundance versus relative position in CDS, where relative position equals absolute position divided by CDS length. Black line and grey boxes show average SNP abundance and 1-, 2-standard deviations as calculated for the last 90% of the CDS. CDS: coding DNA sequence; SD: Shine-Dalgarno sequence.

housekeeping genes in E. coli and invasion genes from S. enterica[82,83], suggesting that there has been little selective pressure against non-synonymous mutations and that these two strains belong to the same clonal complex[84,85].

Despite the strong similarity between GS93–23 and ISP-5461, there are still substantial differences between the two strains. GS93–23 contains 98 genes that are missing in ISP-5461, and ISP-5461 contains 11 unique genes. 66/98 genes unique to GS93–23 are of unknown function. Of genes with functional annotations the largest categories specific to GS93–23 are transcriptional regulators (11/98) and metabolic enzymes (10/98). Of the genes unique to ISP-5461, only a single gene was of unknown function. The largest functional categories for genes unique to ISP-5461 also were transcriptional regulators (3/11) and metabolic enzymes (3/11).

The other two DSS genomes presented here are more divergent from the nearest sequenced relative. Both 3211–3 and S3–4 have two large plasmids that are absent in their closest relatives, S. virginiae NRRL B-1447 and S. katrae NRRL ISP-5550, respectively. These changes alone account for 9 and 7% of the total genome content, respectively. The plasmids in S3–4 are rich in secondary metabolism genes, with four large gene clusters totaling roughly 500 kb of sequence. Besides the plasmid differences, the chromosome of 3211–3 contains 285 large (> 100 bp) insertions compared to B-1447, totaling 609 kb of new sequence, and 309 large deletions totaling 758 kb of sequence lost. In the regions that do align, there are 102,000 SNPs, corresponding to an average sequence identity of 98.7% across the genome. The S3–4 genome lacks a close homolog in the sequence databases. Despite sharing 96.3% sequence identity of the rpoB gene, 26% of the S3–4 genome does not align with the WM6372 sequence.

We next compared the natural product biosynthetic potential for these three strains by analyzing their BGC content. Our closest pair, GS93–23 and ISP-5461, share 26/26 of the high-confidence BGCs and 61/64 'putative' clusters (co-localized clusters of genes that belong to COGs typically found in BGCs, but which lack canonical secondary metabolism signature sequences). The next closest pair, 3211–3 and B-1447, which share 99.7% similarity of the rpoB gene, have in common only 31/38 of the high-

confidence BGC annotations, which is driven mostly by the presence of two plasmids in 3211–3 missing from B-1447. Between S3–4 and WM6372 (96.3% identity of rpoB), 12/28 of high-confidence BGCs are shared, and 27/54 'putative' clusters. These relationships between genetic distance and BGC overlap follow the general trend for rpoB conservation and non-ribosomal peptide synthetase (NRPS) BGC overlap described by Doroghazi et al.[86].

2.3.7 Signaling potential analysis

One possible organization for a highly antagonistic microbial community would have a keystone species that produces a signal to induce antibiotic production in many other community members. The University of Minnesota DSS strain library was assayed for signaling potential using a plate-based phenotypic assay[52] (Kinkel, unpublished data). Strain 3211–3 was selected for whole genome sequencing because it is among the best signalers of antibiosis in our library of DSS isolates. The signaling assay requires dilution of a signaling molecule through solid agar medium, so signaling through cell-cell contact can be ruled out as a mechanism. We looked for genomic features that could explain the signaling promiscuity in 3211–3.

Signaling between *Streptomyces* can be mediated by several well-known classes of hormone-like signaling molecules[87] including γ-butyrolactones[88], furans[89], γ-butenolides[90], SapB[91]-like RiPPs, diamino-bis(hydroxymethyl)-butanediol[92], and diketopiperazines[93]. Signaling can also be mediated by sub-inhibitory concentrations of antibiotics[94–96]. We first looked for the presence of BGCs encoding hormone-like signaling molecules in 3211–3. There are two γ-butyrolactone BGCs in this genome and a SapB BGC, but this number is comparable to other sequenced *Streptomyces*. There is no evidence that 3211–3 produces an unusually diverse set of hormone-like signaling molecules.

A second possibility is that 3211–3 does not produce many diverse hormone-like signaling molecules, but the molecule they do produce can be sensed by many species of *Streptomyces*. There are at least fifteen unique γ-butyrolactone signals produced by the genus, and unfortunately it is not possible to pre-

dict the specific γ-butyrolactone chemical structure from sequence information alone. However, we reasoned γ-butyrolactone biosynthesis genes and receptors that produce/sense the same compound will have a higher degree of sequence similarity than those producing/sensing different compounds (i.e. functionally similar gene clusters would share greater sequence similarity), as this gene cluster does not closely correlate with other phylogeny (Figure 2.4). We performed a CLUSTER-BLAST analysis with the γ-butyrolactone biosynthesis protein ScbA and the receptor AfsR against the set of sequenced *Streptomyces* genomes. Again, we did not see any evidence that 3211–3 produces a more widely-sensed hormone-like signaling molecule.

A third possibility is that 3211–3 is a prolific signaler due to production of sub-inhibitory concentrations of antibiotics (SICA). This genome encodes more 'high-confidence' BGCs than the two genomes from strongly antagonistic DSS isolates (Table 2.1). Among 125 complete *Streptomyces* genomes with antiSMASH 4.1 (Table A.4), the number of high-confidence BGCs in 3211–3 places it in the top 16% in terms of BGC content. Since there is no clear genomic signature that allows us to explain the signaling potential in 3211–3, teasing apart its ability to elicit antibiosis in so many diverse isolates will require future molecular genetic experiments.

2.4 Discussion

Bacteria within the genus *Streptomyces* are ubiquitous in terrestrial soils and marine sediments and have garnered much attention for their ability to produce medicinal natural products. The past decade and a half of genome sequencing efforts[56,86,97] revealed that the majority of natural products encoded in the genomes of *Streptomyces spp.* remain undiscovered and have reinvigorated natural product discovery via genome mining[64,98]. Most genomes deposited in public sequence databases have been sequenced using Illumina short-read technology. The large size, repetitive nature, and high G+C content of *Streptomyces* genomes makes them difficult to fully assemble from short reads, and so roughly 90% of the

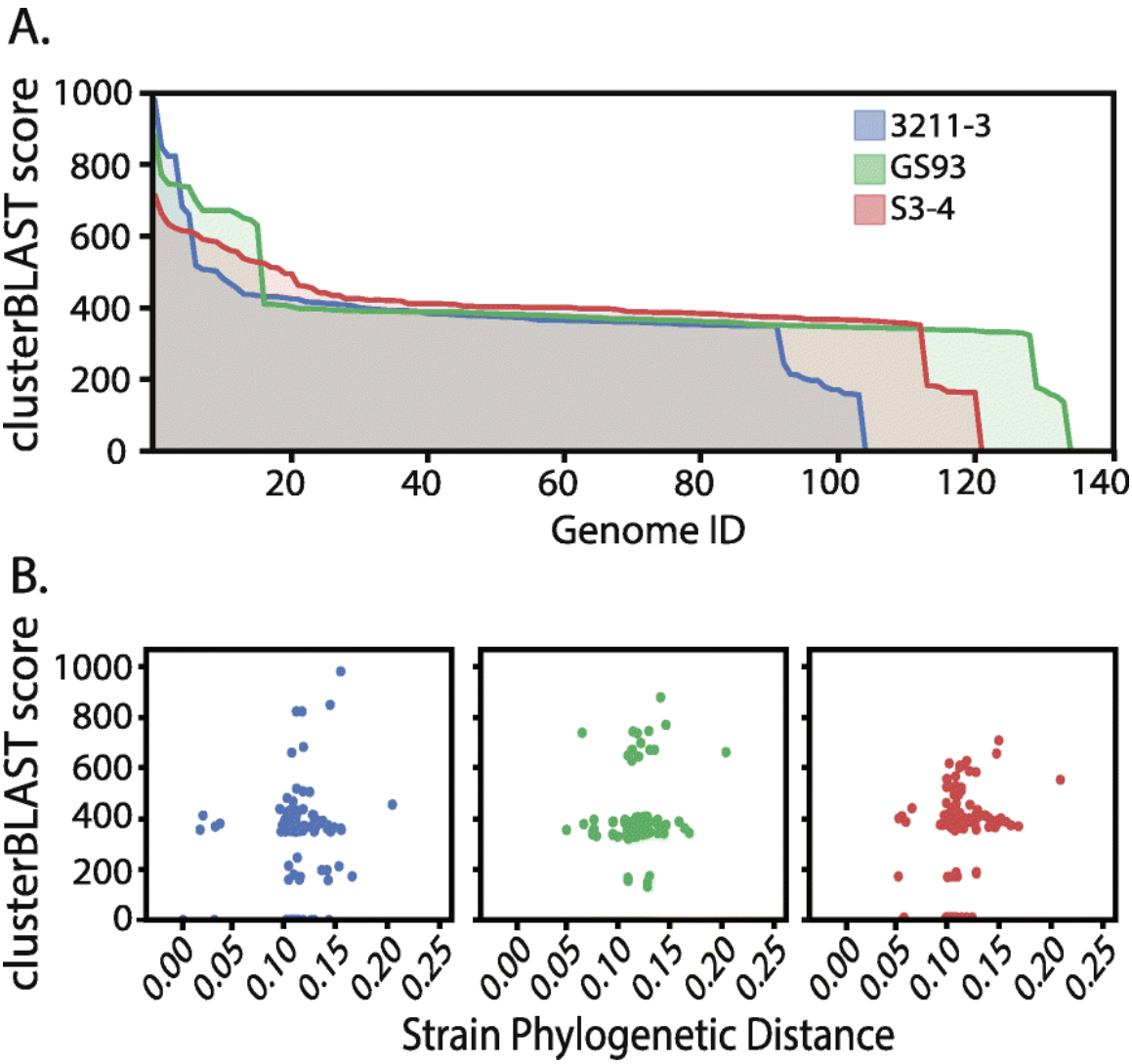


Figure 2.4: Signaling potential analysis of newly sequenced strains. (a) Distribution of homologous gamma-butyrolactone biosynthetic gene clusters throughout 496 *Streptomyces* genomes. Area plots show clusterBLAST scores for 3211–3 (blue), GS93–23 (green), S3–4 (red). Genome ID's for hits have been sorted in order of 3211–3. (b) Relationship between genetic distance and clusterBLAST score. Genetic distances between the three newly sequenced isolates and the genome hits from the signaling potential MultiGeneBLAST analysis genomes were obtained by multilocus alignment of the atpD, gyrB, recA, rpoB, and trpB genes.

available genomes are only available in draft status; typically hundreds of contigs with an average N50 of thousands of bases. With a combination of PacBio and Illumina sequence data, we were able to assemble high-quality genome sequences where the >8 Mb chromosome assembles as a single contig in two strains and as two contigs in the third.

We initially predicted that the increase of genome quality would correspond to an improved ability to identify BGCs that would have been broken up between many small contigs in a short-read only assembly. However, the difference in quality does not appear to effect estimations of natural product biosynthetic potential. For example, in S. lydicus ISP-5461, 26 of the 26 high-confidence BGCs found in GS93–23 were also predicted using the short-read only assembly contigs.

One advantage to generating single-contig genomes using long-read data is the ability to map the chromosomal location of BGCs. In order to help prioritize isolated *Streptomyces* strains for whole-genome sequencing, there have been previous attempts to correlate sequence conservation of phylogenetic markers with BGC conservation between two or more genomes[86]. After sequencing 1000 actinomycete genomes, Metcalf et al. found that a 99% sequence identity between concatenated ribosomal protein sequences correlates with a 73 and 80% conservation of Type I polyketide synthase (PKS) and NRPS clusters, respectively[86]. Our data supports the rapid diversification of secondary metabolite gene clusters, and suggests that this is primarily driven by changes in episomal elements, not by changes to the core genome. This information could make future sequencing campaigns more efficient by limiting sequencing efforts in closely related strains to isolated plasmids.

Bacterial genome organization has been described as mosaic[99–101], referring to the composition of a vertically-inherited (clonally-expanded) backbone interspersed with laterally-transferred mobile elements. Mutations accumulate in clonal complexes between bouts of periodic selection[85]. The genomic comparison of GS93–23 and ISP-5461 suggests that these strains are part of the same clonal complex, despite being isolated 850km apart and several decades removed. Our analysis of the SNP accumulation in relationship to relative location within genes shows a de-enrichment of sequence variation in regions known

to control translation initiation rates. This points to a microevolution of genomes where there is a selection to maintain relative expression levels of genes during clonal expansion. We have previously shown that transfer of multi-gene systems between hosts from the same genus can result in wildly different relative expression levels[102]. These likely result from the accumulation of subtle differences between the transcription/translation machinery and corresponding cis-acting regulatory elements that co-evolve during clonal expansion. Taken together, the importance of maintaining relative expression levels during microevolution and the changes between seemingly closely related species likely contributes to low success rates and low titers during heterologous introduction of BGCs to model host strains[103].

We sequenced the three strains presented here in hopes to gain insight towards the mechanisms and ecology that underlie DSSs. While the sample size is small, there is no indication that the increased antibiosis observed in DSS isolates compared to isolates from non-suppressive soils is due to an increased number of BGCs. Transcriptomic and chemical characterization of these and other DSS isolates is pending. With over 500 species of *Streptomyces* currently recognized[104] and roughly 800 draft *Streptomyces* genomes available in public databases at the time of this study, we were initially surprised by the level of sequence conservation between these strains and previously sequenced genomes. The level of divergence between GS93–23 and ISP-5461 is only ten times greater than clonally-related lab-cultivated strains of E. coli separated by only 20 years of evolution[105]. There are a few possible explanations for this. First, species groups are not expected to be equally abundant. It is likely that the genomes already present in the public databases are those of highly abundant clonal complexes. The similarity between these genomes and extant sequences reflects the fact that no attempts were made to bias our strain selection towards rare *Streptomyces*. A second possibility is that the ecology of DSSs has selected for strains that are also abundant in sequenced collections. This makes sense in light of the experimental data and ecological models that suggest DSSs community members are selected for their antagonistic phenotypes[34]. Likewise, most *Streptomyces* strains whose genomes are in public databases were originally isolated and maintained in collections of drug discovery groups. If this is true, it will suggest that evolution of DSS

isolates occurs on the level of the genome/strain, not the individual genes, contrary to what has been observed in other environments[106]. Strain recruitment is a proposed mechanism of the establishment of disease suppressive soils[107], in which plants support the maintenance of those microbial strains which inhibit phytopathogens. 16S sequencing and denaturing gel electrophoresis of the rhizosphere microbiome of strawberry plants showed that the Actinobacteria community profile was more similar between species of strawberry plant, regardless of site, when compared to oil rape rhizosphere communities[108]. It is not unreasonable, then, to assume that under the dispersal-recruitment model, that ancestral bacterial strains that were beneficial to plant growth would be under similar selective pressures if co-evolving with the same plant species in distant locations.

2.5 Materials and Methods

2.5.1 Preparation of high molecular-weight DNA

The three strains of *Streptomyces* sequenced for this study were obtained from a culture collection maintained by Linda Kinkel at the University of Minnesota. Single colonies are isolated on IWL-4 solid medium and used to inoculate 4 mL liquid cultures in R2YE medium. Following three days of growth, cells are harvested by centrifugation and washed with a 10% sucrose solution. Mycelia are resuspended in 450 µL TSE buffer (15% sucrose, 25 mM Tris, 25 mM EDTA, pH 8) with 5 mg/mL lysozyme and incubated at 37 °C for one hour. Cells are lysed by addition of 225 µL of 2% SDS over a 5 min room temperature incubation. Following a phenol:chloroform extraction (100 µL neutral phenol, 50 µL chloroform), supernatant is transferred to a tube containing 60 µL 3 M sodium acetate and 700 µL isopropanol to precipitate gDNA. DNA is pelleted by centrifugation and resuspended in 500 µL TE buffer (10 mM Tris, 1 mM EDTA, pH 8). To remove RNA, 10 µL RNase (10 mg/ml) is added to the sample and incubated at room temperature for at least 15 min. Next, a second phenol:chloroform extraction (300 µL neutral phenol, 150 µL chloroform) is performed followed by a final extraction with 300 µL chloroform to remove

trace phenol. DNA in the supernatant is precipitated with 50 µL 3 M sodium acetate and 350 µL isopropanol and incubated on ice for 30 min. Final gDNA is resuspended in 150 µL TE buffer and quality is assessed by agarose gel electrophoresis, spectrophotometry, and PicoGreen analysis.

2.5.2 DNA sequencing and assembly

We performed PacBio long-read sequencing using protocols for 20 Kb insert size with BluePippin Size Selection (Saga Science). For each of the three genomic DNA samples, sequencing was performed using P4 chemistry on two SMRT cells and using P6 chemistry on an additional SMRT cell from November 2014 to January 2015. In total, subread filtering from the three SMRT cells yielded 1.26 Gb (S3–4), 1.40 Gb (GS93–23), and 1.18 Gb (3211–3) of sequence data with average read lengths of 6703 kb, 6782 kb, 6478 kb, respectively and N50 values of 9095 kb, 8819 kb, and 8680 kb, respectively.

2.5.3 Short-read sequencing and error correction

Illumina MiSeq sequencing was performed at the UMN Genomics center in March 2015. The three genomic DNA samples were uniquely barcoded and sequenced alongside genomes from unrelated bacteria to account for 30% of a MiSeq lane. Nextera library prep was performed using standard protocols at the University of Minnesota Genomics Center. The 250 nt paired-end reads were mapped to the PacBio-reference genome sequence using Breseq[53] to generate. BAM files. Single-base differences and small indels were corrected using Pilon to generate the final error-corrected genome assembly.

2.5.4 Annotation of genomic features

Prokka[61] is a command line software tool that uses Prodigal[109] for coding DNA sequence (CDS) annotation, RNAmmer[110] for ribosomal RNA annotation, Aragorn[111] for transfer RNA annotation, SignalP[112] for signal leader peptide annotation, and Infernal[113] for non-coding RNA anno-

tation. Each genome was annotated with the Prokka software package using default options and the '–compliant' command to force compliance with GenBank.

Assignment of putative functional categories to CDSs was performed using the BASys[62] web server (https://www.basys.ca/). For each CDS, start position, end position, strand information, and a unique identifier was provided in tabular format to ensure that Prokka-generated annotations would be used for clusters of orthologous genes (COG) assignment in place of the default Glimmer algorithm. The following options were selected for functional assignment by BASys: Gram positive, Linear contig, Bacterial genetic code. Functional assignments of proteins in Table 2.2 were performed with EggNOG-mapper[63]. The following EggNOG-mapper settings were selected: mapping mode was set to DIAMOND[114], taxonomic scope was set to all bacteria, all orthologs were used, and non-electronic gene ontology evidence terms were selected.

2.5.5 Phylogenetic analysis

Streptomyces genomes were obtained from PATRIC (https://www.patricbrc.org/). Nucleotide sequences for molecular phylogeny markers *atpD*, *gyrB*, *recA*, *rpoB*, and *trpB* were extracted. Regions for comparison were identified and concatenated head-to-tail in-frame[115,116]. Multi-sequence alignment of concatenations, and maximum-likelihood tree construction was performed in MEGA7[117]. For the S3–4 subtree phylogeny the *recA* sequence was not available for WM6372 and a four-gene concatenation was used.

2.6 Conclusion

In summary, we have added three high-quality whole genome sequences to the growing number of sequenced *Streptomyces* isolates. Each genome is rich with yet-uncharacterized natural product biosynthetic potential. While genome sequence alone was not sufficient to explain the observed phenotypes of

DSS isolates, it is an important first step to future investigations of gene expression and function.

Supplementary tables and figures can be found in Appendix A.

Chapter 3

Metatranscriptomic analysis of synthetic communities of sympatric *Streptomyces* isolates from disease suppressive soil

The following is a pre-submission version of the article Heinsch, S. C., Song, Z., Kinkel, L. & Smanski, M. J. (2024). Metatranscriptomic analysis of synthetic communities of sympatric *Streptomyces* isolates from disease suppressive soil.

S H, ZS, LK, and MS designed the experiments and performed the analyses. SH and MS wrote the manuscript.

3.1 Summary

Interspecies interactions in soil microbiomes impact the ecology and function of soil in agricultural and natural settings. The result of these interactions, however, remains poorly understood at the level of gene expression. We measured changes in gene expression from ten bacteria isolated from a single sample of disease suppressive soil (DSS). Isolates were grown in 61 different axenic cultures or co-cultures with increasing levels of complexity. We discovered reproducible changes in secondary metabolism spanning across levels of community complexity. These data lay the foundation for a more mechanistic understanding of DSS biology and ecology and their eventual engineering for crop protection.

3.2 Introduction

Healthy soil ecosystems provide crop nourishment, disease suppression, and protection against environmental stress in agriculture[118]. They remediate environmental pollutants in natural and industrial settings[119] and contribute to global geochemical nutrient cycling and carbon sequestration[120]. These beneficial microbial functions are a direct consequence of the collective primary and secondary metabolic activities of the soil community.

Disease-suppressive soils (DSSs) have been identified that mitigate pre-harvest crop loss by providing long-lasting and robust protection against numerous plant pathogens[34]. The beneficial effects of DSSs are provided by their microbiomes, which are enriched in microbial taxa known for natural product biosynthesis (*e.g.*, *Streptomyces* spp. and *Pseudomonas* spp.)[10]. Phenotypic studies of Streptomycetes isolated from DSS or disease-conducive soil (DCS) points to differences in the frequency of antibiosis and signaling interactions within these microbiomes[28,45].

DNA sequencing has enhanced our ability to document the composition of soil microbiomes[10,121] and the putative metabolic functions encoded in the constituent genomes[122,123]. Yet, we have little understanding of the regulation and dynamics of gene expression outside of a few model lab strains[59,124–127].

Streptomycete genomes are among the largest in the bacterial domain and are rich in transcriptional regulatory genes. Lack of insight into the regulation and responsiveness of their primary and secondary metabolic activities to community interactions severely limits our ability to understand, predict, and manage functional capacities of soil microbiomes.

Here we focus on identifying the specific roles of species-species interactions in mediating the transcriptional activities of synthetic microbial communities comprising *Streptomyces* strains. We use RNAseq expression analysis to observe changes in gene expression during co-culture of 2, 4, 8, or 10 strains in an experimental set-up that allows for ecological interactions via nutrient competition and extracellular signaling. We call this type of study 'synthetic metatranscriptomics' to denote that the communities

under observation are pre-defined and substantially less complex than natural soil communities. Our synthetic communities are composed of strains isolated from the same sample of DSS, so their signaling interactions are likely to have co-evolved[34]. We find evidence of conserved signaling responses that are robust to community complexity, that secondary metabolism is more likely to be repressed than induced in communities, and that iron modulation may be a mechanism of interaction between some of the isolates.

3.3 Results

3.3.1 Design of synthetic communities

The ten isolates were all obtained from a single soil core from a control plot at the Cedar Creek Ecosystem Science Reserve, in East Bethel, Minnesota. Strain designations are a five digit code representing sampling location data (See Figure 3.1 for designations and isolation diagram). The first digit represents the plot number, the second digit identifies the sub-grid within the plot, the third digit represents the soil sample ID, and fourth digit indicate the depth at which the sample was isolated, and the fifth digit is a unique identifier for each strain isolated from that same depth. Isolates were screened phenotypically using antibiosis assays[46], signaling-inhibition assays[28], and nutrient utilization (Biolog) assays[46]. Whole genome sequences were obtained using PacBio long read sequencing or a combination of PacBio and Illumina short read sequencing (Supplementary Figure B.1-B.12). The 10 isolates comprise members of 3 clades (Supplementary Figure B.1) and include two sets of closely related "twins" (Supplementary Figure B.2). Out of a total combinatorial space of 1023 possible communities that could be grown as axenic or co-culture from these ten isolates, we selected 61 that spanned different levels of complexity (Figure 3.1b, Supplementary Figure B.13). These included 10 axenic (single strain) cultures, 23 two-member co-cultures, 18 four-member co-cultures, 4 eight-member co-cultures, and a ten-member co-culture. RNA isolated from axenic cultures was barcoded and sequenced individually as well as in

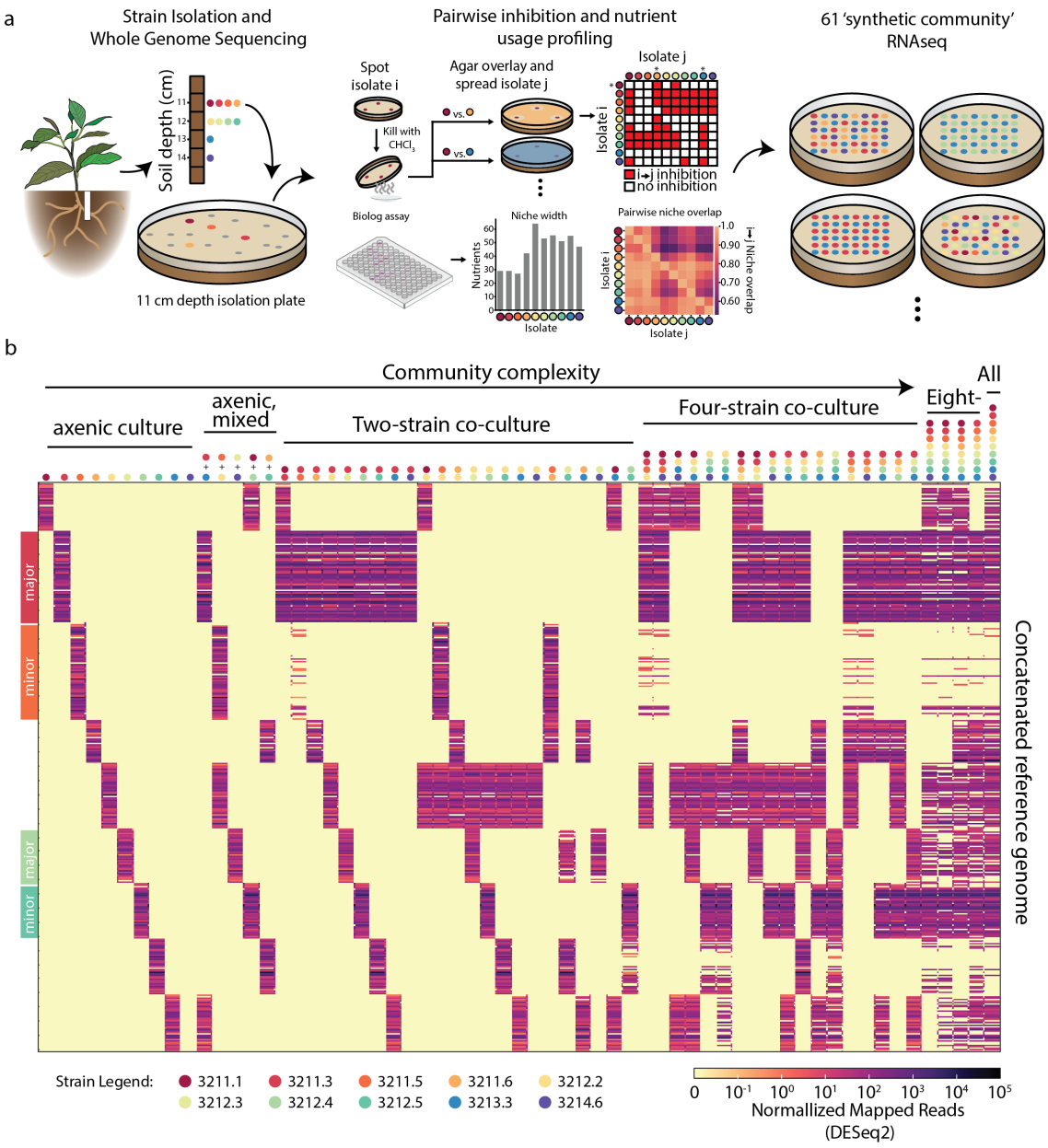


Figure 3.1: Summary of experimental design for synthetic metatranscriptomics. (a) Schematic illustration of strain isolation, inhibition and nutrient usage phenotyping, plate-based co-culture, and RNAseq analysis. (b) Normalized mapped RNAseq reads (normalization described in main text) for each 'synthetic community'. Community composition for each column of data is noted across the top axis. For two pairs of highly similar 'twin' genomes in this study, the major and minor twin (which affects read mapping priorities) is noted on the left axis. Prioritized mapping (see main text) is only applied to communities for which both the major and minor twin are present.

pooled controls. The latter were mixtures of RNA isolated from two different axenic cultures that were combined in equal masses prior to sequencing, allowing us to account for artifacts of sequencing or read-mapping in multi-strain communities. We selected two focal strains, *Streptomyces* sp. 3211.3 and 3212.2, to be overrepresented in the synthetic communities analyzed for this study. Lastly, higher complexity communities are compositions of lower complexity communities (Supplementary Figure B.13). This hierarchical design allows us to track the persistence of low-order signaling interactions as community complexity increases. All communities were grown on solid ISP2 agar medium using a spotting technique to maximize neighbor interactions (see Methods, B.14).

3.3.2 Synthetic community metatranscriptomics

RNA sequencing was performed for the 61 communities by the Joint Genome Institute on an Illumina HiSeq-2500 platform (see Methods) with approximately 1 gigabase of RNAseq data per genome in each community (2 gb per genome for axenic cultures) (Supplementary Figure B.15). We analyzed biological replicates of all axenic cultures to determine a gene-by-gene variance that was used in calculating p-values for differential gene expression across more complex communities. Final RNAseq results, including quality control metrics for library prep, total reads generated per sample, and read quality, is presented for each community in Supplementary Table B.11. Details on read-mapping are reported in the Methods section.

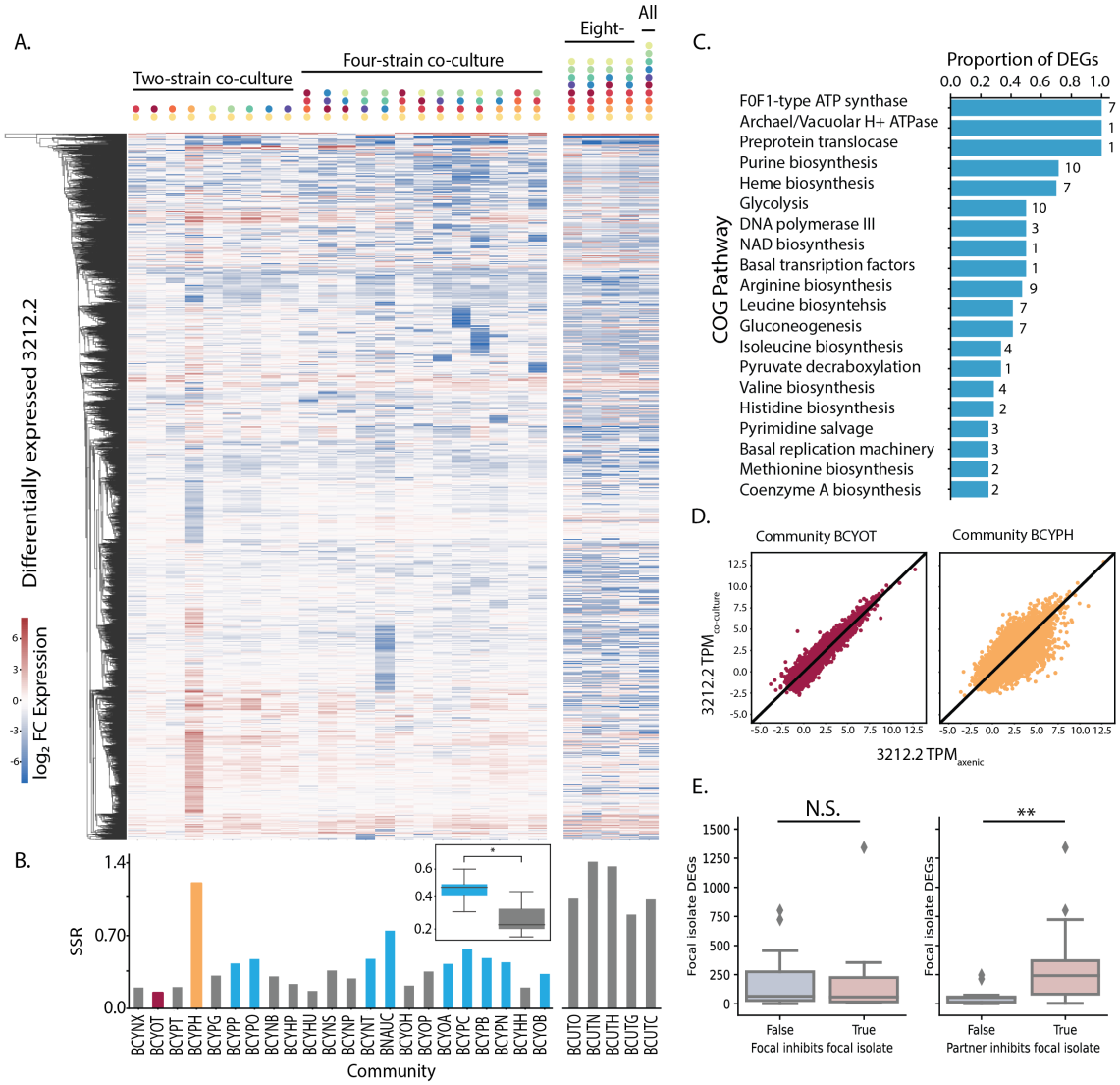
Normalized read-depth across each community is shown in Figure 3.1b (DESeq2 normalized) and Supplementary Figure B.16 (TPM normalized). We observe differences between the average global transcription perturbation between the two- and four-member communities and the eight- and ten-member communities (Supplementary Figure B.17; Kruskal-Wallis test, p-value = 2.589e-12). While some genes appear to have a reproducibly consistent expression level at each level of complexity, we consider the predominant effect on the eight- and ten-member communities to be an artifact of poor read depth for the high complexity communities compared to less complex communities (despite the fact that they were

allocated a proportionally greater number of reads during Illumina sequencing). The source of the poor read depth appears to be due to the presence of a high quantity of ribosomal RNA in the higher order communities (see Supplementary Figure B.18 and Supplementary Table B.11). Because of this, most analyses in this paper are focused on the 56 axenic, two-, and four-member communities.

Replicate axenic cultures were processed independently or as pooled sample to determine if read-mapping in mixed communities was a source of bias (Supplementary Note B.10). We observed that RNAseq results were comparable for axenic cultures processed independently and those from the same strain pooled and sequenced with RNA from other axenic cultures. Therefore, readmapping from mixed communities is not likely a major source of bias. Instead, the R-squared values from Supplementary Figure B.19 more likely represent the stochasticity of gene expression for each strain, which is a well-known characteristic of *Streptomyces* spp. In fact, the stochasticity appears to be a result of reproducibly high variance genes (Supplementary Figure B.20). In this regard, it is interesting and noteworthy that this stochasticity is not uniform from strain to strain.

Figure 3.2 ***(following page)*****:** Transcriptional changes in a focal strain across community complexity. (a) Change in gene expression for genes displaying at least a 2-fold change in at least one community, organized by hierarchical clustering along the vertical axis and by community complexity along the horizontal axis (two-member communities at left to ten-member communities at right). All filtering by expression strength, and hierarchical clustering used only two- and four-member communities. Fold change values for eight- and ten-member communities were included for plotting purposes, but have been ordered by the hierarchical clustering result for the lower-complexity communities. (b) Comparison of Sum of Square Residuals (SSR) for 3212.2 expression in all communities, maroon and gold bars highlight comparisons from (d), blue bars highlight communities that include 3212.4 and/or 3212.5. Inset boxplot comparing those communities against all other 3212.2 communities; * P value < 0.05 by two-tailed independent t-test. (c) Top 20 COG pathways represented in 3212.2 DEGs. Blue bars indicate the proportion of DEGs within a given COG pathway that were found to be significantly differentially expressed in co-culture with 3212.4 and/or 3212.5. (d) Plots of the gene-by-gene expression in axenic (x-axis) versus pairwise (y-axis) communities for the co-culture of 3212.2 with 3211.1 (left, magenta) or 3211.6 (right, yellow). (e) Inhibitory interaction influence on gene expression when grown with a partner inhibited by the focal isolate (left), or with a partner that inhibits the focal isolate (right). Note, in this subpanel Focal isolate refers to the isolate's DEGs that are being counted. The data in this analysis include all of the eight non-minor twin strains. ** P value < 0.01 by two-tailed t-test.

Figure 3.2: (continued)



3.3.3 Focal isolates

We focused the synthetic communities around two isolates, 3211.3 and 3212.2. These isolates were chosen because of their unique inhibitory and nutrient use profiles. 3211.3 is an exquisite inhibitor of other sympatric isolates[28]. In agar overlay assays 3211.3 inhibited all sympatric isolates described in this study except for its near-isogenic strain 3211.5. In the same experiments 3212.2 only inhibited 3211.6. Nutrient usage studies were performed previously[46]. In these studies, each isolate is tested for growth on 95 different carbon and nitrogen sources, resulting in a niche-width (number of nutrients metabolized by a strain). 3211.3 had a relatively narrow niche-width (nutrients = 29), while 3212.2 had a relatively wide niche-width (nutrients = 64). The intensity of inhibition between a pair of isolates was positively correlated with their niche-overlap, or the degree to which they can metabolize the same nutrients.

For each focal isolate we investigated changes in gene expression at both the gene- and genome- level. To explore gene-level changes in expression we performed a differential gene expression analysis using DESeq2[128]. Gene expression for each community was compared to axenic and axenic-mixed controls. Hierarchical clustering of relative expression values revealed several interesting expression patterns across communities (Figure 3.2,A). In general, the relative expression data is dominated by repression. Community BCYPH (3212.2, 3211.6) showed a high percentage of DEGs compared to the other 3212.2 communities. It is worth noting that 3211.6 is the only isolate inhibited by 3212.2. We have no evidence that this is driven by non-biological variables (e.g. sequencing artifact), but we also cannot show that this is a reproducible effect. While there are no true 3212.2-3211.6 experimental replicates, in no other communities featuring both 3212.2 and 3211.6 do we see this striking change.

We quantified the genome-level change in transcription. We defined the genome-level change in transcription as the sum of squared residuals (SSR) for the expression level of all genes in a genome when grown in co-culture versus axenic culture; normalized by the number of genes in the genome (Figure 3.2 B, C). Neither the pairwise community SSR for the focal isolates, nor the SSR for all accompany-

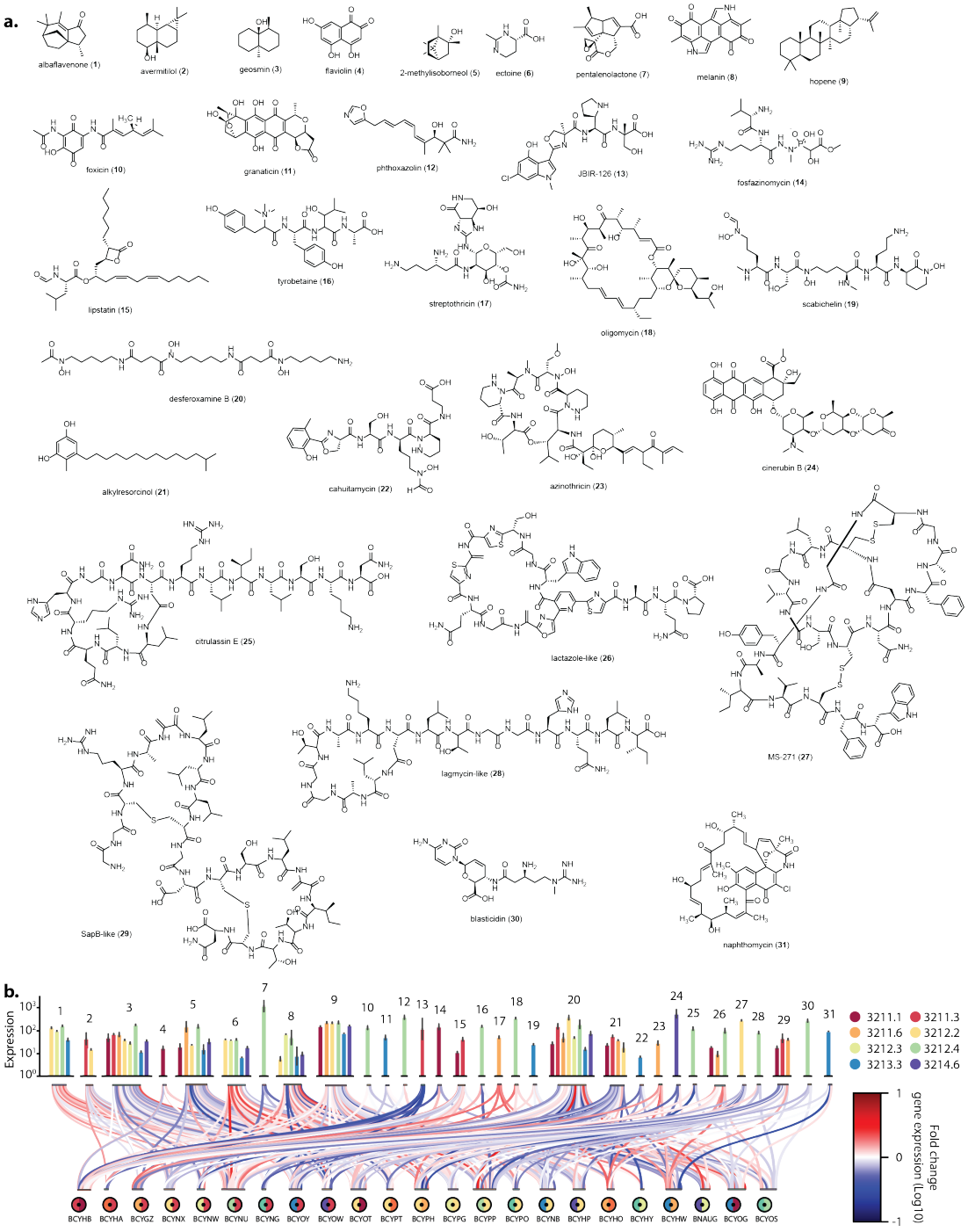
ing strains in pairwise communities with the focal isolates were significantly different (Supplementary Figure B.21). For focal isolate 3212.2 we found that its SSR was significantly higher in communities including 3212.4 and/or 3212.5 (Figure 3.2 B, inset). 3212.2 genes that were significantly differentially expressed in co-culture with 3212.4 and or 3212.5 were identified by statistical analysis (two-tailed t-test, FDR < 0.1) and subjected to COG pathway analysis through JGI IMG/ER (Figure 3.2, D). We find that co-culture with 3212.4 and/or 3212.5 appears to perturb regulation of several primary metabolic pathways including ATP synthesis, purine biosynthesis, heme biosynthesis, glycolysis, and biosynthesis of several amino acids.

3.3.4 Changes in gene expression are influenced by inhibition

We hypothesized that the previously described positive correlation between pairwise niche-overlap and intensity of antibiotic inhibition[28,46] would be mirrored in the expression data. We also investigated the possibility that the increased interactions between symaptric isolates may be evident at an even smaller scale, by correlating isolation depth and differentially expressed genes. We found no significant correlation between the number of an isolate's differentially expressed genes (DEGs) in a pairwise community and the pairwise niche-overlap, or isolation depth distance (Supplementary Figure B.22, A-C). We did, however, find that there is a significant difference between the number of an isolate's DEGs in a pairwise community, and whether or not the isolate is inhibited by its partner (Figure 3.2, E).

Figure 3.3 ***(following page)*****:** Impact of community interactions on expression of known-compound BGCs. (a) Chemical structure of compounds encoded in genomes of synthetic community members with 100% conservation of gene content to previously-characterized BGCs. (b, top) DESeq2-normalized median expression of biosynthetic genes in axenic (single strain) cultures from BGCs corresponding to compounds in (a). Bars are color coded by strain, and 'minor twin' strains are excluded. Bars show the mean of two independent replicates and error bars are SEM. (b, bottom) Colored lines below the bar graphs connect each compound/strain to the two-member communities in which it is found. Lines are colored to indicate the log10 fold-change in gene expression from axenic to two-member communities. Two-member communities are labeled with 5-letter IDs and colored according to community composition using the same color-legend as the bar graph above.

Figure 3.3: (continued)



3.3.5 Impact of community interactions on secondary metabolism

BGCs were annotated for the ten genomes using antiSMASH v5.0[129], and manually curated to indicate which clusters have been previously described (Supplementary Tables B.1-B.10). We were conservative in our identification of these 'known BGCs', requiring 100% conservation of biosynthetic genes compared to the characterized paralogs. One minor exception is for RIPPs, including lactazole and SapB, where we considered BGCs to be known even if they had a slightly different primary amino acid sequence for the core peptide that would result in minor structural differences. Out of a total of 458 BGCs annotated by antiSMASH in the ten strains, 109 are considered 'known'. These correspond to 31 natural products shown in Figure 3.3a. We quantified the expression of known BGCs in each axenic and mixed culture (Figure 3.3b). Transcription levels in axenic cultures span a range of two orders of magnitude, but we do not know how this corresponds to production titer. Viewing the fold change in expression between axenic and two-member mixed cultures reveals several interesting patterns (Figure 3.3b). For example, the nonribosomal peptide JBIR-126 (also known as tambromycin, compound 13) is strongly repressed in each of five two-member mixed communities. On the contrary, the antibiotic streptothricin (compound 17) is strongly induced in each of four two-member mixed communities.

To provide a high-level view of how community composition affects expression of all predicted BGCs, we examined the fold change in specialized metabolite biosynthetic genes in each community. For each predicted BGC, fold change in expression of the genes annotated as 'core biosynthetic' or 'additional biosynthetic' (violet in Figure 3.4A) was compared to the axenic controls (Figure 3.4A,B). To simplify the analysis (Figure 3.4B), points representing each BGC in each community report the arithmetic mean expression and arithmetic mean adjusted p-value for all 'core' and 'additional' biosynthetic genes, and are colored corresponding to the strain containing each BGC. In this plot, each BGC is represented multiple times, once for each pairwise or higher-level community in which it is present. We observe several interesting trends. First, BGCs are more likely to be repressed than activated by commu-

nity interactions. There are 53 instances where BGCs are upregulated, while 131 instances where BGCs are downregulated (chi-squared, p<0.001). The 184 instances of significant differential regulation represent 73 unique BGCs across 8 strains. Since BGCs tend to be located in the chromosomal arms, this data agrees with, but is not sufficient to fully explain, an observation that genes in chromosome arms are more likely to be repressed than genes in the core genome[130,131] (Figure 3.4C). We find that there are significantly more expression events in the arms compared to the core genome. Of these expression events, repression is the dominant regulatory mode. In particular, it is repression events in the arms of the chromosome that dominate all other expression events.

To better understand how BGC expression varied within each genome, we generated a hierarchically clustered heatmap of the expression level of each BGC in each community (Supplementary Figures B.23-B.32). In these heatmaps, BGCs are sorted based on their average expression with the most expressed BGCs on top and the least expressed on the bottom. The variance in BGC expression across communities is plotted at right. There are a few noteworthy observations from these analyses. First, there is an obvious bias towards high variance in the lowly expressed gene clusters. This results from these genes being close to the limit of detection in our RNAseq experiment, and likely explains some of the preponderance of repressed BGCs in the fold-change analysis of Figure 3.4. The higher variance for lowly expressed genes is seen across community complexities, and so is not an artifact of difficult read mapping in the most complex communities. There are several interesting strain-to-strain differences that are observed in these figures. First, strains differ markedly in their robustness to changing community interactions. For example, 3211.6 contains 52 annotated BGCs, and only two of these have a variance of expression greater than 0.25 (Figure B.26). In contrast, 3211.1 has a similar number of BGCs, yet 20 of 54 have a variance of expression greater than 0.25 (Figure B.23).

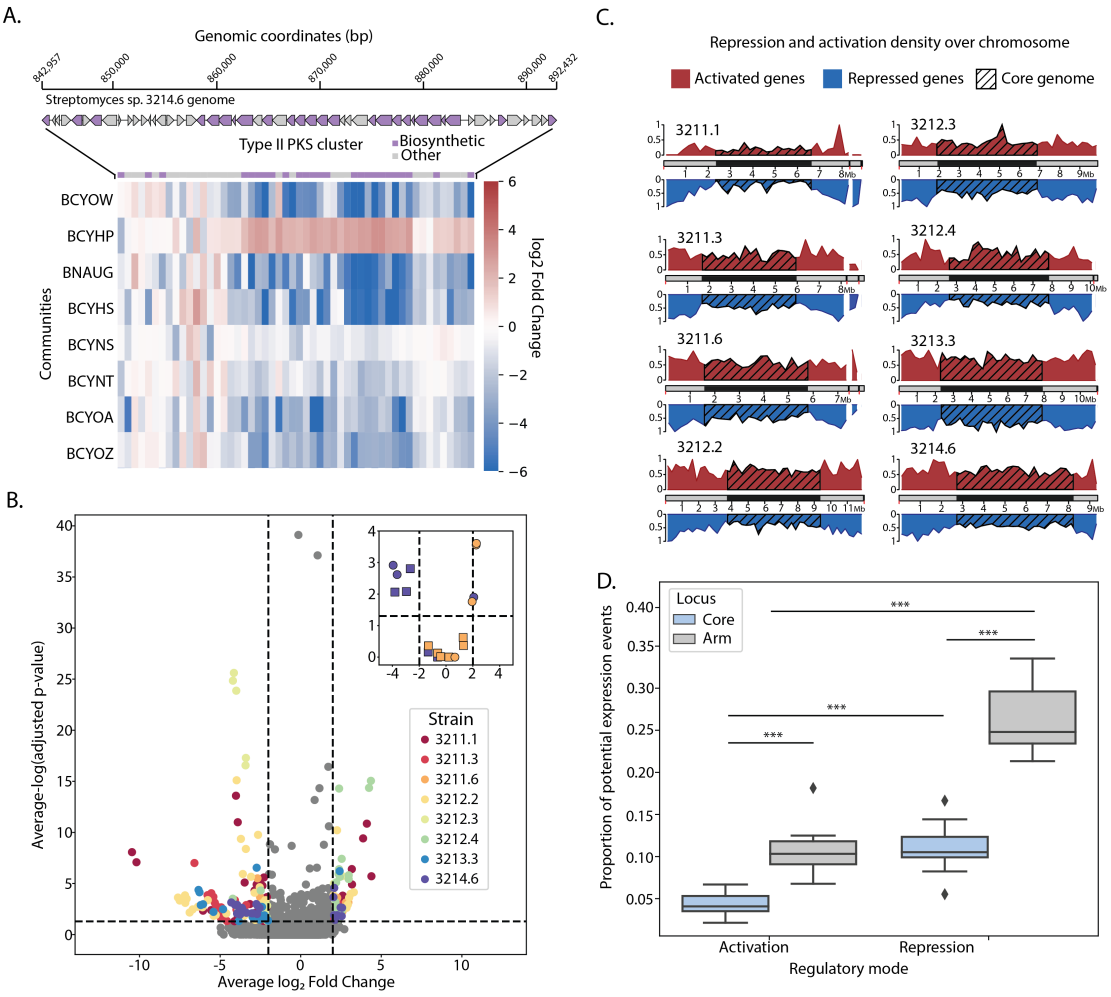


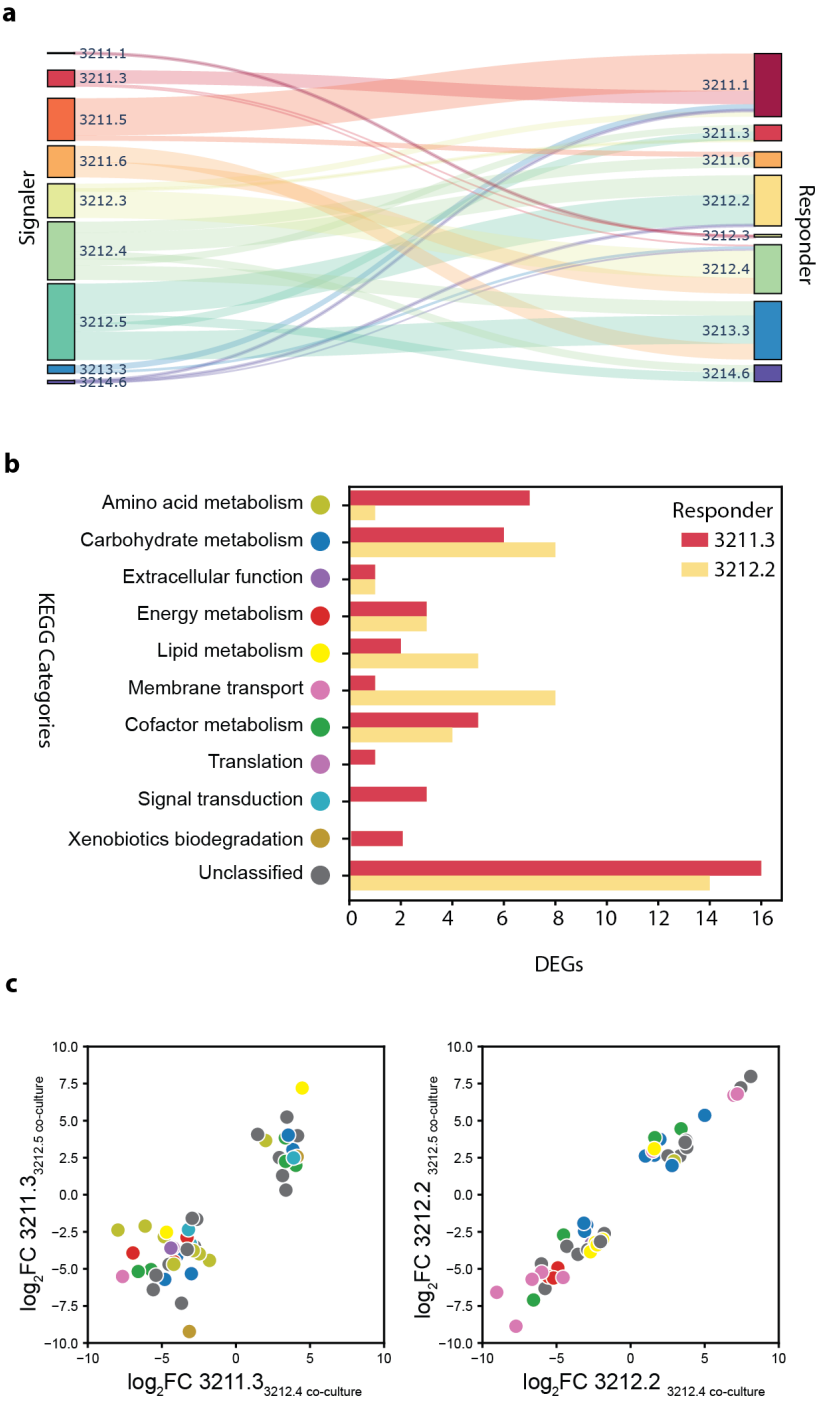
Figure 3.4: Changes in secondary metabolic gene expression. Unique activation of an unknown Type II PKS BGC in 3214.6 (A). Differential gene expression (log2FC) of biosynthetic (purple) and non-biosynthetic (grey) coding sequences found within the BGC. Differential gene expression of all predicted BGCs in the study (B). Each data point represents the average log2FC and average -log(adjusted pvalue) of a BGC's biosynthetic genes in a single community. Vertical dashed lines indicate substantial change in expression (abs(log2FC) > 2). Horizontal dashed line indicates the cutoff for significance (-log(adjusted pvalue) = 1.3). Inset shows the same BGC-level differential expression metrics from two- (circle) and four-member (square) communities, for the unknown Type II PKS BGC from 3214.6 (purple) and the streptothricin cluster from 3211.6 (orange). Regulatory mode density (C). Plots show the density (proportion of expression events that are either activation (log2FC > 1, red) or repression (log2FC < -1, blue). Red markers indicate beginning of chromosome or plasmid. Shaded black region on chromosome, and black hashbox region of density plots show the predicted position of the core genome. Regulatory mode dominance in different regions of the chromosome (D). The proportion of potential expression events (number of genes within a region * the number of communities in which the isolate appears) that are realized by either activation or repression within the chromosomal core (blue) or arms (grey); *** two-sided P value < 0.001 using an independent t-test.

3.3.6 Community-level gene-expression patterns impact primary metabolism

Leveraging the hierarchical structure of the community strain-composition, we explored persistent gene-expression patterns at the two-, and four-member levels of community complexity. We identified gene expression events that seemed to be associated with the presence or absence of a strain for all strains in the dataset. In this analysis we only consider the gene expression events in the eight non-twin strains (responders). We do, however, include the minor twins, 3211.5 and 3212.5, as possible signalers of persistent gene expression events. Note that persistent patterns in major twins in response to minor twins was excluded from this analysis. We find that in several cases the signaler that leads to the highest number of persistent expression events in the responder is 3212.4 or 3212.5 (Figure 3.5, a). We reasoned that because 3212.4 and 3212.5 tend to be associated with the largest response in other strains, and if that response is triggered by the same mechanism, then we may see commonalities within the persistent expression event genes between responders. To simplify our analysis we focused on primary metabolic genes within the two focal isolates 3211.3 and 3212.2. We submitted the list of 3212.4/3212.5 'signaled' persistent DEGs to JGI IMG/ER for KEGG category functional analysis (Supplementary Tables B.12, B.13). We find that for both 3211.3 and 3212.2, several genes related to carbohydrate, energy, and cofactor metabolism were differentially regulated. Amino acid metabolism, signal transduction, translation, and xenobiotic degradation functions were uniquely differentially expressed in 3211.3. Lipid metabolism and membrane transport showed noticeably higher DEG count relative to 3211.3. In both 3211.3 and 3212.2, the highest number of DEGs were unclassified (Figure 3.5, b).

Figure 3.5 ***(following page)*****:** Persistent expression events. Persistent expression differentially expressed genes (DEGs) in Responder when co-cultured with Signaler (A). Bar height (right) represents the number of unique DEGs that are persistently upregulated or downregulated when the signaler strain is present. Bar height (left) indicates the number of unique differential expression events that are associated with the signaler. Width of the connecting curves represents the number of unique expression events associated with the corresponding signaler-responder pair. Connecting curve color indicates signaler isolate. Differentially expressed primary metabolic genes in 3211.3 and 3212.2 (B). (C) Data points show the change in gene expression in primary metabolic genes when 3211.3 (left), and 3212.2 (right) are grown in pairwise co-culture with 3212.4 (horizontal axis) and 3212.5 (vertical axis). Data point color corresponds to predicted KEGG category (bottom) for each gene.

Figure 3.5: (continued)



We investigated the expression levels of the primary metabolic genes (Figure 3.5, c) to see if any pathways were commonly regulated between the strains. We found several genes related to nitrogen regulation and amino acid metabolism were downregulated in both strains. In 3211.3 both beta and gamma subunits of the urease enzyme were downregulated. Similarly, a five gene operon encoding urea transport proteins; as well as nitrite reductase large and small subunits (Ga0334680_112673, 112674) were downregulated in 3212.2.

Carbohydrate metabolism also appeared to be impacted by the co-cultures. In 3211.3, acetyl-CoA synthetase (Ga0334680_11996), pyruvate dehydrogenase E1 component alpha subunit (Ga0334680_11997), and succinate dehydrogenase subunit B (Ga0334680_116596); as well as *paaA* (Ga0334680_114027) and *paaC* (Ga0334680_114025) from the phenylacetate degradation operon were downregulated, while gluconolactonase (Ga0334680_113772) and cellulose synthase (Ga0334680_11386) were upregulated. In 3212.2, we found that three genes (pyruvate phosphate dikinase (Ga0181101_1146) 2767713238, menaquinone-dependent protoporphyrinogen oxidase (Ga0181101_1152) 2767713244, glyceraldehyde 3-phosphate dehydrogenase (Ga0181101_1158) 2767713250) which appear to belong to a cluster of operons were all downregulated, while catalase (Ga0181101_4031) was upregulated.

Fatty acid biosynthesis featured several downregulated genes in both strains (3211.3: *plsC* (Ga0334680_111778); 3212.2: fabF (Ga0181101_7146), *fabH* (Ga0181101_7148), and S-malonyltransferase (Ga0181101_7149).

3.4 Discussion

The inhibition and niche-width profiles of the two focal isolates align with the proposed co-evolutionary framework for DSS establishment[34]. 3212.2 and 3211.3 were both isolated from the same small region of soil from a wheat mono-culture plot. This suggests that the nutrient pool available to isolates was low-diversity, relative to a polyculture plot. The focal isolates exemplify two different ecological

strategies in a nutrient-diversity limited environment. While 3211.3 lacks the ability to metabolize a large number of nutrients, it is capable of inhibiting all other strains in this study. 3212.2, on the other hand, inhibits only one strain (3211.6), but is capable of metabolizing the largest number of nutrients. We found that primary metabolism within the focal strains was impacted, especially so, when grown in pairwise co-culture with either 3212.4 or 3212.5. The genes impacted seem to suggest that both nitrogen and carbohydrate metabolism is being perturbed by the presence of the partner isolates. Tala et al.[132] recently described a regulon in *Streptomyces ambofaciens* ATCC 23877 that is composed of many of the same genes and pathways. The authors inactivated *pirA*, a non-heme iron-binding protein, and combined transcriptomics, proteomics, and metabolite profiling to determine which pathways were affected. They suggest that PirA modulates fatty acid beta-oxidation in response to detection of reactive oxygen species. By limiting entry into the citric acid cycle, and upregulating catalase genes, the authors suggest that the regulon protects against DNA damage. We identified pirin-like genes in 3211.3 and 3212.2 (Supplementary Table B.14), and performed hierarchical clustering, combining expression values from these potential regulators of primary metabolism. We find that some, but not all, of the identified primary metabolic genes may be under direct regulation by PirA in 3212.2. Strain 3212.2 showed a subset of genes clustering around pirin-like gene (Ga0181101_5334) (Supplementary Figures B.34), while there was no clear clustering of genes around the pirin-like genes from 3211.3 (Supplementary Figure B.35). Since pirins are iron binding, and we hypothesized that a lack of iron may be leading to the changes in gene expression. Mey et al.[133] showed that a pirin-like protein was negatively regulated by Fur, a metal-dependent DNA-binding protein, in *Vibrio*. Recently, Secgin et al.[134] described a Fur protein present in *Streptomyces clavuligerus*, that is involved in iron homeostasis. We analyzed the association between *fur* homologues and the primary metabolic genes using the same workflow used with the pirin-like homologues. We identifed eight *fur* homologues in 3211.3 and 3212.2 (Supplementary Figure B.15). In this case, hierarchical clustering (Supplementary Figures B.36, B.37) suggests that only one of the homologues (3211.3: Ga0334680_115082) appears to cluster with the primary metabolic DEGs in a

3212.4/3212.5 dependent manner. Perhaps unsurprisingly, the literature features many publications investigating the link between iron availability and changes in *Streptomyces* secondary metabolism[135–139], community dynamics[137,140], and pathogen inhibition[141,142]. Both 3212.4 and 3212.5 genomes encode siderophore BGCs (4 and 3, respectively). Of particular interest is candidate cluster 40 (Supplementary Figure B.7), which shows high and differential expression in several communities (Supplementary Figure B.29).Further studies are needed to confirm that these changes in primary metabolic gene expression are indeed modulated by iron, and if so, the elements within 3212.4 and 3212.5 that are responsible.

The results of this study shed new light on what has been shown about *Streptomyces* genome organization and how it correlates with gene expression. The *Streptomyces* chromosome has a conserved core encompassing a bidirectional origin of replication, featuring genes that are required for vegetative growth. The chromosomal arms are composed of auxillary genes, many of which are found within BGCs, thought to be conditionally adaptive in response to changing stressors imposed upon the cell. In lab conditions, gene expression tends to be concentrated more on the core during vegetative growth, with expression levels decreasing for many core genes after transitioning to stationary phase. The genes of the chromosomal arms tend to be lowly expressed throughout development. Indeed, a great deal of work has been invested in activation of silent BGCs[98]. One technique that has been useful in activating cryptic BGCs is co-culture[143]. However, in our study, we show that the chromosomal arms, inlcuding BGCs, tend to be repressed with respect to axenic culture. One possible explanation for this observation is that these sympatric isolates have evolved BGC repression as a mode of community establishment.

Our final experimental design was largely guided by overcoming technical hurdles. We initially sought to perform all co-cultures in liquid media by inoculating equal numbers of each synthetic community member. However, dilution plating after five days and scoring community composition by colony morphology revealed that a single community member often dominated the liquid cultures (data not shown). Similarly, performing co-culture on agar plates in lawns grown from an equally mixed inocula resulted in a single species dominating. Both observations could be a combined effect of different

growth rates and direct antagonism by antibiotic production. Regardless, these co-culture methods would have substantially impaired our ability to get sufficient RNAseq read depth for the outcompeted strains. We circumvented this problem by developing the plate-based community assay described above wherein axenic colonies are physically separated from other strains on the agar medium. This limited the disparity in species growth during the five day incubations and we routinely saw 1 cm diameter colonies for each strain. Our ability to map similar numbers of RNAseq reads back to each genome supports the notion that each strain in a given community was represented by approximately the same number of cells at the time of RNA isolation. A major shortcoming of this approach is that we generated spatial differences within colonies, but because of the *en masse* RNA extraction method, we are unable to resolve spatial differences in gene expression. For example, in Supplementary Figure B.14, there are several colonies that appear to have a directionally dependent sporulation phenotype, that may be influenced by the community member to which the differential sporulation phenotype 'points'. By isolating RNA en masse from a plate, we are in essence mixing gene expression events from several distinct chemical/biological environments and reporting the average response for a plate. Future work that uses more precise sectioning of plates prior to RNA extraction could provide better insight to specific interactions in these communities. Another drawback of our co-culture method is that we lack the ability to observe community interactions that result from direct cell-cell contact.

While the breadth of data collected for this study captured many combinations of isolates, it falls short in terms of the replicate data collected. According to standard practices for differential expression analyses[144], a minimum of three replicates should be available for each condition. In our case, we have duplicate data for the axenic controls, and only singleton data for the experimental groups. While the lack of replicates limits the types of questions that can be asked, the intention of this pilot study was to identify communities and gene-expression patterns that would be investigated more thoroughly in future projects. We also note that the necessity of using a modified mapping strategy due to the high genetic identity between the twin pairs is not ideal. Given the low-cost of DNA sequencing, any studies

featuring a small number of strains, should begin with whole genome sequencing of all isolates that will be included. In cases where a high identity pair is required, we advise simulating the transcriptome, attempting mixed community mapping, and modeling which genes are impacted by the high-identity. In our case, we found that the genes most effected are highly conserved genes including tRNAs, ribosomal proteins, and transposases.

3.5 Methods

3.5.1 Strain isolation

The strains presented in this study were all isolated from a pathogen-suppressive plot in the Cedar Creek Ecosystem Science Reserve (CCESR), East Bethel, MN. To target sympatric isolates, a soil corer (10 cm x 1 cm) was used to extract a single core of soil. From that core, 1 g of soil from 11 cm, 12 cm, 13 cm, and 14 cm depth were isolated. Soil samples were maintained at 12 °C until processing. To process the soil samples, each sample was dried overnight between two layers of sterile cheesecloth. After drying, samples were suspended in 10 mL phosphate buffered saline (0.5 M K_2HPO_4) and shaken for 1 hour on a reciprocal shaker at 4 °C. Suspensions were dilution plated on ISP3 and incubated at 28 °C for 7 days. 9-10 colonies displaying morphology characteristic of *Streptomyces* were randomly selected. Spore stocks for each isolate were prepared in 20% glycerol and maintained at -80 °C.

3.5.2 Strain growth and general microbiological methods

For DNA extraction, isolates were grown at 28°C in a shaking incubator in 3 mL R2YE (per liter: 103 g sucrose, 0.25 g K_2SO_4, 10.12 g $MgCl_2{\cdot}6H_2O$, 10 g glucose, 0.1 g casamino acids, 5.0 g yeast extract, 5.73 g N-[Tris(hydroxymethyl)methyl]-2-aminoethanesulfonic acid, 5 mL 1 M NaOH, 10 mL 0.5% KH_2PO_4, 20 mL $CaCl_2$, 15 mL 20% proline, 2 mL trace element solution (per liter: 40 mg $ZnCl_2$, 200 mg $FeCl_3{\cdot}6H_2O$, 10 mg $CuCl_2{\cdot}2H_2O$, 10 mg $MnCl_2{\cdot}4H_2O$, 10 mg $Na_2B_4O_7{\cdot}10H_2O$, 10 mg

$(NH_4)6Mo_7O_{24}{\cdot}4H_2O$). For RNA extraction and metabolite extraction isolates were grown at 28°C for 3 days on ISP2 agar (4.0 g yeast extract, 10.0 g malt extract, 4.0 g dextrose, 20.0 g agar). IPS2 agar plates were allowed to dry at room temperature for 24 hours before plating.

3.5.3 Genome sequencing and annotation

Genomic DNA was extracted from 100 µL of packed cell pellet using the phenol-chloroform method[145]. Samples were submitted to the JGI for sequencing as part of the 1K Actinomyces sequencing project. Due to low-yields of high molecular weight genomic DNA from 3211.3 and 3211.5, these isolates were sequenced using an experimental pipeline for low input samples on the PacBio Sequel platform. The assemblies from this pipeline were highly discontinuous (13 and 15 contigs for 3211.3 and 3211.5 respectively). Because of this we used a 3-contig 3211.3 assembly produced using a hybrid PacBio-Illumina approach, previously published by our lab[122]. The other eight isolates were sequenced with the PacBio RSII platform. Genomes were assembled and annotated using the JGI IMG Annotation Pipeline (v.4.15.2). A more complete description of genome sequencing and analysis will be published elsewhere. BGCs were predicted for each genome using antiSMASH 5[129].

3.5.4 Multilocus phylogeny

Sequences for the house keeping genes *atpD*, *gyrB*, *recA*, *rpoB*, *trpB*, as well as the *16S* SSU rRNA were extracted from each genome. Partial nucleotide sequences for each locus were concatenated in-frame[116]. These concatenations as well as those from the *Streptomyces* PUBMLST database[146] were used in a multisequence alignment using MUSCLE v3.8.31[147]. The resulting alignment was used to build a phylogenetic tree using FastTree 2[148] version 2.1.8 SSE3 using generalized time-reversible models and CAT approximation, which was drawn using Archaeopteryx version 0.9928 `http://www.phylosoft.org/archaeopteryx/`.

3.5.5 Culture and RNA preparation

Spore suspensions (10^7 spores/mL) were plated in 1 uL spots, 1 cm apart. Plating designs for 4-, 8-, and 10-member communities that minimized the variance in distance between all possible strain pairs in the community were selected from 10,000 iterations of random designs. Spore suspensions were plated in a 6 x 6 grid arrangement on a 10 cm circular petri dish. Total number of spots were standardized so that the number of times all strains would appear in a community was equal (e.g. 8-member community, 4 spots per member, 32 spots total). In cases where spot total was less than 36, outer corner spots were omitted. Spore suspensions were spotted such that the resulting colonies did not come into physical contact, ensuring that changes in gene expression due to co-culture were more likely responses to changes in diffusible molecules. After a 72 hour incubation at 28 °C biomass was scraped en masse from plates using a sterile plastic spatula, immediately transferred to RNALater (Thermo Fisher), and stored at 4 °C until RNA extraction, for no longer than 1 week. RNA was extracted using the MoBio PowerMicrobiome kit. Mechanical disruption of the samples was modified slightly from the protocol. A 0.25 inch diameter ceramic ball (MP Biomedical), 500 μL phenol-chloroform (1:1), and 26 μL β-mercaptoethanol were added to the bead disruption tube. The tube was heated to 58°C. Bacterial samples were removed from RNAlater,allowed to drain on three layers of KimWipes, and immeidately transferred to the preheated mechanical disruption tube. Cells were homogenized by bead beating. For artificial mix controls equimolar concentrations of RNA from two distinct axenic cultures were combined. Artificial-mix strain combinations were selected by maximizing 16S divergence. 61 RNA samples were submitted to JGI for library preparation and sequencing.

3.5.6 RNA sequencing

Paired-end sequencing was performed with the Illumina HiSeq-2500 system. Communities were assigned to lanes such that read bias would be reduced (Supplementary Figure B.15). Controls included all

axenic cultures, as well as a pairwise mixture of RNA isolates from axenic cultures. Pairs of isolates were chosen for each artificial mix to minimize 16S sequence similarity. Lanes were assigned such that each isolate would achieve an estimated 100X coverage of its transcriptome in each condition. Communities with higher complexity therefore received a proportionally greater allocation of sequencing resources.

3.5.7 Transcriptomic analysis

Reads were trimmed and Illumina adapters removed using Trimmomatic[149]. Trimmed reads were aligned to references containing predicted coding sequences (CDS) using BBmap[150]. In cases where none of the high genomic similarity isolates (twins) were present reads were mapped to a concatenation of CDS from all ten genomes. In cases where a twin was present, we arbitrarily defined one as the least important twin, and two reference sequences were used in the mapping: (1) a concatenation of all CDS except from the least important twins, and (2) a concatenation of all CDS from the important twin and least important twin. For example, community BCUTC contains strains 3211.3, 3211.5, 3212.4, and 3212.5. In this example the first reference (2a) was a concatenation of all CDS from the six isolates with little genomic similarity (< 88% average nucleotide identity) plus CDS from 3211.3 and 3212.4; while the second reference (2b) was a concatenation of 3211.3, 3211.5, 3212.4, and 3212.5. The counts from the custom references are then compiled by removing all counts from reference 2b except for those from the least important twin, and then assigning the counts to the least important twin in the counts for reference 2a.

Two pairs of strains (3211.3-3211.5, and 3212.4-3212.5, referred to below as 'twins') described in this study have high-similarity genomes (Supplementary Figure B.2) despite their unique behavior in phenotypic assays. To minimize the effect of ambiguously mapped reads on the analysis, we employed three mapping strategies based on the genomic identity of a community's constituents. The first strategy is straightforward and was used for communities lacking any of the four twin strains. Paired-end reads from the RNAseq analysis were mapped to a single reference sequence created by concatenating all ten

genomes using the BBmap algorithm with the following mapping parameters:(ambig=toss, strictmaxindel=4, minid=0.9). The second strategy was used for communities that contain only one stain from a twin-pair. Reads were mapped (using the same parameters) to a 'drop-out' concatenated reference file in which the genome from the absent twin was not included. This minimized the number of ambiguously mapping reads from the present twin. The third, and most complex, mapping strategy was needed for communities containing both strains from a twin-pair. Strains 3211.3 and 3212.4 are denoted 'major twins' and strains 3211.5 and 3212.5 are denoted 'minor twins' in the sentences that follow (Figure 3.1). This mapping strategy comprises two separate steps. First, RNAseq reads from the community were mapped to a 'drop-out' concatenated reference sequence lacking the genome of the minor twins. This mapping provided the mapped reads for all but the minor twin genomes. Separately, the RNAseq reads from the community were mapped with higher stringency to a short two-genome concatenation of just the major and minor twin genomes. Sequencing reads mapping unambiguously to the minor twin genome were obtained from this step. Finally, reads mapped to each genome were normalized as per-genome TPM. Figure 3.1 shows the normalized mapped reads for each genome across 61 1-, 2-, 4-, 8-, or 10-membered communities.

3.5.8 Differential gene expression

Raw counts were processed by DESeq2. For each genome, the normalized counts from each experimental community were compared with the axenic control and the artificial mix control for that genome. Fold change in gene expression was calculated, and associated p-values were determined to assess the statistical significance of the observed changes. To control for the false discovery rate (FDR) due to multiple hypothesis testing, p-values were adjusted using the Benjamini-Hochberg procedure[151] (BH-procedure). An FDR threshold of 0.1 was set for determining statistically significant differentially expressed genes.

3.5.9 Hierarchical clustering of gene expression data

Hierarchical clustering was conducted using the clustermap function from the Seaborn library in Python. Default settings were used (distance metric: Euclidean, linkage method: average).

3.5.10 BGC expression analysis

An in-house python script was used to extract the relative expression for all predicted BGCs. Genes were then filtered by their antiSMASH classification as either 'biosynthetic', or 'biosynthetic-additional'. The arithmetic mean expression and arithmetic mean adjusted p-value was computed for the filtered set of genes.

3.5.11 Persistent expression analysis

For each isolate pair (isolate A, isolate B), we partitioned relative expression data for isolate A into two groups, (i) with, and (ii) without isolate B. We performed a Student's t-test, followed by multiple-test correction of the p-values using the BH-procedure (see Methods: Differential gene expression). Genes with significant p-values ($p < 0.05$) were considered to be 'persistent'.

3.5.12 Data visualization

All data manipulation and plotting was performed using Python, Pandas[152], and Seaborn[153]. All statistical tests utilized functions from the scipy[154] Python package.

Supplementary tables and figures can be found in Appendix B.

Chapter 4

Simulation Modeling to Compare High-Throughput, Low-Iteration Optimization Strategies for Metabolic Engineering

The following is a reprint of the article Heinsch, S. C., Das, S. R., & Smanski, M. J. (2018). Simulation Modeling to Compare High-Throughput, Low-Iteration Optimization Strategies for Metabolic Engineering. *Frontiers in microbiology*, 9, 313.

Article hyperlink:

https://doi.org/10.3389/fmicb.2018.00313

SH, SD, and MS designed the experiments and performed the analyses. SH and MS wrote the manuscript.

4.1 Summary

Increasing the final titer of a multi-gene metabolic pathway can be viewed as a multivariate optimization problem. While numerous multivariate optimization algorithms exist, few are specifically designed to accommodate the constraints posed by genetic engineering workflows. We present a strategy for optimizing expression levels across an arbitrary number of genes that requires few design-build-test iterations. We compare the performance of several optimization algorithms on a series of simulated expression land-

scapes. We show that optimal experimental design parameters depend on the degree of landscape ruggedness. This work provides a theoretical framework for designing and executing numerical optimization on multi-gene systems.

4.2 Introduction

Biotechnology applications that require the coordinated expression of dozens of genes have the potential to meet current and future needs for energy generation, production of medicinal or commodity chemicals, biosynthesis of functional biomaterials, and living biosensors[155]. Moving these complex systems between alternative host species, for example a microbial host amenable to industrial scale-up, is difficult[103]. A major challenge is optimizing the expression levels of each required gene to maximize final output and minimize toxicity to the host cell[64,156–158]. Technical capabilities now exist for building and testing 1000s of unique genetic constructs in parallel[157,159–161]. Further, numerous improvements have been made in our ability to quantitatively control individual gene expression levels in the most commonly used organisms for industrial fermentation[64,81,162–170]. Leveraging both of these capabilities will enable high-throughput optimization strategies that rationally improve productivity and yield in less time than low-throughput trial-and-error approaches[157].

Several strategies have been proposed for genetic optimization (Figure 4.1). In the 'multivariate modular metabolic engineering' approach, the combinatorial design space is reduced by grouping pathway genes into operons based on previous knowledge (e.g., enzyme kinetics, branching of pathway, etc.)[171,172]. The reduced combinatorial space can be elucidated empirically. For instance, this strategy was used to improve taxadiene titers ~15,000-fold in E. coli[171]. In another example of modular multivariate optimization, Xu et al.[173] modified the expression levels of three modules comprising nine genes involved in fatty-acid synthesis to improve fatty-acid titers 20-fold. Recently combinatorial RBS libraries designed using biophysical models[81] have been implemented in high-throughput via multi-

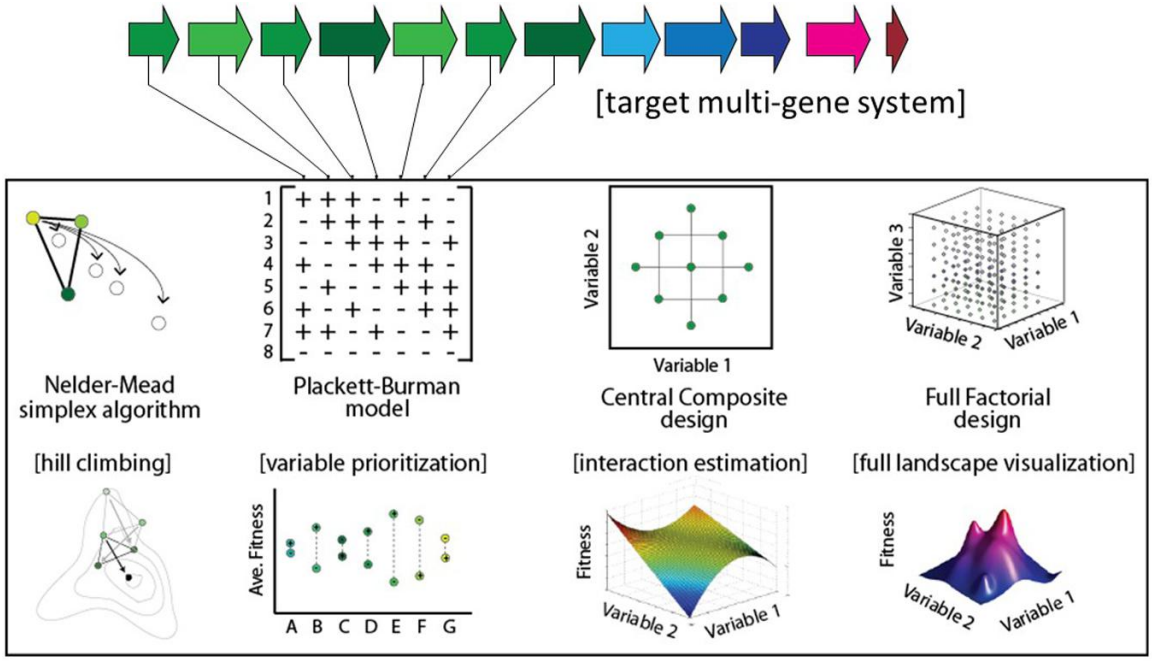


Figure 4.1: Select optimization strategies for multi-gene biological systems.

plexed automated genome engineering[159] to improve isopropanol titers 1.5-fold[174], and NADPH regeneration rates 25-fold[81,175]. Alternatively, algorithmic optimization is possible using a Design of Experiments (DOE) approach. For example, the fractional factorial 'Yates algorithm' was used to co-optimize both gene expression and media conditions in a single experiment, resulting in an approximately fivefold improvement in 6-aminocaproic acid titer (9–48 mg/L) in E. coli[176]. Lastly, linear regression is an effective approach for predicting improved expression levels of a multi-gene metabolic pathway, following a small sampling of the combinatorial design space[177,178]. Previously, linear regression was shown to be capable of predicting relative titers of intermediates within engineered variants of the violacein pathway[177], and more recently regression modeling was used to increase violacein titers 3.2-fold[179].

The ability of any global search algorithm to predict optimal expression levels depends on the ruggedness of the 'fitness landscape'[177,180]. Smooth landscapes arise when variables are independent of each

other and lend themselves well to linear regression approaches. However, if the landscape is rugged, with multiple local optima separated by valleys[181], rational optimization methods will not be as effective[182]. Fitness landscape analyses performed on a library of nitrogen fixation gene clusters suggests that complex multi-gene systems can be moderately rugged and will not lend themselves to linear regression (Smanski, unpublished).

Numerical optimization refers to a set of techniques aimed at identifying a local or global maximum (or minimum) in a fitness landscape. A common goal for numerical optimization methods is to find the maximum with the smallest amount of computational resources, which normally correlate to the number of sampled points. For metabolic engineering, this corresponds to the number of alternative genetic designs that would have to be designed, built, and tested. In a recent comparison of numerical optimization algorithms, variations of the DIRECT search algorithm performed well[182]. The DIRECT method balances local and global searching strategies. It was designed specifically with engineering optimization in mind, where time or resource costs associated with running experiments calls for methods with efficient use of function evaluations[183]. Unfortunately, methods that seek to optimize the efficiency of function evaluations do not distinguish between the number of iterations and the number of function evaluations per iteration. This distinction is important for genetic engineering projects. Increasing the throughput of a single design-build-test cycle can typically be done at a small fraction of the cost compared to increasing the number of design-build-test cycle iterations.

Here, we describe and model an approach to genetic optimization that combines (i) the quantification of fitness landscape ruggedness with (ii) a high-throughput, low-iteration optimization algorithm for improving genetic design. We show that the optimization parameters should be tailored for each system based on fitness landscape ruggedness. Finally, we compare the performance of this approach to several alternative hill-climbing algorithms.

4.3 Results

4.3.1 Assessing the Ruggedness of a Multivariate Expression Landscape

We began by creating three model landscapes for testing optimization algorithms (Figure 4.4A). The 3D landscapes simulate a two-gene system, where the X- and Y-dimensions represent the expression levels of the two genes, and the Z-dimension represents the measured performance of the system (e.g., the product titer for a metabolic system). Most metabolic pathways are more complex than this, but we chose to model a two-gene system because the progress and results of the algorithm are easily visualized. The algorithms described in this study can be easily adapted to higher-dimensional space.

We first aimed to establish a metric for determining the ruggedness of a gene expression landscape based on Kauffman's N–K method[184,185]. In the N–K method, N refers to the number of component parts and K is the order of interaction. When K = 0, the system variables behave independently, and the landscape is expected to be smooth. The maximal value of K is N-1, which would represent a system where the optimal level of any variable depends on the setting of all other variables. This would produce a rugged landscape. A LAA allows one to estimate the average ruggedness of a landscape using sampled data points[186,187]. LAAs have been performed in biology to problems of RNA folding and protein structure/function, but not to multi-gene expression analyses. A key difference in these types of problems is that the permutable variables in macromolecular optimization problems are discrete, whereas gene expression level is a continuous variable. We have slightly modified previous LAAs to account for this difference. For each model landscape, we sampled 40,000 points in the X,Y coordinate space to evaluate f(x,y). The autocorrelation compares the average variance for pairs of data points within a given Euclidian distance on the (X,Y) plane to the average variance for the landscape as a whole. On smooth landscapes, the variance of f(x,y) for two points located near each other in the (x,y) plane is expected to be small. The variance will approach the average landscape variance as distance between two points increases. The plotted landscape autocorrelation, ($1-\sigma 2$d=bin(x)$\sigma 2$landscape), is approximately 1 for very

close points and approaches 0 as the distance between compared datapoints increases. The rate at which this landscape autocorrelation value decreases is related to landscape ruggedness, with more rugged landscapes dropping off more rapidly (Figure 4.4B). We quantify landscape ruggedness by comparing landscape autocorrelation plots to the equation: f(x)=(1−xN)(1−kN)x and solving for k. The model smooth, medium, and rugged landscapes generated for testing optimization algorithms have k values of 0.832, 1.07, and 2.07, respectively. For empirical optimization of metabolic pathways, we envision that the actual landscape ruggedness would be measured with a seed library of diverse expression cassettes. Our model landscapes are in the same range of ruggedness as seen in multigene metabolic pathways for which pathway productivity is measured under combinatorial expression levels[157,171,177].

4.3.2 High-Throughput, Low-Iteration Optimization Algorithms

We next developed a set of numerical optimization algorithms that are designed with the technical aspects of metabolic engineering in mind. Namely, the algorithms search the multivariate expression space with very large sampling libraries, but low numbers of iterations. As a comparison, a 20-gene synthetic nitrogen fixation pathway was recently improved using five iterations, each with approximately 100 alternative genetic designs.

Each optimization algorithm follows a similar order of operations. An initial set of (x,y) coordinate points are sampled and their fitness is evaluated using the landscape function, f(x,y). The subset of points with the greatest fitness (i.e., the 'parents') are used to determine the center point and shape of the next set of samples (Figure 4.2). The algorithm parameters are listed in Figure 4.2 and include the number of samples taken in each generation, the area of the multidimensional expression space sampled, and the fraction of sampled points carried forward as parents for the next iteration. In each case, we sample a defined area using Sobol sequences. Sobol sequences provide a quasi-random distribution of a search space and provide more even coverage of the space than a random Gaussian sampling.

Three unique optimization algorithms were tested that differ in how the new sampling space is de-

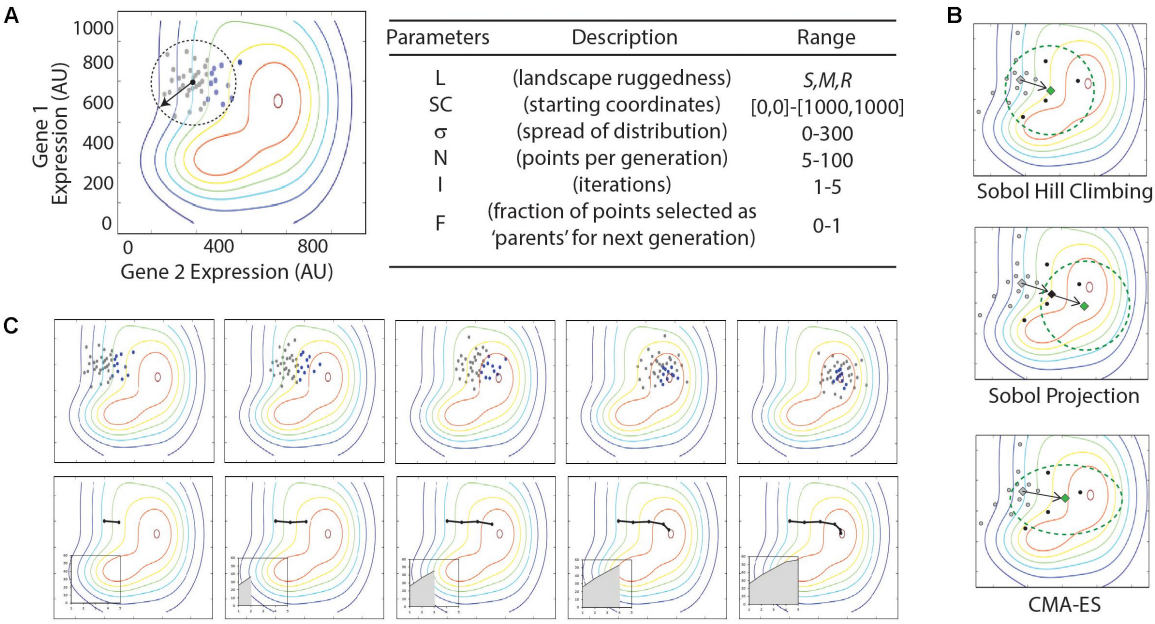


Figure 4.2: Illustration of optimization algorithms used in this study. (A) Illustration and table of parameters in Sobol Hill Climbing algorithm. The vector and dashed circle denote the spread of sample points (σ) from the starting coordinates (SC). Gray and black dots show the sampled points per iteration (N). Black dots represent the fraction of points selected as parents for the next generation (F). Parameters are listed in the table with approximate ranges of parameter values explored in the current study. (B) Top panels show results from a single simulation experiment with the following parameter values: L = smooth; SC = [300,700]; σ = x; N = xx; I = 5; F = 0.x. Each iteration is shown from left to right. Bottom panel shows the route taken by the optimization algorithm, with black line tracing location of center-points for Sobol sampling. Insets show increase of fitness (z-axis) through each iteration. (C) Summaries of triplicate simulations on three distinct parameter sets. Graphs represent optimization routes as described in (B), with triplicate simulations represented as black, red, and blue lines. Parameters are given below each graph in the format of [L; SC; σ; N; I; F].

termined for each iteration (Figure 4.2). The most simple method, which we call 'Sobol Hill Climbing,' takes the geometric center of the high-fitness parent points in the (x,y) plane and uses that as the center point for the next iteration of Sobol sampling (Figure 4.2A). The 'Sobol Projection' algorithm draws a vector from the center of the sampled space through the geometric center of the high-fitness parent points. If the distance [in the (x,y) plane] between those two points is d, the center of the next generation of sampled points is along that vector 2×d away from the previous center (Figure 4.2B). The Sobol Projection algorithm has the advantage of moving faster in an uphill direction with each generation, but it will also over-shoot the global maximum more easily than the Sobol Hill Climbing algorithm. The last and most complex algorithm uses the covariance matrix adaptation evolution strategy (CMA-ES; Figure 4.2B)[188]. This algorithm differs from Sobol Hill Climbing in two important ways. First, the center point for the next iteration is determined by the weighted average of the high-fitness parent points, with weights determined by fitness value. Second, the shape of the sampling space is adjusted with each iteration. While the first two algorithms always search with a Sobol sequence following an N-dimensional standard normal distribution, the CMA-ES algorithm adjusts both the size and shape of the sampled area, according to the size and shape of the distribution of high-fitness parent points.

We evaluate the performance of an algorithm by tracking the fitness of the center point for each of the first five iterations (Figure 4.2C). The area under this curve represents the performance of the algorithm. In this way, the performance reflects both the fitness value attained and how quickly the algorithm arrived at that fitness value. We run each algorithm five times with identical parameters and record the standard deviation of the performance metric. This gives a measure for how reliably the algorithm can be expected to perform.

4.3.3 Parameter Optimization for Each Algorithm

Parameters such as number of points sampled per iteration or the number of iterations are likely to be determined by the time and resources available for expression optimization efforts. Parameters affect-

ing the distribution of sampled points and the fraction of sampled points used as parents for the next iteration do not change the cost of a given design-build-test iteration, but can greatly influence the optimization results. We simulated each optimization algorithm using a range of parameter values for σ and F. For each combination of parameters, we simulated five optimizations and score both the average fitness and the standard deviation, as measures of performance and reliability, respectively.

Results from the survey of parameter combinations for the three search algorithms are shown in Figure 4.3. Not surprisingly, each algorithm performed best on the smoothest landscape, both in terms of the gain in fitness and in the reliability. The Sobol Hill Climbing algorithm (Figure 4.3A) generally worked best when each iteration sampled a disperse set of points (large σ value) and only a small fraction of sampled points (small F-value) were used to seed the next generation. For medium and rugged landscapes, the algorithm was less reliable at values of $F < 0.2$. This was not observed for the smoothest landscape.

The Sobol Projection algorithm (Figure 4.3B) performed slightly better than the Sobol Hill Climbing method, particularly on more rugged landscapes. Notably, this algorithm was more sensitive to the fraction of kept values (F). Low F-values resulted in a substantial decrease in fitness as well as an increase in noise. Both the Sobol Projection and Sobol Hill Climbing algorithms showed a prominent loss of reliability (high standard deviation) on the medium-ruggedness landscape when the sampling range was approximately 100 units, even at intermediate F-values. At these parameter values, the optimization algorithm tended to get trapped in one local optimum, which was determined stochastically at an early iteration.

The CMA-ES optimization strategy (Figure 4.3C) performed substantially worse than the others in the conditions tested, both in terms of fitness values attained and in the reliability. It routinely found the global maximum in the smoothest landscape, but not as quickly as the other two algorithms. For the medium and rugged landscape, it rarely found the global maximum in the first five iterations. When the CMA-ES algorithm was allowed to run for more iterations, it routinely found the global maximum

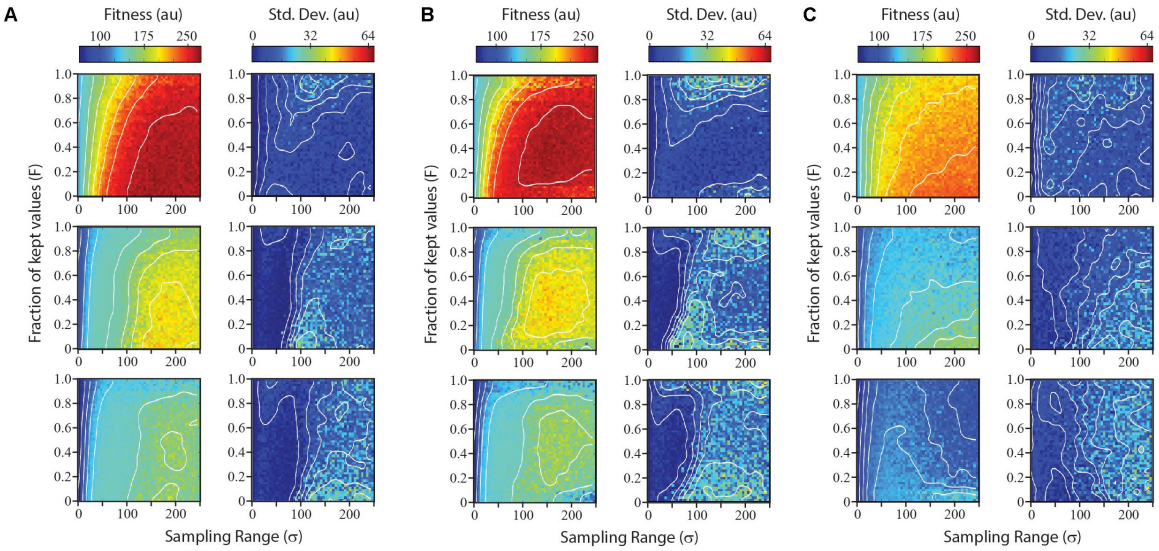


Figure 4.3: Performance and reliability of numerical optimization algorithms across parameter space. The Sobol hill climbing algorithm (A), Sobol projection algorithm (B), and CMA-ES algorithm (C) are compared. For each algorithm, left plots show mean performance from five independent simulations at each parameter combination for optimizations run on a smooth landscape (top), medium landscape (middle), and rugged landscape (bottom). Right plots show reliability of algorithm for each parameter combination, measured as the standard deviation of performance over the five independent simulations.

(data not shown).

4.4 Discussion

The topology of landscapes connecting sequence space to biological phenotypes impacts the evolution of biological systems[185]. This has been shown through a combination of theoretical and experimental work, but primarily at the level of single proteins or RNA molecules[187,189]. Smooth landscapes occur when the variables behave independently. Systems with smooth Mt. Fuji-like landscapes lend themselves to simple optimization approaches[181]. In a system comprising perfectly independent variables, each variable could be optimized separately and the optimum of each variable combined to locate the global maximum. However, in rugged or partially rugged landscapes, interactions among variables can create several local maxima or minima that will confuse optimization efforts[181]. In a recent comparison, prob-

lem dimensionality and non-smoothness decreased the performance of all optimization algorithms[182] tested.

Modern DNA synthesis and assembly capabilities allow for the design, construction, and evaluation of large libraries of multi-gene systems[157,176,190]. This enables evolution-landscape analyses that connect expression levels over each gene in the system with overall system performance. The ruggedness of multi-gene expression landscapes has never been rigorously analyzed, but is important for the performance of optimization algorithms. Linear regression optimizations require that the landscape is smooth and devoid of sub-optima[177]. However, we have observed moderate ruggedness in the multivariate expression landscape of the nitrogen fixation gene cluster[157]. Landscape ruggedness in multi-gene systems can arise from several scenarios. It is possible that the landscape is rugged because of interactions between the final protein products. For example, for multi-protein complexes, optimal system performance might occur at a particular stoichiometry of component parts[157]. In this case the optimal level of each component is not fixed, but depends on the expression levels of other components in the system. A second mechanism for landscape ruggedness in multi-gene systems, which can be considered an 'apparent ruggedness' arises from genetic context effects[191]. These genetic context effects are often unintended consequences that arise from manipulating expression levels of different genes that are in close proximity in the DNA sequence. For example, strong transcription of one gene can attenuate the expression of a neighboring, reverse-oriented transcript via several possible mechanisms[192]. Apparent ruggedness caused by genetic context effects will diminish the efficacy of linear regression and other methods that assume a smooth landscape. Whether the ruggedness of a gene expression landscape comes from interactions of gene products, or genetic context effects that produce a lot of noise when sampling a multidimensional expression space, the impact on optimization strategies is similar. The global optimum on smooth landscapes can be found through conservative searches that continuously walk uphill. Rough landscapes require a less conservative approach where a fraction of the sampling resources are used to search for other local maxima.

We have presented a set of analyses that first assess landscape ruggedness and then optimize the landscape using a limited number of high-throughput iterations. We show that landscape ruggedness affects optimal parameter settings during a multigene optimization strategy. As the landscape topology is a characteristic of the system being optimized, it will not be tunable (as it was with our model landscapes). However, knowledge of the ruggedness can guide the engineer to select appropriate parameter values such as the sampling range and the fraction of sampled points used to guide the next iteration. Smooth landscapes tolerate optimization strategies that cast a broad net over the sampling space and use information from only a small number of sampled points to direct the next round of sampling. Conversely, optimization of more rugged landscapes benefits, both in terms of performance and reliability, from sampling less broadly and using information from roughly 40% of the sampled space to direct the next round of sampling. We did not assess whether the benefit of improved optimization parameters outweighs the cost of performing an initial sampling of variable space to quantify ruggedness. Such a cost/benefit analysis would be highly specific to the system being optimized.

Landscape ruggedness assessments are likely only valid in the local neighborhood of variable space. Rugged fitness landscapes can appear smooth across small search spaces, and empirically derived fitness landscapes tend to be asymmetric[193]. Because of this, it is important to reassess local ruggedness in optimizations that drift far from the original starting point. While not included in the models tested here, it would be useful to continuously update the ruggedness quantification with each round of sampling. This could be done using points sampled during optimization efforts and would not require any additional experimental steps.

The modeling we have performed in this study optimizes over a landscape with two independent variables (X and Y axes; representing the gene expression from two different genes), and one dependent variable (Z axis; representing system fitness). We chose a simple system for ease of visualization of how the algorithm functions to climb in three-dimensional space. Each of the components of our work flow will work equally well for any N-dimensional optimization. For example, a 10-gene metabolic pathway

would contain 10 independent variables representing expression levels of each gene and an 11th dependent variable corresponding to the final titer of the molecule of interest. Because we ran our simulation experiments on a relatively low-dimensional space, we decreased the number of sampled points per iteration accordingly. For an 8–12 gene metabolic pathway, an analogous experiment would require 100–200 sampled points per iteration. This scale is in line with recently demonstrated capabilities[157].

4.5 Materials and Methods

4.5.1 Creation of Model Multivariate Landscapes

We created three model multivariate landscapes on which to test the optimization algorithms in this study. The landscapes were made by summing multiple three-dimensional Gaussian surfaces, the equations for which are given in Supplementary Files 'surface_matrix-low.py,' 'surface_matrix-med.py,' and 'surface_matrix-high.py' for the smooth, medium, and rugged landscapes, respectively. Each model landscape was designed with different levels of ruggedness by varying the X- and Y-dimensional spread of each sub-peak. The height and location in the X–Y coordinate plane of each sub-peak were maintained in each model landscape. Three-dimensional graphics of each landscape are shown in Figure 4.4.

4.5.2 Quantification of Model Landscapes

Forty thousand coordinate (X,Y) points were sampled from each model landscape in a square-grid pattern (200 × 200 points) and evaluated to determine the Z-value at each location. For all possible pairwise combination of points, two values were recorded: (i) the Euclidian distance between the pairs of points in the X–Y plane, and (ii) the squared difference between the two Z-values. Next, all pairwise comparisons were binned based on Euclidian distance into bins from 0–100, 100–200, ...600–700. The average variance for each bin was calculated by taking the mean of the squared differences for pairs of points in that bin. For the landscape autocorrelation analysis (LAA), we plot:

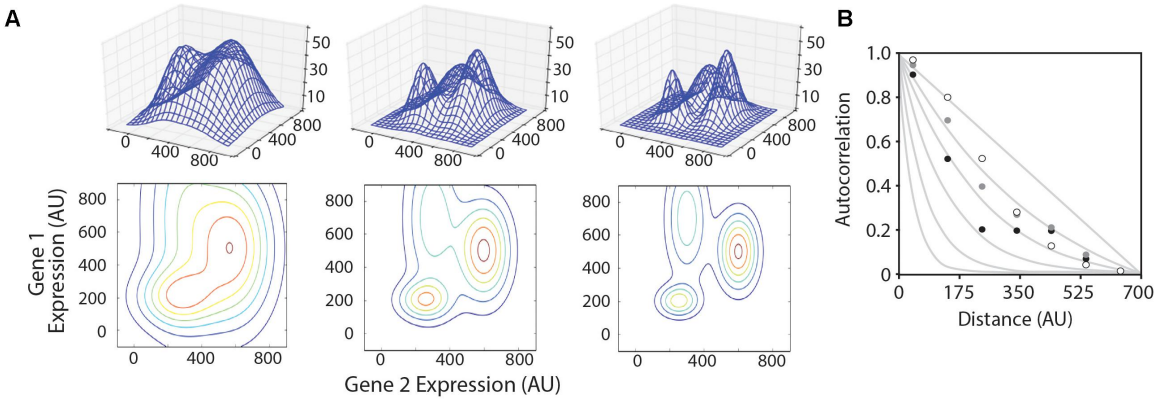


Figure 4.4: Model landscapes and ruggedness analysis. (A) Three model landscapes described in the text as 'smooth' (left), 'medium' (center), and 'rugged' (right) are shown as three-dimensional wire surfaces (top) and two-dimensional contour maps (bottom). X- and Y-axes represent hypothetical expression levels of two genes in a multi-gene system, and Z-axis represents system performance. (B) Autocorrelation function plotted for smooth (white circles), medium (gray circles), and rugged (black circles) landscapes compared to hypothetical traces based on NK-model, where N = 700 and K = 0, 1, 2, 4, 8, and 16 (gray lines, from right to left).

$$LA = (1 - \frac{\sigma^2_{d=bin(x)}}{\sigma^2_{landscape}})$$

where σ^2 landscape is the random variance for the landscape. This was approximated using the pairs of points for which the Euclidian distance is between 600 and 700, as distances greater than 700 are constrained by the size of the search space (1000 × 1000 grid), leading to less pairs sampled at greater distances. Landscape ruggedness was quantified by plotting lines from the function:

$$f(x) = (1 - \frac{x}{N})(1 - \frac{k}{N})^x$$

for N = 700 and determining the best-fit value of k by the non-linear least squares method in R (Version 3.3.3, R Core Team, 2017).

4.5.3 Simulation Algorithm for Optimizing on Model Landscapes

A series of python scripts were created to sample a quasi-random distribution of points around a defined starting coordinate, evaluate the fitness (Z-value) for each sampled point, and determine the center point for the next round of sampling, and iterate this process. These are included as Supplementary Files 'SobolHillClimb.py,' 'SobolHillClimbWithProjection.py,' and 'SobolHillClimb-CMA-ES.py.' For each algorithm, parameters that must be specified include the starting coordinates, the Sobol range radius (a measure for how broad of an area is sampled with each iteration), the number of dimensions, the number of designs to evaluate per iteration, and the fraction of top-performing designs to use in calculating the center point for the subsequent iteration. The three algorithms differ in how each iteration of sampled points is generated. In the most basic algorithm, the center point is the geometric center of top-performing designs. In the 'projection' algorithm, the new center point is projected twofold along a vector connecting the previous center point and the center of the top-performing designs. In the CMA-ES strategy[188], the center point is generated as described for the basic algorithm, but the subsequent quasi-random sampling is perturbed to preference sampling in the same direction as the vector connecting the previous center point to the next center point.

4.6 Conclusion

We propose an integrated strategy for metabolic pathway engineering that combines landscape analysis with a multivariate optimization algorithm. An initial autocorrelation analysis provides a quantitative measure of the ruggedness of the adaptive landscape. This ruggedness metric is used to guide an appropriate selection of parameters during the iterative optimization process. Of the three optimization strategies simulated in this study, the Sobol Projection method gave the best performance on several model landscapes. Further work is needed to validate this strategy using an experimental system.

Chapter 5

Conclusion

In this dissertation, I have presented contributions to *Streptomyces* biology, microbial ecology, as well as metabolic engineering. We answered some of the questions we set out to answer, and we have set the stage for asking the next questions that will drive research in these areas. Early in my graduate studies, we published what, at the time, were some the highest quality, in terms of continuity and per-base quality, genomes available for not only genus *Streptomyces*, but non-model streptomycetes. Much of the *Streptomyces* genomics, transcriptomics, proteomics, and metabolomics work available in the literature has been motivated by (i) the basic science desire to understand the genus as a whole, as well as (ii) the application of specific *Streptomyces* species to the manufacture and discovery of novel small molecule natural products. In both cases, resources are focused (rightfully so) on studying model streptomycetes. In the case of the earliest developmental biology focused studies in *Streptomyces*, *Streptomyces coelicolor* was studied, in part, due to its ability to produce the blue-pigmented antibiotic actinorhodin. This easily observed phenotype facilitated pre-genomics era genetics studies. If we hope to understand the complex ecologies that exist within soil, we ought to devote resources to the study of diverse isolates. We learned that neither biosynthetic gene cluster (BGC) abundance, nor BGC uniqueness, nor presence of canonical signaling compound biosynthetic genes explained the unique ecological features of the *Streptomyces* isolates we sequenced. This work suggested that the answers to what makes this trio of isolates unique from

an interspecies communication perspective may not be readily interpretable out of the context in which they are observed. This of course was on a small sample size. Given sufficient examples of genotypically diverse isolates, with similar phenotype, modern computational methods like deep learning may be able to suggest genomic signatures to investigate.

Several contributions were made through the synthetic metatranscriptomics experiment. This work set out to explore questions that could not be answered by simply looking at the genomes of a handful of isolates. Although it has been six years since we set out on this project, there is still, to the best of our knowledge, no other study of this scale that focuses on how gene expression changes within defined *Streptomyces* communities, at multiple levels of community complexity. Some findings surprised us, such as not finding a single BGC whose expression clearly correlated with community composition, or that gene repression was the dominant mode of regulation in communities, and that this repression was pronounced in the chromosomal arms, which are known to house the majority of secondary metabolic related genes in *Streptomyces*. We also found hints to the modulation of iron as a potential mechanism driving interactions between some groups of isolates.

One of the contributions of this work that has already shaped how we study *Streptomyces* community gene expression, has been our experience dealing with the technical short-comings of the experiment. For instance, the data from this study was sufficient to validate the metatranscriptomics approach to studying community-level gene expression, and our struggle with limited replicates has influenced the design of a more focused four-member *Streptomyces* metatranscriptomics study in collaboration with the University of Manchester, England.

While no formal study was performed, a large contribution to the field of *Streptomyces* genomics was made as a result of a sequencing project with the Joint Genome Institute (JGI). Our collaboration with JGI provided 76, high-quality, non-model *Streptomyces* whole-genome sequences. The isolates sequenced were selected for their unique ecologies and habitats from which they were isolated. All 76 genomes were sequenced entirely using PacBio technology. This technology captures not only the nu-

cleotide identity of each residue in the genome, but so too the methylation/modification state of each nucleotide.

5.1 Future directions

While there was no indisputable evidence of a single BGC driving community interactions, we did discover several BGCs that are conditionally expressed in some communities. Moreover, some of these BGCs are yet to be characterized. It would be a worthwhile endeavor for to investigate these BGCs, and take them through full characterization of the structure of their molecular products. Our approach to studying gene expression within communities could be used as a tool for the induction, and characterization, of large or complex BGCs. Most BGCs are not expressed in lab conditions; known to the field as 'cryptic' or 'silent' BGCs. While methods exist to clone and heterologously express large segments of DNA[194–196], they are sensitive to fragment size, and often require laborious methods. Additionally, heterologous expression of BGCs relies on a degree of independence, in that all required inputs (i.e. precursors) must either be packaged in the BGC, or present in the heterologous host genome. Community-level induction of cryptic BGCs, within the native host, alleviates these issues. Further, while not guaranteed, the metatranscriptomics data from the community may be used in conjunction with phenotypic data (e.g. competition assays, inhibition assays, etc.), and molecular structure, to suss out the function of the natural product in the interaction.

That the modified primary metabolic genes in the focal isolate pairs with *Streptomyces sp.* 3212.4 and *Streptomyces sp.* 3212.5 largely agree with what is known about changes in carbon metabolism, nitrogen regulation, and oxidative stress response is promising lead for future studies. Given that a great deal has already been characterized about these regulons, a more novel study may be one focused on the genetic component of the change inducing partner isolates *Streptomyces sp.* 3212.4 and *Streptomyces sp.* 3212.5. In Chapter 2 I provided suggestions for targets. A simple first experiment would be deletion studies in

which the siderophore clusters are knocked out, and changes in morphological phenotype in the focal isolates are analyzed.

There is a wealth of data available within the 76 genomes sequenced via JGI. Future researchers interested in natural product discovery would do well to begin their search within these isolates. With approximately 30 canonical BGCs per genome, there are over 2000 BGCs that could be explored. A good first step would be performing basic genomic analyses (genomic similarity) to determine redundancy, followed by predicting BGCs and determining their similarity within the 76 isolates, as well as to already characterized BGCs. From there I can provide no specific advice as the mode of characterization (fermentation, heterologous expression, refactoring, deletion studies, regulator engineering) depends heavily on the nature (size, biosynthetic class, regulatory logic) of the BGC itself.

To the best of our knowledge, there has been no direct application in the literature of our simulation based metabolic engineering strategy. Since 2018 there have been substantial advances in the field of machine learning, and within deep learning in particular. There have been several examples of applying deep learning to the optimization of multi-gene pathways[197,198]. One very useful application of deep learning in this regard is as an imputation tool. The performance of deep learning models relies on a substantially sized training set. Some deep learning architectures, generative adversarial networks (GANs) in particular, would be a promising addition to the simulation framework. Should one choose to implement our approach, I recommend that they investigate the possibility of using a GAN to generate additional data points to feed the simulation model.

At this point there is no evidence in the field that a generalized set of rules to establish and maintain disease-suppressive soils exists. Given the biological, chemical, and physical complexity of both the soil and rhizosphere microbiome, great advances in high-throughput genotyping, phenotyping, and culturing will be necessary if we hope to fully understand the mechanisms that drive DSS. The work featured in this dissertation highlights two key aspects that I believe will set the theme for the development of this field. Engineering DSS, beyond inoculating with high-levels of a single isolate, will require looking

beyond what is encoded in a handful of genomes, and forming a deep understanding of community interactions.

References

[1] FAO. *The future of food and agriculture – Alternative pathways to 2050*. 2018. ISBN 9789251301586.

[2] David Tilman, Christian Balzer, Jason Hill, and Belinda L. Befort. Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences of the United States of America*, 108:20260–20264, 2011. ISSN 00278424. doi: 10.1073/pnas.1116437108.

[3] Deepak K. Ray, Nathaniel D. Mueller, Paul C. West, and Jonathan A. Foley. Yield trends are insufficient to double global crop production by 2050. *PLoS ONE*, 8, 2013. ISSN 19326203. doi: 10.1371/journal.pone.0066428.

[4] Serge Savary, Laetitia Willocquet, Sarah Jane Pethybridge, Paul Esker, Neil McRoberts, and Andy Nelson. The global burden of pathogens and pests on major food crops. *Nature Ecology and Evolution*, 3:430–439, 2019. ISSN 2397334X. doi: 10.1038/s41559-018-0793-y.

[5] E. C. Oerke. Crop losses to pests. *Journal of Agricultural Science*, 144:31–43, 2006. ISSN 00218596. doi: 10.1017/S0021859605005708.

[6] Mikhail A. Beketov, Ben J. Kefford, Ralf B. Schäfer, and Matthias Liess. Pesticides reduce regional biodiversity of stream invertebrates. *Proceedings of the National Academy of Sciences of the United States of America*, 110:11039–11043, 2013. ISSN 00278424. doi: 10.1073/pnas.1305618110.

[7] Sebastian Stehle and Ralf Schulz. Agricultural insecticides threaten surface waters at the global scale. *Proceedings of the National Academy of Sciences of the United States of America*, 112:5750–5755, 2015. ISSN 10916490. doi: 10.1073/pnas.1500232112.

[8] George Francis Atkinson. *Some diseases of cotton*. Agricultural Experiment Station of the Agricultural and Mechanical College, 1892.

[9] A W Henry. The natural microflora of the soil in relation to the foot-rot problem of wheat. *Canadian Journal of Research*, 4:69–77, 1 1931. ISSN 1923-4287. doi: 10.1139/cjr31-006. doi: 10.1139/cjr31-006.

[10] Rodrigo Mendes, Marco Kruijt, Irene De Bruijn, Ester Dekkers, Menno Van Der Voort, Johannes HM Schneider, Yvette M Piceno, Todd Z DeSantis, Gary L Andersen, Peter AHM Bakker, et al. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science*, 332(6033):1097–1100, 2011.

[11] David J. Newman and Gordon M. Cragg. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products*, 83:770–803, 2020. ISSN 15206025. doi: 10.1021/acs.jnatprod.9b01285.

[12] Kang Zhou, Kangjian Qiao, Steven Edgar, and Gregory Stephanopoulos. Distributing a metabolic pathway among a microbial consortium enhances production of natural products. *Nature biotechnology*, 33(4):377–383, 2015.

[13] Trevor G Johnston, Shuo-Fu Yuan, James M Wagner, Xiunan Yi, Abhijit Saha, Patrick Smith, Alshakim Nelson, and Hal S Alper. Compartmentalized microbes and co-cultures in hydrogels for on-demand bioproduction and preservation. *Nature Communications*, 11(1):563, 2020.

[14] Cunhu Wang, Yanjun Li, Mingjia Li, Kefei Zhang, Wenjing Ma, Lei Zheng, Hanyu Xu, Baofeng Cui, Ran Liu, Yongqing Yang, Yongjia Zhong, and Hong Liao. Functional assembly of root-associated microbial consortia improves nutrient efficiency and yield in soybean. *Journal of Integrative Plant Biology*, 63(6):1021–1035, 2021. ISSN 1744-7909. doi: 10.1111/jipb.13073. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jipb.13073.

[15] Ashok Kumar, Bihari Ram Maurya, and Richa Raghuwanshi. The microbial consortium of indigenous rhizobacteria improving plant health, yield and nutrient content in wheat (*Triticum aestivum*). *Journal of Plant Nutrition*, 44(13):1942–1956, August 2021. ISSN 0190-4167. doi: 10.1080/01904167.2021.1884706. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01904167.2021.1884706.

[16] Anuj Rana, Baljeet Saharan, Lata Nain, Radha Prasanna, and Yashbir S. Shivay. Enhancing micronutrient uptake and yield of wheat through bacterial PGPR consortia. *Soil Science and Plant Nutrition*, 58(5):573–582, October 2012. ISSN 0038-0768. doi: 10.1080/00380768.2012.716750. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00380768.2012.716750.

[17] Ali Raza, Ali Hassan, Waheed Akram, Tehmina Anjum, Zill-e-Huma Aftab, and Basharat Ali. Seed coating with the synthetic consortium of beneficial *Bacillus* microbes improves seedling growth and manages *Fusarium* wilt disease. *Scientia Horticulturae*, 325:112645, February 2024. ISSN 0304-4238. doi: 10.1016/j.scienta.2023.112645.

[18] Roeland L. Berendsen, Gilles Vismans, Ke Yu, Yang Song, Ronnie de Jonge, Wilco P. Burgman, Mette Burmølle, Jakob Herschend, Peter A. H. M. Bakker, and Corné M. J. Pieterse. Disease-induced assemblage of a plant-beneficial bacterial consortium. *The ISME Journal*, 12(6):1496–1507, June 2018. ISSN 1751-7370. doi: 10.1038/s41396-018-0093-1. Number: 6 Publisher: Nature Publishing Group.

[19] Sangay Tshewang, Zed Rengel, Kadambot HM Siddique, and Zakaria M Solaiman. Microbial consortium inoculant increases pasture grasses yield in low-phosphorus soil by influencing root morphology, rhizosphere carboxylate exudation and mycorrhizal colonisation. *Journal of the Science of Food and Agriculture*, 102(2):540–549, 2022. ISSN 1097-0010. doi: 10.1002/jsfa.11382. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jsfa.11382.

[20] Jay Ram Lamichhane, David Camilo Corrales, and Elias Soltani. Biological seed treatments promote crop establishment and yield: a global meta-analysis. *Agronomy for Sustainable Development*, 42(3):45, May 2022. ISSN 1773-0155. doi: 10.1007/s13593-022-00761-z.

[21] Jos M. Raaijmakers and Mark Mazzola. Soil immune responses. *Science*, 352(6292):1392–1393, June 2016. doi: 10.1126/science.aaf3252. Publisher: American Association for the Advancement of Science.

[22] Jay Shendure, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss, and Robert H. Waterston. DNA sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, October 2017. ISSN 1476-4687. doi: 10.1038/nature24286. Number: 7676 Publisher: Nature Publishing Group.

[23] Bilal Aslam, Madiha Basit, Muhammad Atif Nisar, Mohsin Khurshid, and Muhammad Hidayat Rasool. Proteomics: Technologies and Their Applications. *Journal of Chromatographic Science*, 55(2):182–196, February 2017. ISSN 0021-9665. doi: 10.1093/chromsci/bmw167.

[24] Xiaojing Liu and Jason W. Locasale. Metabolomics: A Primer. *Trends in Biochemical Sciences*, 42 (4):274–284, April 2017. ISSN 0968-0004. doi: 10.1016/j.tibs.2017.01.004. Publisher: Elsevier.

[25] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, Carla Porto, Amina Bouslimani, Alexey V Melnik, Michael J Meehan, Wei-Ting Liu, Max Crüsemann, Paul D Boudreau, Eduardo Esquenazi, Mario Sandoval-Calderón, Roland D Kersten, Laura A Pace, Robert A Quinn, Katherine R Duncan, Cheng-Chih Hsu, Dimitrios J Floros, Ronnie G Gavilan, Karin Kleigrewe, Trent Northen, Rachel J Dutton, Delphine Parrot, Erin E Carlson, Bertrand Aigle, Charlotte F Michelsen, Lars Jelsbak, Christian Sohlenkamp, Pavel Pevzner, Anna Edlund, Jeffrey McLean, Jörn Piel, Brian T Murphy, Lena Gerwick, Chih-Chuang Liaw, Yu-Liang Yang, Hans-Ulrich Humpf, Maria Maansson, Robert A Keyzers, Amy C Sims, Andrew R Johnson, Ashley M Sidebottom, Brian E Sedio, Andreas Klitgaard, Charles B Larson, Cristopher A Boya P, Daniel Torres-Mendoza, David J Gonzalez, Denise B Silva, Lucas M Marques, Daniel P Demarque, Egle Pociute, Ellis C O'Neill, Enora Briand, Eric J N Helfrich, Eve A Granatosky, Evgenia Glukhov, Florian Ryffel, Hailey Houson, Hosein Mohimani, Jenan J Kharbush, Yi Zeng, Julia A Vorholt, Kenji L Kurita, Pep Charusanti, Kerry L McPhail, Kristian Fog Nielsen, Lisa Vuong, Maryam Elfeki, Matthew F Traxler, Niclas Engene, Nobuhiro Koyama, Oliver B Vining, Ralph Baric, Ricardo R Silva, Samantha J Mascuch, Sophie Tomasi, Stefan Jenkins, Venkat Macherla, Thomas Hoffman, Vinayak Agarwal, Philip G Williams, Jingqui Dai, Ram Neupane, Joshua

Gurr, Andrés M C Rodríguez, Anne Lamsa, Chen Zhang, Kathleen Dorrestein, Brendan M Duggan, Jehad Almaliti, Pierre-Marie Allard, Prasad Phapale, Louis-Felix Nothias, Theodore Alexandrov, Marc Litaudon, Jean-Luc Wolfender, Jennifer E Kyle, Thomas O Metz, Tyler Peryea, Dac-Trung Nguyen, Danielle VanLeer, Paul Shinn, Ajit Jadhav, Rolf Müller, Katrina M Waters, Wenyuan Shi, Xueting Liu, Lixin Zhang, Rob Knight, Paul R Jensen, Bernhard Ø Palsson, Kit Pogliano, Roger G Linington, Marcelino Gutiérrez, Norberto P Lopes, William H Gerwick, Bradley S Moore, Pieter C Dorrestein, and Nuno Bandeira. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnology*, 34:828–837, 8 2016. ISSN 1087-0156. doi: 10.1038/nbt.3597.

[26] Ilija Dukovski, Djordje Bajić, Jeremy M. Chacón, Michael Quintin, Jean C. C. Vila, Snorre Sulheim, Alan R. Pacheco, David B. Bernstein, William J. Riehl, Kirill S. Korolev, Alvaro Sanchez, William R. Harcombe, and Daniel Segrè. A metabolic modeling platform for the computation of microbial ecosystems in time and space (COMETS). *Nature Protocols*, 16(11):5030–5082, November 2021. ISSN 1750-2799. doi: 10.1038/s41596-021-00593-3. Number: 11 Publisher: Nature Publishing Group.

[27] Christopher E. Lawson, William R. Harcombe, Roland Hatzenpichler, Stephen R. Lindemann, Frank E. Löffler, Michelle A. O'Malley, Héctor García Martín, Brian F. Pfleger, Lutgarde Raskin, Ophelia S. Venturelli, David G. Weissbrodt, Daniel R. Noguera, and Katherine D. McMahon. Common principles and best practices for engineering microbiomes. *Nature Reviews Microbiology*, 9 2019. ISSN 1740-1526. doi: 10.1038/s41579-019-0255-9.

[28] Daniel C. Schlatter, Zewei Song, Patricia Vaz-Jauri, and Linda L. Kinkel. Inhibitory interaction networks among coevolved *Streptomyces* populations from prairie soils. *PLOS ONE*, 14: e0223779, 10 2019. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0223779.

[29] Jack A. Lewis and G. C. Papavizas. Biocontrol of plant diseases: the approach for tomorrow. *Crop Protection*, 10:95–105, 1991. ISSN 02612194. doi: 10.1016/0261-2194(91)90055-V.

[30] Charles Morrow Wilson. *Roots: Miracles Below.* Doubleday & Co., 1968.

[31] Milton N Schroth and Joseph G Hancock. Disease-suppressive soil and root-colonizing bacteria. *Science*, 216(4553):1376–1381, 1982.

[32] William E Finch-Savage and George W Bassel. Seed vigour and crop establishment: extending performance beyond adaptation. *Journal of experimental botany*, 67(3):567–591, 2016.

[33] Julien Papaïx, Jeremy J. Burdon, Jiasui Zhan, and Peter H. Thrall. Crop pathogen emergence and evolution in agro-ecological landscapes. *Evolutionary Applications*, 8:385–402, 2015. ISSN 17524571. doi: 10.1111/eva.12251.

[34] Linda L. Kinkel, Matthew G. Bakker, and Daniel C. Schlatter. A coevolutionary framework for managing disease-suppressive soils. *Annual Review of Phytopathology*, 49:47–67, 2011. ISSN 0066-4286. doi: 10.1146/annurev-phyto-072910-095232.

[35] Blanca B Landa, Dmitri M Mavrodi, Linda S Thomashow, and David M Weller. Interactions between strains of 2,4-diacetylphloroglucinol-producing *Pseudomonas fluorescens* in the rhizosphere of wheat. *Phytopathology*, 93:982–994, 2003. ISSN 0031-949X. doi: 10.1094/PHYTO.2003.93.8.982.

[36] Claude Alabouvette, Philippe Lemanceau, and Christian Steinberg. Recent advances in the biological control of Fusarium wilts. *Pesticide Science*, 37:365–373, 1993. ISSN 0031613X. doi: 10.1002/ps.2780370409.

[37] J D Menzies. Occurence and transfer of a biological factor in soil that suppresses potato scab. *Phytopathology*, 1959.

[38] Mark Mazzola. Mechanisms of natural soil suppressiveness to soilborne diseases. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 81:557–564, 2002. ISSN 00036072. doi: 10.1023/A:1020557523557.

[39] H. Murakami, S. Tsushima, and Y. Shishido. Soil suppressiveness to clubroot disease of chinese cabbage caused by *Plasmodiophora brassicae*. *Soil Biology and Biochemistry*, 32:1637–1642, 2000. ISSN 00380717. doi: 10.1016/S0038-0717(00)00079-1.

[40] David M Weller, Jos M Raaijmakers, Brian B McSpadden Gardener, and Linda S Thomashow. Microbial populations responsible for specific soil suppressiveness to plant pathogens. *Annual review of phytopathology*, 40:309–348, 2002. ISSN 0066-4286. doi: 10.1146/annurev.phyto.40.030402.110010.

[41] Mark Mazzola. Assessment and management of soil microbial community structure for disease suppression. *Annual review of phytopathology*, 42:35–59, 2004. ISSN 0066-4286. doi: 10.1146/annurev.phyto.42.040803.140408.

[42] Leonardo De La Fuente, Blanca B Landa, and David M Weller. Host crop affects rhizosphere colonization and competitiveness of 2,4-diacetylphloroglucinol-producing *Pseudomonas fluorescens*. *Phytopathology*, 96:751–762, 2006. ISSN 0031-949X. doi: 10.1094/PHYTO-96-0751.

[43] Céline Janvier, François Villeneuve, Claude Alabouvette, Véronique Edel-Hermann, Thierry Mateille, and Christian Steinberg. Soil health through soil disease suppression: Which strategy from descriptors to indicators? *Soil Biology and Biochemistry*, 39:1–23, 2007. ISSN 00380717. doi: 10.1016/j.soilbio.2006.07.001.

[44] Janet L Schottel, Kyoko Shimizu, and Linda L Kinkel. Relationships of in vitro pathogen inhibition and soil colonization to potato scab biocontrol by antagonistic *Streptomyces spp. Biological Control*, 20:102–112, 2001. ISSN 10499644. doi: 10.1006/bcon.2000.0893.

[45] Matthew G. Bakker, Lindsey Otto-Hanson, A. J. Lange, James M. Bradeen, and Linda L. Kinkel. Plant monocultures produce more antagonistic soil *Streptomyces* communities than high-diversity

plant communities. *Soil Biology and Biochemistry*, 65:304–312, 2013. ISSN 00380717. doi: 10.1016/j.soilbio.2013.06.007.

[46] Linda L. Kinkel, Daniel C. Schlatter, Kun Xiao, and Anita D. Baines. Sympatric inhibition and niche differentiation suggest alternative coevolutionary trajectories among streptomycetes. *ISME Journal*, 8:249–256, 2014. ISSN 17517362. doi: 10.1038/ismej.2013.175.

[47] BE Paulsrud. Characterization of antagonistic *Streptomyces spp.* from potato scab-suppressive soils and evaluation of their biological potential against potato and non-potato pathogens. pages 1–72, 1996.

[48] EW Buxton and JB Kendrick Jr. A method of isolating *Pythium spp.* and *Fusarium oxysporum* from soil. *Annals of Applied Biology*, pages 215–221, 1963.

[49] Kun Xiao, Linda L. Kinkel, and Deborah A. Samac. Biological control of *Phytophthora* root rots on alfalfa and soybean with *Streptomyces*. *Biological Control*, 23:285–295, 2002. ISSN 10499644. doi: 10.1006/bcon.2001.1015.

[50] Jerry F Franklin, Caroline S Bledsoe, James T Callahan, Jerry F Franklin, Caroline S Bledsoe, and James T Callahan. Contributions of the long-term ecological research program provide crucial comparative analyses. 40:509–523, 2008.

[51] Adil Essarioui, Nicholas LeBlanc, Harold C Kistler, and Linda L Kinkel. Plant community richness mediates inhibitory interactions and resource competition between *Streptomyces* and *Fusarium* populations in the rhizosphere. *Microbial Ecology*, 74:157–167, 2017. ISSN 1432-184X. doi: 10.1007/s00248-016-0907-5.

[52] Patricia Vaz Jauri and Linda L. Kinkel. Nutrient overlap, genetic relatedness and spatial origin influence interaction-mediated shifts in inhibitory phenotype among *Streptomyces spp.* *FEMS Microbiology Ecology*, 2014. ISSN 15746941. doi: 10.1111/1574-6941.12389.

[53] Daniel E Deatherage and Jeffrey E Barrick. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods in molecular biology (Clifton, N.J.)*, 1151:165–88, 2014. ISSN 1940-6029. doi: 10.1007/978-1-4939-0554-6_12.

[54] Bruce J. Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K. Young, and Ashlee M. Earl. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, 9:e112963, 11 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0112963.

[55] James Harrison and David J. Studholme. Recently published *Streptomyces* genome sequences. *Microbial Biotechnology*, 7:373–380, 2014. ISSN 17517915. doi: 10.1111/1751-7915.12143.

[56] S. D. Bentley, K. F. Chater, A.-M. Cerdeño-Tárraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C.-H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M.-A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill, and D. A. Hopwood. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, 417: 141–147, 5 2002. ISSN 0028-0836. doi: 10.1038/417141a.

[57] Juan Pablo Gomez-Escribano, Jean Franco Castro, Valeria Razmilic, Govind Chandra, Barbara Andrews, Juan a. Asenjo, and Mervyn J. Bibb. The *Streptomyces leeuwenhoekii* genome: de novo sequencing and assembly in single contigs of the chromosome, circular plasmid psle1 and linear plasmid psle2. *BMC genomics*, 16:485, 2015. ISSN 1471-2164. doi: 10.1186/s12864-015-1652-8.

[58] Haruo Ikeda, Jun Ishikawa, Akiharu Hanamoto, Mayumi Shinose, Hisashi Kikuchi, Tadayoshi Shiba, Yoshiyuki Sakaki, Masahira Hattori, and Satoshi Ōmura. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nature Biotechnology*, 21:526–531, 2003. ISSN 10870156. doi: 10.1038/nbt820.

[59] Nestor Zaburannyi, Mariia Rabyk, Bohdan Ostash, Victor Fedorenko, and Andriy Luzhetskyy. Insights into naturally minimised *Streptomyces albus* j1074 genome. *BMC Genomics*, 15:97, 2014. ISSN 1471-2164. doi: 10.1186/1471-2164-15-97.

[60] Christian Rückert, Andreas Albersmeier, Tobias Busche, Sebastian Jaenicke, Anika Winkler, Ólafur H. Fridjónsson, Gudmundur Óli Hreggvidsson, Christophe Lambert, Daniel Badcock, Kristel Bernaerts, Jozef Anne, Anastassios Economou, and Jörn Kalinowski. Complete genome sequence of *Streptomyces lividans* tk24. *Journal of Biotechnology*, 199:21–22, 2015. ISSN 18734863. doi: 10.1016/j.jbiotec.2015.02.004.

[61] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30:2068–2069, 7 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu153.

[62] Gary H Van Domselaar, Paul Stothard, Savita Shrivastava, Joseph A Cruz, AnChi Guo, Xiaoli Dong, Paul Lu, Duane Szafron, Russ Greiner, and David S Wishart. Basys: a web server for automated bacterial genome annotation. *Nucleic acids research*, 33:W455–9, 7 2005. ISSN 1362-4962. doi: 10.1093/nar/gki593.

[63] Jaime Huerta-Cepas, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C Walter, Thomas Rattei, Daniel R Mende, Shinichi Sunagawa, Michael Kuhn, Lars Juhl Jensen, Christian von Mering, and Peer Bork. eggnog 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, 44:D286–D293, 1 2016. ISSN 0305-1048. 10.1093/nar/gkv1248.

[64] Michael J. Smanski, Hui Zhou, Jan Claesen, Ben Shen, Michael A. Fischbach, and Christopher A. Voigt. Synthetic biology to access and expand nature's chemical diversity. *Nature Reviews Microbiology*, 14:135–149, 2016. ISSN 1740-1526. doi: 10.1038/nrmicro.2015.24.

[65] Tilmann Weber, Kai Blin, Srikanth Duddela, Daniel Krug, Hyun Uk Kim, Robert Bruccoleri, Sang Yup Lee, Michael A Fischbach, Rolf Müller, Wolfgang Wohlleben, et al. antismash 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic acids research*, 43(W1):W237–W243, 2015.

[66] Jiang Wang, Y. Yu, Kexuan Tang, Wen Liu, Xinyi He, X. Huang, and Zixin Deng. Identification and analysis of the biosynthetic gene cluster encoding the thiopeptide antibiotic cyclothiazomycin in *Streptomyces hygroscopicus* 10-22. *Applied and environmental microbiology*, 76: 2335–2344, 2010. ISSN 10985336. doi: 10.1128/AEM.01790-09.

[67] Carlos Olano, Cristina Gómez, María Pérez, Martina Palomino, Antonio Pineda-Lucena, Rodrigo J. Carbajo, Alfredo F. Braña, Carmen Méndez, and José A. Salas. Deciphering biosynthesis of the rna polymerase inhibitor streptolydigin and generation of glycosylated derivatives. *Chemistry and Biology*, 16:1031–1044, 2009. ISSN 10745521. doi: 10.1016/j.chembiol.2009.09.015.

[68] Nathan A. Magarvey, Brad Haltli, Min He, Michael Greenstein, and John A. Hucul. Biosynthetic pathway for mannopeptimycins, lipoglycopeptide antibiotics active against drug-resistant gram-positive pathogens. *Antimicrobial Agents and Chemotherapy*, 50:2167–2177, 2006. ISSN 00664804. doi: 10.1128/AAC.01545-05.

[69] Anthony W. Goering, Ryan A. McClure, James R. Doroghazi, Jessica C. Albright, Nicole A. Haverland, Yongbo Zhang, Kou-San Ju, Regan J. Thomson, William W. Metcalf, and Neil L. Kelleher. Metabologenomics: Correlation of microbial gene clusters with metabolites drives discovery of a nonribosomal peptide with an unusual amino acid monomer. *ACS Central Science*, 2:99–108, 2016. ISSN 2374-7943. doi: 10.1021/acscentsci.5b00331.

[70] Sylvie Lautru, Robert J Deeth, Lianne M Bailey, and Gregory L Challis. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nature Chemical Biology*, 1: 265–269, 2005. ISSN 1552-4450. doi: 10.1038/nchembio731.

[71] Chieh M. Wang and David E. Cane. Biochemistry and molecular genetics of the biosynthesis of the earthy odorant methylisoborneol in *Streptomyces coelicolor*. *Journal of the American Chemical Society*, 130:8908–8909, 2008. ISSN 00027863. doi: 10.1021/ja803639g.

[72] Chitose Maruyama, Junya Toyoda, Yasuo Kato, Miho Izumikawa, Motoki Takagi, Kazuo Shin-ya, Hajime Katano, Takashi Utagawa, and Yoshimitsu Hamano. A stand-alone adenylation domain forms amide bonds in streptothricin biosynthesis. *Nature Chemical Biology*, 8:791–797, 2012. ISSN 1552-4450. doi: 10.1038/nchembio.1040.

[73] Francisco Barona-Gómez, Ursula Wong, Anastassios E. Giannakopulos, Peter J. Derrick, and Gregory L. Challis. Identification of a cluster of genes that directs desferrioxamine biosynthesis in *Streptomyces coelicolor* m145. *Journal of the American Chemical Society*, 126:16282–16283, 2004. ISSN 00027863. doi: 10.1021/ja045774k.

[74] Jiaoyang Jiang, Xiaofei He, and David E Cane. Biosynthesis of the earthy odorant geosmin by a bifunctional *Streptomyces coelicolor* enzyme. *Nature Chemical Biology*, 3:711–715, 2007. ISSN 1552-4450. doi: 10.1038/nchembio.2007.29.

[75] Gabriele Siedenburg and Dieter Jendrossek. Squalene-hopene cyclases. *Applied and Environmental Microbiology*, 77(12):3905–3915, 2011. doi: 10.1128/AEM.00300-11.

[76] Shinya Kodani, Michael E Hudson, Marcus C Durrant, Mark J Buttner, Justin R Nodwell, and Joanne M Willey. The sapb morphogen is a lantibiotic-like peptide derived from the product of the developmental gene *ramS* in *Streptomyces coelicolor*. *Proceedings of the National Academy of Sciences*, 101(31):11448–11453, 2004.

[77] Naomi Ofer, Marina Wishkautzan, Michael Meijler, Ying Wang, Alexander Speer, Michael Niederweis, and Eyal Gur. Ectoine biosynthesis in *Mycobacterium smegmatis*. *Applied and Environmental Microbiology*, 78:7483–7486, 2012. ISSN 00992240. doi: 10.1128/AEM.01318-12.

[78] KC Gupta and IC Chopra. *Streptomyces katrae* - a new species of *Streptomyces* isolated from soil. *Indian Journal of Microbiology*, 3:1–4, 1963.

[79] W E Grundy, Alma L Whitman, Elbina G Rdzok, E J Rdzok, Marjorie E Hanes, J C Sylvester, et al. Actithiazic acid. i. microbiological studies. *Antibiotics & Chemotherapy*, 2:399–408, 1952.

[80] C Deboer, A Dietz, GM Savage, and WS Silver. Streptolydigin, a new antimicrobial antibiotic. i. biologic studies of streptolydigin. *Antibiotics annual*, 3:886–892, 1955.

[81] Howard M Salis, Ethan a Mirsky, and Christopher a Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nature biotechnology*, 27:946–950, 2009. ISSN 1087-0156. doi: 10.1038/nbt.1568.

[82] Fu-Sheng Wang, Thomas S Whittam, and Robert K Selander. Evolutionary genetics of the isocitrate dehydrogenase gene (icd) in *Escherichia coli* and *Salmonella enterica*. *Journal of bacteriology*, 179(21):6551–6559, 1997.

[83] E Fidelma Boyd, Jia Li, Howard Ochman, and Robert K Selander. Comparative genetics of the inv-spa invasion gene complex of *Salmonella enterica*. *Journal of bacteriology*, 179(6):1985–1991, 1997.

[84] Edward J. Feil, Jessica E. Cooper, Hajo Grundmann, D. Ashley Robinson, Mark C. Enright, Tony Berendt, Sharon J. Peacock, John Maynard Smith, Michael Murphy, Brian G. Spratt, Catrin E. Moore, and Nicholas P.J. Day. How clonal is *Staphylococcus aureus*? *Journal of Bacteriology*, 185:3307–3316, 2003. ISSN 00219193. doi: 10.1128/JB.185.11.3307-3316.2003.

[85] Edward J. Feil. Small change: Keeping pace with microevolution. *Nature Reviews Microbiology*, 2:483–495, 2004. ISSN 17401526. doi: 10.1038/nrmicro904.

[86] James R Doroghazi, Jessica C Albright, Anthony W Goering, Kou-San Ju, Robert R Haines, Konstantin a Tchalukov, David P Labeda, Neil L Kelleher, and William W Metcalf. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nature Chemical Biology*, 10, 9 2014. ISSN 1552-4450. doi: 10.1038/nchembio.1659.

[87] Joanne M. Willey and Alisa A. Gaskell. Morphogenetic signaling molecules of the streptomycetes. *Chemical Reviews*, 111:174–187, 2011. ISSN 00092665. doi: 10.1021/cr1000404.

[88] Eriko Takano. γ-butyrolactones: *Streptomyces* signalling molecules regulating antibiotic production and differentiation. *Current opinion in microbiology*, 9(3):287–294, 2006.

[89] Tatsuo Haneishi, Akira Terahara, Kiyoshi Hamano, and Mamoru Arai. New antibiotics, methylenomycins a and b. *The Journal of Antibiotics*, 27:400–407, 1974. ISSN 0021-8820. doi: 10.7164/antibiotics.27.400.

[90] Kenji Arakawa, Naoto Tsuda, Akihiro Taniguchi, and Haruyasu Kinashi. The butenolide signaling molecules srb1 and srb2 induce lankacidin and lankamycin production in *Streptomyces rochei*. *ChemBioChem*, 13(10):1447–1457, 2012.

[91] J Guijarro, R Santamaria, A Schauer, and R Losick. Promoter determining the timing and spatial localization of transcription of a cloned *Streptomyces coelicolor* gene encoding a spore-associated polypeptide. *Journal of bacteriology*, 170:1895–901, 4 1988. ISSN 0021-9193. doi: 10.1128/JB.170.4.1895-1901.1988.

[92] Eliseo Recio, Angel Colinas, Angel Rumbero, Jesús F Aparicio, and Juan F Martín. Pi factor, a novel type quorum-sensing inducer elicits pimaricin production in *Streptomyces natalensis*. *The Journal of biological chemistry*, 279:41586–93, 10 2004. ISSN 0021-9258. doi: 10.1074/jbc.M402340200.

[93] Matthew T.G. Holden, Siri Ram Chhabra, Rocky De Nys, Paul Stead, Nigel J. Bainton, Philip J. Hill, Mike Manefield, Naresh Kumar, Maurice Labatte, Dacre England, Scott Rice, Mike Givskov, George P.C. Salmond, Gordon S.A.B. Stewart, Barrie W. Bycroft, Staffan Kjelleberg, and Paul Williams. Quorum-sensing cross talk: isolation and chemical characterization of cyclic dipeptides from *Pseudomonas aeruginosa* and other gram-negative bacteria. *Molecular Microbiology*, 33:1254–1266, 3 2002. ISSN 0950382X. doi: 10.1046/j.1365-2958.1999.01577.x.

[94] Diego Romero, Matthew F. Traxler, Daniel López, and Roberto Kolter. Antibiotics as signal molecules, 2011. ISSN 00092665.

[95] Julian Davies, George B Spiegelman, and Grace Yim. The world of subinhibitory antibiotic concentrations. *Current opinion in microbiology*, 9(5):445–453, 2006.

[96] Grace Yim, Helena Huimi Wang, and Julian Davies. Antibiotics as signalling molecules. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362:1195–1200, 7 2007. ISSN 09628436. doi: 10.1098/RSTB.2007.2044.

[97] Peter Cimermancic, Marnix H. Medema, Jan Claesen, Kenji Kurita, Laura C. Wieland Brown, Konstantinos Mavrommatis, Amrita Pati, Paul A. Godfrey, Michael Koehrsen, Jon Clardy, Bruce W. Birren, Eriko Takano, Andrej Sali, Roger G. Linington, and Michael A. Fischbach. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, 158:412–421, 7 2014. ISSN 0092-8674. doi: 10.1016/J.CELL.2014.06.034.

[98] Peter J. Rutledge and Gregory L. Challis. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nature Reviews Microbiology*, 13:509–523, 2015. ISSN 1740-1526. doi: 10.1038/nrmicro3496.

[99] H Chiapello, I Bourgait, F Sourivong, G Heuclin, A Gendrault-Jacquemard, M-A Petit, and M El Karoui. Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops. *BMC Bioinformatics*, 6:171, 7 2005. ISSN 14712105. doi: 10.1186/1471-2105-6-171.

[100] Mohammed Sebaihia, Brendan W Wren, Peter Mullany, Neil F Fairweather, Nigel Minton, Richard Stabler, Nicholas R Thomson, Adam P Roberts, Ana M Cerdeño-Tárraga, Hongmei Wang, Matthew TG Holden, Anne Wright, Carol Churcher, Michael A Quail, Stephen Baker, Nathalie Bason, Karen Brooks, Tracey Chillingworth, Ann Cronin, Paul Davis, Linda Dowd, Audrey Fraser, Theresa Feltwell, Zahra Hance, Simon Holroyd, Kay Jagels, Sharon Moule, Karen Mungall, Claire Price, Ester Rabbinowitsch, Sarah Sharp, Mark Simmonds, Kim Stevens, Louise Unwin, Sally Whithead, Bruno Dupuy, Gordon Dougan, Bart Barrell, and Julian Parkhill. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nature Genetics*, 38:779–786, 7 2006. ISSN 1061-4036. doi: 10.1038/ng1830.

[101] R A Welch, V Burland, G Plunkett, P Redford, P Roesch, D Rasko, E L Buckles, S-R Liou, A Boutin, J Hackett, D Stroud, G F Mayhew, D J Rose, S Zhou, D C Schwartz, N T Perna, H L T Mobley, M S Donnenberg, and F R Blattner. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 99:17020–4, 12 2002. ISSN 0027-8424. doi: 10.1073/pnas.252529799.

[102] Michael J Smanski, Jeffrey Casper, Ryan M Peterson, Zhiguo Yu, Scott R Rajski, and Ben Shen. Expression of the platencin biosynthetic gene cluster in heterologous hosts yielding new platencin congeners. *Journal of natural products*, 75(12):2158–2167, 2012.

[103] Ute Galm and Ben Shen. Expression of biosynthetic gene clusters in heterologous hosts for natural product production and combinatorial biosynthesis, 2006. ISSN 1746-0441.

[104] Encyclopedia of Life. Encyclopedia of life. http://eol.org, 2016. Accessed: 15 January 2016.

[105] Olivier Tenaillon, Jeffrey E. Barrick, Noah Ribeck, Daniel E. Deatherage, Jeffrey L. Blanchard, Aurko Dasgupta, Gabriel C. Wu, Sébastien Wielgoss, Stéphane Cruveiller, Claudine Médigue, Dominique Schneider, and Richard E. Lenski. Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, 536:165–170, 2016. ISSN 14764687. doi: 10.1038/nature18959.

[106] B Jesse Shapiro, Sonia C Timberlake, Gitta Szabó, Martin F Polz, and Eric J Alm. Population genomics of early differentiation of bacteria. *Science*, 336:48–51, 2012. doi: 10.1126/science.1218198.

[107] R James Cook, Linda S Thomashow, David M Weller, Debbie Fujimoto, Mark Mazzola, Gita Bangera, and Dal-Soo Kim. Molecular mechanisms of defense by rhizobacteria against root disease. *Proceedings of the National Academy of Sciences*, 92(10):4197–4201, 1995.

[108] Rodrigo Costa, Monika Götz, Nicole Mrotzek, Jana Lottmann, Gabriele Berg, and Kornelia Smalla. Effects of site and plant species on rhizosphere community structure as revealed by molecular analysis of microbial guilds. *FEMS Microbiology Ecology*, 56:236–249, 5 2006. ISSN 01686496. doi: 10.1111/j.1574-6941.2005.00026.x.

[109] Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, 12 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-119.

[110] Karin Lagesen, Peter Hallin, Einar Andreas Rødland, Hans-Henrik Stærfeldt, Torbjørn Rognes, and David W. Ussery. Rnammer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35:3100–3108, 5 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm160.

[111] Dean Laslett and Bjorn Canback. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic acids research*, 32(1):11–16, 2004.

[112] Jannick Dyrløv Bendtsen, Henrik Nielsen, Gunnar von Heijne, and Søren Brunak. Improved prediction of signal peptides: Signalp 3.0. *Journal of Molecular Biology*, 340:783–795, 7 2004. ISSN 0022-2836. doi: 10.1016/J.JMB.2004.05.028.

[113] Eric P Nawrocki and Sean R Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.

[114] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature Methods*, 12:59–60, 1 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3176.

[115] David P Labeda, JR Doroghazi, K-S Ju, and WW Metcalf. Taxonomic evaluation of *Streptomyces albus* and related species using multilocus sequence analysis and proposals to emend the description of *Streptomyces albus* and describe *Streptomyces pathocidini sp.* nov. *International journal of systematic and evolutionary microbiology*, 64(Pt 3):894, 2014.

[116] Yinping Guo, Wen Zheng, Xiaoying Rong, and Ying Huang. A multilocus phylogeny of the *Streptomyces griseus 16S* rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics. *International journal of systematic and evolutionary microbiology*, 58(1):149–159, 2008.

[117] Sudhir Kumar, Glen Stecher, Koichiro Tamura, and Joel Dudley. Mega7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets downloaded from. *Mol. Biol. Evol*, 33:1870–1874, 2016. doi: 10.1093/molbev/msw054.

[118] Maddur Puttaswamy Raghavendra, Aralakuppe Narayana Santhoshkannada, M P Raghavendra, and A N Santhoshkannada. *Role of Rhizomicrobiome in Maintaining Soil Fertility and Crop Production*. Springer, Cham, 2020. doi: 10.1007/978-3-030-44364-1_19.

[119] Juhi Sharma, Jyoti Goutam, Yogesh Kumar Dhuriya, and Divakar Sharma. *Bioremediation of Industrial Pollutants*. Springer, Singapore, 2021. doi: 10.1007/978-981-15-7455-9_1.

[120] Rattan Lal. Soil carbon sequestration to mitigate climate change. *Geoderma*, 123(1-2):1–22, 2004.

[121] Rodrigo Mendes, Paolina Garbeva, and Jos M. Raaijmakers. The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiology Reviews*, 37:634–663, 9 2013. ISSN 0168-6445. doi: 10.1111/1574-6976.12028.

[122] Stephen C. Heinsch, Szu Yi Hsu, Lindsey Otto-Hanson, Linda Kinkel, and Michael J. Smanski. Complete genome sequences of *Streptomyces spp.* isolated from disease-suppressive soils. *BMC Genomics*, 20:13–16, 2019. ISSN 14712164. doi: 10.1186/s12864-019-6279-8.

[123] Govind Chandra and Keith F. Chater. Developmental biology of *Streptomyces* from the perspective of 100 actinobacterial genome sequences. *FEMS Microbiology Reviews*, 38:345–379, 5 2014. doi: 10.1111/1574-6976.12047.

[124] Yujin Jeong, Ji Nu Kim, Min Woo Kim, Giselda Bucca, Suhyung Cho, Yeo Joon Yoon, Byung Gee Kim, Jung Hye Roe, Sun Chang Kim, Colin P. Smith, and Byung Kwan Cho. The dynamic transcriptional and translational landscape of the model antibiotic producer *Streptomyces coelicolor* A3(2). *Nature Communications 2016 7:1*, 7:1–11, 6 2016. ISSN 2041-1723. doi: 10.1038/ncomms11605.

[125] Marcha L Gatewood, Patricia Bralley, M Ryan Weil, and George H Jones. RNA-Seq and RNA immunoprecipitation analyses of the transcriptome of *Streptomyces coelicolor* identify substrates for RNase III. *Journal of bacteriology*, 194(9):2228–2237, 2012.

[126] Yunzi Luo, Lu Zhang, Katherine W Barton, and Huimin Zhao. Systematic identification of a panel of strong constitutive promoters from *Streptomyces albus*. *ACS synthetic biology*, 4(9):1001–1010, 2015.

[127] Woori Kim, Soonkyu Hwang, Namil Lee, Yongjae Lee, Suhyung Cho, Bernhard Palsson, and Byung Kwan Cho. Transcriptome and translatome profiles of *Streptomyces* species in different growth phases. *Scientific Data 2020 7:1*, 7:1–12, 5 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0476-9.

[128] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:1–21, 12 2014. ISSN 1474760X. doi: 10.1186/S13059-014-0550-8/FIGURES/9.

[129] Kai Blin, Simon Shaw, Katharina Steinke, Rasmus Villebro, Nadine Ziemert, Sang Yup Lee, Marnix H. Medema, and Tilmann Weber. antismash 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research*, 47:W81–W87, 7 2019. ISSN 0305-1048. doi: 10.1093/NAR/GKZ310.

[130] Nitsara Karoonuthaisiri, David Weaver, Jianqiang Huang, Stanley N. Cohen, and Camilla M. Kao. Regional organization of gene expression in *Streptomyces coelicolor*. *Gene*, 353:53–66, 6 2005. ISSN 0378-1119. doi: 10.1016/J.GENE.2005.03.042.

[131] Virginia S. Lioy, Jean Noël Lorenzi, Soumaya Najah, Thibault Poinsignon, Hervé Leh, Corinne Saulnier, Bertrand Aigle, Sylvie Lautru, Annabelle Thibessard, Olivier Lespinet, Pierre Leblond, Yan Jaszczyszyn, Kevin Gorrichon, Nelle Varoquaux, Ivan Junier, Frédéric Boccard, Jean Luc Pernodet, and Stéphanie Bury-Moné. Dynamics of the compartmentalized *Streptomyces* chromosome during metabolic differentiation. *Nature Communications 2021 12:1*, 12:1–14, 9 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25462-1.

[132] Adelfia Talà, Fabrizio Damiano, Giuseppe Gallo, Eva Pinatel, Matteo Calcagnile, Mariangela Testini, Daniela Fico, Daniela Rizzo, Alberto Sutera, Giovanni Renzone, Andrea Scaloni, Gianluca De Bellis, Luisa Siculella, Giuseppe Egidio De Benedetto, Anna Maria Puglia, Clelia Peano, and Pietro Alifano. Pirin: A novel redox-sensitive modulator of primary and secondary metabolism in *Streptomyces*. *Metabolic Engineering*, 48:254–268, July 2018. ISSN 1096-7176. doi: 10.1016/j.ymben.2018.06.008.

[133] Alexandra R. Mey, Elizabeth E. Wyckoff, Vanamala Kanukurthy, Carolyn R. Fisher, and Shelley M. Payne. Iron and Fur Regulation in *Vibrio cholerae* and the Role of Fur in Virulence. *Infection and Immunity*, 73(12):8167–8178, December 2005. ISSN 0019-9567. doi: 10.1128/IAI.73.12.8167-8178.2005.

[134] Büşra Abanoz-Seçgin, Çiğdem Otur, Sezer Okay, and Aslıhan Kurt-Kızıldoğan. The regulatory role of Fur-encoding *SCLAV_3199* in iron homeostasis in *Streptomyces clavuligerus*. *Gene*, 878: 147594, August 2023. ISSN 0378-1119. doi: 10.1016/j.gene.2023.147594.

[135] Yaqing Cheng, Renjun Yang, Mengya Lyu, Shiwei Wang, Xingchao Liu, Ying Wen, Yuan Song, Jilun Li, and Zhi Chen. IdeR, a DtxR Family Iron Response Regulator, Controls Iron Homeostasis, Morphological Differentiation, Secondary Metabolism, and the Oxidative Stress Re-

sponse in *Streptomyces avermitilis*. *Applied and Environmental Microbiology*, 84(22):e01503–18, October 2018. doi: 10.1128/AEM.01503-18. Publisher: American Society for Microbiology.

[136] Sedef Tunca, Carlos Barreiro, Juan-José R. Coque, and Juan F. Martín. Two overlapping antiparallel genes encoding the iron regulator DmdR1 and the Adm proteins control siderophore and antibiotic biosynthesis in *Streptomyces coelicolor* A3(2). *The FEBS Journal*, 276 (17):4814–4827, 2009. ISSN 1742-4658. doi: 10.1111/j.1742-4658.2009.07182.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1742-4658.2009.07182.x.

[137] Evan M. F. Shepherdson and Marie A. Elliot. Cryptic specialized metabolites drive *Streptomyces* exploration and provide a competitive advantage during growth with other microbes. *Proceedings of the National Academy of Sciences*, 119(40):e2211052119, October 2022. doi: 10.1073/pnas.2211052119. Publisher: Proceedings of the National Academy of Sciences.

[138] Natalia M. Vior, Carlos Olano, Ignacio García, Carmen Méndez, and José A. Salas. Collismycin A biosynthesis in *Streptomyces sp.* CS40 is regulated by iron levels through two pathway-specific regulators. *Microbiology*, 160(3):467–478, 2014. ISSN 1465-2080. doi: 10.1099/mic.0.075218-0. Publisher: Microbiology Society,.

[139] Masashi Ueki, Ryuichiro Suzuki, Satoshi Takamatsu, Hiroshi Takagi, Masakazu Uramoto, Haruo Ikeda, and Hiroyuki Osada. Nocardamin production by *Streptomyces avermitilis*. *Actinomycetologica*, 23(2):34–39, 2009.

[140] Stephanie E Jones, Christine A Pham, Matthew P Zambri, Joseph Mckillip, Erin E Carlson, and Marie A Elliot. *Streptomyces* volatile compounds influence exploration and microbial community dynamics by altering iron availability. 2019. doi: 10.1128/mBio.

[141] Jiarui Zeng, Ting Xu, Lidan Cao, Chunyi Tong, Xuan Zhang, Dingyi Luo, Shuping Han, Pei Pang, Weibin Fu, Jindong Yan, Xuanming Liu, and Yonghua Zhu. The Role of Iron Competition in the Antagonistic Action of the Rice Endophyte *Streptomyces sporocinereus* OsiSh-2 Against the Pathogen *Magnaporthe oryzae*. *Microbial Ecology*, 76(4):1021–1029, November 2018. ISSN 1432-184X. doi: 10.1007/s00248-018-1189-x.

[142] Scott A. Jarmusch, Diego Lagos-Susaeta, Emtinan Diab, Oriana Salazar, Juan A. Asenjo, Rainer Ebel, and Marcel Jaspars. Iron-meditated fungal starvation by lupine rhizosphere-associated and extremotolerant *Streptomyces sp.* S29 desferrioxamine production. *Molecular Omics*, 17(1): 95–107, 2021. doi: 10.1039/D0MO00084A. Publisher: Royal Society of Chemistry.

[143] Shotaro Hoshino, Hiroyasu Onaka, and Ikuro Abe. Activation of silent biosynthetic pathways and discovery of novel secondary metabolites in actinomycetes by co-culture with mycolic acid-containing bacteria. *Journal of Industrial Microbiology and Biotechnology*, 46:363–374, 3 2019. ISSN 1367-5435. doi: 10.1007/S10295-018-2100-Y.

[144] Rory Stark, Marta Grzelak, and James Hadfield. RNA sequencing: the teenage years. *Nature Reviews Genetics 2019 20:11*, 20:631–656, 7 2019. ISSN 1471-0064. doi: 10.1038/s41576-019-0150-2.

[145] Tobias Kieser, Mervyn J. Bibb, Mark J. Buttner, K. F. Chater, D. A. Hopwood, and John Innes Foundation. *Practical Streptomyces genetics*. John Innes Foundation, 2000. ISBN 9780708406236.

[146] Keith A. Jolley, James E. Bray, and Martin C.J. Maiden. Open-access bacterial population genomics: Bigsdb software, the pubmlst.org website and their applications. *Wellcome Open Research*, 3:124, 2018. ISSN 2398502X. doi: 10.12688/WELLCOMEOPENRES.14826.1.

[147] Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32:1792–1797, 3 2004. ISSN 0305-1048. doi: 10.1093/NAR/GKH340.

[148] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5, 3 2010. ISSN 19326203. doi: 10.1371/journal.pone.0009490.

[149] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[150] Brian Bushnell. BBMap: a fast, accurate, splice-aware aligner, 2014.

[151] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[152] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

[153] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6 (60):3021, 2021. doi: 10.21105/joss.03021.

[154] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[155] Michael Fischbach and Christopher a Voigt. Prokaryotic gene clusters: a rich toolbox for synthetic biology. *Biotechnology journal*, 5:1277–96, 12 2010. ISSN 1860-7314. doi: 10.1002/biot.201000181.

[156] Jeong Wook Lee, Dokyun Na, Jong Myoung Park, Joungmin Lee, Sol Choi, and Sang Yup Lee. Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nature Chemical Biology*, 8:536–546, 2012. ISSN 15524469. doi: 10.1038/nchembio.970.

[157] Michael J Smanski, Swapnil Bhatia, Dehua Zhao, YongJin Park, Lauren B A Woodruff, Georgia Giannoukos, Dawn Ciulla, Michele Busby, Johnathan Calderon, Robert Nicol, D Benjamin Gordon, Douglas Densmore, and Christopher a Voigt. Functional optimization of gene clusters by combinatorial design and assembly. *Nature Biotechnology*, 32:1241–1249, 11 2014. ISSN 1087-0156. doi: 10.1038/nbt.3063.

[158] Jens Nielsen and Jay D. Keasling. Engineering cellular metabolism. *Cell*, 164:1185–1197, 2016. ISSN 10974172. doi: 10.1016/j.cell.2016.02.004.

[159] Harris H. Wang, Farren J. Isaacs, Peter A. Carr, Zachary Z. Sun, George Xu, Craig R. Forest, and George M. Church. Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, 460:894–898, 2009. ISSN 00280836. doi: 10.1038/nature08187.

[160] Yongbo Yuan, Jing Du, and Huimin Zhao. Customized optimization of metabolic pathways by combinatorial transcriptional engineering. *Methods in molecular biology (Clifton, N.J.)*, 985: 177–209, 2013. ISSN 19406029. doi: 10.1007/978-1-62703-299-5_10.

[161] Ran Chao, Shekhar Mishra, Tong Si, and Huimin Zhao. Engineering biological systems using automated biofoundries. *Metabolic Engineering*, 42:98–108, 2017. ISSN 10967184. doi: 10.1016/j.ymben.2017.06.003.

[162] Ahmad S. Khalil, Timothy K. Lu, Caleb J. Bashor, Cherie L. Ramirez, Nora C. Pyenson, J. Keith Joung, and James J. Collins. A synthetic biology framework for programming eukaryotic transcription functions. *Cell*, 150:647–658, 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.05.045.

[163] Sriram Kosuri, Daniel B. Goodman, Guillaume Cambray, Vivek K. Mutalik, Yuan Gao, Adam P. Arkin, Drew Endy, and George M. Church. Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 110:14024–14029, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1301301110.

[164] Vivek K. Mutalik, Joao C. Guimaraes, Guillaume Cambray, Colin Lam, Marc Juul Christoffersen, Quynh Anh Mai, Andrew B. Tran, Morgan Paull, Jay D. Keasling, Adam P. Arkin, and Drew Endy. Precise and reliable gene expression via standard transcription and translation initiation elements. *Nature Methods*, 10:354–360, 2013. ISSN 15487091. doi: 10.1038/nmeth.2404.

[165] Alec A.K. Nielsen, Thomas H. Segall-Shapiro, and Christopher A. Voigt. Advances in genetic circuit design: Novel biochemistries, deep part mining, and precision gene expression. *Current Opinion in Chemical Biology*, 17:878–892, 2013. ISSN 13675931. doi: 10.1016/j.cbpa.2013.10.003.

[166] Theresa Siegl, Bogdan Tokovenko, Maksym Myronovskyi, and Andriy Luzhetskyy. Design, construction and characterisation of a synthetic promoter library for fine-tuned gene expression in actinomycetes. *Metabolic Engineering*, 19:98–106, 2013. ISSN 10967176. doi: 10.1016/j.ymben.2013.07.006.

[167] Amin Espah Borujeni, Anirudh S. Channarasappa, and Howard M. Salis. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Research*, 42:2646–2659, 2014. ISSN 03051048. doi: 10.1093/nar/gkt1139.

[168] Chaoxian Bai, Yang Zhang, Xuejin Zhao, Yiling Hu, Sihai Xiang, Jin Miao, Chunbo Lou, and Lixin Zhang. Exploiting a precise design of universal synthetic modular regulatory elements to unlock the microbial natural products in *Streptomyces*. *Proceedings of the National Academy of Sciences of the United States of America*, 112:12181–6, 9 2015. ISSN 1091-6490. doi: 10.1073/pnas.1511027112.

[169] Heidi Redden and Hal S. Alper. The development and characterization of synthetic minimal yeast promoters. *Nature Communications*, 6:1–9, 2015. ISSN 20411723. doi: 10.1038/ncomms8810.

[170] Alexander J. Diaz de Arce, William L. Noderer, and Clifford L. Wang. Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. *Nucleic Acids Research*, 46:985–994, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx1114.

[171] Parayil Kumaran Ajikumar, Wen-Hai Xiao, Keith E J Tyo, Yong Wang, Fritz Simeon, Effendi Leonard, Oliver Mucha, Too Heng Phon, Blaine Pfeifer, and Gregory Stephanopoulos. Isoprenoid pathway optimization for taxol precursor overproduction in *Escherichia coli*. *Science (New York, N.Y.)*, 330:70–74, 2010. ISSN 0036-8075. doi: 10.1126/science.1191652.

[172] Bradley Walters Biggs, Brecht De Paepe, Christine Nicole S Santos, Marjan De Mey, and Parayil Kumaran Ajikumar. Multivariate modular metabolic engineering for pathway and strain optimization. *Current Opinion in Biotechnology*, 29:156–162, 2014. ISSN 18790429. doi: 10.1016/j.copbio.2014.05.005.

[173] Peng Xu, Qin Gu, Wenya Wang, Lynn Wong, Adam G.W. Bower, Cynthia H. Collins, and Mattheos A.G. Koffas. Modular optimization of multi-gene pathways for fatty acids production in *E. coli*. *Nature Communications*, 4:1408–1409, 2013. ISSN 20411723. doi: 10.1038/ncomms2425.

[174] Liya Liang, Rongming Liu, Andrew D. Garst, Thomas Lee, Violeta Sànchez i. Nogué, Gregg T. Beckham, and Ryan T. Gill. Crispr enabled trackable genome engineering for isopropanol production in *Escherichia coli*. *Metabolic Engineering*, 41:1–10, 2017. ISSN 10967184. doi: 10.1016/j.ymben.2017.02.009.

[175] Chiam Yu Ng, Iman Farasat, Costas D Maranas, and Howard M Salis. Rational design of a synthetic entner–doudoroff pathway for improved and controllable nadph regeneration. *Metabolic engineering*, 29:86–96, 2015.

[176] Hui Zhou, Brenda Vonk, Johannes a. Roubos, Roel a.L. Bovenberg, and Christopher a. Voigt. Algorithmic co-optimization of genetic constructs and growth conditions: application to 6-aca, a potential nylon-6 precursor. *Nucleic Acids Research*, page gkv1071, 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv1071.

[177] Michael E. Lee, Anil Aswani, Audrey S. Han, Claire J. Tomlin, and John E. Dueber. Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. *Nucleic Acids Research*, 41:10668–10678, 2013. ISSN 03051048. doi: 10.1093/nar/gkt809.

[178] Iman Farasat, Manish Kushwaha, Jason Collens, Michael Easterbrook, Matthew Guido, and Howard M Salis. Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria. *Molecular systems biology*, 10:731, 1 2014. ISSN 1744-4292.

[179] Peng Xu, Elizabeth Anne Rizzoni, Se Yeong Sul, and Gregory Stephanopoulos. Improving metabolic pathway efficiency by statistical model-based multivariate regulatory metabolic engineering. *ACS Synthetic Biology*, 6:148–158, 2017. ISSN 21615063. doi: 10.1021/acssynbio.6b00187.

[180] Erik Pitzer and Michael Affenzeller. A comprehensive survey on fitness landscape analysis. *Recent advances in intelligent engineering systems*, pages 161–191, 2012.

[181] Philip A. Romero and Frances H. Arnold. Exploring protein fitness landscapes by directed evolution, 12 2009. ISSN 14710072.

[182] Luis Miguel Rios and Nikolaos V. Sahinidis. Derivative-free optimization: A review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56:1247–1293, 2013. ISSN 09255001. doi: 10.1007/s10898-012-9951-y.

[183] Donald R Jones, C Floudas, and P Pardalos. Encyclopedia of optimization. *DIRECT global optimization*, pages 725–735, 2001.

[184] Edward D. Weinberger. Local properties of kauffman's n-k model: A tunably rugged energy landscape. *Physical Review A*, 44:6399–6413, 1991. ISSN 10502947. doi: 10.1103/PhysRevA.44.6399.

[185] Stuart A Kauffman. *The origins of order: Self-organization and selection in evolution*. Oxford University Press, USA, 1993.

[186] Edward D. Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biological Cybernetics*, 63:325–336, 1990. ISSN 03401200. doi: 10.1007/BF00202749.

[187] Walter Fontana, Peter F. Stadler, Erich G. Bornberg-Bauer, Thomas Griesmacher, Ivo L. Hofacker, Manfred Tacker, Pedro Tarazona, Edward D. Weinberger, and Peter Schuster. RNA folding and combinatory landscapes, 1993. ISSN 1063651X.

[188] Nikolaus Hansen. The cma evolution strategy: a comparing review. *Towards a new evolutionary computation: Advances in the estimation of distribution algorithms*, pages 75–102, 2006.

[189] Alan S Perelson and Catherine A Macken. Protein evolution on partially correlated landscapes. *Proceedings of the National Academy of Sciences*, 92(21):9657–9661, 1995.

[190] Todd S. Freestone and Huimin Zhao. Combinatorial pathway engineering for optimized production of the anti-malarial fr900098. *Biotechnology and Bioengineering*, 2015. ISSN 00063592. doi: 10.1002/bit.25719.

[191] Stefano Cardinale and Adam Paul Arkin. Contextualizing context for synthetic biology - identifying causes of failure of synthetic biological systems, 2012. ISSN 18606768.

[192] Jennifer An Brophy and Christopher A Voigt. Antisense transcription as a tool to tune gene expression. *Molecular systems biology*, 12:854, 2016. ISSN 1744-4292. doi: 10.15252/msb.20156540.

[193] Thomas J. DeWitt and Jin Yoshimura. The fitness threshold model: Random environmental change alters adaptive landscapes. *Evolutionary Ecology*, 12:615–626, 1998. ISSN 02697653. doi: 10.1023/A:1006564911480.

[194] Natalay Kouprina and Vladimir Larionov. Transformation-associated recombination (TAR) cloning for genomics studies and synthetic biology. *Chromosoma*, 125:621–632, 2016.

[195] Wenjun Jiang, Xuejin Zhao, Tslil Gabrieli, Chunbo Lou, Yuval Ebenstein, and Ting F Zhu. Cas9-Assisted Targeting of CHromosome segments CATCH enables one-step targeted cloning of large gene clusters. *Nature communications*, 6(1):8101, 2015.

[196] Gaoyan Wang, Zhiying Zhao, Jing Ke, Yvonne Engel, Yi-Ming Shi, David Robinson, Kerem Bingol, Zheyun Zhang, Benjamin Bowen, Katherine Louie, et al. CRAGE enables rapid activation of biosynthetic gene clusters in undomesticated bacteria. *Nature microbiology*, 4(12):2498–2510, 2019.

[197] Ashty S Karim, Quentin M Dudley, Alex Juminaga, Yongbo Yuan, Samantha A Crowe, Jacob T Heggestad, Shivani Garg, Tanus Abdalla, William S Grubbe, Blake J Rasor, et al. In vitro prototyping and rapid optimization of biosynthetic enzymes for cell design. *Nature Chemical Biology*, 16(8):912–919, 2020.

[198] Zak Costello and Hector Garcia Martin. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *NPJ systems biology and applications*, 4(1): 1–14, 2018.

[199] Stefan Kurtz, Adam Phillippy, Arthur L. Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L. Salzberg. Versatile and open software for comparing large genomes. *Genome biology*, 5:1–9, 1 2004. ISSN 14656914. doi: 10.1186/GB-2004-5-2-R12/FIGURES/3.

[200] Kenji Ueda, Ken Ichi Oinuma, Go Ikeda, Kuniaki Hosono, Yasuo Ohnishi, Sueharu Horinouchi, and Teruhiko Beppu. AmfS, an extracellular peptidic morphogen in *Streptomyces griseus*. *Journal of Bacteriology*, 184:1488–1492, 2002. ISSN 00219193. doi: 10.1128/JB.184.5.1488-1492.2002.

[201] Masanori Funabashi, Nobutaka Funa, and Sueharu Horinouchi. Phenolic lipids synthesized by type III polyketide synthase confer penicillin resistance on *Streptomyces griseus*. *The Journal of biological chemistry*, 283:13983–13991, 5 2008. ISSN 0021-9258 (Print). doi: 10.1074/jbc.M710461200.

[202] Jia Zeng, Richard Decker, and Jixun Zhan. Biochemical characterization of a type III polyketide biosynthetic gene cluster from *Streptomyces toxytricini*. *Applied Biochemistry and Biotechnology*, 166:1020–1033, 2012. ISSN 02732289. doi: 10.1007/s12010-011-9490-x.

[203] Jiangtao Gao, Kou San Ju, Xiaomin Yu, Juan E. Velasquez, Subha Mukherjee, Jaeheon Lee, Changming Zhao, Bradley S. Evans, James R. Doroghazi, William W. Metcalf, and Wilfred A. Van Der Donk. Use of a phosphonate methyltransferase in the identification of the fosfazinomycin biosynthetic gene cluster. *Angewandte Chemie - International Edition*, 53:1334–1337, 2014. ISSN 14337851. doi: 10.1002/anie.201308363.

[204] Tingli Bai, Daozhong Zhang, Shuangjun Lin, Qingshan Long, Yemin Wang, Hongyu Ou, Qianjin Kang, Zixin Deng, Wen Liu, and Meifeng Tao. Operon for biosynthesis of lipstatin, the beta-lactone inhibitor of human pancreatic lipase. *Applied and environmental microbiology*, 80:7473–7483, 12 2014. ISSN 1098-5336 (Electronic). doi: 10.1128/AEM.01765-14.

[205] Shohei Hayashi, Taro Ozaki, Shumpei Asamizu, Haruo Ikeda, Satoshi Omura, Naoya Oku, Yasuhiro Igarashi, Hiroshi Tomoda, and Hiroyasu Onaka. Genome mining reveals a minimum gene set for the biosynthesis of 32-membered macrocyclic thiopeptides lactazoles. *Chemistry & biology*, 21:679–688, 5 2014. ISSN 1879-1301 (Electronic). doi: 10.1016/j.chembiol.2014.03.008.

[206] Hosein Mohimani, Roland D. Kersten, Wei Ting Liu, Mingxun Wang, Samuel O. Purvine, Si Wu, Heather M. Brewer, Ljiljana Pasa-Tolic, Nuno Bandeira, Bradley S. Moore, Pavel A.

Pevzner, and Pieter C. Dorrestein. Automated genome mining of ribosomal peptide natural products. *ACS Chemical Biology*, 9:1545–1551, 7 2014. ISSN 15548937. doi: 10.1021/CB500199H.

[207] Wayne K.W. Chou, Immacolata Fanizza, Takuma Uchiyama, Mamoru Komatsu, Haruo Ikeda, and David E. Cane. Genome mining in *Streptomyces avermitilis*: Cloning and characterization of *sav-76*, the synthase for a new sesquiterpene, avermitilol. *Journal of the American Chemical Society*, 132:8850–8851, 7 2010. ISSN 00027863. doi: 10.1021/JA103087W.

[208] Jan Bursy, Anne U. Kuhlmann, Marco Pittelkow, Holger Hartmann, Mohamed Jebbar, Antonio J. Pierik, and Erhard Bremer. Synthesis and uptake of the compatible solutes ectoine and 5-hydroxyectoine by *Streptomyces coelicolor* A3(2) in response to salt and heat stresses. *Applied and Environmental Microbiology*, 74:7286, 12 2008. ISSN 00992240. doi: 10.1128/AEM.00768-08.

[209] Tin Wein Yu, Yuemao Shen, Robert McDaniel, Heinz G. Floss, Chaitan Khosla, David A. Hopwood, and Bradley S. Moore. Engineered biosynthesis of novel polyketides from *Streptomyces* spore pigment polyketide synthases. *Journal of the American Chemical Society*, 120:7749–7759, 8 1998. ISSN 00027863. doi: 10.1021/JA9803658.

[210] Nobutaka Funa, Masanori Funabashi, Yasuo Ohnishi, and Sueharu Horinouchi. Biosynthesis of hexahydroxyperylenequinone melanin via oxidative aryl coupling by cytochrome P-450 in *Streptomyces griseus*. *Journal of Bacteriology*, 187:8149–8155, 2005. doi: 10.1128/JB.187.23.8149-8155.2005.

[211] Miho Izumikawa, Teppei Kawahara, Noritaka Kagaya, Hideki Yamamura, Masayuki Hayakawa, Motoki Takagi, Masahito Yoshida, Takayuki Doi, and Kazuo Shin-ya. Pyrrolidine-containing peptides, JBIR-126, -148, and -149, from *Streptomyces sp.* nbrc 111228. *Tetrahedron Letters*, 56:5333–5336, 9 2015. ISSN 0040-4039. doi: 10.1016/J.TETLET.2015.07.080.

[212] Yanhua Du, Yemin Wang, Tingting Huang, Meifeng Tao, Zixin Deng, and Shuangjun Lin. Identification and characterization of the biosynthetic gene cluster of polyoxypeptin a, a potent apoptosis inducer. *BMC Microbiology*, 14:1–12, 2014. ISSN 14712180. doi: 10.1186/1471-2180-14-30.

[213] Satoshi Takamatsu, Xin Lin, Ayako Nara, Mamoru Komatsu, David E Cane, and Haruo Ikeda. Characterization of a silent sesquiterpenoid biosynthetic pathway in *Streptomyces avermitilis* controlling epi-isozizaene albaflavenone biosynthesis and isolation of a new oxidized epi-isozizaene metabolite. *Microbial Biotechnology*, 4(2):184–191, 2011.

[214] Zhi Feng, Yasushi Ogasawara, Satoshi Nomura, and Tohru Dairi. Biosynthetic gene cluster of a d-tryptophan-containing lasso peptide, MS-271. *ChemBioChem*, 19:2045–2048, 10 2018. ISSN 1439-7633. doi: 10.1002/CBIC.201800315.

[215] Martha C. Cone, Xihou Yin, Laura L. Grochowski, Morgan R. Parker, and T. Mark Zabriskie. The blasticidin s biosynthesis gene cluster from *Streptomyces griseochromogenes*: Sequence analysis, organization, and initial characterization. *ChemBioChem*, 4:821–828, 2003. ISSN 14394227. doi: 10.1002/cbic.200300583.

[216] Dian Anggraini Suroto, Shigeru Kitani, Masayoshi Arai, Haruo Ikeda, and Takuya Nihira. Characterization of the biosynthetic gene cluster for cryptic phthoxazolin a in *Streptomyces avermitilis*. *PLOS ONE*, 13:e0190973, 1 2018. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0190973.

[217] Haruo Ikeda, Kazuo Shin-Ya, and Satoshi Omura. Genome mining of the *Streptomyces avermitilis* genome and development of genome-minimized hosts for heterologous expression of biosynthetic gene clusters. *Journal of Industrial Microbiology and Biotechnology*, 41:233–250, 2 2014. ISSN 1367-5435. doi: 10.1007/S10295-013-1327-X.

[218] Elizabeth I Parkinson, James H Tryon, Anthony W Goering, Kou-San Ju, Ryan A McClure, Jeremy D Kemball, Sara Zhukovsky, David P Labeda, Regan J Thomson, Neil L Kelleher, William W Metcalf, and Carl R Woese. Discovery of the tyrobetaine natural products and their biosynthetic gene cluster via metabologenomics. *ACS Chem. Biol*, 13:14, 2018. doi: 10.1021/acschembio.7b01089.

[219] Jonathan I. Tietz, Christopher J. Schwalen, Parth S. Patel, Tucker Maxson, Patricia M. Blair, Hua Chia Tai, Uzma I. Zakai, and Douglas A. Mitchell. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nature Chemical Biology 2017 13:5*, 13:470–478, 2 2017. ISSN 1552-4469. doi: 10.1038/nchembio.2319.

[220] Anja Greule, Marija Marolt, Denise Deubel, Iris Peintner, Songya Zhang, Claudia Jessen-Trefzer, Christian De Ford, Sabrina Burschel, Shu Ming Li, Thorsten Friedrich, Irmgard Merfort, Steffen Lüdeke, Philippe Bisel, Michael Müller, Thomas Paululat, and Andreas Bechthold. Wide distribution of foxicin biosynthetic gene clusters in *Streptomyces* strains - an unusual secondary metabolite with various properties. *Frontiers in Microbiology*, 8:221, 2 2017. ISSN 1664302X. doi: 10.3389/FMICB.2017.00221/BIBTEX.

[221] Charles N. Tetzlaff, Zheng You, David E. Cane, Satoshi Takamatsu, Satoshi Omura, and Haruo Ikeda. A gene cluster for biosynthesis of the sesquiterpenoid antibiotic pentalenolactone in *Streptomyces avermitilis*. *Biochemistry*, 45:6179–6186, 5 2006. ISSN 00062960. doi: 10.1021/BI060419N.

[222] Shinya Kodani, Joanna Bicz, Lijiang Song, Robert J. Deeth, Mayumi Ohnishi-Kameyama, Mitsuru Yoshida, Kozo Ochi, and Gregory L. Challis. Structure and biosynthesis of scabichelin, a novel tris-hydroxamate siderophore produced by the plant pathogen *Streptomyces scabies* 87.22. *Organic & Biomolecular Chemistry*, 11:4686–4694, 6 2013. ISSN 14770520. doi: 10.1039/C3OB40536B.

[223] Yingying Wu, Qianjin Kang, Yuemao Shen, Wenjin Su, and Linquan Bai. Cloning and functional analysis of the naphthomycin biosynthetic gene cluster in *Streptomyces sp.* cs. *Molecular BioSystems*, 7:2459–2469, 8 2011. ISSN 1742206X. doi: 10.1039/C1MB05036B.

[224] Koji Ichinose, David J. Bedford, Diethild Tornus, Andreas Bechthold, Maureen J. Bibb, W. Peter Revill, Heinz G. Floss, and David A. Hopwood. The granaticin biosynthetic gene cluster of *Streptomyces violaceoruber* Tu22: Sequence analysis and expression in a heterologous host. *Chemistry and Biology*, 5:647–659, 1998. ISSN 10745521. doi: 10.1016/S1074-5521(98)90292-7.

[225] Sung Ryeol Park, Ashootosh Tripathi, Jianfeng Wu, Pamela J. Schultz, Isaiah Yim, Thomas J. McQuade, Fengan Yu, Carl Johan Arevang, Abraham Y. Mensah, Giselle Tamayo-Castillo, Chuanwu Xi, and David H. Sherman. Discovery of cahuitamycins as biofilm inhibitors derived from a convergent biosynthetic pathway. *Nature Communications 2016 7:1*, 7:1–11, 2 2016. ISSN 2041-1723. doi: 10.1038/ncomms10710.

[226] Roland D. Kersten, N. Ziemert, David J. Gonzalez, Brendan M. Duggan, Victor Nizet, Pieter C. Dorrestein, and Bradley S. Moore. Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 110:E4407–E4416, 11 2013. ISSN 00278424. doi: 10.1073/PNAS.1315492110.

Appendix A

Supplemental materials for Complete genome sequences of *Streptomyces spp.* isolated from disease-suppressive soils

A.1 Predicted biosynthetic gene clusters for the three *Streptomyces* isolates

Table A.1: *Streptomyces sp.* GS93-23 gene clusters

Accession	Cluster	Predicted class	Start	End	Our Annotation
CP019457.1	1	Lassopeptide	195123	264181	
	2	Thiopeptide-Lantipeptide	439594*	453423*	Cyclothiazomycin [66]
	3	Nrps-T1pks	531139	609114	
	4	Lantipeptide	708680	731292	
	5	Nrps	908141	965139	
	6	Bacteriocin-Cf_saccharide	1059184	1101127	
	7	T2pks	1121146	1169639	
	8	Nrps-Terpene-T1pks	1505293*	1593855*	Streptolydigin [67]
	9	Siderophore	2053362*	2058673*	Desferrioxamine B [73]
	10	Ectoine	2142175*	2145592*	Ectoine [77]
	11	Lassopeptide	3346247	3368813	

Table A.1: *Streptomyces sp.* GS93-23 gene clusters

Accession	Cluster	Predicted class	Start	End	Our Annotation
	12	T1pks-Cf_fatty_acid	3696168	3743881	
	13	Lantipeptide-Cf_fatty_acid	4195732	4231551	
	14	Terpene	4654286	4676523	
	15	Nrps	4707417*	4750551*	Mannopeptimycin[68]
	16	Siderophore	6255261	6270018	
	17	T3pks-Nrps	6456534	6541453	
	18	Butyrolactone	6519962	6541453	
	19	Bacteriocin	6571918	6607981	
	20	Terpene	6988730	7035712	
	21	Terpene	7161797*	7173404*	Hopene[75]
	22	Terpene	7394231	7415328	
	23	Melanin-Nrps-T1pks	7578517	7647049	
	24	Butyrolactone	7845349	7866865	
	25	Bacteriocin	7991854	8035675	
	26	Nrps-Cf_saccharide-Terpene	7991854	8111863	

Table A.2: *Streptomyces sp.* 3211-3 gene clusters

Accession	Cluster	Predicted class	Start	End	Our Annotation
CP020039.1	1	Thiopeptide-Lantipeptide	42452	86704	
	2	Nrps-Ectoine-Cf_fatty_acid	50628	162596	
	3	Butyrolactone	92595	162596	
	4	Terpene	164334	212168	
	5	Bacteriocin	490923	501096	
	6	Ectoine	650075*	653448*	Ectoine[77]
	7	T2pks	734087	776581	

Table A.2: *Streptomyces sp.* 3211-3 gene clusters

Accession	Cluster	Predicted class	Start	End	Our Annotation
	8	Nrps	1237647*	1272586*	Tambromycin[69]
	9	Nrps-T1pks	1365295*	1385246*	Coelichelin[70]
	10	Butyrolactone	1434091	1490762	
	11	Cf_fatty_acid-Nrps-T1pks	1434091	1545568	
	12	Siderophore	3095812*	3103099*	Desferrioxamine B[73]
	13	Phosphonate-Ladderane-Cf_fatty_acid-T1pks	3504301	3594577	
	14	Nucleoside	3533889	3597786	
	15	Butyrolactone	4231389	4272166	
	16	Lantipeptide-Cf_fatty_acid-Arylpolyene	5096360	5139726	
	17	Siderophore	6031316	6046390	
	18	Nrps-T1pks	6134059	6251759	
	19	Bacteriocin	6326616	6357733	
	20	Terpene	6468007	6536887	
	21	T1pks-Cf_fatty_acid	6488707	6536887	
	22	Cf_saccharide-T1pks	6488707	6599667	
	23	Terpene	6888835*	6897732*	Hopene[75]
	24	T1pks-Otherks	6942937	7014751	
	25	Nrps	7053970	7140641	
	26	Nrps	7234813	7285532	
	27	Lantipeptide	7399629	7428343	
	28	Terpene	7486707	7507804	
	29	Lantipeptide	7582565*	7589118*	SapB[76]
	30	Terpene	7646639*	7648894*	2-methylisoborneol[71]

Table A.2: *Streptomyces sp.* 3211-3 gene clusters

Accession	Cluster	Predicted class	Start	End	Our Annotation
	31	Terpene-Melanin	7741551	7798205	
	32	Siderophore	7922498	7944290	
	33	T3pks	7965814	8015579	
	34	Other	8044240	8084893	
	35	Other	8071720	8115631	
	36	Nrps	8153326	8230599	
CP020040.1	37	Lantipeptide	275029	297872	
	38	Nrps	317542	369043	

Table A.3: *Streptomyces sp.* S3-4 gene clusters

Accession	Cluster	Predicted class	Start	End	Our Annotation
CP020042.1	1	Other	293884	334687	
	2	Other	318352	362332	
	3	Otherks-T1pks	415409	467555	
	4	Other	460056	500778	
	5	Nrps	503717	547195	
	6	Other	599623	641515	
	7	T2pks	774236	816742	
	8	T1pks	810028	855541	
	9	Thiopeptide-Lantipeptide	1246647	1274503	
	10	Nrps-T1pks-Cf_fatty_acid	2543542	2608861	
	11	Siderophore	2890650*	2897998*	Desferrioxamine B [73]
	12	Siderophore	5727419	5741847	

Table A.3: *Streptomyces sp.* S3-4 gene clusters

Accession	Cluster	Predicted class	Start	End	Our Annotation
	13	Bacteriocin	5989611	6000969	
	14	Terpene	6095043	6117265	
	15	Nrps	6256178	6324749	
	16	Terpene	6456916*	6472585*	Hopene[75]
	17	Terpene	6625217	6646326	
	18	Nrps	6755726*	6783618*	Streptothricin[72]
	19	Terpene	6634300*	6636326*	2-methylisoborneol[71]
	20	Melanin-Terpene	6898352	6956176	
	21	T3pks-Siderophore	6999210	7042095	
	22	Lantipeptide	7197530*	7203981*	SapB[76]
	23	Other	7244572	7288504	
	24	Bacteriocin	7441958	7475954	
CP020043.1	25	Nrps-T1pks-Melanin	37745	168930	
	26	Nrps	157767	209377	
	27	T1pks-Nrps-T2pks-Butyrolactone	205255	348805	
CP020044.1	28	T1pks	1	188553	

A.2 Cluster abundance for 125 Complete *Streptomyces* genomes

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces rapamycinicus NRRL 5491	1343740.8	49	1	1	0	70.60	12700734	10393
Streptomyces bingchenggensis BCW-1	749414.3	48	1	1	0	70.75	11936683	10313
Streptomyces sp. 11-1-2	1851167.4	47	2	1	1	70.85	11655206	10366
Streptomyces autolyticus strain CGMCC0516	75293.3	46	8	1	7	71.21	10184660	8515
Streptomyces hygroscopicus strain XM201	1912.5	46	1	1	0	70.75	12012215	10639
Streptomyces sp. TLI_053	1855352.4	45	1	1	0	73.63	9900053	8757
Streptomyces malaysiensis strain DSM 41137	92644.3	44	2	1	1	71.09	10744231	8968

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces sp. PBH53	1577075.3	43	1	0	0	72.73	9153597	8401
Streptomyces olivoreticuli strain ATCC 31159	284034.3	40	1	1	0	71.11	8809793	7923
Streptomyces roseochromogenus subsp. oscitans DS 12.976	1352936.5	40	2	1	1	68.80	9782257	9636
Streptomyces sp. CB01881	2078691.4	38	1	1	0	73.14	11344947	9001
Streptomyces sp. 769	1262452.3	38	2	1	1	71.60	10338286	9177
Streptomyces hygroscopicus subsp. jinggangensis TL01	1203460.3	38	3	1	2	42.71	10077952	9146
Streptomyces albireticuli strain SMD11	1940.3	38	1	1	0	72.82	8144417	7301

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces hygroscopicus subsp. jinggangensis 5008	1133850.2	38	3	1	2	71.84	10383684	9901
Streptomyces sp. CFMR 7	1649184.3	38	2	1	1	72.03	8307279	7407
Streptomyces clavuligerus strain F1D-5	1901.15	38	3	1	2	72.53	8059125	7323
Streptomyces sp. ICC1	2099583.3	37	1	1	0	72.02	9034319	8909
Streptomyces albulus ZPM	1434306.3	37	1	1	0	72.13	9784577	9148
Streptomyces albolongus strain YIM 101047	68173.3	37	1	1	0	71.98	8027788	7261
Streptomyces avermitilis MA-4680	227882.9	37	2	1	1	70.70	9119895	8106

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces armeniacus strain ATCC 15676	83291.4	36	1	1	0	72.37	8083249	7273
Streptomyces sp. Sge12 strain Sge12	1972846.3	36	2	1	1	72.17	8110698	7511
Streptomyces silaceus strain ACCC40021	545123.5	36	1	1	0	72.09	8625867	7813
Streptomyces griseus subsp. griseus NBRC 13350	455632.4	36	1	1	0	72.20	8545929	7294
Streptomyces albulus strain NK660	68570.5	35	2	1	1	72.32	9372401	8793
Streptomyces sp. ICC4	2099584.3	35	1	1	0	72.03	9010404	8878
Streptomyces fradiae strain NKZ-259	1906.17	35	1	1	0	72.13	8081458	7431
Streptomyces lydicus A02	1403539.3	35	1	1	0	70.70	9300149	8888
Streptomyces sp. P3	2135430.3	35	1	1	0	71.37	9851971	9315

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces albus DSM 41398	1888.4	35	1	1	0	72.64	8384669	6923
Streptomyces luteoverticillatus strain CGMCC 15060	66425.3	35	1	1	0	72.05	7367863	6832
Streptomyces formicae strain KY5	1616117.3	34	1	1	0	71.38	9611874	8393
Streptomyces albus strain BK3-25	1888.11	34	1	1	0	72.64	8308430	7171
Streptomyces lincolnensis strain NRRL 2936	1915.4	34	1	1	0	71.01	10319054	9514
Streptomyces hygroscopicus subsp. limoneus KCTC 1717	264445.3	34	2	2	0	71.96	10537932	9983
Streptomyces anulatus strain ATCC 11523	1892.7	34	2	1	1	71.72	8847108	8036

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces fulvissimus DSM 40593	1303692.3	34	1	1	0	71.50	7905758	7081
Streptomyces clavuligerus strain F613-1	1901.9	34	2	1	1	72.60	7590758	6641
Streptomyces sp. NEAU-S7GS2	2202000.4	34	2	1	1	70.78	9687439	9008
Streptomyces lydicus strain WYEC 108	47763.27	34	1	1	0	70.80	9125666	8388
Streptomyces pactum strain KLBMP 5084	68249.6	34	1	1	0	72.41	8180260	7562
Streptomyces sp. CdTB01	1725411.3	33	2	1	1	71.53	10191567	9718
Streptomyces sp. 2323.1	1938841.3	33	1	1	0	71.19	8212455	7369
Streptomyces puniciscabiei strain TW1S1	164348.7	33	1	1	0	71.09	9698948	9166
Streptomyces sp. S8 strain S8	1837283.3	33	2	1	1	72.29	7601864	6837

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces griseoviridis strain F1-27	45398.5	33	1	1	0	72.38	8963414	8148
Streptomyces sp. Mg1	465541.12	32	4	1	3	72.13	8716193	8218
Streptomyces collinus Tu 365	1214242.5	32	3	1	2	72.55	8377286	7336
Streptomyces cyaneogriseus NMWT 1	477245.3	32	1	1	0	72.86	7762396	6922
Streptomyces bacillaris strain ATCC 15855	68179.3	32	1	1	0	71.95	7888441	7125
Streptomyces sp. HNM0039	2174846.3	31	1	1	0	72.46	7289495	6755
Streptomyces cavourensis strain TJ430	67258.5	31	1	1	0	72.12	7613757	6798

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces cavourensis strain 1AS2a	67258.4	31	1	1	0	72.13	7600475	6788
Streptomyces sp. ZFG47	2184053.3	31	2	1	1	70.76	10239272	9684
Streptomyces davawensis JCM 4913	1214101.3	31	2	1	1	70.59	9555950	8696
Streptomyces chartreusis NRRL 3882 strain NRRL3882	1079985.1	31	1	1	0	71.23	8983317	8293
Streptomyces sp. Go-475	2072505.3	31	1	1	0	71.96	8570609	7865
Streptomyces leeuwenhoekii NRRL B-24963	1437453.6	31	3	1	2	72.68	8122491	7432

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces violaceoruber strain S21	1935.7	31	1	1	0	72.65	7916045	7235
Streptomyces incarnatus strain NRRL8089	665007.5	30	4	0	3	71.48	8897465	8452
Streptomyces venezuelae	54571.11	30	1	1	0	71.75	9034396	8487
Streptomyces niveus NCIMB 11891	1352941.4	30	5	1	4	69.44	8726876	8170
Streptomyces lincolnensis strain LC-G	1915.7	30	1	1	0	71.06	9513637	8737
Streptomyces sp. Tue 6075	1661694.4	30	1	1	0	71.57	7931832	7175
Streptomyces vietnamensis GIM4.0001	362257.4	30	2	1	1	71.99	9153777	8292
Streptomyces venezuelae strain NRRL B-65442	54571.16	30	2	0	1	72.42	8380320	7771

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces sp. GSSD-12	2282738.3	30	1	1	0	71.19	8454852	7430
Streptomyces venezuelae ATCC 10712	953739.5	30	1	1	0	72.40	8226158	7409
Streptomyces gilvosporeus strain F607	553510.3	29	1	1	0	70.95	8482298	7969
Streptomyces coelicolor A3(2)	100226.15	29	3	1	2	72.00	9054847	8325
Streptomyces bottropensis ATCC 25435	1054862.11	29	1	0	0	71.19	8955726	8324
Streptomyces venezuelae strain ATCC 15439	54571.1	28	1	1	0	71.74	9054831	8457
Streptomyces niveus strain SCSIO 3406	193462.5	28	1	1	0	70.46	7990492	7301
Streptomyces laurentii strain ATCC 31255	39478.4	28	1	1	0	72.30	8032664	7709

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces lunaelactis strain MM109	1535768.3	28	3	1	2	69.75	8570191	8269
Streptomyces sp. CNQ-509	444103.5	28	1	1	0	73.07	8039333	7243
Streptomyces sp. SCSIO 03032 strain SCSIO 03032	1109743.5	28	1	1	0	73.52	6287975	5677
Streptomyces sp. PAMC26508	1265601.4	28	2	1	1	71.06	7630245	6806
Streptomyces sp. S063	2005885.3	28	1	1	0	71.49	7614683	7330
Streptomyces spongiicola strain HNM0071	1690221.3	27	1	1	0	72.45	7180417	6703
Streptomyces globisporus strain TFH56	1908.1	27	3	1	2	71.54	7666521	7248
Streptomyces sp. fd1-xmd strain fd1-xmd	1812480.3	27	1	1	0	72.51	7929999	7460

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces sp. S10(2016)	1783515.3	27	1	1	0	71.26	9083372	8480
Streptomyces aureofaciens strain DM-1	1894.18	27	2	1	1	72.57	7076402	6652
Streptomyces sp. TN58	234612.4	27	1	1	0	72.30	7585034	7194
Streptomyces lydicus strain GS93	47763.1	27	1	1	0	72.01	8243179	7497
Streptomyces lividans TK24	457428.16	27	1	1	0	72.24	8345283	7749
Streptomyces flavogriseus ATCC 33331	591167.6	27	3	1	2	71.00	7656104	6866
Streptomyces alboflavus strain MDJK44	67267.5	27	1	1	0	72.09	9622415	9185
Streptomyces sp. MOE7 strain MOE7	1961713.3	26	1	1	0	71.99	8399509	7802
Streptomyces sp. CMB-StM0423	2059884.3	26	1	1	0	73.14	8029398	7153

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces globosus strain LZH-48 strain soil	68209.3	26	3	1	2	73.65	7535750	6974
Streptomyces sp. 3124.6 strain 3124.6	1882757.3	26	1	1	0	70.67	9375551	8748
Streptomyces lydicus strain 103	47763.9	26	1	1	0	72.22	8201357	7502
Streptomyces ambofaciens ATCC 23877	278992.5	26	2	1	1	72.19	8393598	7953
Streptomyces sp. PVA 94-07	1223307.4	26	3	1	2	73.05	7106149	6359
Streptomyces sp. GBA 94-10	1218177.5	26	2	1	1	72.96	7224475	6566
Streptomyces sp. 2114.2	1881022.3	26	1	1	0	72.11	8620263	8008
Streptomyces atratus strain SCSIO_ZH16	1893.7	26	1	1	0	69.59	9641288	9294

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces sp. CC0208	2306165.3	25	1	1	0	70.59	9320089	8728
Streptomyces glaucescens GLA.O	1907.4	25	2	1	1	72.91	7623774	6719
Streptomyces ambofaciens strain DSM 40697	1889.1	25	1	1	0	72.33	8137876	7413
Streptomyces griseorubiginosus strain 3E-1	67304.8	25	1	1	0	70.94	9512378	8665
Streptomyces sp. WAC00288	2094021.4	24	5	1	4	72.73	7852859	7532
Streptomyces sp. XZHG99	2049881.3	24	3	1	2	69.92	8710171	8663
Streptomyces rubrolavendulae strain MJM4426	285473.5	24	1	1	0	74.78	6543262	5844
Streptomyces sp. CCM_MD2014	1561022.5	24	1	1	0	72.13	8274043	7689

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces actuosus strain ATCC 25421	1885.4	23	1	1	0	72.53	8145579	7618
Streptomyces sp. SirexAA-E	862751.12	22	1	1	0	71.75	7414440	6808
Streptomyces sp. CB09001	2083284.3	22	1	1	0	71.95	7787608	7303
Streptomyces parvulus strain 2297	146923.5	22	2	1	1	72.73	7766531	7056
Streptomyces peucetius subsp. caesius ATCC 27952	316280.3	22	1	1	0	70.59	8023114	7759
Streptomyces albus J1074	457425.27	22	1	1	0	73.32	6841649	6120
Streptomyces koyangensis strain VK-A60T	188770.3	21	1	1	0	73.03	7220839	6695
Streptomyces sp. FR-008	206662.6	21	3	1	2	73.32	7258031	6897

Table A.4: Cluster abundance for 125 Complete *Streptomyces* genomes

Organism	PATRIC Accession	Clusters	Contigs	Chromosomes	Plasmids	GC Content	Genome Size (bp)	CDS
Streptomyces sp. 452	1827580.3	21	1	1	0	71.91	7641029	7171
Streptomyces sampsonii strain KJ40	42239.3	21	1	1	0	73.39	7070328	6359
Streptomyces albus strain SM254	1888.8	21	1	1	0	73.34	7170504	6473
Streptomyces pristinaespiralis strain HCCB 10218	38300.4	21	1	1	0	71.54	8532592	7700
Streptomyces sp. ETH9427	2211357.3	20	3	1	2	72.00	7978031	7803
Streptomyces xiamenensis strain MCCC 1A01550	408015.6	20	1	1	0	72.02	5961401	5556
Streptomyces nodosus ATCC 14899	40318.3	20	1	1	0	70.80	7714110	6875
Streptomyces sp. 4F	1751294.3	19	1	1	0	72.28	8047771	7510
Streptomyces sp. CLI2509	1984801.3	18	2	1	1	73.30	7232701	6733

A.3 Supplementary Methods

A.3.1 Cluster abundance comparison

Unannotated sequences for 125 complete status Streptomyces genomes were downloaded from the PATRIC database12. Sequences were annotated with antiSMASH 4.1. Detection of putative clusters was disabled. A Python script was used to count the number of cluster features in the resulting genbank files. The cluster abundance data was combined with genome statistics from the PATRIC database including contig, chromosome, and plasmid counts, GC content, genome size, and CDS counts.

A.3.2 Signaling potential analysis

MultiGeneBlast13 was used to query the 496 genomes from the phylogenetic analysis dataset. Query sequences were ScbA and AfsR homologue amino acid sequences from γ-butyrolactone BGCs from 3211-3, S3-4, or GS93-23. Genome IDs were sorted in descending order of clusterBLAST score from 3211-3. Identity scores from the phylogenetic analysis were plotted against clusterBLAST scores.

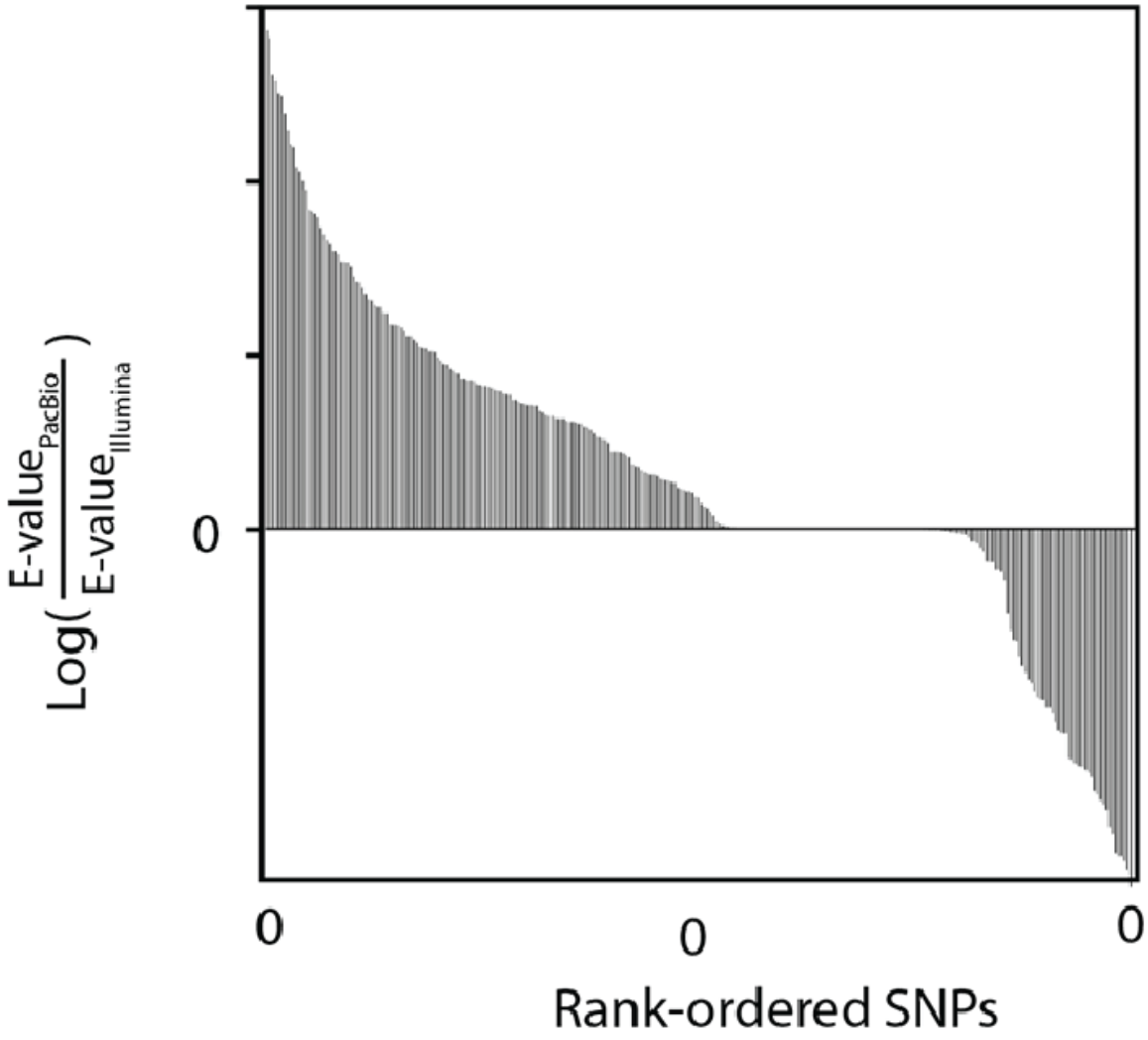


Figure A.1: Indel comparison of Illumina polished vs. PacBio only assemblies.

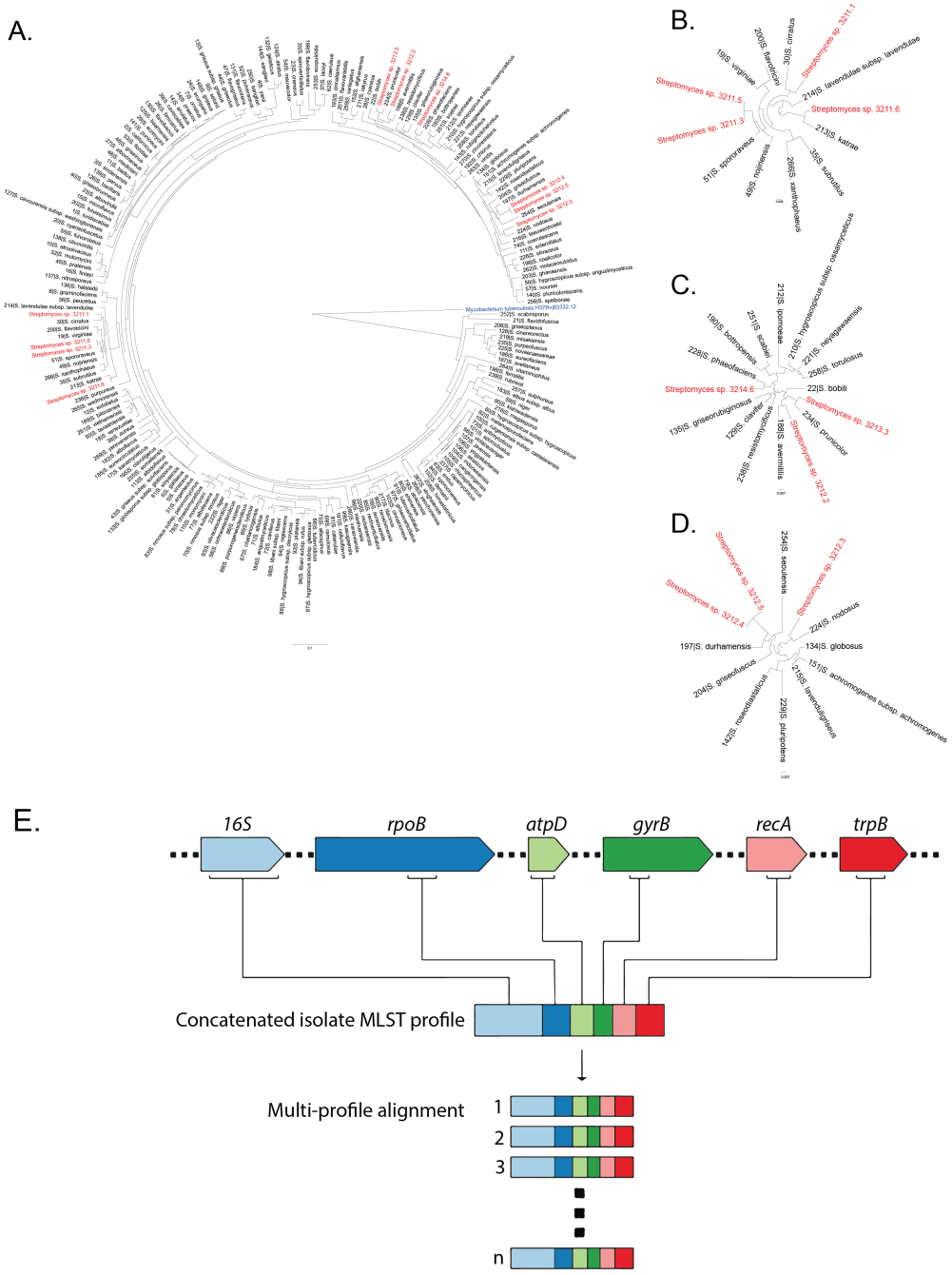
Appendix B

Supplemental materials for Metatranscriptomic analysis of synthetic communities of sympatric Streptomyces isolates from disease suppressive soil

B.1 Molecular phylogeny of ten sympatric *Streptomyces* isolates compared to type strains

Figure B.1 ***(following page)*:** Molecular phylogeny of ten sympatric *Streptomyces* isolates compared to type strains. (a) Rooted phylogenetic tree showing *Streptomyces* type strains (black) from PubMLST and new strains from this study (red). Mycobacterium tuberculosis was used as an outgroup. The tree was produced using FastTree. Scale bar shows the radial distance corresponding to a 10% change in sequence identity. (b-d) Higher resolution unrooted phylogenetic trees showing the most similar type strains to our ten DSS isolates. (e) Multi-locus sequence typing approach for molecular phylogeny.

Figure B.1: (continued)



B.2 Average nucleotide identity matrix of ten sympatric *Streptomyces* isolates

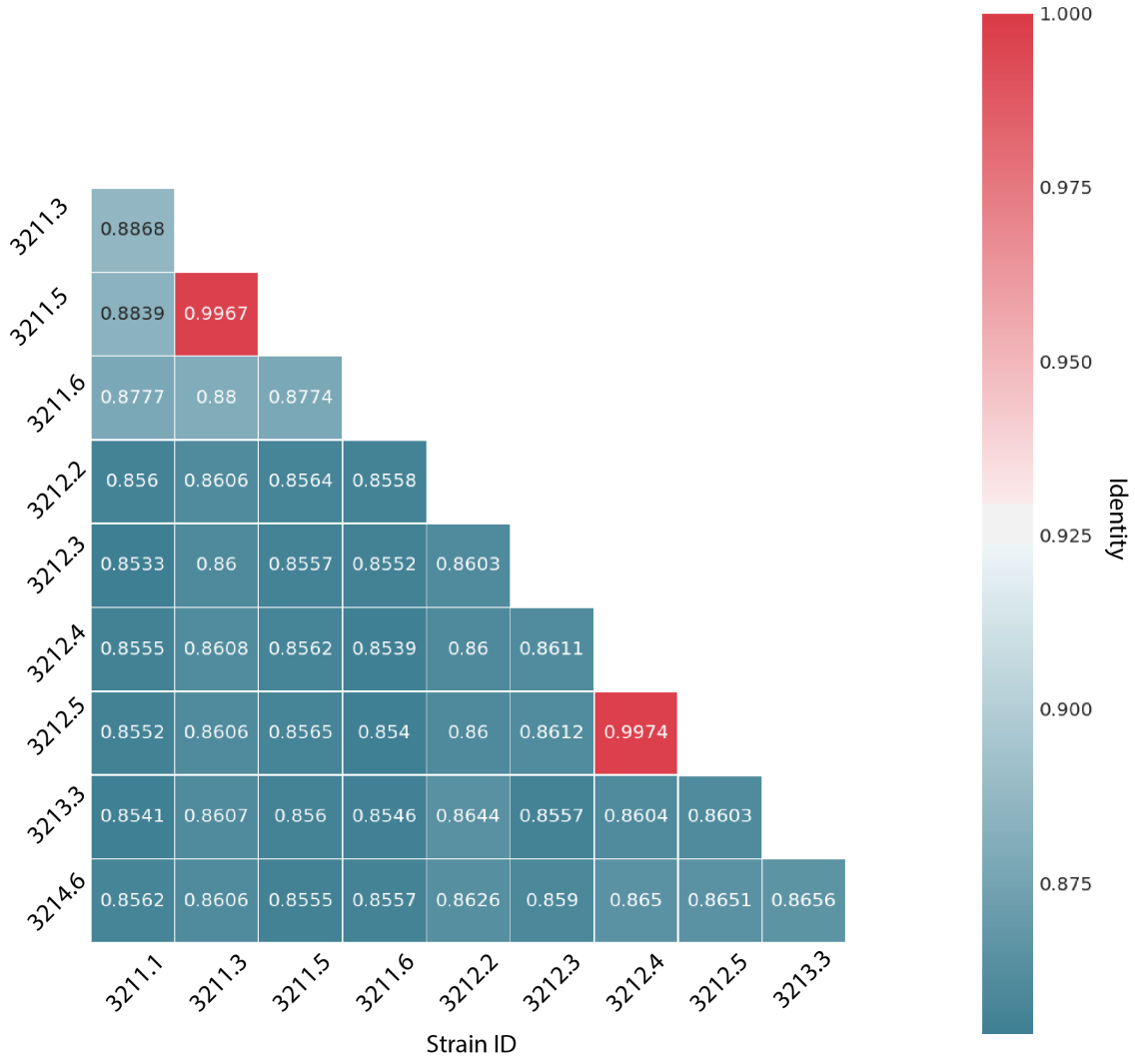


Figure B.2: Average nucleotide identity matrix of ten sympatric *Streptomyces* isolates. Identity values are based on pairwise whole-genome alignments performed using the dnadiff algorithm from the MUMmer suite[199].

B.3 Genome maps of the ten sympatric *Streptomyces* isolates

Figure B.3 *(following page):* Genome map of *Streptomyces sp.* 3211.1. Outer black ring represent assembled contigs. Genomic position (Mb) indicated along contigs. Tracks from outside to inside represent the following data. Putative BGC annotations, colored by predicted biosynthetic class (track 1). Coding sequences on forward and reverse strands (tracks 2 and 3, respectively), colored by broad NOG functional categories. Coding sequences predicted to encode proteins containing signal peptides. Upper and lower halves of signal peptide markers are colored if they are upregulated (red, log2FC > 1), or downregulated (blue, log2FC < -1) in at least one community (track 4). Moving average of log2FC within a 10,000 bp window, incrementing by 1,000 bp (variable number of tracks depending on isolate, starting at track 5). Average log2FC values are colored red (positive) or blue (negative), and appear dark if values are outside the range [-1, 1]. Inner most tracks, from outside in represent G + C content and G + C skew, respectively. Each is shown for two window sizes: 10 Kb (dark blue above average, dark orange below average) and 1 Mb (light blue above average, light orange below average). Only the 10 Kb resolution data is shown for plasmids.

Figure B.3: (continued)

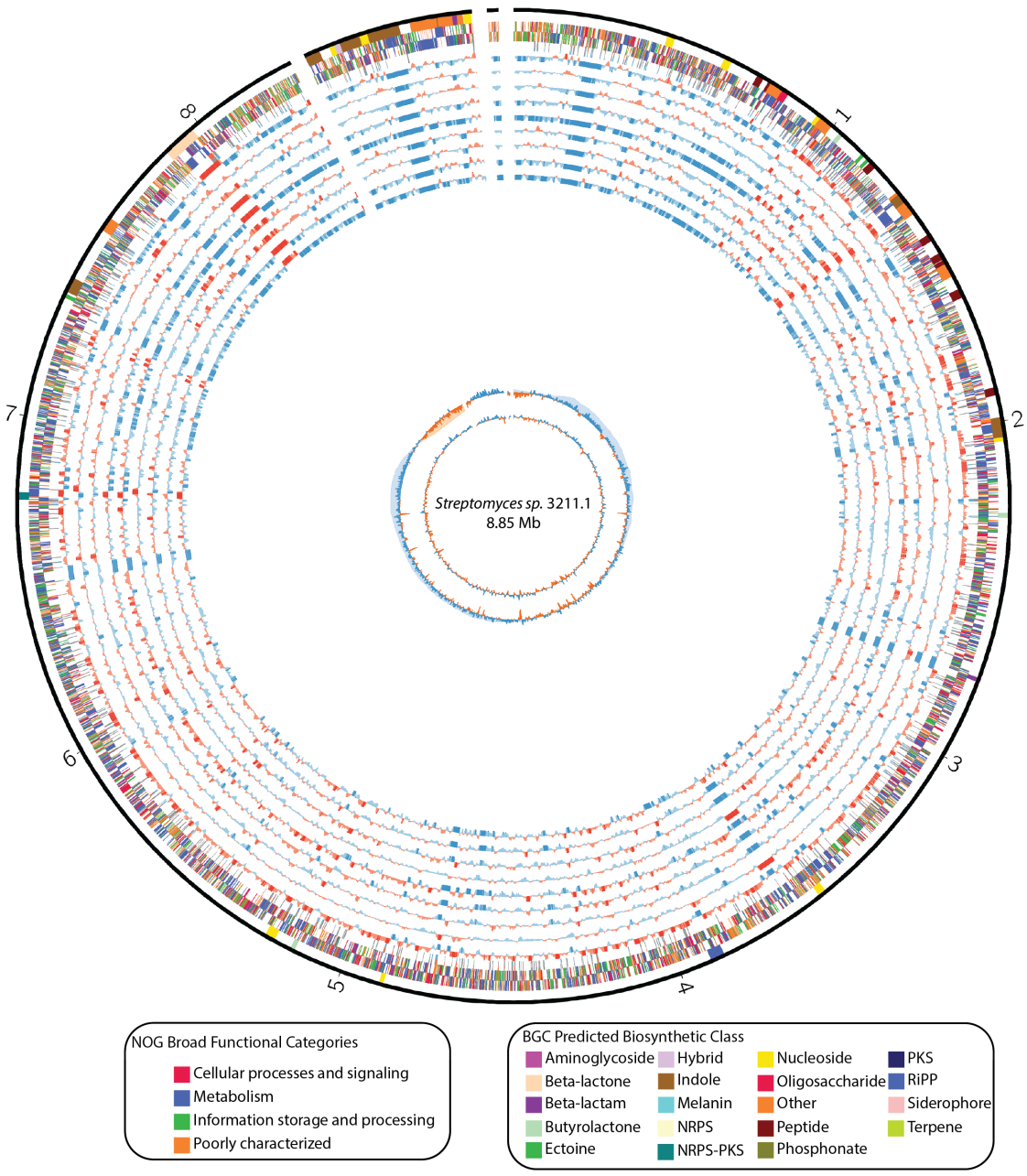


Figure B.4 ***(following page)*****:** Genome map of *Streptomyces sp.* 3211.3. Outer black ring represent assembled contigs. Genomic position (Mb) indicated along contigs. Tracks from outside to inside represent the following data. Putative BGC annotations, colored by predicted biosynthetic class (track 1). Coding sequences on forward and reverse strands (tracks 2 and 3, respectively), colored by broad NOG functional categories. Coding sequences predicted to encode proteins containing signal peptides. Upper and lower halves of signal peptide markers are colored if they are upregulated (red, log2FC > 1), or downregulated (blue, log2FC < -1) in at least one community (track 4). Inner most tracks, from outside in represent G + C content and G + C skew, respectively. Each is shown for two window sizes: 10 Kb (dark blue above average, dark orange below average) and 1 Mb (light blue above average, light orange below average). Only the 10 Kb resolution data is shown for plasmids.

Figure B.4: (continued)

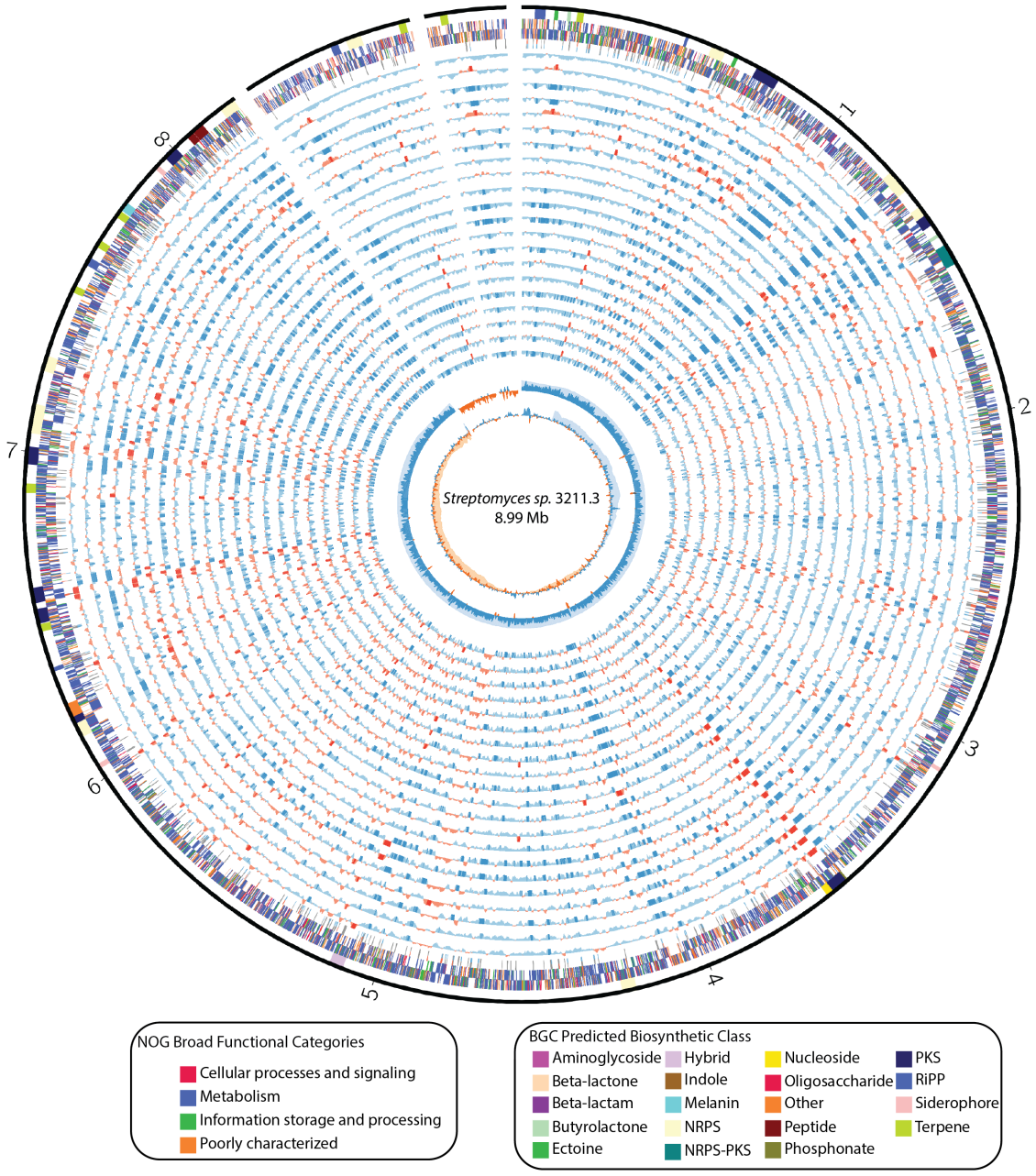


Figure B.5 ***(following page):*** Genome map of *Streptomyces sp.* 3211.5. Outer black ring represent assembled contigs. Genomic position (Mb) indicated along contigs. Tracks from outside to inside represent the following data. Putative BGC annotations, colored by predicted biosynthetic class (track 1). Coding sequences on forward and reverse strands (tracks 2 and 3, respectively), colored by broad NOG functional categories. Coding sequences predicted to encode proteins containing signal peptides. Upper and lower halves of signal peptide markers are colored if they are upregulated (red, log2FC > 1), or downregulated (blue, log2FC < -1) in at least one community (track 4). Moving average of log2FC within a 10,000 bp window, incrementing by 1,000 bp (variable number of tracks depending on isolate, starting at track 5). Average log2FC values are colored red (positive) or blue (negative), and appear dark if values are outside the range [-1, 1]. Inner most tracks, from outside in represent G + C content and G + C skew, respectively. Each is shown for two window sizes: 10 Kb (dark blue above average, dark orange below average) and 1 Mb (light blue above average, light orange below average). Only the 10 Kb resolution data is shown for plasmids.

Figure B.5: (continued)

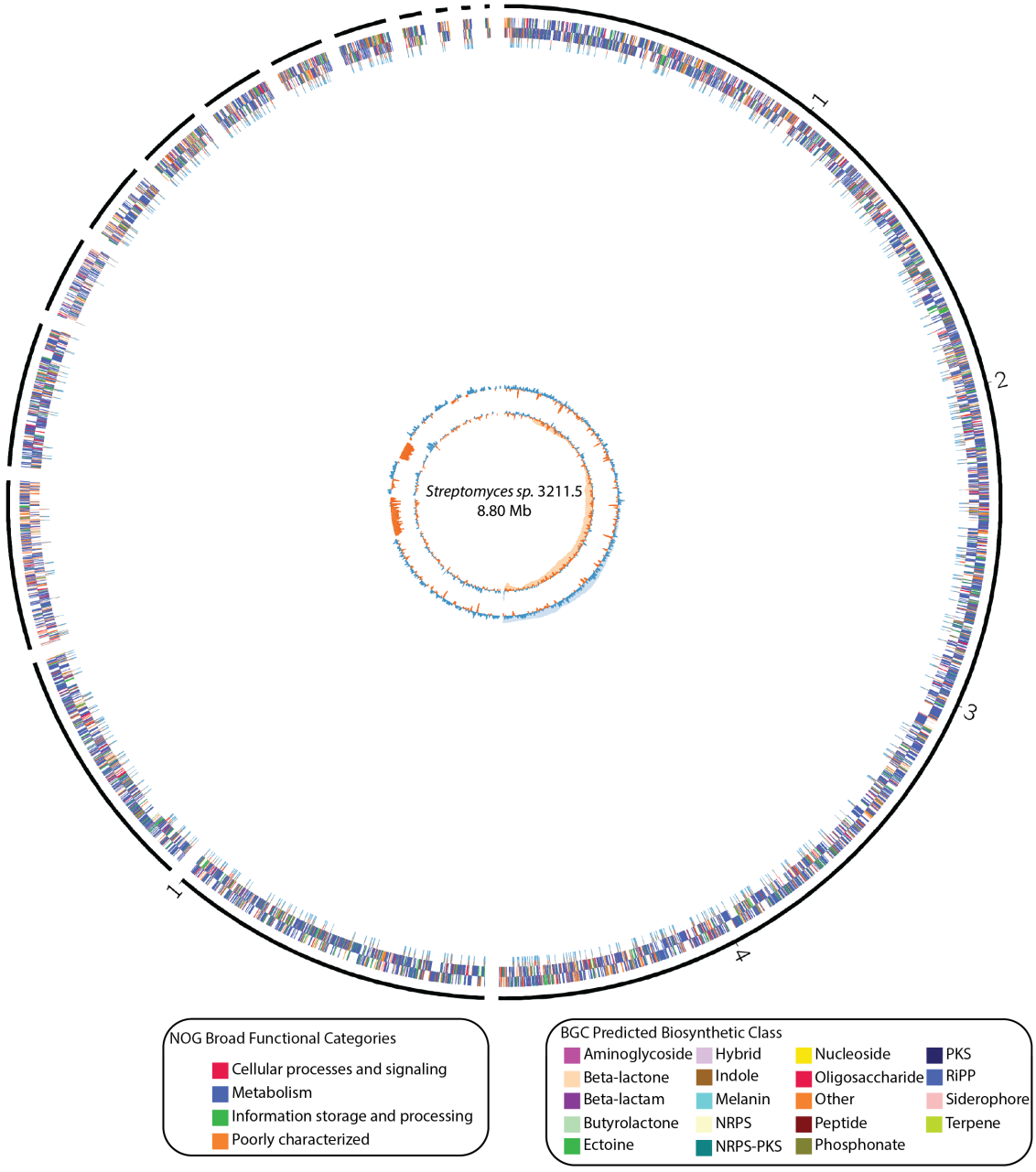


Figure B.6 *(following page)*: Genome map of *Streptomyces sp.* 3211.6. Outer black ring represent assembled contigs. Genomic position (Mb) indicated along contigs. Tracks from outside to inside represent the following data. Putative BGC annotations, colored by predicted biosynthetic class (track 1). Coding sequences on forward and reverse strands (tracks 2 and 3, respectively), colored by broad NOG functional categories. Coding sequences predicted to encode proteins containing signal peptides. Upper and lower halves of signal peptide markers are colored if they are upregulated (red, log2FC > 1), or downregulated (blue, log2FC < -1) in at least one community (track 4). Moving average of log2FC within a 10,000 bp window, incrementing by 1,000 bp (variable number of tracks depending on isolate, starting at track 5). Average log2FC values are colored red (positive) or blue (negative), and appear dark if values are outside the range [-1, 1]. Inner most tracks, from outside in represent G + C content and G + C skew, respectively. Each is shown for two window sizes: 10 Kb (dark blue above average, dark orange below average) and 1 Mb (light blue above average, light orange below average). Only the 10 Kb resolution data is shown for plasmids.

Figure B.6: (continued)

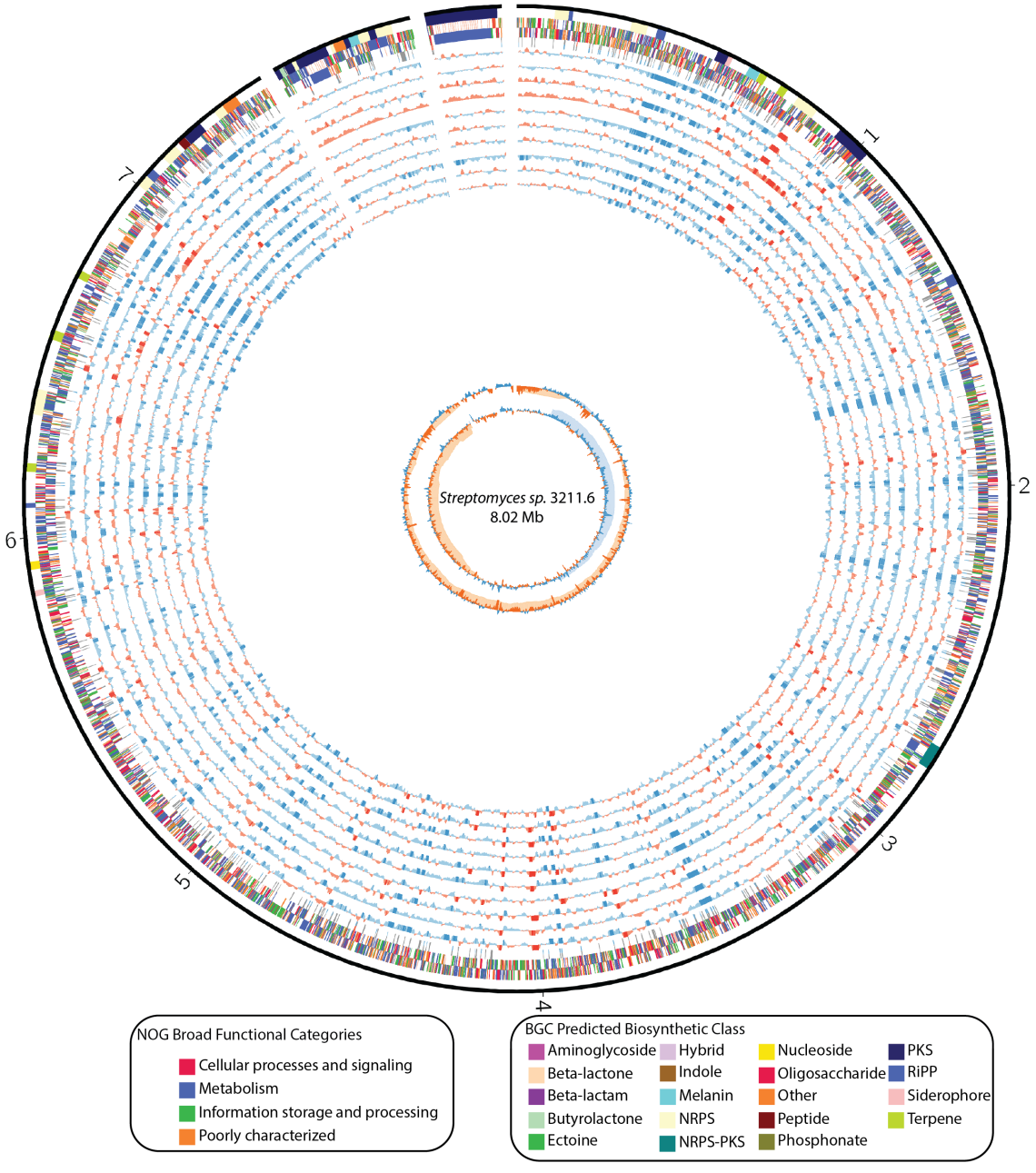


Figure B.7 ***(following page):*** Genome map of *Streptomyces sp.* 3212.2. Outer black ring represent assembled contigs. Genomic position (Mb) indicated along contigs. Tracks from outside to inside represent the following data. Putative BGC annotations, colored by predicted biosynthetic class (track 1). Coding sequences on forward and reverse strands (tracks 2 and 3, respectively), colored by broad NOG functional categories. Coding sequences predicted to encode proteins containing signal peptides. Upper and lower halves of signal peptide markers are colored if they are upregulated (red, log2FC > 1), or downregulated (blue, log2FC < -1) in at least one community (track 4). Moving average of log2FC within a 10,000 bp window, incrementing by 1,000 bp (variable number of tracks depending on isolate, starting at track 5). Average log2FC values are colored red (positive) or blue (negative), and appear dark if values are outside the range [-1, 1]. Inner most tracks, from outside in represent G + C content and G + C skew, respectively. Each is shown for two window sizes: 10 Kb (dark blue above average, dark orange below average) and 1 Mb (light blue above average, light orange below average). Only the 10 Kb resolution data is shown for plasmids.

Figure B.7: (continued)

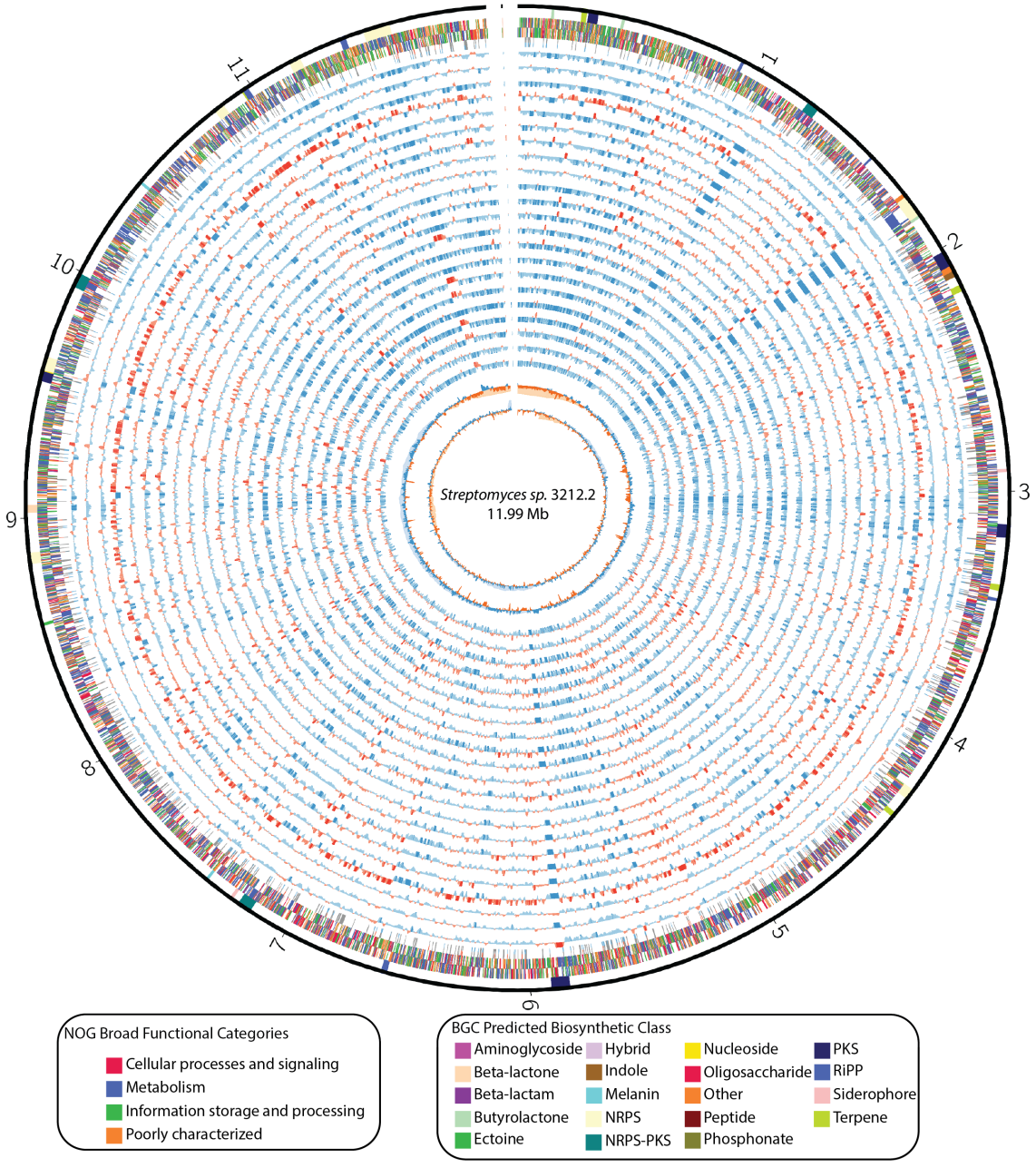


Figure B.8 *(following page):* Genome map of *Streptomyces sp.* 3212.3. Outer black ring represent assembled contigs. Genomic position (Mb) indicated along contigs. Tracks from outside to inside represent the following data. Putative BGC annotations, colored by predicted biosynthetic class (track 1). Coding sequences on forward and reverse strands (tracks 2 and 3, respectively), colored by broad NOG functional categories. Coding sequences predicted to encode proteins containing signal peptides. Upper and lower halves of signal peptide markers are colored if they are upregulated (red, log2FC > 1), or downregulated (blue, log2FC < -1) in at least one community (track 4). Moving average of log2FC within a 10,000 bp window, incrementing by 1,000 bp (variable number of tracks depending on isolate, starting at track 5). Average log2FC values are colored red (positive) or blue (negative), and appear dark if values are outside the range [-1, 1]. Inner most tracks, from outside in represent G + C content and G + C skew, respectively. Each is shown for two window sizes: 10 Kb (dark blue above average, dark orange below average) and 1 Mb (light blue above average, light orange below average). Only the 10 Kb resolution data is shown for plasmids.

Figure B.8: (continued)

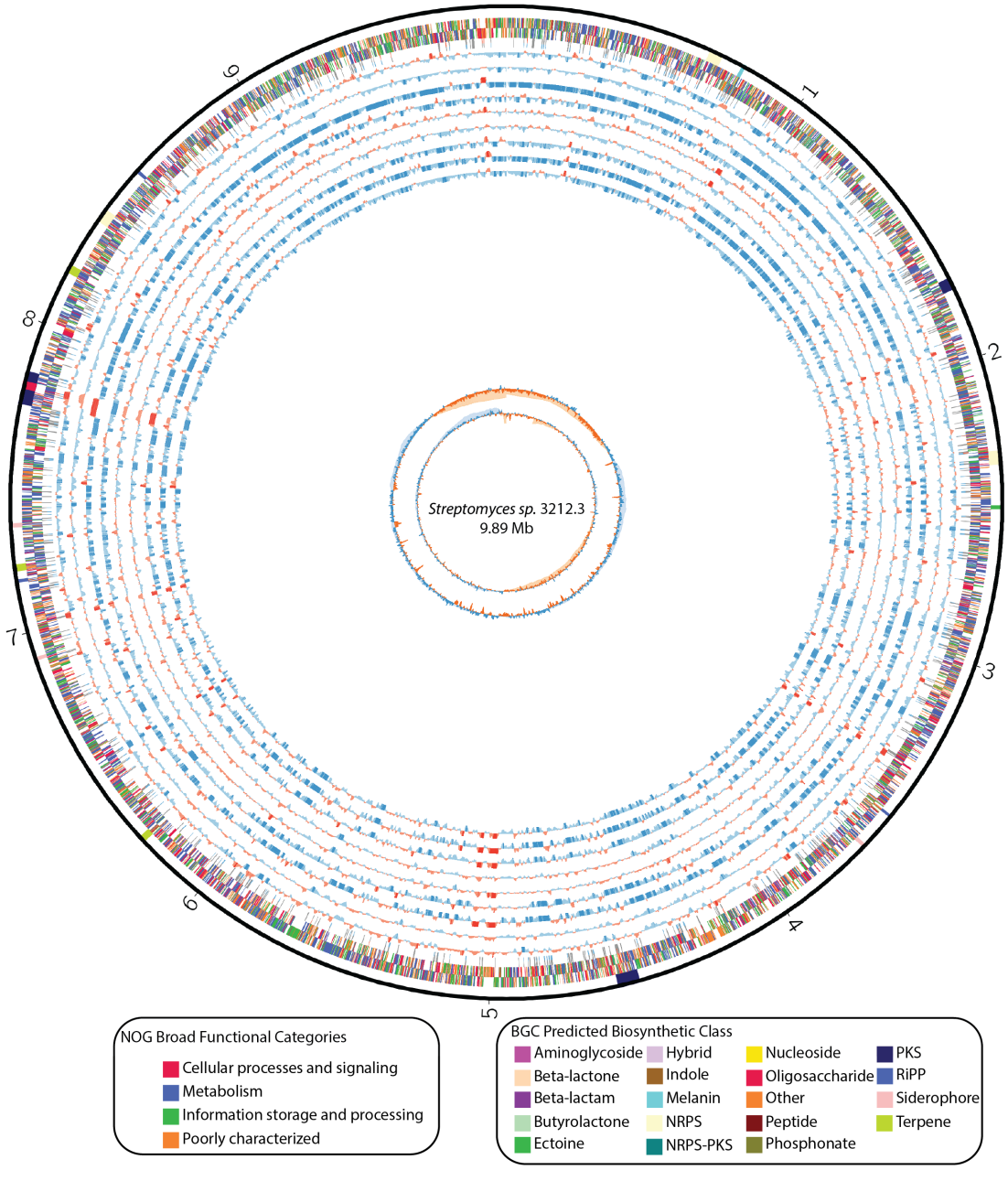


Figure B.9 *(following page):* Genome map of *Streptomyces sp.* 3212.4. Outer black ring represent assembled contigs. Genomic position (Mb) indicated along contigs. Tracks from outside to inside represent the following data. Putative BGC annotations, colored by predicted biosynthetic class (track 1). Coding sequences on forward and reverse strands (tracks 2 and 3, respectively), colored by broad NOG functional categories. Coding sequences predicted to encode proteins containing signal peptides. Upper and lower halves of signal peptide markers are colored if they are upregulated (red, log2FC > 1), or downregulated (blue, log2FC < -1) in at least one community (track 4). Moving average of log2FC within a 10,000 bp window, incrementing by 1,000 bp (variable number of tracks depending on isolate, starting at track 5). Average log2FC values are colored red (positive) or blue (negative), and appear dark if values are outside the range [-1, 1]. Inner most tracks, from outside in represent G + C content and G + C skew, respectively. Each is shown for two window sizes: 10 Kb (dark blue above average, dark orange below average) and 1 Mb (light blue above average, light orange below average). Only the 10 Kb resolution data is shown for plasmids.

Figure B.9: (continued)

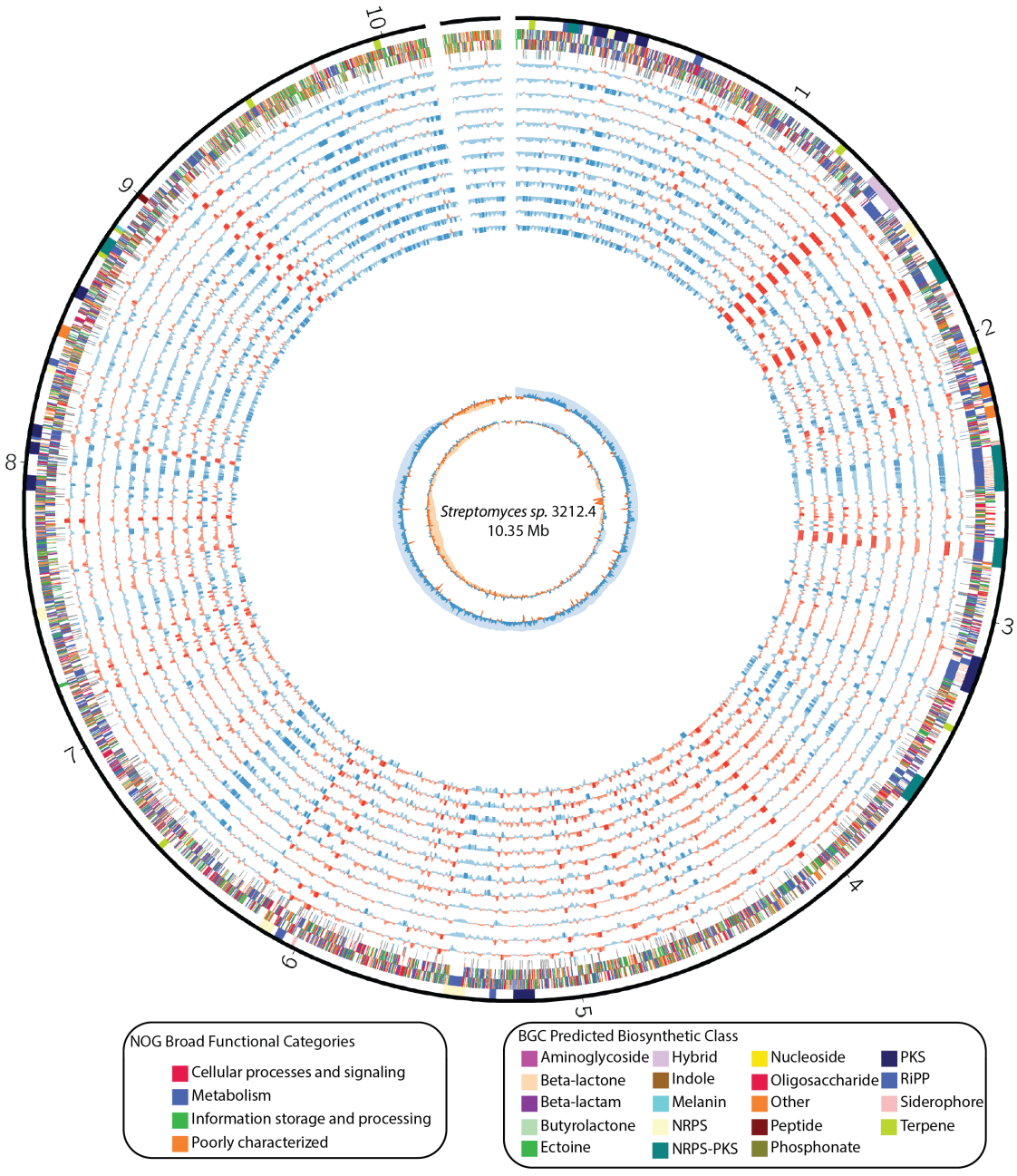


Figure B.10 ***(following page):*** Genome map of *Streptomyces sp.* 3212.5. Outer black ring represent assembled contigs. Genomic position (Mb) indicated along contigs. Tracks from outside to inside represent the following data. Putative BGC annotations, colored by predicted biosynthetic class (track 1). Coding sequences on forward and reverse strands (tracks 2 and 3, respectively), colored by broad NOG functional categories. Coding sequences predicted to encode proteins containing signal peptides. Upper and lower halves of signal peptide markers are colored if they are upregulated (red, log2FC > 1), or downregulated (blue, log2FC < -1) in at least one community (track 4). Inner most tracks, from outside in represent G + C content and G + C skew, respectively. Each is shown for two window sizes: 10 Kb (dark blue above average, dark orange below average) and 1 Mb (light blue above average, light orange below average). Only the 10 Kb resolution data is shown for plasmids.

Figure B.10: (continued)

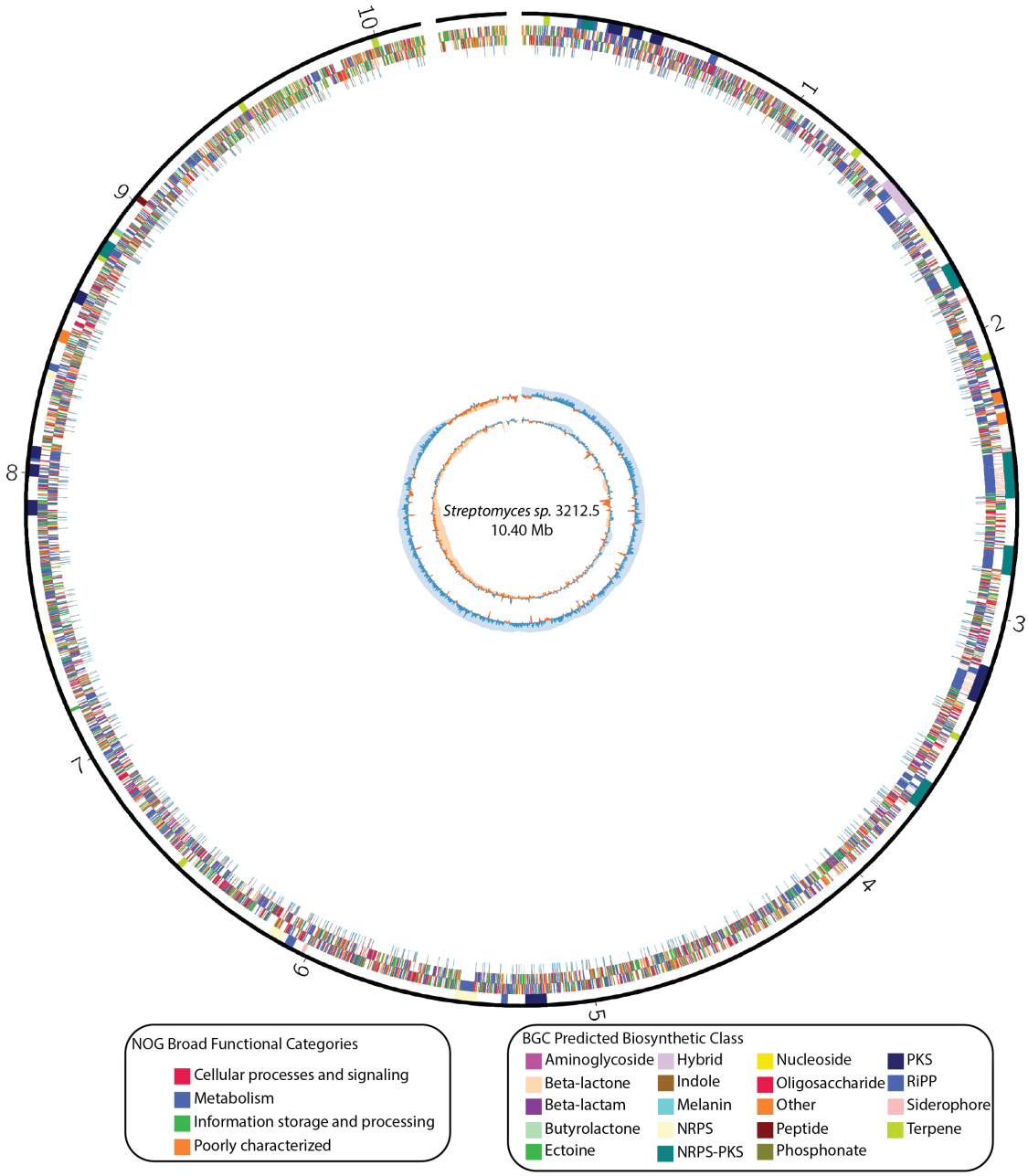


Figure B.11 ***(following page)*****:** Genome map of *Streptomyces sp.* 3213.3. Outer black ring represent assembled contigs. Genomic position (Mb) indicated along contigs. Tracks from outside to inside represent the following data. Putative BGC annotations, colored by predicted biosynthetic class (track 1). Coding sequences on forward and reverse strands (tracks 2 and 3, respectively), colored by broad NOG functional categories. Coding sequences predicted to encode proteins containing signal peptides. Upper and lower halves of signal peptide markers are colored if they are upregulated (red, log2FC > 1), or downregulated (blue, log2FC < -1) in at least one community (track 4). Moving average of log2FC within a 10,000 bp window, incrementing by 1,000 bp (variable number of tracks depending on isolate, starting at track 5). Average log2FC values are colored red (positive) or blue (negative), and appear dark if values are outside the range [-1, 1]. Inner most tracks, from outside in represent G + C content and G + C skew, respectively. Each is shown for two window sizes: 10 Kb (dark blue above average, dark orange below average) and 1 Mb (light blue above average, light orange below average). Only the 10 Kb resolution data is shown for plasmids.

Figure B.11: (continued)

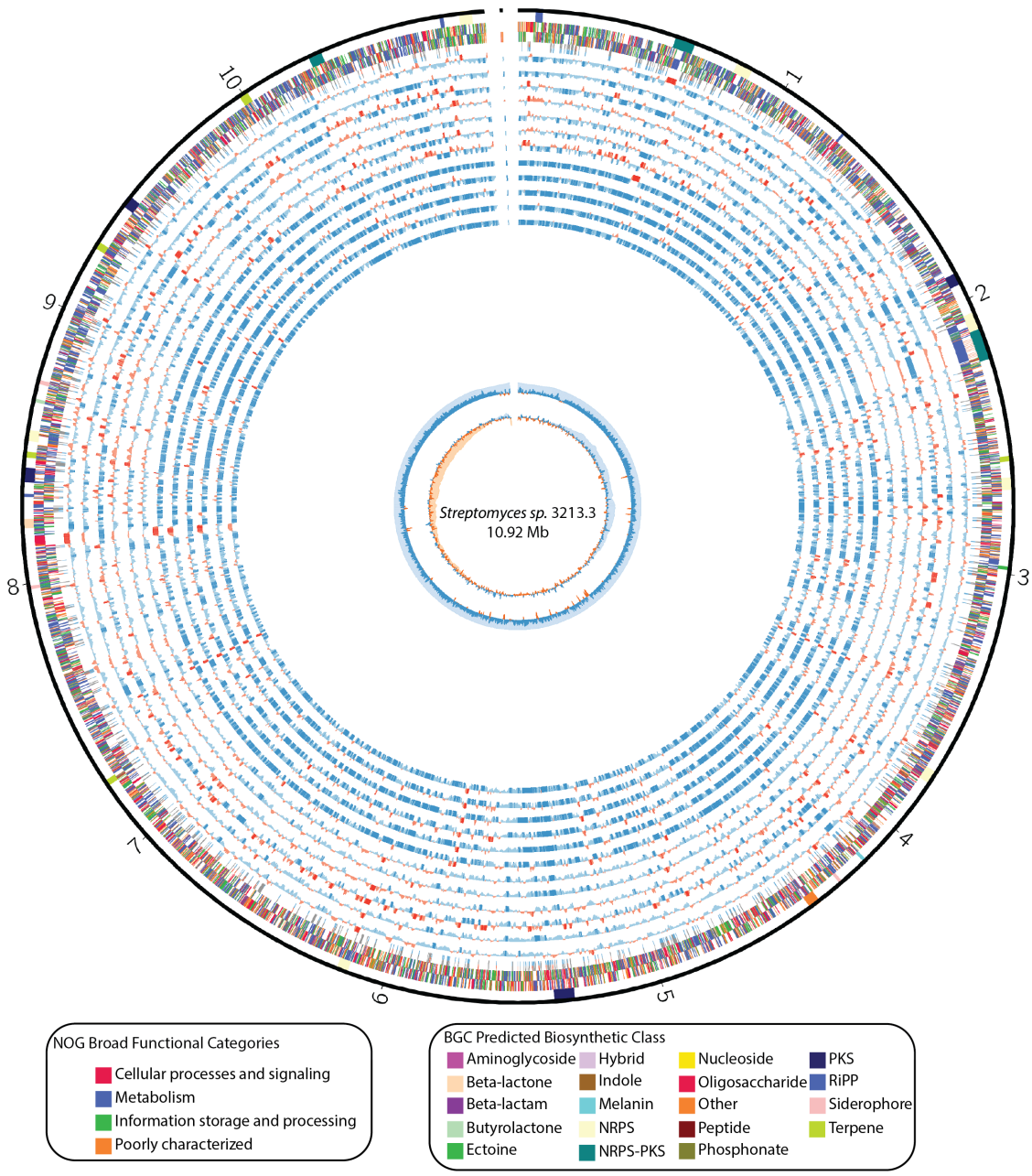
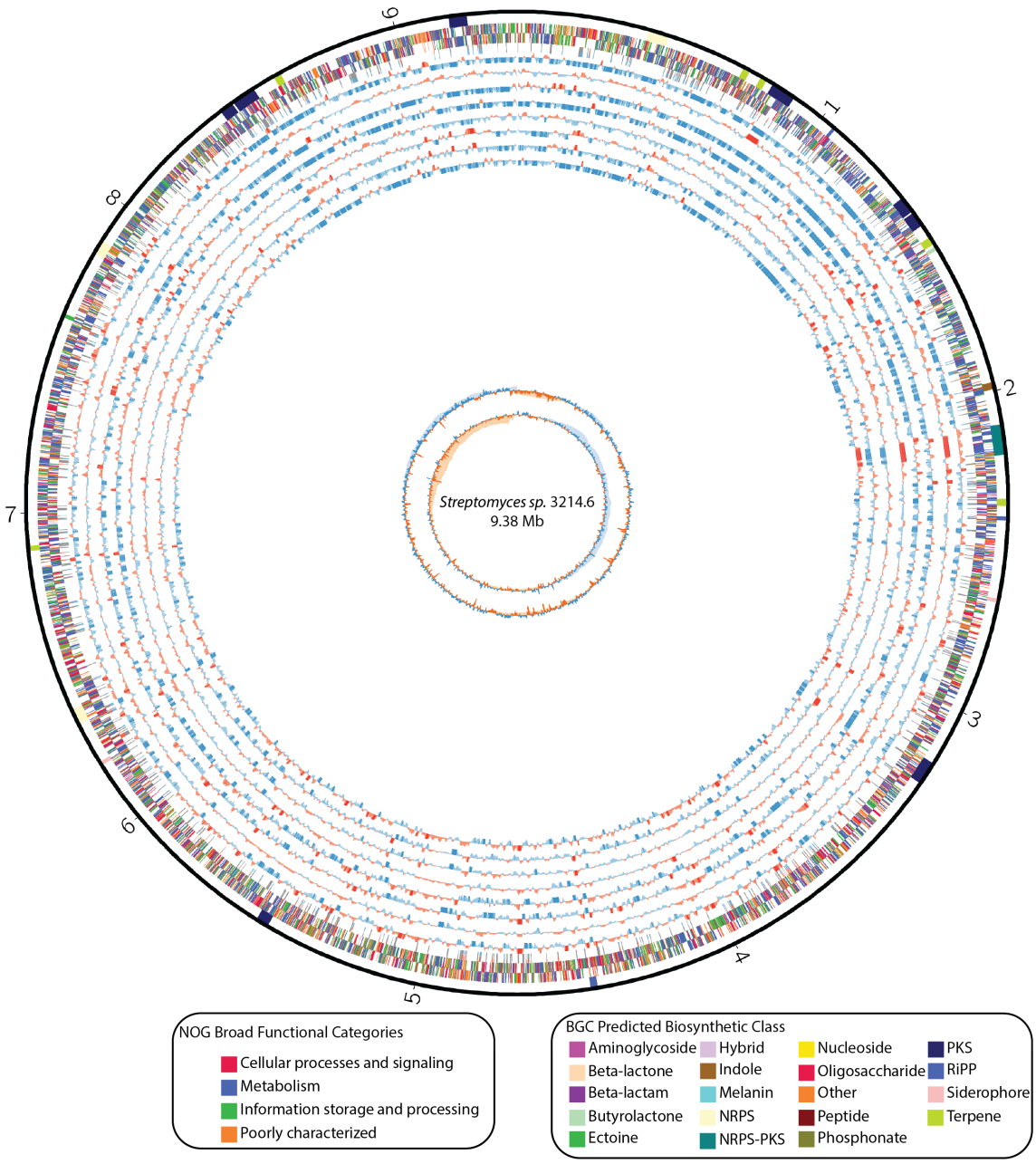


Figure B.12 ***(following page)*****:** Genome map of *Streptomyces sp.* 3214.6. Outer black ring represent assembled contigs. Genomic position (Mb) indicated along contigs. Tracks from outside to inside represent the following data. Putative BGC annotations, colored by predicted biosynthetic class (track 1). Coding sequences on forward and reverse strands (tracks 2 and 3, respectively), colored by broad NOG functional categories. Coding sequences predicted to encode proteins containing signal peptides. Upper and lower halves of signal peptide markers are colored if they are upregulated (red, log2FC > 1), or downregulated (blue, log2FC < -1) in at least one community (track 4). Moving average of log2FC within a 10,000 bp window, incrementing by 1,000 bp (variable number of tracks depending on isolate, starting at track 5). Average log2FC values are colored red (positive) or blue (negative), and appear dark if values are outside the range [-1, 1]. Inner most tracks, from outside in represent G + C content and G + C skew, respectively. Each is shown for two window sizes: 10 Kb (dark blue above average, dark orange below average) and 1 Mb (light blue above average, light orange below average). Only the 10 Kb resolution data is shown for plasmids.

Figure B.12: (continued)



B.4 Hierarchical composition of synthetic communities

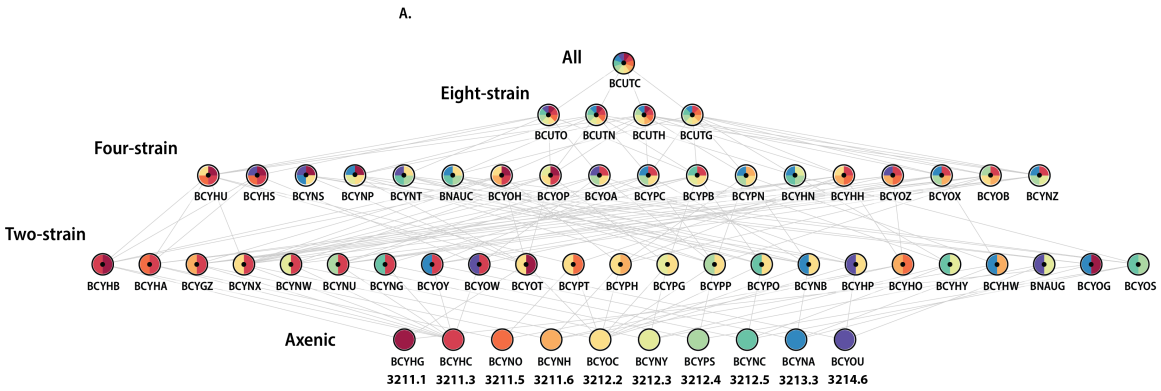


Figure B.13: Hierarchical composition of synthetic communities. Network graph showing color-coded composition of each community analyzed in this study. Communities are arrayed vertically by community complexity with most complex on top and axenic cultures on the bottom. Five-letter codes are community identifiers used to track RNAseq reads. Short-hand abbreviations for strains are given below the axenic nodes (e.g. *Streptomyces sp.* 3211.1 is labeled as '3211.1'). Edges connect nodes when all members from the lower complexity community are present in the higher complexity community.

B.5 Images of example synthetic communities.

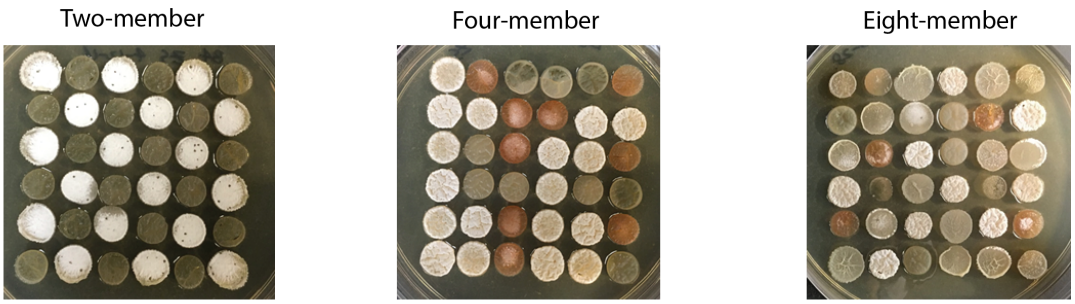


Figure B.14: Example synthetic communities. Two-member (Left), four-member (Middle), eight-member (Right) communities. Each spot is inoculated with 1 uL of a 10^7/mL spore glycerol stock. Medium is ISP2. Images were taken after 72-hour, 30 incubation.

B.6 RNAseq lane allocation and relative loading amounts.

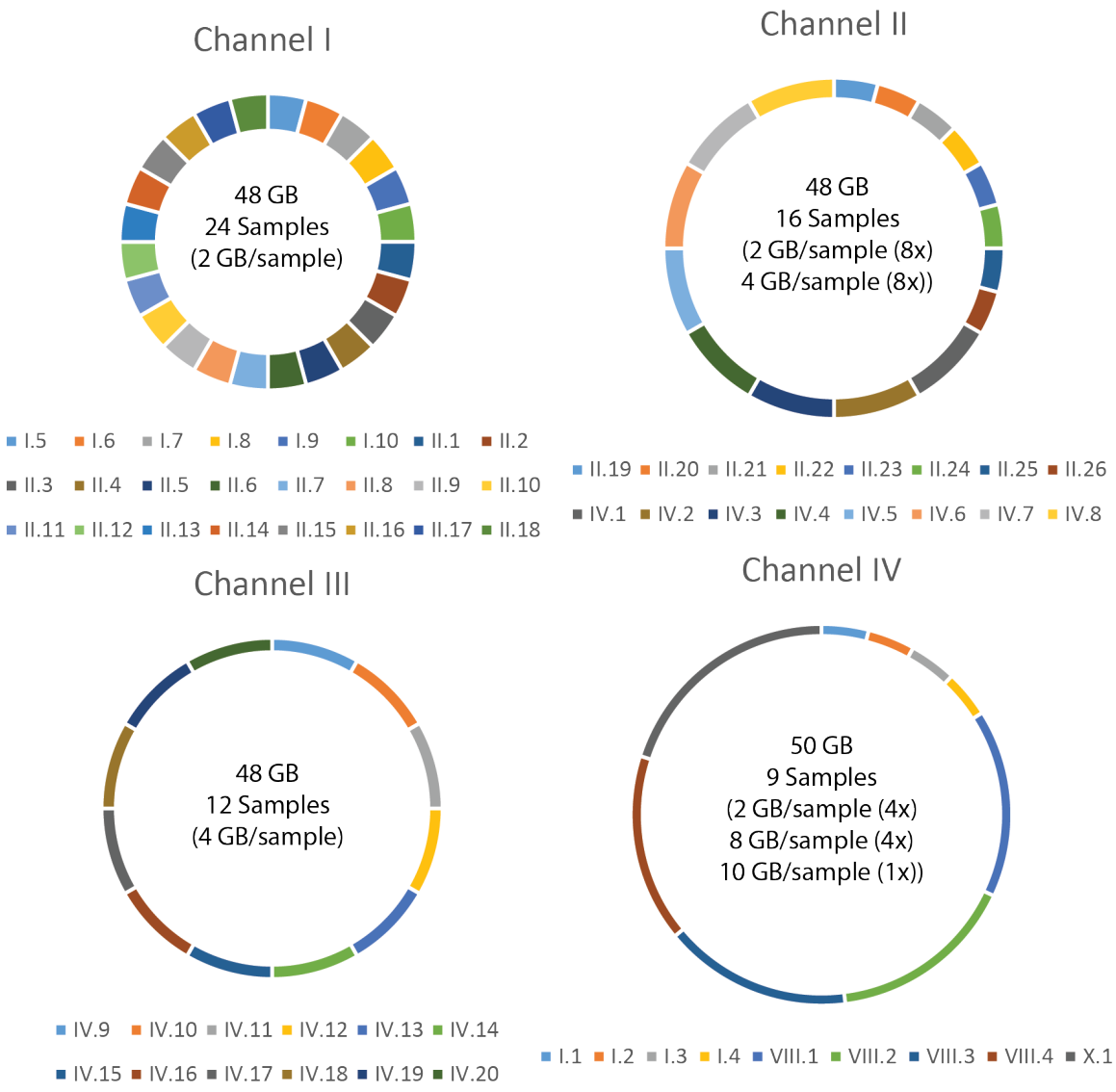


Figure B.15: Allocation of RNAseq samples across 4 Channels to obtain sufficient read depth. Four Illumina Channels were used for sequencing, with approximately 50 Gb of sequencing data expected for each channel. Molar ratio of mixed samples is represented by size of fragment in the donut chart, with larger molar equivalents used for more complex communities, proportionally to give approximately the same read depth per genome. Legend gives community names, with roman numerals denoting the number of genomes present in each community

B.7 Mapped RNAseq reads normalized by adjusted tpm

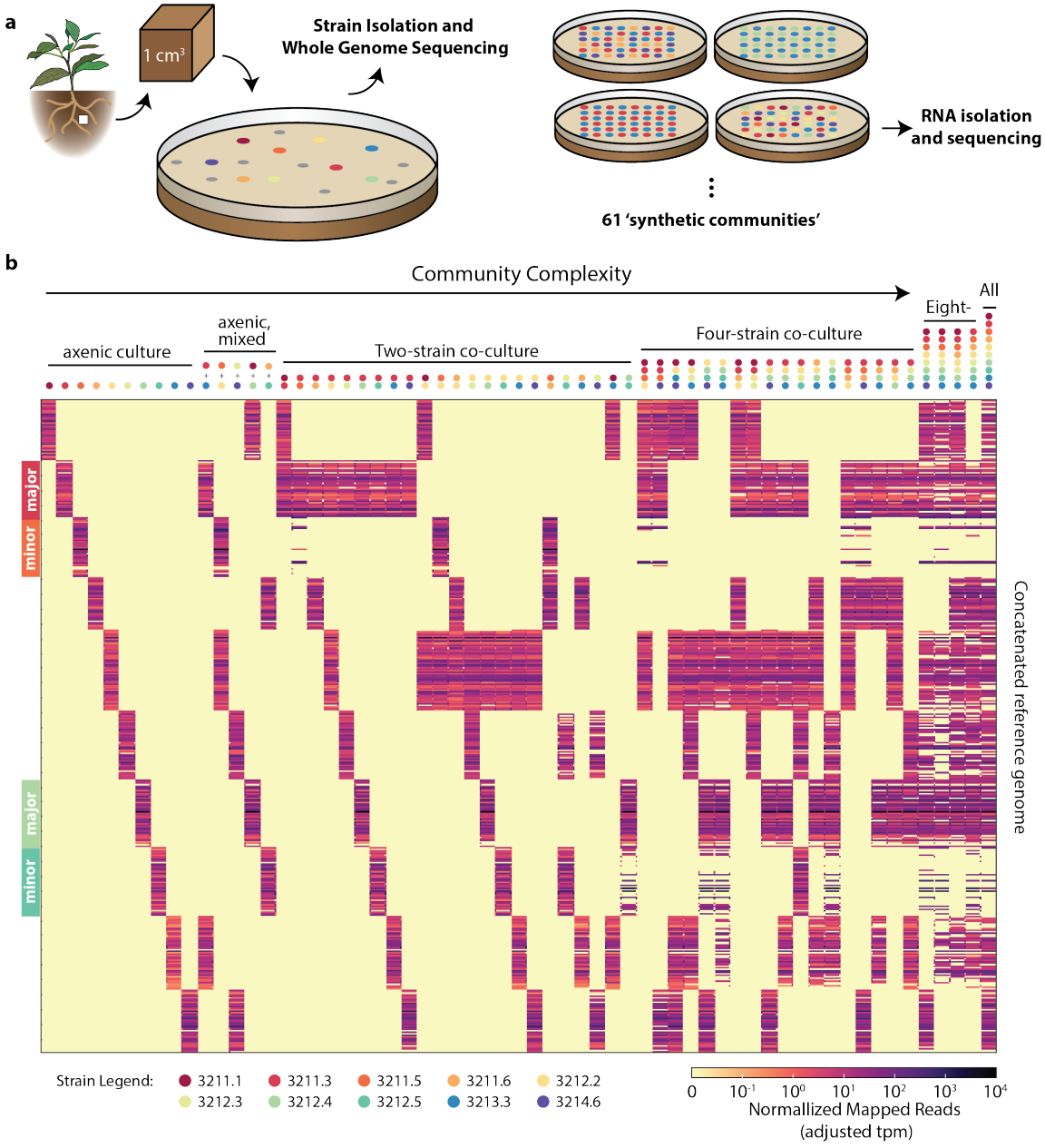


Figure B.16: Data is equivalent to that reported in main text Figure 1b, but normalized via adjusted tpm (transcripts per million). The adjustment is to normalize different samples to the same total read abundance prior to coloring the plot to control for small changes in total read depth between samples.

B.8 Global transcription perturbation as a function of community complexity

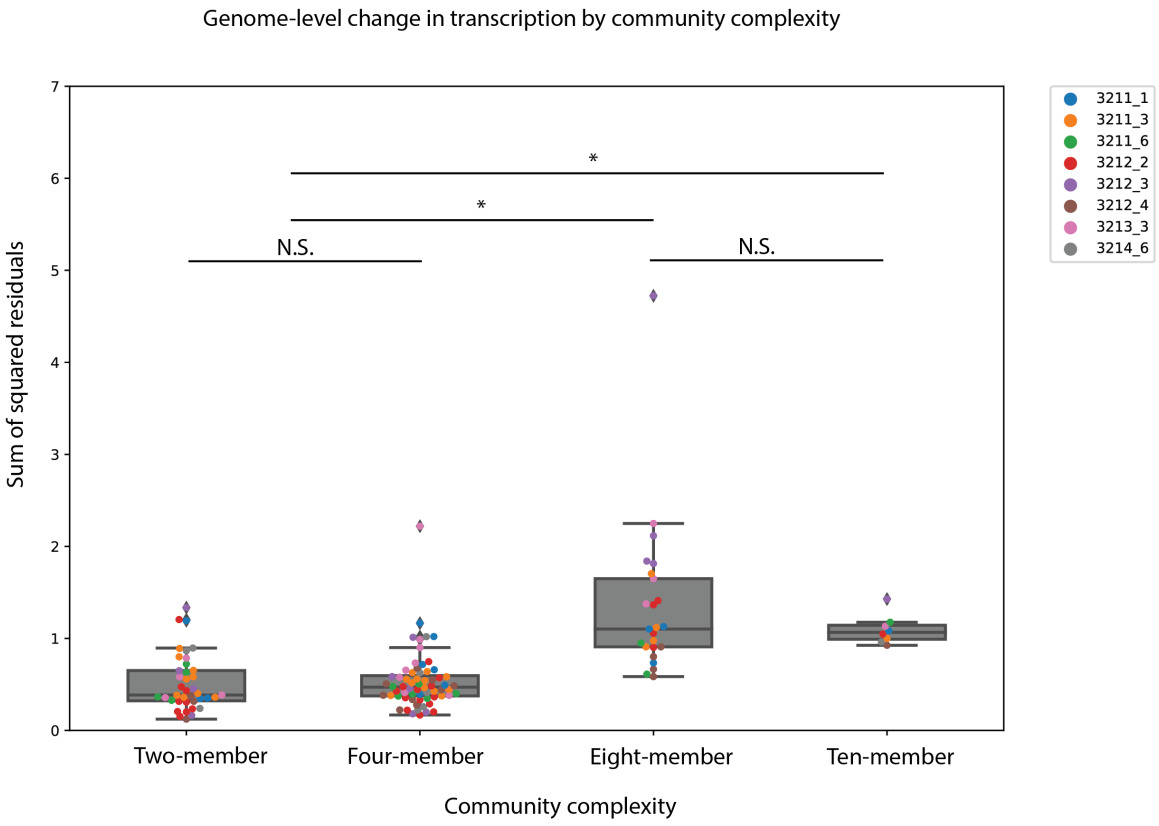


Figure B.17: The sum of square residuals (vertical axis) for plots of axenic versus community RNAseq results for each genome in each community are plotted against community complexity (horizontal axis). Data points are color-coded by genome according to the legend in the upper right and summarized by box plots.Medians between complexity levels were significantly different (Kruskal-Wallis, H=56.985, p=2.589e-12). Stars above lines indicate level of significance (Dunn's test, Bonferroni adjusted p-value) between medians of connected samples(*=p<0.001).

B.9 Percent of reads classified as rRNA at each community complexity level

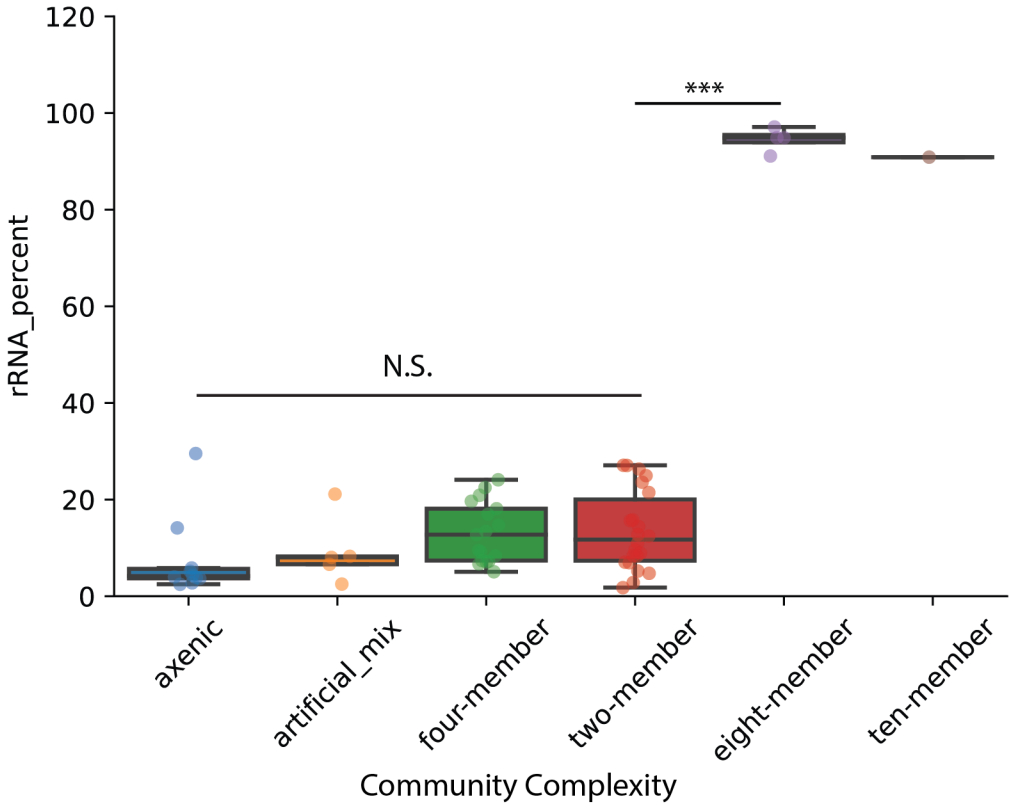


Figure B.18: Percent of reads classified as rRNA at each community complexity level. Data points represent percent of reads for a single community. Data points are grouped by community complexity level and summarized by box plots. Mean rRNA_percent between eight-member and two-member communities was significant (Student's t-test, *** $p<0.0001$). Differences between all lower complexity communities was non-significant at $p<0.05$.

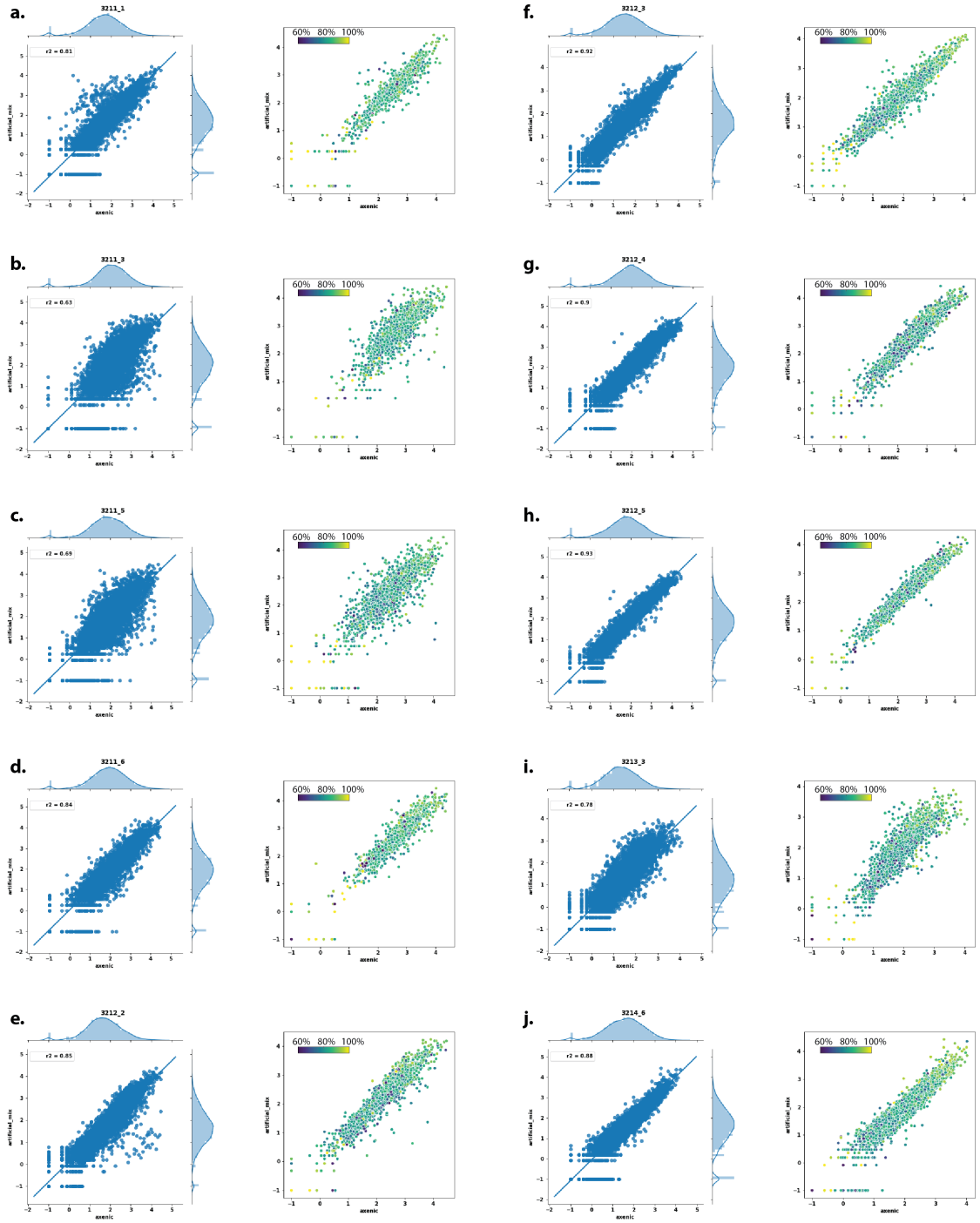
B.10 Supplementary note: Read-mapping is not a major source of bias

Replicate RNAseq experiments for axenic cultures were processed in different ways to identify possible artifacts in the read-mapping pipeline that emerge from mapping reads from multiple strains to a concatenated metagenome. For one replicate, the isolated RNA was kept separate from other samples during library prep, but was pooled with samples containing different barcodes prior to sequencing. For the second replicate, RNA isolated from two unique axenic cultures was mixed prior to library prep. This led to reads mapping to two different genomes (Figure 3.1, 'axenic, mixed' columns) despite the fact that the strains were grown separately. Correlation of mapped read counts between the non-mixed axenic and the mixed axenic samples was strong (Supplementary Figure B.19). If read-mapping from the mixed axenic RNA samples were biasing gene expression results, we would expect the most variable expression results between the non-mixed axenic and the mixed axenic to be for genes that have a high-sequence identity homolog in the genome of the mixed partner. This is not the case (right plots in Supplementary Figure B.19, B.33), so read-mapping from mixed community strains does not appear to be a major source of bias in our experiment.

B.11 Comparing gene expression between axenic and artificially mixed samples

Figure B.19 ***(following page):*** Comparing gene expression between axenic and artificially mixed samples. Log10 expression of each gene in the axenic (x-axis) and artificially mixed sample (y-axis) for all genes (left plots), and for just those with homologs in the mixed partner genome (right plots). In right plots, dot hue denotes percent identity between that gene and the closest homolog in the genome that it was artificially mixed with prior to measuring the expression level (y-axis). (a-j) Different strains, as labeled above the left plot for each subpanel.

Figure B.19: (continued)



B.12 High variance in gene expression in 3211.3 and 3211.5

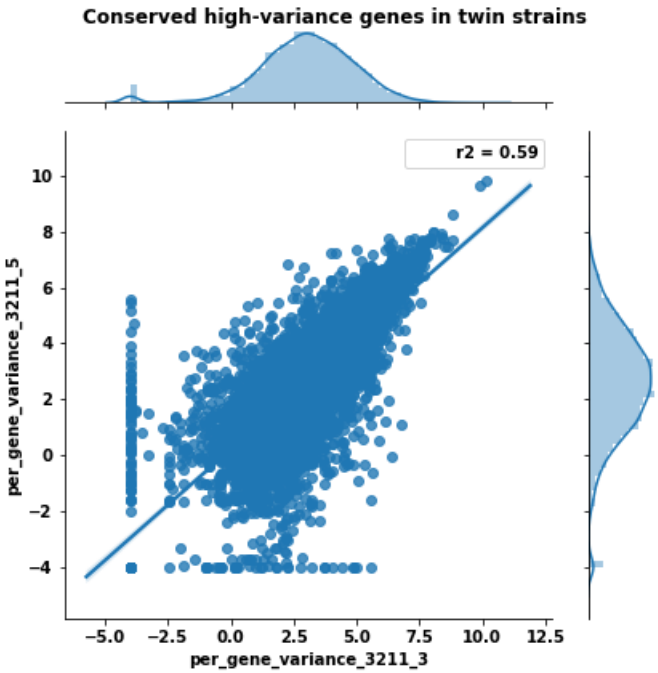


Figure B.20: High variance in gene expression in 3211.3 and 3211.5 is due to reproducibly variant homolog genes, not noise in measurement. Plot shows the per-gene variance (in axenic versus artificial mixed RNAseq experiments) for 3211.3 (x-axis) versus that of it's ortholog in 3211.5 (y-axis)

B.13 Global changes in transcription in pairwise communities.

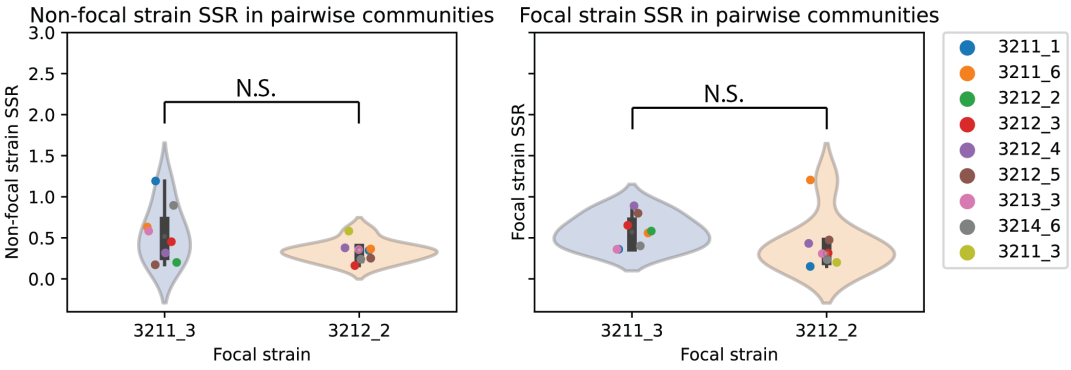


Figure B.21: Global changes in transcription in pairwise communities. SSR for all non-focal isolate strains when in co-culture with focal strains 3211.3 or 3212.2 (Left). SSR for focal isolate strains when in co-culture with non-focal strains (Right).

B.14 Correlation between between DEGs and ecological factors

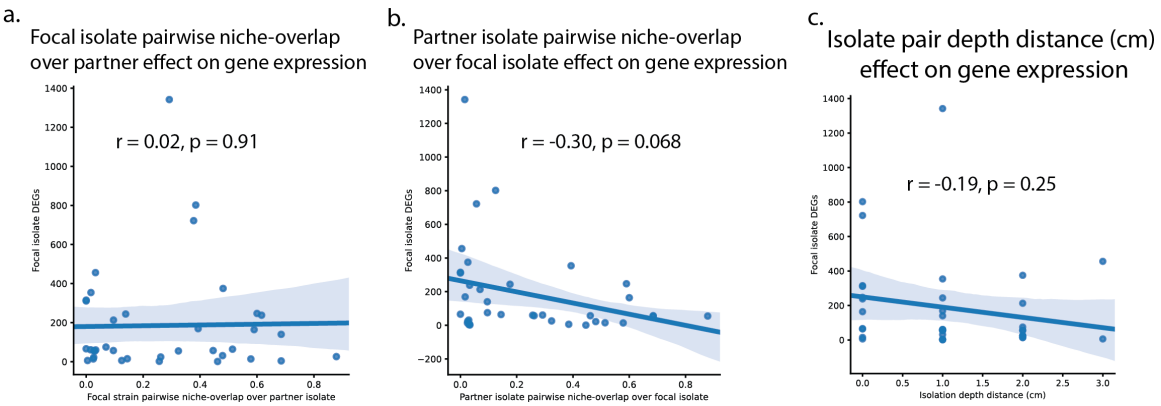
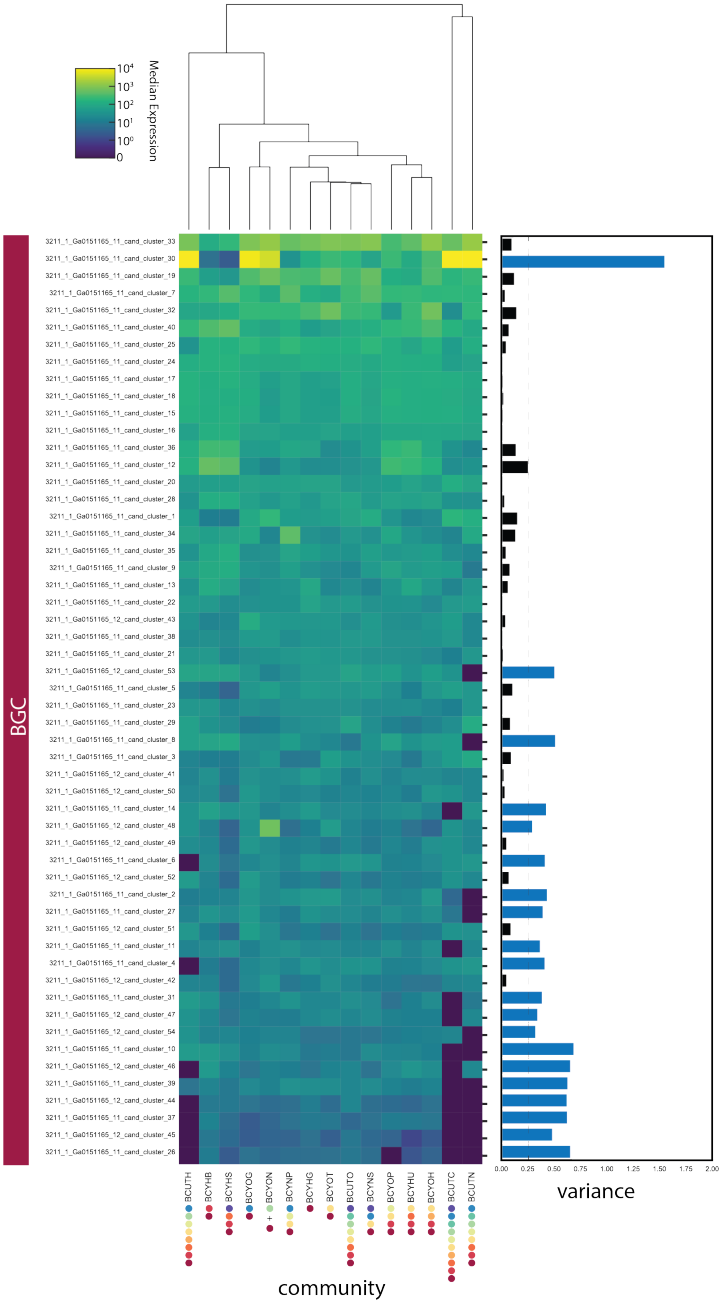


Figure B.22: Correlation of number of differentially expressed genes with ecological factors. Number of differentialy expressed genes in an isolate in pairwise co-culture compared to pairwise niche-overlap over partner (a), partner's pairwise niche-overlap over focal isolate (b), and isolation depth distance between focal and partner isolates (c). Note that in this case, focal does not refer to only to 3211.3 or 3212.2, but rather to any of the eight isolates for which we are measuring the number of DEGs.

B.15 Differential BGC expression for the ten *Streptomyces* isolates

Figure B.23 ***(following page)*****:** Differential BGC expression for strain 3211.1. Heatmap of median DESeq2-normalized reads for genes annotated as 'core biosynthesis' or 'additional biosynthesis' genes in each antiSMASH5.0-identified BGC. Rows (BGCs) are ordered by the row median expression level, with highly expressed BGCs on top. Columns (communities) are clustered based on the pattern of BGC expression and labeled with names and species color code across the x-axis. The bar graph at right shows the variance across each row, with BGCs that have a variance above the arbitrary cutoff of 0.25 highlighted in blue.

Figure B.23: (continued)



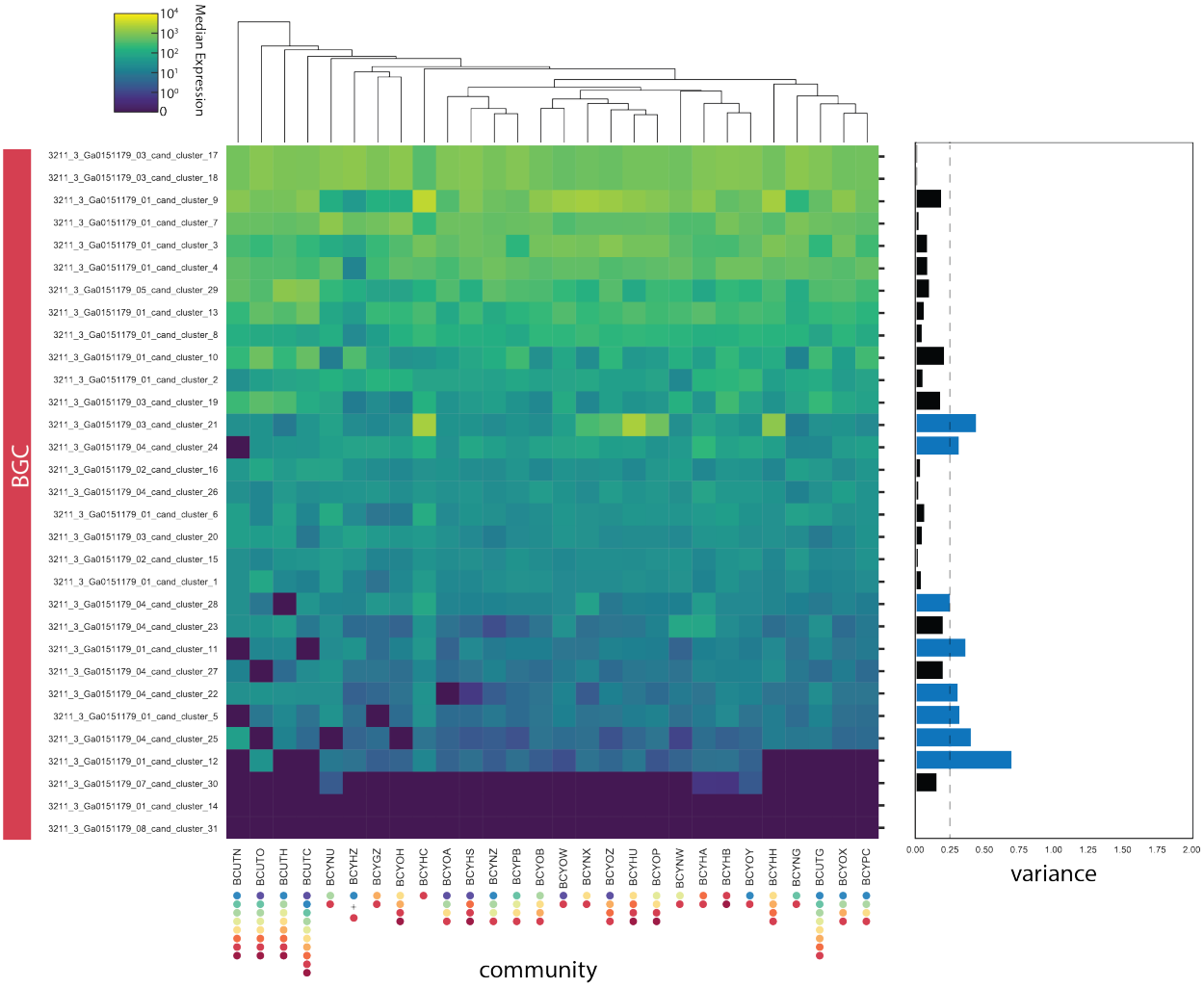


Figure B.24: Differential BGC expression for strain 3211.3. Heatmap of median DESeq2-normalized reads for genes annotated as 'core biosynthesis' or 'additional biosynthesis' genes in each antiSMASH5.0-identified BGC. Rows (BGCs) are ordered by the row median expression level, with highly expressed BGCs on top. Columns (communities) are clustered based on the pattern of BGC expression and labeled with names and species color code across the x-axis. The bar graph at right shows the variance across each row, with BGCs that have a variance above the arbitrary cutoff of 0.25 highlighted in blue.

Figure B.25 ***(following page)*****:** Differential BGC expression for strain 3211.5. Heatmap of median DESeq2-normalized reads for genes annotated as 'core biosynthesis' or 'additional biosynthesis' genes in each antiSMASH5.0-identified BGC. Rows (BGCs) are ordered by the row median expression level, with highly expressed BGCs on top. Columns (communities) are clustered based on the pattern of BGC expression and labeled with names and species color code across the x-axis. The bar graph at right shows the variance across each row, with BGCs that have a variance above the arbitrary cutoff of 0.25 highlighted in blue. Note that 3211.5 is a 'minor twin', which specially impacts its read mapping in communities with 3211.3.

Figure B.25: (continued)

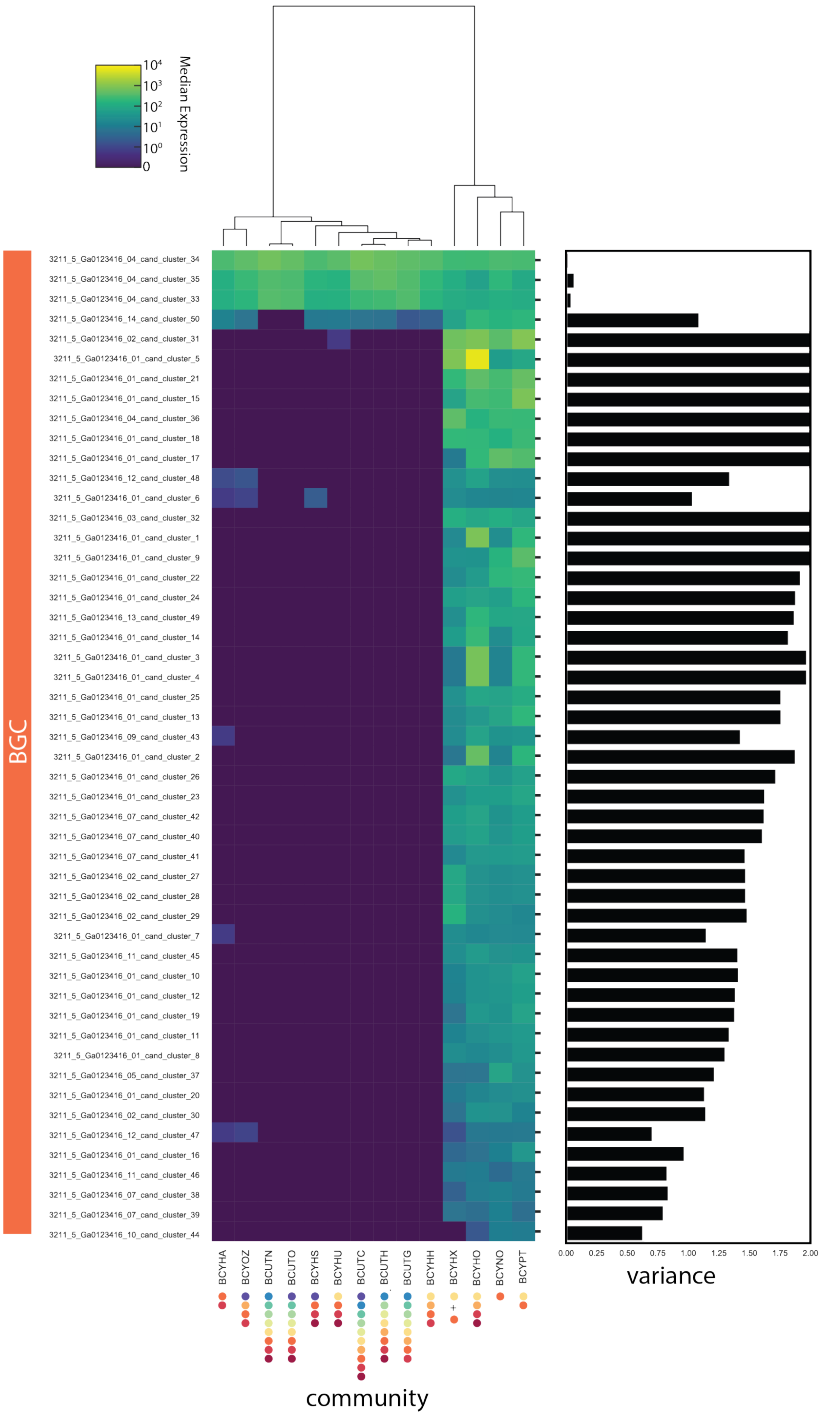
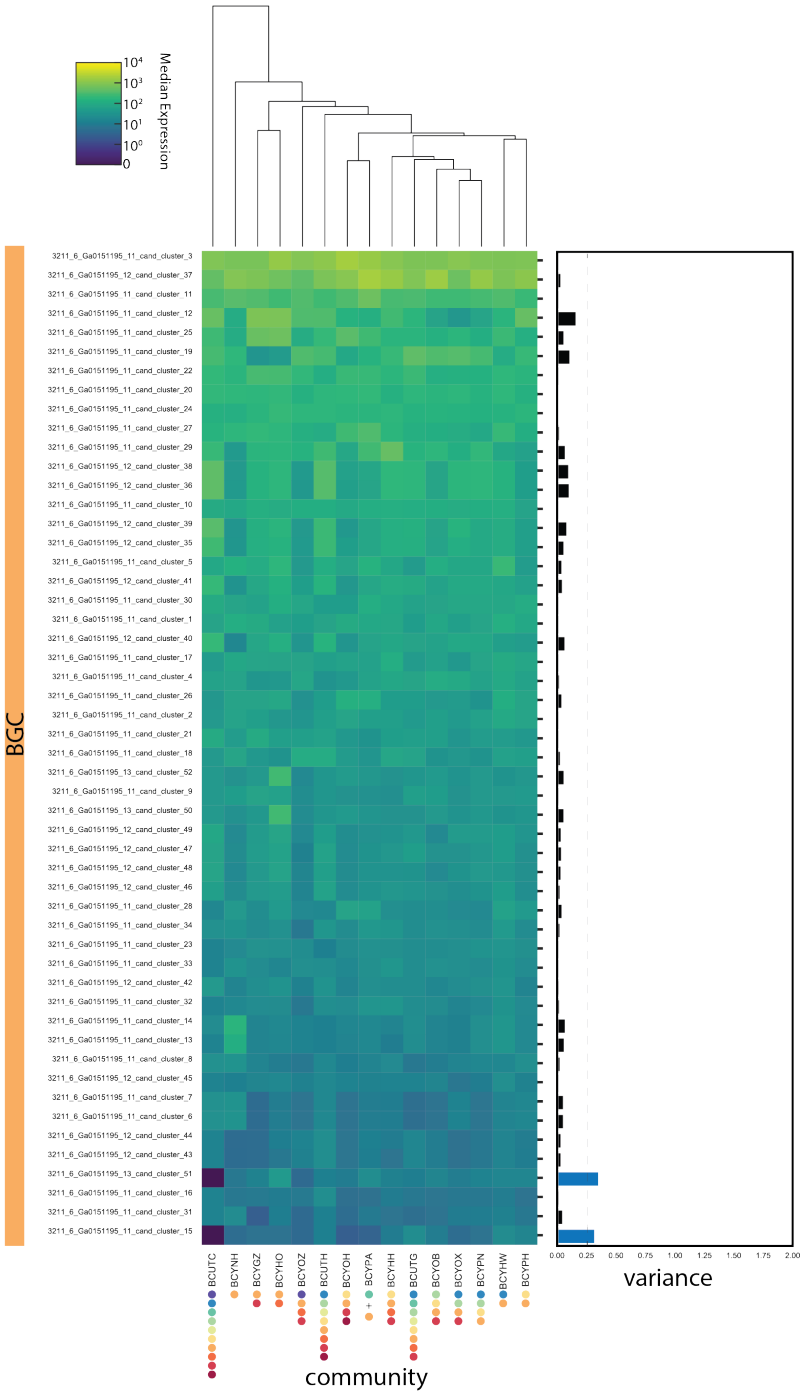


Figure B.26 ***(following page)*****:** Differential BGC expression for strain 3211.6. Heatmap of median DESeq2-normalized reads for genes annotated as 'core biosynthesis' or 'additional biosynthesis' genes in each antiSMASH5.0-identified BGC. Rows (BGCs) are ordered by the row median expression level, with highly expressed BGCs on top. Columns (communities) are clustered based on the pattern of BGC expression and labeled with names and species color code across the x-axis. The bar graph at right shows the variance across each row, with BGCs that have a variance above the arbitrary cutoff of 0.25 highlighted in blue.

Figure B.26: (continued)



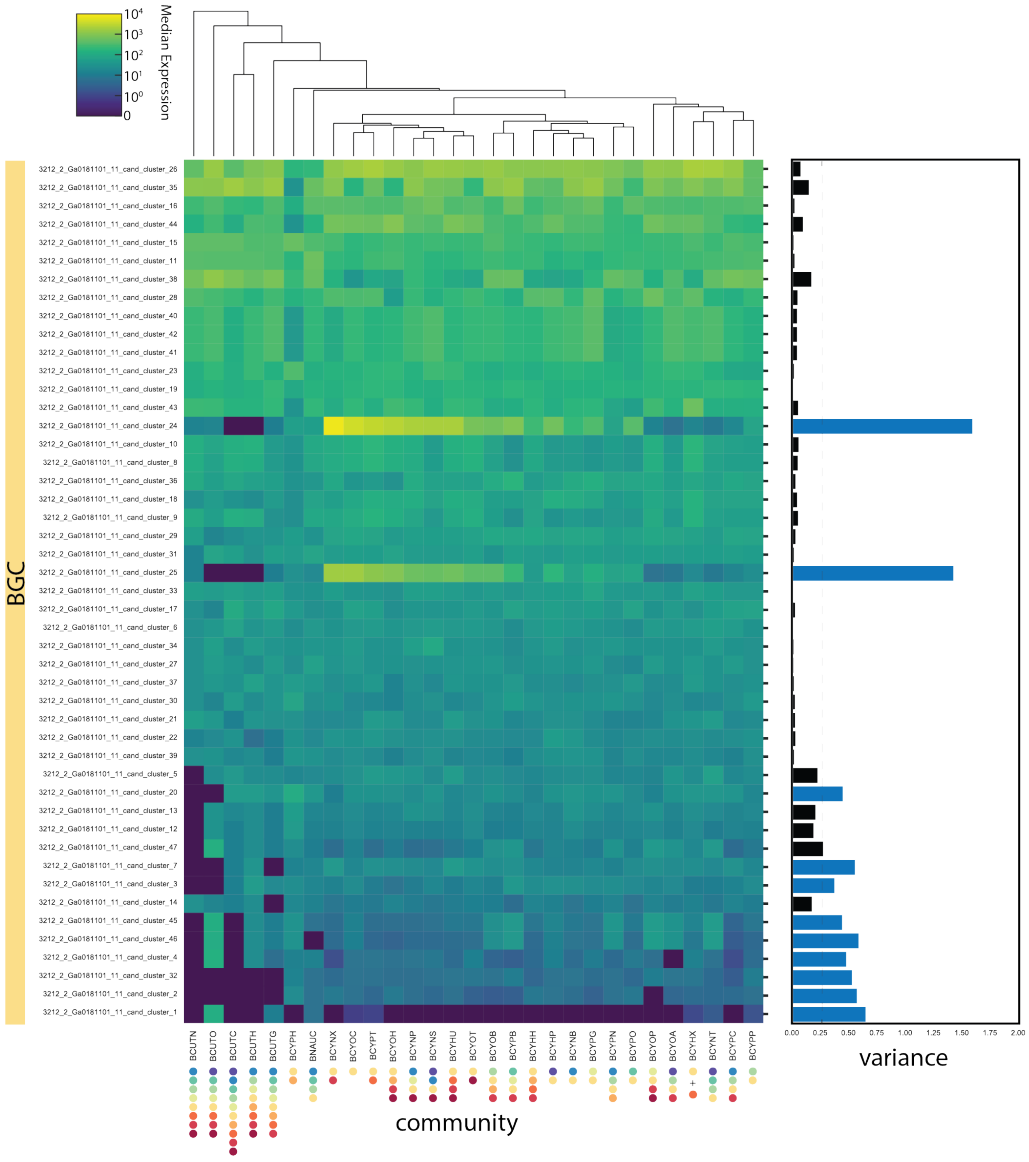


Figure B.27: Differential BGC expression for strain 3212.2. Heatmap of median DESeq2-normalized reads for genes annotated as 'core biosynthesis' or 'additional biosynthesis' genes in each antiSMASH5.0-identified BGC. Rows (BGCs) are ordered by the row median expression level, with highly expressed BGCs on top. Columns (communities) are clustered based on the pattern of BGC expression and labeled with names and species color code across the x-axis. The bar graph at right shows the variance across each row, with BGCs that have a variance above the arbitrary cutoff of 0.25 highlighted in blue.

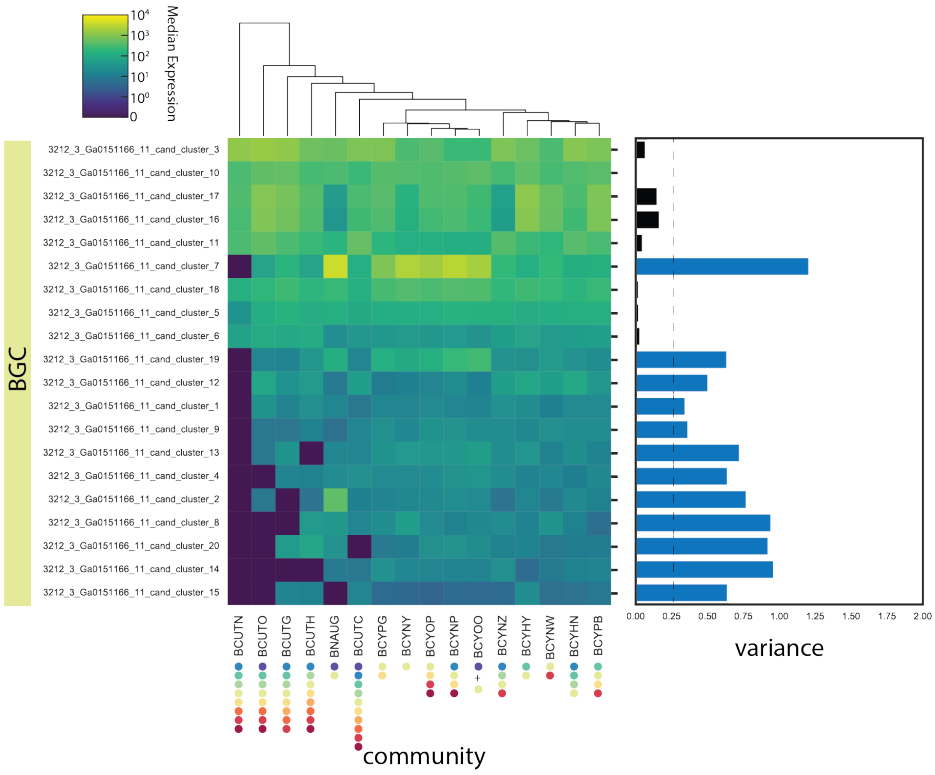


Figure B.28: Differential BGC expression for strain 3212.3. Heatmap of median DESeq2-normalized reads for genes annotated as 'core biosynthesis' or 'additional biosynthesis' genes in each antiSMASH5.0-identified BGC. Rows (BGCs) are ordered by the row median expression level, with highly expressed BGCs on top. Columns (communities) are clustered based on the pattern of BGC expression and labeled with names and species color code across the x-axis. The bar graph at right shows the variance across each row, with BGCs that have a variance above the arbitrary cutoff of 0.25 highlighted in blue.

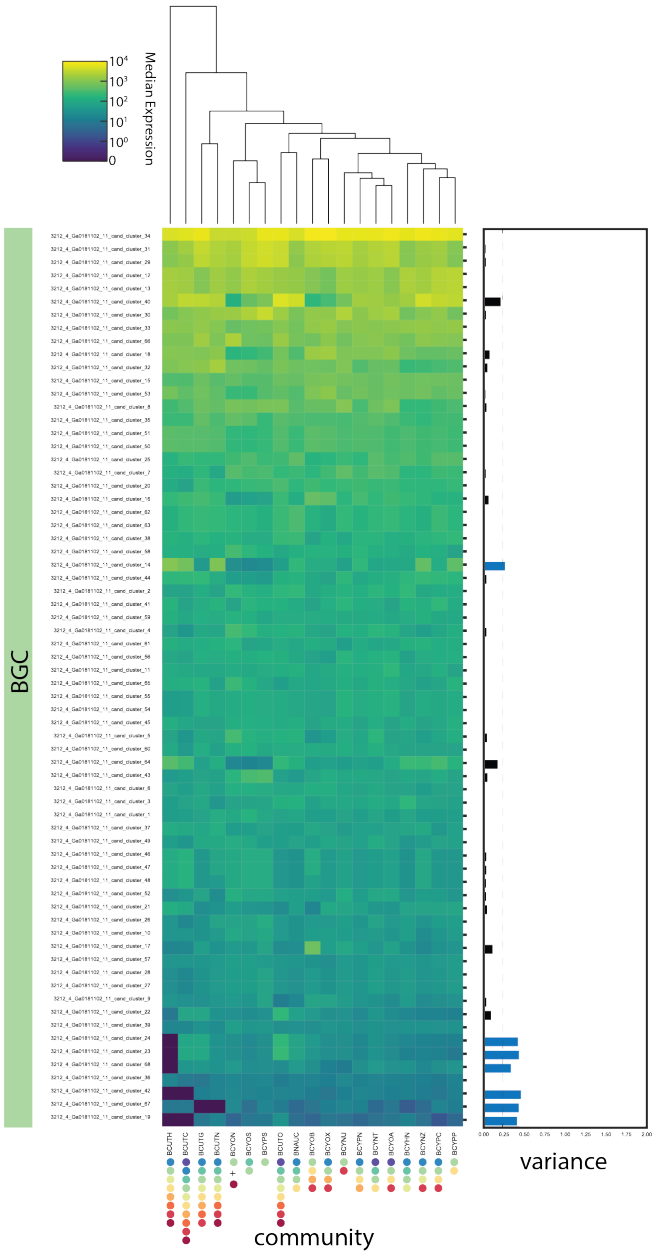
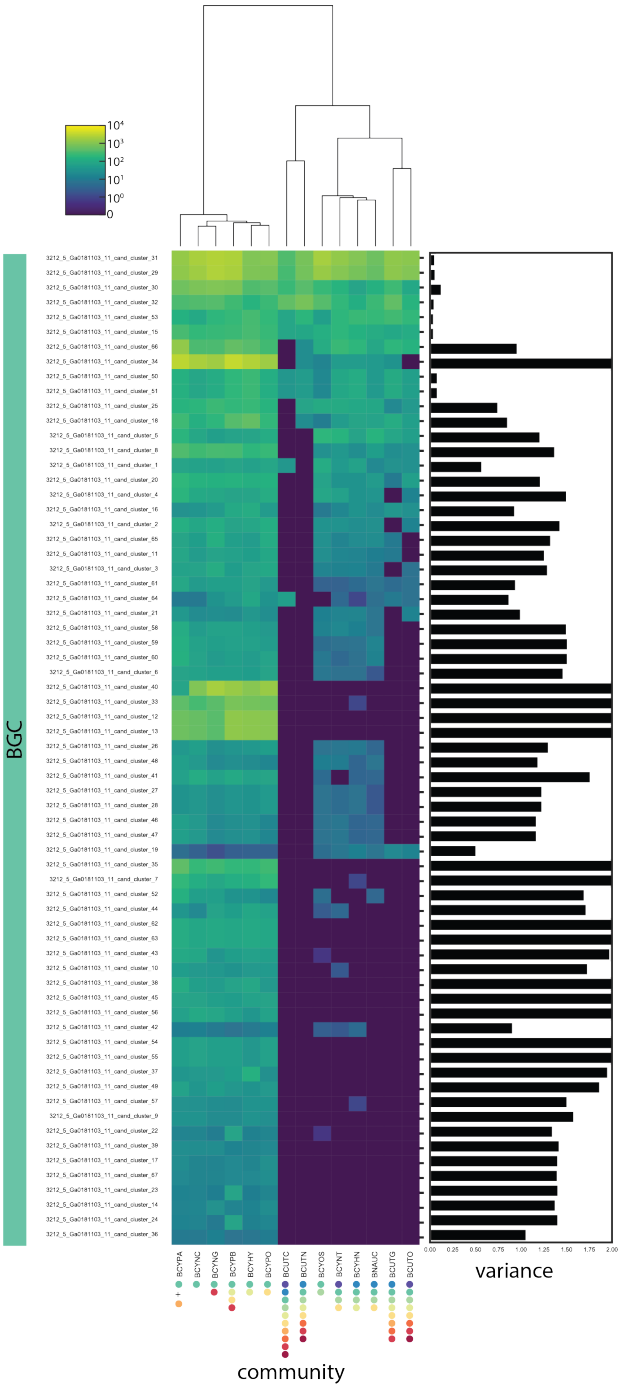


Figure B.29: Differential BGC expression for strain 3212.4. Heatmap of median DESeq2-normalized reads for genes annotated as 'core biosynthesis' or 'additional biosynthesis' genes in each antiSMASH5.0-identified BGC. Rows (BGCs) are ordered by the row median expression level, with highly expressed BGCs on top. Columns (communities) are clustered based on the pattern of BGC expression and labeled with names and species color code across the x-axis. The bar graph at right shows the variance across each row, with BGCs that have a variance above the arbitrary cutoff of 0.25 highlighted in blue.

Figure B.30 ***(following page)*****:** Differential BGC expression for strain 3212.5. Heatmap of median DESeq2-normalized reads for genes annotated as 'core biosynthesis' or 'additional biosynthesis' genes in each antiSMASH5.0-identified BGC. Rows (BGCs) are ordered by the row median expression level, with highly expressed BGCs on top. Columns (communities) are clustered based on the pattern of BGC expression and labeled with names and species color code across the x-axis. The bar graph at right shows the variance across each row, with BGCs that have a variance above the arbitrary cutoff of 0.25 highlighted in blue. Note that 3212.5 is a 'minor twin', which specially impacts its read mapping in communities with 3212.4.

Figure B.30: (continued)



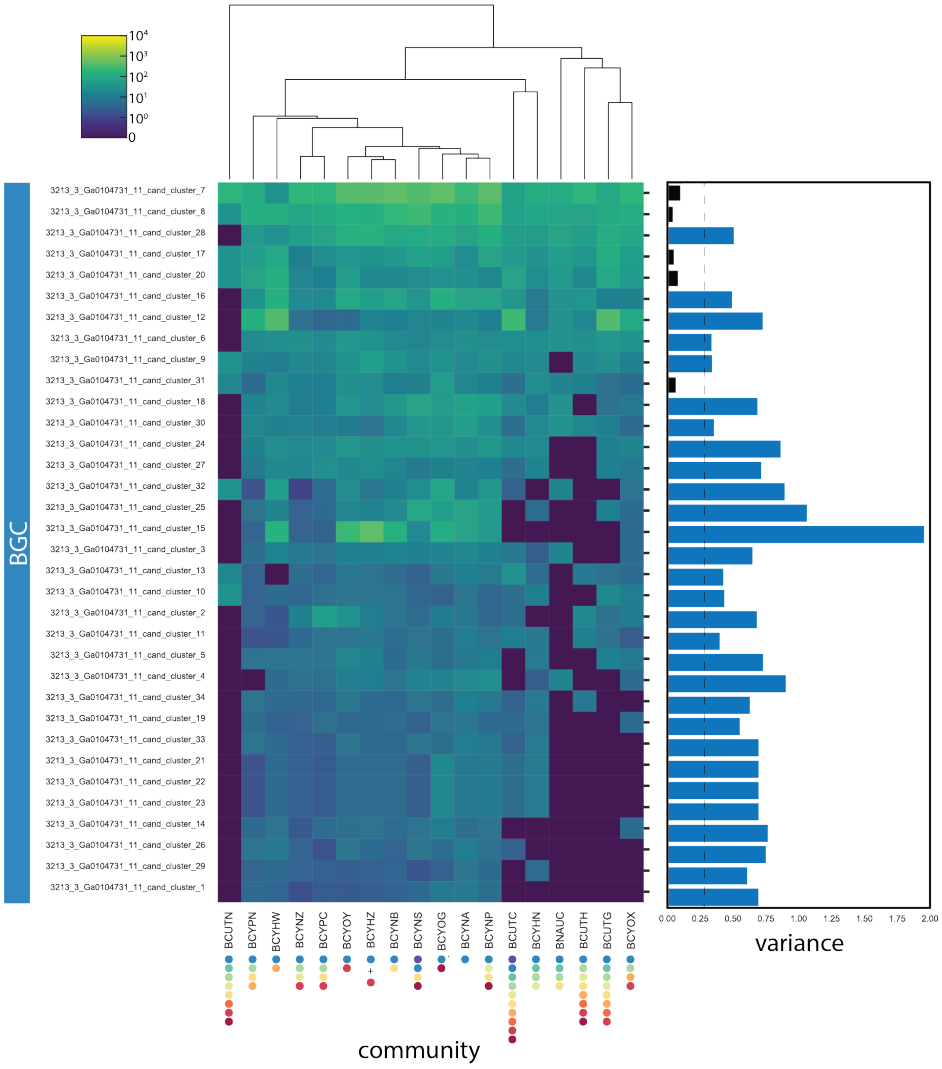


Figure B.31: Differential BGC expression for strain 3213.3. Heatmap of median DESeq2-normalized reads for genes annotated as 'core biosynthesis' or 'additional biosynthesis' genes in each antiSMASH5.0-identified BGC. Rows (BGCs) are ordered by the row median expression level, with highly expressed BGCs on top. Columns (communities) are clustered based on the pattern of BGC expression and labeled with names and species color code across the x-axis. The bar graph at right shows the variance across each row, with BGCs that have a variance above the arbitrary cutoff of 0.25 highlighted in blue.

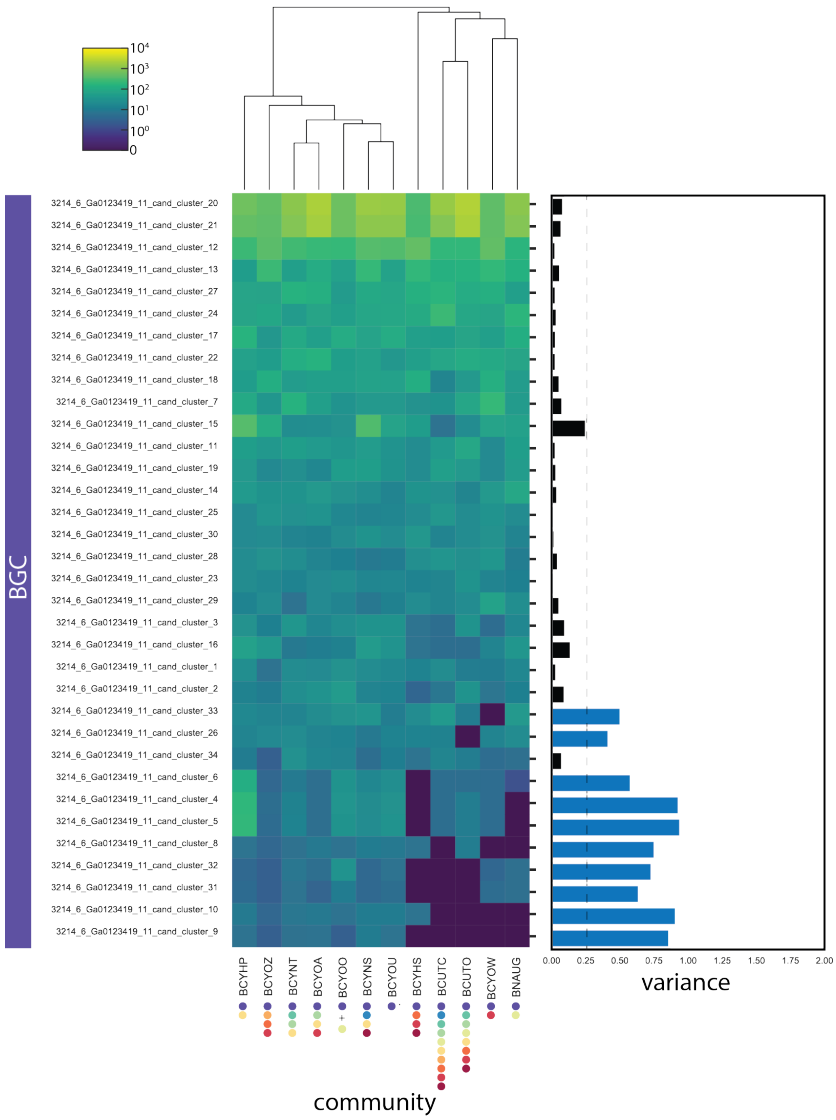


Figure B.32: Differential BGC expression for strain 3214.6. Heatmap of median DESeq2-normalized reads for genes annotated as 'core biosynthesis' or 'additional biosynthesis' genes in each antiSMASH5.0-identified BGC. Rows (BGCs) are ordered by the row median expression level, with highly expressed BGCs on top. Columns (communities) are clustered based on the pattern of BGC expression and labeled with names and species color code across the x-axis. The bar graph at right shows the variance across each row, with BGCs that have a variance above the arbitrary cutoff of 0.25 highlighted in blue.

B.16 Curated high-confidence BGCs.

Table B.1: *Streptomyces* sp. 3211.1 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
NZ_QXDI01000001.1	1	453025	474984	lanthipeptide	Amfs-like[200]
	2	629952	652167	lanthipeptide	
	3	737599	756179	terpene	
	4	773313	841430	terpene,hglE-KS,T1PKS,CDPS	
	5	773313	793241	terpene	2-methylisoborneol[71]
	6	781902	833516	hglE-KS,T1PKS	
	7	820714	841430	CDPS	
	9	944727	979516	thiopeptide,LAP	
	10	956999	996626	T3PKS	Alkylresorcinol[201]
	11	1026657	1038551	siderophore	
	12	1114635	1124302	melanin	
	13	1135348	1144459	melanin	
	14	1149433	1169578	terpene	
	16	1264478	1290936	T3PKS	
	17	1272538	1338908	NRPS	
	18	1301918	1343228	ladderane	
	19	1417109	1437458	terpene	
	20	1479375	1499808	terpene	
	22	1502516	1528812	terpene	Geosmin[74]

Table B.1: *Streptomyces* sp. 3211.1 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	23	1515560	1556039	T3PKS	Flaviolin [202]
	24	1596966	1623245	terpene	Hopene [75]
	25	1900365	1922653	terpene	Geosmin [74]
	27	2021905	2045207	lanthipeptide	
	28	2046326	2056479	bacteriocin	
	29	2265987	2279096	siderophore	
	30	2752641	2763513	butyrolactone	
	31	3517382	3539830	lassopeptide	
	32	3863355	3904209	phosphonate	Fosfazinomycin [203]
	33	4872710	4884635	bacteriocin	
	34	5147766	5159550	siderophore	Desferrioxamine B [73]
	35	5212543	5242610	thiopeptide,LAP	
	36	6756739	6777980	amglyccycl	
	37	7360439	7370867	melanin	
	38	7378099	7421698	NRPS-like	
	39	7583333	7624385	T3PKS	Flaviolin [202]
	40	7874540	7973010	NRPS,T1PKS	
NZ_QXDI01000002.1	41	51	44563	NRPS	
	43	75894	98449	lanthipeptide	

Table B.1: *Streptomyces* sp. 3211.1 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	44	90771	131394	other	
	45	108416	152498	NRPS	
	46	141527	182327	NRPS-like	Lipstatin [204]
	47	168146	212222	thiopeptide,LAP	Lactazole-like [205]
	48	192347	285615	NRPS	
	50	319460	430947	T1PKS	
	51	397318	453325	NRPS	
	52	400639	473187	T2PKS	
	53	444141	455142	butyrolactone	
	54	476922	497503	linaridin	

Table B.2: *Streptomyces* sp. 3211.3 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
NZ_JACHIY010000001.1	1	595269	650126	T1PKS	
	2	760241	771589	bacteriocin	
	3	882099	903758	terpene	Geosmin [74]
	4	967938	1012316	T1PKS	
	5	1376991	1428603	T1PKS,hglE-KS	
	6	1496783	1554439	NRPS	
	7	1652165	1699387	NRPS	
	8	1900563	1921659	terpene	
	9	1990266	2013045	lanthipeptide	SapB [76,206]
	10	2196676	2207061	melanin	
	11	2336355	2347868	siderophore	
	12	2379671	2420731	T3PKS	Alkylresorcinol [201]
	13	2468097	2488653	CDPS	
	14	2638178	2649136	butyrolactone	
NZ_JACHIY010000001.2	15	1319679	1362503	NRPS-like,butyrolactone	
	16	1343337	1354349	butyrolactone	
NZ_JACHIY010000001.3	17	398418	449481	T1PKS,NRPS,betalactone	
	18	404130	449481	NRPS,betalactone	
	19	477499	488469	butyrolactone	

Table B.2: *Streptomyces* sp. 3211.3 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	20	519346	564024	T1PKS	
	21	671580	714905	NRPS	
NZ_JACHIY010000001.4	22	42799	72865	thiopeptide,LAP	Lactazole-like[205]
	23	74768	129534	NRPS,ectoine	
	24	166511	187550	terpene	Avermitilol[207]
	25	493111	503391	bacteriocin	
	26	574516	617901	NRPS-like	Lipstatin[204]
	27	648215	658618	ectoine	Ectoine[208]
	28	721280	793773	T2PKS	Spore pigment[209]
NZ_JACHIY010000001.5	29	206937	218717	siderophore	Desferrioxamine B[73]
NZ_JACHIY010000001.7	30	1	22957	phosphonate	
NZ_JACHIY010000001.8	31	190132	217966	phosphonate	

Table B.3: *Streptomyces* sp. 3211.5 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
NZ_JACHIX010000001.1	2	403085	444713	phosphonate	
	3	409192	450451	ladderane	
	4	416159	461588	T1PKS	
	5	458099	478461	nucleoside	
	7	1112066	1123079	butyrolactone	
	8	1979242	2020405	arylpolyene	
	9	2910828	2925902	siderophore	
	11	3017219	3080783	NRPS	
	12	3060934	3115792	T1PKS	
	13	3082242	3123054	other	
	14	3225906	3237255	bacteriocin	
	15	3347764	3369424	terpene	Geosmin [74]
	16	3373103	3414566	T1PKS	
	17	3433603	3477982	T1PKS	
	18	3758462	3785345	terpene	Hopene-like [75]
	19	3842656	3894269	T1PKS,hglE-KS	
	20	3934474	4020105	NRPS	
	21	4117830	4165053	NRPS	
	22	4366690	4386330	terpene	

Table B.3: *Streptomyces* sp. 3211.5 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	23	4455931	4478711	lanthipeptide	SapB[76,206]
	24	4516760	4537427	terpene	2-methylisoborneol[71]
	25	4621073	4641978	terpene	
	26	4645920	4672727	melanin	Melanin[210]
NZ_JACHIX010000002.1	28	5566	55333	NRPS	
	29	35934	69217	T1PKS	
	30	100679	111650	butyrolactone	
	31	139667	202167	betalactone,NRPS,T1PKS	
NZ_JACHIX010000003.1	32	4871	16652	siderophore	Desferrioxamine B[73]
NZ_JACHIX010000004.1	33	52651	75464	lanthipeptide	
	34	100956	146634	NRPS	
	35	262687	280533	terpene	
	36	451800	472837	terpene	
NZ_JACHIX010000005.1	37	323312	393137	NRPS	JBIR-126[211]
NZ_JACHIX010000007.1	38	42622	72689	thiopeptide,LAP	Lactazole-like[205]
	39	75128	128272	NRPS,ectoine	
	40	138218	205840	butyrolactone,NRPS	
	41	138218	149177	butyrolactone	
	42	142899	205840	NRPS	

Table B.3: *Streptomyces* sp. 3211.5 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
NZ_JACHIX010000009.1	43	99320	171814	T2PKS	Spore pigment [209]
NZ_JACHIX010000010.1	44	81411	91692	bacteriocin	
NZ_JACHIX010000011.1	45	45424	58615	siderophore	
	46	88740	129801	T3PKS	Alkylresorcinol [201]
NZ_JACHIX010000012.1	47	177	10581	ectoine	Ectoine [208]
	48	40894	66976	NRPS-like	Lipstatin [204]
NZ_JACHIX010000013.1	49	5606	34464	NRPS-like	
NZ_JACHIX010000014.1	50	3922	24479	CDPS	

Table B.4: *Streptomyces* sp. 3211.6 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
NZ_RBXC01000001.1	2	102271	143223	NRPS	
	3	139794	149108	bacteriocin	
	4	307994	350505	NRPS-like	
	5	378567	401206	lanthipeptide	SapB [76,206]
	7	548498	589562	T3PKS	Alkylresorcinol [201]
	8	579034	590484	siderophore	
	9	641136	670883	melanin	
	10	673298	692965	terpene	
	11	737380	757535	terpene	2-methylisoborneol [71]
	12	792295	852497	NRPS	Streptothricin [72]
	14	933020	1005526	T2PKS	Spore pigment [209]
	15	983961	1028310	T1PKS	
	16	1417984	1447557	thiopeptide,LAP	Lactazole-like [205]
	17	2723117	2787104	T1PKS,NRPS	
	18	3070277	3081356	siderophore	Desferrioxamine B [73]
	19	5845263	5859692	siderophore	
	20	5916545	5936922	nucleoside	
	21	6081167	6092462	bacteriocin	
	22	6177082	6198277	terpene	Geosmin [74]

Table B.4: *Streptomyces* sp. 3211.6 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	23	6329094	6395885	NRPS	
	24	6532950	6557565	terpene	Hopene[75]
	25	6708479	6729588	terpene	
	27	7002116	7042788	NRPS-like	
	28	7034387	7057790	LAP	
	29	7098695	7142161	NRPS	
	30	7155121	7175831	CDPS	
	31	7177516	7230392	T1PKS,hglE-KS	
	33	7280120	7324100	NRPS-like	
	34	7307716	7348519	other	
NZ_RBXC01000002.1	36	1	66485	T2PKS	
	37	13547	24542	butyrolactone	
	38	26953	70515	NRPS-like	
	39	28958	101485	T2PKS	
	40	51031	89022	NRPS	
	41	62783	149209	T1PKS	
	43	161263	204838	NRPS	
	44	171355	211537	other	
	45	200687	244308	T1PKS	

Table B.4: *Streptomyces* sp. 3211.6 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	46	214564	278853	other,melanin	
	47	239813	297931	NRPS	Azinothrycin-like[212]
	48	269164	327742	T1PKS	Azinothrycin-like[212]
	49	287867	331118	NRPS	
NZ_RBXC01000003.1	51	1	37237	NRPS-like	
	52	81	190935	T1PKS	

Table B.5: *Streptomyces* sp. 3212.2 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
NZ_QEOG00000000.1	1	135031	144112	butyrolactone	
	2	253605	273612	terpene	Avermitilol[207]
	3	280035	318789	T3PKS	Alkylresorcinol[201]
	4	415219	424609	butyrolactone	
	5	913857	924168	bacteriocin	
	6	1221383	1268401	NRPS-like,T1PKS	
	7	1548084	1555788	bacteriocin	
	9	1738241	1774636	other	
	10	1754957	1849531	NRPS	
	11	1842647	1853125	butyrolactone	
	13	2014785	2084904	T2PKS	Spore pigment[209]
	14	2077880	2118839	other	
	15	2100899	2123493	indole	
	16	2161155	2185678	terpene	Hopene-like[75]
	17	2910898	2923135	siderophore	
	18	3129748	3180873	hglE-KS,T1PKS	
	19	3369149	3391145	terpene	Geosmin[74]
	20	3422956	3434317	bacteriocin	
	21	3631951	3646674	siderophore	

Table B.5: *Streptomyces* sp. 3212.2 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	22	4250767	4293431	NRPS	
	23	4364311	4383561	terpene	Albaflavenone[213]
	24	5846121	5918603	T2PKS	
	25	5882249	5893355	butyrolactone	
	26	6571220	6593767	lassopeptide	MS-271[214]
	27	7160208	7216404	NRPS,T1PKS	
	28	7243967	7255739	siderophore	Desferrioxamine B[73]
	29	7366107	7376182	melanin	Melanin[210]
	30	8573179	8583583	ectoine	Ectoine[208]
	31	8820900	8863061	NRPS-like	
	32	9022946	9052394	betalactone	
	34	9543276	9584340	T3PKS	
	35	9582699	9603061	nucleoside	
	36	9586720	9640695	NRPS	
	37	9937118	9992637	NRPS,T1PKS	
	38	10417568	10427951	melanin	Melanin-like[210]
	39	10825928	10866764	NRPS-like	
	41	10955792	10976850	terpene	2-methylisoborneol[71]
	42	10962051	10983965	lanthipeptide	

Table B.5: *Streptomyces* sp. 3212.2 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	43	11174189	11219026	NRPS	
	44	11392057	11414420	lanthipeptide	
	45	11489906	11594564	NRPS	
	46	11489906	11545656	NRPS	
	47	11532646	11594564	NRPS	

Table B.6: *Streptomyces* sp. 3212.3 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
NZ_QTTM01000001.1	1	668339	711608	NRPS	
	2	668630	677460	bacteriocin	
	3	780568	790951	melanin	Melanin [210]
	4	1733998	1772465	T3PKS	
	5	2308158	2350850	NRPS-like	
	6	2483438	2493836	ectoine	Ectoine [208]
	7	3547130	3555661	bacteriocin	
	8	3670589	3681869	siderophore	Desferrioxamine B [73]
	9	4516705	4587123	T2PKS	Spore pigment [209]
	10	6234626	6255590	terpene	Albaflavenone [213]
	11	6908517	6918649	siderophore	
	12	7163237	7172554	bacteriocin	
	13	7199284	7220754	terpene	Geosmin [74]
	14	7338959	7352054	siderophore	
	15	7734334	7775392	T3PKS	
	16	7768708	7841211	T2PKS	
	17	7783718	7807946	oligosaccharide	
	18	8178453	8205183	terpene	
	19	8369245	8413129	NRPS-like	

Table B.6: *Streptomyces* sp. 3212.3 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	20	8566024	8576239	bacteriocin	

Table B.7: *Streptomyces* sp. 3212.4 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
NZ_RBIY01000001.1	1	47498	68199	terpene	
	3	164014	185496	lanthipeptide	
	4	179653	227178	NRPS-like,T1PKS	Blasticidin [215]
	5	220176	230907	bacteriocin	
	7	265342	287071	lanthipeptide	
	8	272135	323146	T1PKS	
	9	320007	360981	NRPS	
	10	344328	389158	T1PKS	
	11	417704	460207	T1PKS	
	13	630789	653479	lanthipeptide	Informatipeptin-like [206]
	14	648280	656439	bacteriocin	
	15	1211039	1237106	terpene	Hopene [75]
	16	1364866	1507091	NRPS,terpene	
	17	1562418	1614761	NRPS	
	18	1715607	1802468	T1PKS,NRPS	
	19	1843746	1855703	siderophore	
	20	2048497	2069839	terpene	Geosmin [74]
	21	2097577	2108235	bacteriocin	
	23	2173620	2217871	T1PKS	

Table B.7: *Streptomyces* sp. 3212.4 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	24	2185126	2225069	other	
	25	2229129	2240985	bacteriocin	
	27	2242349	2293387	NRPS	
	28	2256915	2297177	other	
	30	2375270	2417055	NRPS	
	31	2379702	2550843	T1PKS,NRPS	
	32	2380828	2391522	siderophore	
	33	2712367	2812921	transAT-PKS,transAT-PKS-like,NRPS	Phthoxazolin [216]
	34	3126384	3254278	T1PKS	Oligomycin [217]
	35	3377452	3398257	terpene	Albaflavenone [213]
	36	3580933	3677051	T1PKS,aminocoumarin,NRPS	
	37	5161391	5233915	T2PKS	Spore pigment [209]
	38	5294935	5315837	LAP	
	39	5400146	5476245	NRPS	
	40	6009640	6020097	siderophore	Desferrioxamine B [73]
	41	6053776	6083697	thiopeptide,LAP	Lactazole [205]
	42	6100688	6146323	NRPS,melanin	Melanin [210]
	43	6567692	6587879	terpene	
	44	7221143	7231547	ectoine	Ectoine [208]

Table B.7: *Streptomyces* sp. 3212.4 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	45	7460295	7501710	NRPS-like	
	47	7906940	7954181	hglE-KS	
	48	7916666	7958945	T1PKS	
	49	8032026	8073090	T3PKS	
	51	8075841	8129830	NRPS	Tyrobetaine[218]
	52	8082883	8095428	lassopeptide	Citrulassin E[219]
	53	8093225	8134220	PKS-like	
	55	8316911	8364187	NRPS	
	56	8343697	8366433	lassopeptide	lagmysin-like[219]
	57	8445212	8486585	other	
	58	8588568	8632532	T1PKS	
	60	8751862	8771230	terpene	
	61	8768790	8824417	NRPS,T1PKS	Foxicin-like[220]
	63	8854359	8872879	terpene	2-methylisoborneol[71]
	64	8864946	8875323	melanin	Melanin[210]
	65	8989142	9009852	CDPS	
	66	9482908	9503143	terpene	Pentalenolactone[221]
	67	9736740	9750061	siderophore	
	68	9965234	9986130	terpene	

Table B.8: *Streptomyces* sp. 3212.5 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
NZ_QXCZ01000001.1	1	75749	96475	terpene	
	3	191026	212939	lanthipeptide	
	4	207096	254594	NRPS-like,T1PKS	Blasticidin [215]
	5	248430	258432	bacteriocin	
	7	292942	314536	lanthipeptide	
	8	298253	349703	T1PKS	
	9	347571	388546	NRPS	
	10	371893	416723	T1PKS	
	11	446100	489064	T1PKS	
	13	659636	682326	lanthipeptide	Informatipeptin-like [206]
	14	677127	685304	bacteriocin	
	15	1246610	1272677	terpene	Hopene [75]
	16	1400167	1542463	NRPS,terpene	
	17	1597842	1650180	NRPS	
	18	1751007	1838762	T1PKS,NRPS	
	19	1880023	1891725	siderophore	
	20	2084138	2105481	terpene	Geosmin [74]
	21	2133034	2143695	bacteriocin	
	23	2209231	2253488	T1PKS	

Table B.8: *Streptomyces* sp. 3212.5 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	24	2220742	2260686	other	
	25	2264746	2276602	bacteriocin	
	27	2277966	2329027	NRPS	
	28	2292532	2332817	other	
	30	2410969	2446188	NRPS	
	31	2415401	2586578	T1PKS,NRPS	
	32	2416527	2427179	siderophore	
	33	2748087	2848638	transAT-PKS,transAT-PKS-like,NRPS	Phthoxazolin [216]
	34	3167372	3295257	T1PKS	Oligomycin [217]
	35	3418535	3439340	terpene	Albaflavenone [213]
	36	3607146	3703264	T1PKS,aminocoumarin,NRPS	
	37	5165099	5237623	T2PKS	Spore pigment [209]
	38	5298681	5319583	LAP	
	39	5403874	5479943	NRPS	
	40	6012745	6024514	siderophore	Desferrioxamine B [73]
	41	6057354	6087275	thiopeptide,LAP	Lactazole [205]
	42	6104278	6149973	NRPS,melanin	Melanin [210]
	43	6519952	6540139	terpene	
	44	7170953	7181357	ectoine	Ectoine [208]

Table B.8: *Streptomyces* sp. 3212.5 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	45	7410126	7451510	NRPS-like	
	47	7856696	7903937	hglE-KS	
	48	7866326	7908702	T1PKS	
	49	7989900	8030964	T3PKS	
	51	8033715	8088149	NRPS	Tyrobetaine[218]
	52	8040757	8053302	lassopeptide	Citrulassin E[219]
	53	8051099	8092094	PKS-like	
	55	8330522	8377809	NRPS	
	56	8357308	8380044	lassopeptide	lagmysin-like[219]
	57	8458821	8500194	other	
	58	8602295	8646258	T1PKS	
	60	8768023	8787391	terpene	
	61	8784951	8840578	NRPS,T1PKS	Foxicin-like[220]
	63	8870612	8889132	terpene	2-methylisoborneol[71]
	64	8881199	8891576	melanin	Melanin[210]
	65	9005593	9026303	CDPS	
	66	9478188	9498421	terpene	Pentalenolactone[221]
	67	9983526	10004422	terpene	

Table B.9: *Streptomyces* sp. 3213.3 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
NZ_FNTK01000001.1	1	64478	89630	lanthipeptide	
	2	572104	643436	NRPS,T1PKS	
	3	809297	866703	NRPS	
	4	1258200	1266037	bacteriocin	
	5	1898605	1939669	T3PKS	
	6	2050852	2112803	NRPS	Scabichelin [222]
	7	2114194	2225915	T1PKS,NRPS-like	Naphthomycin [223]
	8	2579103	2596731	terpene	
	9	2655852	2699162	NRPS-like	
	10	2973183	2983587	ectoine	Ectoine [208]
	11	3752217	3798843	NRPS	
	12	4149041	4159484	melanin	Melanin [210]
	13	4263060	4273691	siderophore	Desferrioxamine B [73]
	14	4368015	4408635	other	
	15	5313694	5386182	T2PKS	Granaticin [224]
	16	6134891	6173904	NRPS-like	
	17	7224190	7245017	terpene	Albaflavenone [213]
	18	7980551	7991233	siderophore	
	19	8198105	8230139	betalactone	

Table B.9: *Streptomyces* sp. 3213.3 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	20	8309576	8320079	bacteriocin	
	22	8361333	8404679	T1PKS	
	23	8365048	8411902	hglE-KS	
	24	8449103	8469505	terpene	Geosmin [74]
	25	8508830	8547266	NRPS-like	
	26	8648889	8659815	butyrolactone	
	27	8714937	8728045	siderophore	
	28	9238172	9262918	terpene	Hopene [75]
	29	9417317	9458319	T1PKS	Cahuitamycins [225]
	30	9984434	10015384	terpene	2-methylisoborneol [71]
	31	10266471	10313439	T1PKS,NRPS-like	
	32	10754493	10766421	bacteriocin	
	33	10824765	10869717	NRPS	
	34	10827317	10847679	nucleoside	

Table B.10: *Streptomyces* sp. 3214.6 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
NZ_LT670819.1	1	406396	467863	NRPS	
	2	717880	738204	terpene	2-methylisoborneol[71]
	3	777309	796209	terpene	
	5	819136	891437	T2PKS	
	6	854195	895451	ladderane	
	7	1044196	1052363	bacteriocin	
	9	1345338	1390772	hglE-KS	
	10	1351972	1395351	T1PKS	
	11	1401531	1447137	T1PKS	
	12	1490206	1514631	terpene	Hopene[75]
	13	1523959	1534894	butyrolactone	
	14	1972107	1993234	indole	
	15	2103262	2195964	NRPS,PKS-like,T1PKS	
	16	2179962	2192806	siderophore	
	17	2331538	2352920	terpene	Geosmin[74]
	18	2385836	2397266	bacteriocin	
	19	2637161	2648748	siderophore	
	20	3178681	3250644	T2PKS	Cinerubin B[226]
	21	3192151	3233257	PKS-like	Cinerubin B[226]

Table B.10: *Streptomyces* sp. 3214.6 gene clusters.

Accession	Candidate Cluster	Start	End	Class	Annotation
	22	4441812	4462390	linaridin	
	23	5484205	5522895	T3PKS	
	24	6191478	6203247	siderophore	Desferrioxamine B [73]
	25	6335111	6382259	NRPS	
	26	6341479	6351964	melanin	Melanin [210]
	27	6886048	6900357	terpene	
	28	7611727	7622131	ectoine	Ectoine [208]
	29	7828855	7872721	NRPS-like	
	30	8408302	8449366	T3PKS	
	31	8456269	8528784	T2PKS	Spore pigment [209]
	32	8459645	8483201	blactam	
	33	8600711	8622597	terpene	
	34	9166946	9221365	hglE-KS	

B.17 RNAseq quality control information.

Table B.11: Quality control data for RNAseq experiment.

Instrument Type	Read Length (bp)	Read Count	Percent reads greater than Q30 average	Average Base Quality	Insert Size (bp)	Insert Size Mode	Library Name	Strain Composition	Community Type
HiSeq−2500 1TB	2 × 151	96803806	93.95	36.26 +-5.23	188 +-52	158	BCUTN	A,B,C,E,F,G,H,I	eight-member
HiSeq−2500 1TB	2 × 151	92187722	94.06	36.28 +-5.18	184 +-53	158	BCUTO	A,B,C,E,F,G,H,J	eight-member
HiSeq−2500 1TB	2 × 151	80247832	94.07	36.28 +-5.17	183 +-53	158	BCUTG	B,C,D,E,F,G,H,I	eight-member
HiSeq−2500 1TB	2 × 151	95956410	93.55	36.17 +-5.42	188 +-53	158	BCUTH	A,B,C,D,E,F,G,I	eight-member
HiSeq−2500 1TB	2 × 151	72522564	94.07	36.29 +-5.15	187 +-51	158	BCUTC	A,B,C,D,E,F,G,H,I,J	ten-member
HiSeq−2500 1TB	2 × 151	17745672	95.15	36.56 +-4.67	219 +-43	242	BCYNW	B,F	two-member
HiSeq−2500 1TB	2 × 151	21388510	95.37	36.62 +-4.5	210 +-44	165	BCYNS	A,E,I,J	four-member
HiSeq−2500 1TB	2 × 151	16042742	95.34	36.62 +-4.52	216 +-43	242	BCYOA	B,E,G,J	four-member
HiSeq−2500 1TB	2 × 151	19553048	95.41	36.62 +-4.59	212 +-43	207	BCYPT	E,C	two-member
HiSeq−2500 1TB	2 × 151	17307620	94.8	36.5 +-4.76	212 +-44	207	BCYOC	E	axenic
HiSeq−2500 1TB	2 × 151	15743040	94.48	36.41 +-4.92	211 +-43	165	BCYHO	C,D	two-member
HiSeq−2500 1TB	2 × 151	15284256	95.09	36.57 +-4.7	201 +-46	175	BCYNU	B,G	two-member
HiSeq−2500 1TB	2 × 151	18169290	95.29	36.63 +-4.5	212 +-43	243	BCYOZ	B,C,D,J	four-member
HiSeq−2500 1TB	2 × 151	21999416	95.34	36.64 +-4.5	199 +-45	165	BCYHG	A	axenic
HiSeq−2500 1TB	2 × 151	17017072	94.89	36.53 +-4.66	212 +-44	207	BCYNB	E,I	two-member
HiSeq−2500 1TB	2 × 151	19892392	95.41	36.65 +-4.49	211 +-44	207	BCYHW	D,I	two-member
HiSeq−2500 1TB	2 × 151	17429116	95.15	36.58 +-4.58	219 +-42	233	BCYNY	F	axenic
HiSeq−2500 1TB	2 × 151	19783716	95.36	36.63 +-4.51	211 +-43	216	BCYPS	G	axenic
HiSeq−2500 1TB	2 × 151	22178248	95.18	36.58 +-4.63	215 +-43	207	BCYNZ	B,F,G,I	four-member
HiSeq−2500 1TB	2 × 151	16442692	95.06	36.59 +-4.61	203 +-44	165	BCYGZ	B,D	two-member
HiSeq−2500 1TB	2 × 151	15649768	95.29	36.6 +-4.6	208 +-45	213	BCYHN	F,G,H,I	four-member
HiSeq−2500 1TB	2 × 151	19074446	95.39	36.63 +-4.5	215 +-43	207	BCYPN	D,E,G,I	four-member
HiSeq−2500 1TB	2 × 151	20890644	95.45	36.63 +-4.49	208 +-44	165	BCYOG	A,I	two-member
HiSeq−2500 1TB	2 × 151	17520046	95.08	36.55 +-4.64	215 +-43	243	BCYOW	B,J	two-member
HiSeq−2500 1TB	2 × 151	16315780	95.03	36.54 +-4.69	214 +-43	213	BCYPO	E,H	two-member
HiSeq−2500 1TB	2 × 151	19518268	94.51	36.4 +-4.99	213 +-44	207	BCYHP	E,J	two-member
HiSeq−2500 1TB	2 × 151	18072140	95.12	36.56 +-4.68	211 +-43	207	BCYPP	E,G	two-member
HiSeq−2500 1TB	2 × 151	16611992	95.12	36.56 +-4.69	209 +-44	207	BCYHZ	B,I	artificial_mix
HiSeq−2500 1TB	2 × 151	16913648	95.09	36.56 +-4.68	215 +-43	203	BCYOU	J	axenic
HiSeq−2500 1TB	2 × 151	18880498	95.06	36.54 +-4.72	214 +-43	207	BCYHH	B,C,D,E	four-member
HiSeq−2500 1TB	2 × 151	18313496	95.06	36.55 +-4.67	212 +-44	207	BCYPH	E,D	two-member
HiSeq−2500 1TB	2 × 151	20625332	94.77	36.46 +-4.84	218 +-43	255	BCYOO	F,J	artificial_mix
HiSeq−2500 1TB	2 × 151	18418550	95.16	36.57 +-4.73	205 +-46	243	BCYHC	B	axenic
HiSeq−2500 1TB	2 × 151	13775902	94.74	36.5 +-4.78	204 +-44	165	BCYHS	A,B,C,J	four-member
HiSeq−2500 1TB	2 × 151	16621540	94.92	36.52 +-4.75	214 +-43	223	BCYNG	B,H	two-member
HiSeq−2500 1TB	2 × 151	15693470	94.89	36.52 +-4.74	211 +-44	165	BCYNP	A,E,F,I	four-member
HiSeq−2500 1TB	2 × 151	21237540	95.09	36.52 +-4.76	215 +-43	224	BCYOT	E,A	two-member
HiSeq−2500 1TB	2 × 151	22842010	94.67	36.45 +-4.85	220 +-43	255	BCYNA	I	axenic
HiSeq−2500 1TB	2 × 151	18440260	94.72	36.44 +-4.95	209 +-44	207	BCYNT	E,G,H,J	four-member
HiSeq−2500 1TB	2 × 151	20756406	95.02	36.52 +-4.77	217 +-42	207	BCYPG	E,F	two-member
HiSeq−2500 1TB	2 × 151	17037204	95.09	36.55 +-4.68	211 +-43	207	BCYOY	B,I	two-member
HiSeq−2000 1TB	2 × 151	25531936	92.5	35.91 +-5.48	220 +-42	242	BCYNH	D	axenic
HiSeq−2000 1TB	2 × 151	24365958	92.73	35.96 +-5.48	212 +-43	207	BCYPB	B,E,F,H	four-member
HiSeq−2000 1TB	2 × 151	23375470	92.29	35.9 +-5.56	213 +-43	207	BCYNO	C	axenic
HiSeq−2000 1TB	2 × 151	28523566	92.98	36.02 +-5.35	216 +-42	216	BCYNC	H	axenic
HiSeq−2000 1TB	2 × 151	25304504	92.61	35.94 +-5.46	221 +-42	243	BCYOH	A,B,D,E	four-member
HiSeq−2000 1TB	2 × 151	23843578	92.43	35.92 +-5.52	215 +-42	213	BCYHY	F,H	two-member
HiSeq−2000 1TB	2 × 151	22070200	92.24	35.87 +-5.68	194 +-45	165	BCYHB	B,A	two-member
HiSeq−2000 1TB	2 × 151	23409842	93.13	36.07 +-5.3	212 +-43	207	BCYOB	B,D,E,G	four-member
HiSeq−2000 1TB	2 × 151	20279708	92.3	35.9 +-5.52	211 +-43	207	BCYHX	C,E	artificial_mix
HiSeq−2000 1TB	2 × 151	24847918	92.92	36.02 +-5.29	214 +-43	165	BCYOP	A,B,E,F	four-member
HiSeq−2000 1TB	2 × 151	29463066	92.64	35.95 +-5.46	216 +-42	217	BCYOS	G,H	two-member
HiSeq−2000 1TB	2 × 151	21183650	92.97	36.03 +-5.34	205 +-45	187	BCYPA	D,H	artificial_mix
HiSeq−2000 1TB	2 × 151	26833476	92.03	35.79 +-5.69	221 +-41	243	BCYOX	B,D,G,I	four-member
HiSeq−2000 1TB	2 × 151	19346058	92.81	36.01 +-5.36	208 +-44	207	BCYPC	B,E,G,I	four-member
HiSeq−2000 1TB	2 × 151	22325996	92.45	35.93 +-5.49	212 +-44	207	BCYNX	B,E	two-member
HiSeq−2000 1TB	2 × 151	22851880	92.52	35.92 +-5.49	205 +-44	165	BCYHU	A,B,C,E	four-member
HiSeq−2000 1TB	2 × 151	21910994	92.05	35.83 +-5.65	210 +-43	165	BCYON	A,G	artificial_mix
HiSeq−2000 1TB	2 × 151	22150068	92.63	35.96 +-5.47	206 +-45	207	BCYHA	B,C	two-member

Strain legend for community composition column: 3211.1 (A), 3211.3 (B), 3211.5 (C), 3211.6 (D), 3212.2 (E), 3212.3 (F), 3212.4 (G), 3212.5 (H), 3213.3 (I), 3214.6 (J)

B.18 Lack of correlation between gene expression variance and gene homology

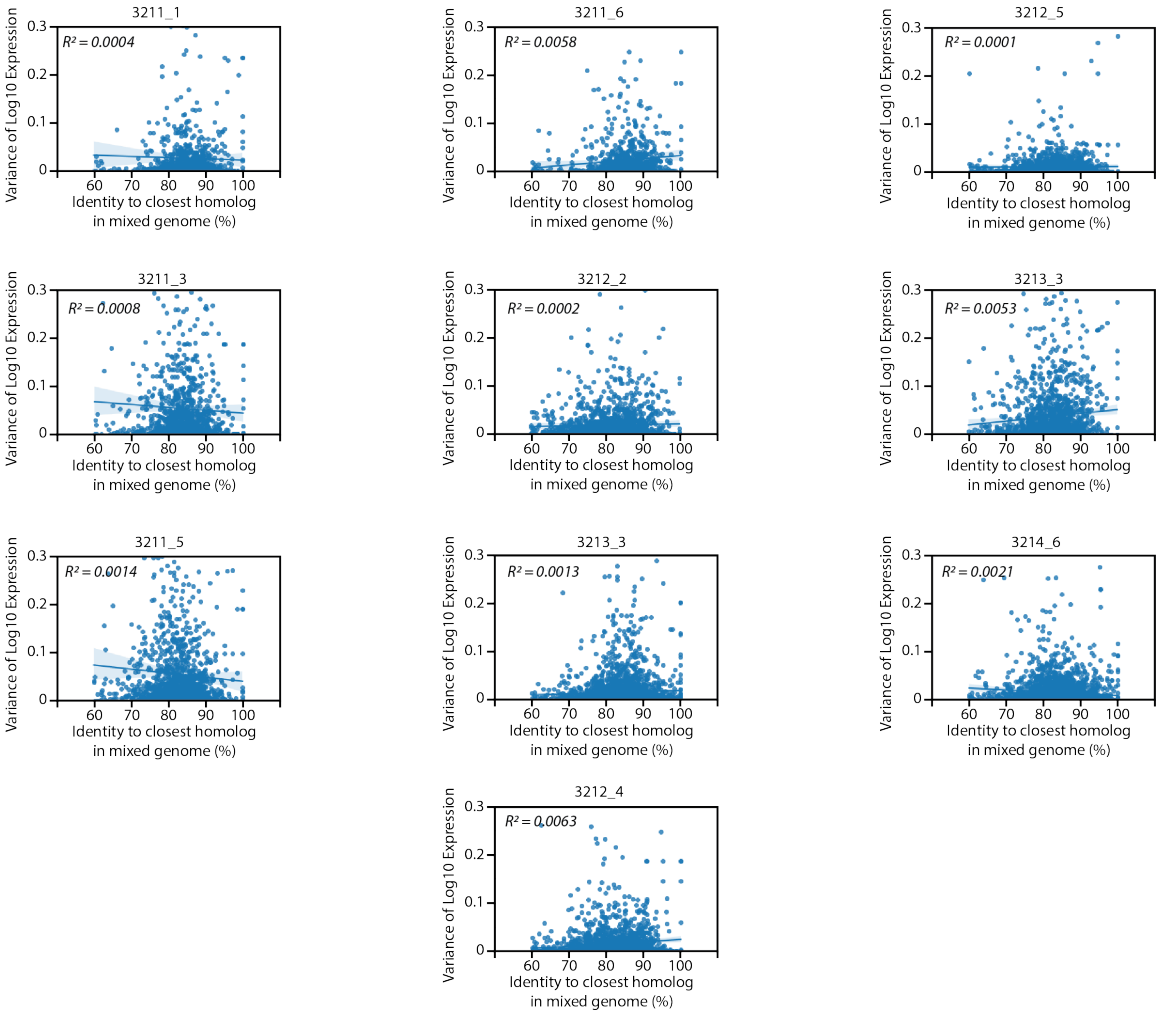


Figure B.33: Measured variance in gene expression is not explained by reads mapping to homologous sequences in other genomes. The per-gene variance is plotted as a function of the percent identity with the closest homolog in the genome to which it was paired/mixed for RNAseq and read mapping. Only genes with a homolog of greater than 60% identity are plotted.

B.19 Differentially expressed primary metabolic genes from 3211.3 and 3212.2 in response to co-culture with 3212.4 and 3212.5

Table B.12: *Streptomyces* sp. 3211.3 primary metabolic DEGs.

Gene ID	Gene Product Name	Direction
Ga0334680_111231	tryptophan halogenase	Dn
Ga0334680_111349	deoxyribodipyrimidine photo-lyase type I	Dn
Ga0334680_111456	urease subunit beta	Dn
Ga0334680_111457	urease subunit gamma	Dn
Ga0334680_111480	alkaline phosphatase D	Dn
Ga0334680_111777	glycerophosphoryl diester phosphodiesterase	Dn
Ga0334680_111778	1-acyl-sn-glycerol-3-phosphate acyltransferase	Dn
Ga0334680_112423	glutamine synthetase	Dn
Ga0334680_112673	nitrite reductase (NADH) large subunit	Dn
Ga0334680_112674	nitrite reductase (NADH) small subunit	Dn
Ga0334680_112683	aryl-alcohol dehydrogenase-like predicted oxidoreductase	Up
Ga0334680_11284	2-dehydropantoate 2-reductase	Up
Ga0334680_113040	[SSU ribosomal protein S5P]-alanine acetyltransferase	Dn
Ga0334680_113041	uroporphyrinogen-III synthase	Dn
Ga0334680_113591	two-component system sensor histidine kinase SenX3	Dn
Ga0334680_113632	5'-nucleotidase	Up
Ga0334680_113634	polyphosphate kinase	Dn
Ga0334680_113639	phosphate ABC transporter ATP-binding protein (PhoT family)	Dn
Ga0334680_113772	gluconolactonase	Up
Ga0334680_11386	cellulose synthase (UDP-forming)	Up
Ga0334680_113985	methionyl-tRNA synthetase	Dn
Ga0334680_114025	ring-1,2-phenylacetyl-CoA epoxidase subunit PaaC	Dn

Table B.12: *Streptomyces* sp. 3213.3 gene clusters.

Gene ID	Gene Product Name	Direction
Ga0334680_114027	ring-1,2-phenylacetyl-CoA epoxidase subunit PaaA	Dn
Ga0334680_11410	p-hydroxybenzoate 3-monooxygenase	Dn
Ga0334680_114312	lysyl-tRNA synthetase class 2	Dn
Ga0334680_11448	methyltransferase family protein	Up
Ga0334680_114752	glycosyltransferase involved in cell wall biosynthesis	Dn
Ga0334680_114754	sialic acid synthase SpsE	Dn
Ga0334680_114807	acetolactate synthase-1/2/3 large subunit	Up
Ga0334680_114812	ribonucleotide reductase beta subunit family protein with ferritin-like domain	Up
Ga0334680_114813	3-oxoacyl-[acyl-carrier-protein] synthase II	Up
Ga0334680_114815	enoyl-[acyl-carrier-protein] reductase [NADH]	Up
Ga0334680_114857	cellulose synthase/poly-beta-1,6-N-acetylglucosamine synthase-like glycosyltransferase	Dn
Ga0334680_115503	UTP–GlnB (protein PII) uridylyltransferase GlnD	Dn
Ga0334680_115836	anthraniloyl-CoA monooxygenase	Up
Ga0334680_115988	ribosome biogenesis GTPase	Up
Ga0334680_116102	two-component system sensor histidine kinase UhpB	Up
Ga0334680_116547	D-inositol-3-phosphate glycosyltransferase	Dn
Ga0334680_116566	hypothetical protein	Dn
Ga0334680_116596	succinate dehydrogenase subunit B	Dn
Ga0334680_11710	D-alanyl-D-alanine carboxypeptidase	Up
Ga0334680_117143	tyrosinase	Dn
Ga0334680_117360	D-alanyl-D-alanine carboxypeptidase	Dn
Ga0334680_117439	3-phytase	Dn
Ga0334680_11771	ferredoxin–NADP+ reductase	Up

Table B.12: *Streptomyces* sp. 3213.3 gene clusters.

Gene ID	Gene Product Name	Direction
Ga0334680_11996	acetyl-CoA synthetase	Dn
Ga0334680_11997	pyruvate dehydrogenase E1 component alpha subunit	Dn

Table B.13: *Streptomyces* sp. 3212.2 primary metabolic DEGs.

Gene ID	Gene Product Name	Direction
Ga0181101_0010	aldehyde dehydrogenase (NAD+)	Up
Ga0181101_0230	glyceraldehyde-3-phosphate dehydrogenase (NAD+)	Up
Ga0181101_0535	glycine cleavage system H protein	Dn
Ga0181101_10000	glyceraldehyde 3-phosphate dehydrogenase	Up
Ga0181101_10209	3-oxoacyl-[acyl-carrier-protein] synthase III	Dn
Ga0181101_10213	long-chain acyl-CoA synthetase	Dn
Ga0181101_1137	alcohol dehydrogenase	Dn
Ga0181101_1146	pyruvate phosphate dikinase	Dn
Ga0181101_1152	menaquinone-dependent protoporphyrinogen oxidase	Dn
Ga0181101_1158	glyceraldehyde 3-phosphate dehydrogenase	Dn
Ga0181101_1526	putative N-acetyltransferase (TIGR04045 family)	Dn
Ga0181101_2997	carbon-monoxide dehydrogenase small subunit	Dn
Ga0181101_3280	NADPH-dependent sulfite reductase flavoprotein alpha-component	Dn
Ga0181101_3340	NitT/TauT family transport system substrate-binding protein	Up
Ga0181101_3793	ABC-2 type transport system permease protein	Up
Ga0181101_3794	ABC-2 type transport system ATP-binding protein	Up
Ga0181101_3837	methylmalonyl-CoA mutase /isobutyryl-CoA mutase large subunit	Up
Ga0181101_4031	catalase	Up
Ga0181101_4416	peptide/nickel transport system substrate-binding protein	Dn
Ga0181101_4625	Zn-dependent protease with chaperone function	Dn
Ga0181101_5036	putative serine esterase DUF676	Dn
Ga0181101_5639	ABC-2 type transport system ATP-binding protein	Up
Ga0181101_5670	GTP cyclohydrolase II	Up
Ga0181101_6441	NNP family nitrate/nitrite transporter-like MFS transporter	Dn
Ga0181101_6442	uroporphyrinogen-III synthase	Dn

Table B.13: *Streptomyces* sp. 3212.2 gene clusters.

Gene ID	Gene Product Name	Direction
Ga0181101_6468	DNA-binding IclR family transcriptional regulator	Up
Ga0181101_6479	4-hydroxyphenylpyruvate dioxygenase	Up
Ga0181101_7002	zinc transport system substrate-binding protein	Up
Ga0181101_7043	assimilatory nitrite reductase (NAD(P)H) large subunit precursor	Dn
Ga0181101_7117	carbohydrate ABC transporter substrate-binding protein (CUT1 family)	Dn
Ga0181101_7146	3-oxoacyl-[acyl-carrier-protein] synthase II	Dn
Ga0181101_7148	3-oxoacyl-[acyl-carrier-protein] synthase III	Dn
Ga0181101_7149	[acyl-carrier-protein] S-malonyltransferase	Dn
Ga0181101_7344	oleandomycin transport system permease protein	Up
Ga0181101_7345	oleandomycin transport system ATP-binding protein	Up
Ga0181101_8382	ABC-2 type transport system permease protein	Up
Ga0181101_8383	ABC-2 type transport system ATP-binding protein	Up
Ga0181101_8586	acetyl-CoA acyltransferase	Up
Ga0181101_9443	urea ABC transporter ATP-binding protein	Dn
Ga0181101_9444	urea ABC transporter ATP-binding protein	Dn
Ga0181101_9445	urea ABC transporter membrane protein	Dn
Ga0181101_9446	urea ABC transporter membrane protein	Dn
Ga0181101_9447	urea-binding protein	Dn
Ga0181101_9508	tyrosinase	Up

B.20 Identification of *pirin*-like homologues and *fur* homologues in 3211.3 and 3212.2

The 'genome blast' tool available on the Joint Genome Institute IMG-ER website was used to find homologues of the pirin-like *pirA* and *fur*. Four query sequences for *pirA* were selected from the *Streptomyces ambofaciens* ATCC 23877 genome assembly (NCBI accession: GCF_001267885.1). One *fur* homologue (*SCLAV_3199* (NCBI accession: EFG08270.1)[132]) was used. BLASTp was used to query against the 3211.3 and 3212.2 genomes.

pirA query sequences:

```
>WP_053127613.1 pirin family protein [Streptomyces ambofaciens]
MSNVETNPVAVRCGPAADAGPPDTAPRVEVLAPRDVPLGGPRAMTVRRTLPQRSRTLIGAWCFADHYGPDDVARTGGMDVAPHP
HTGLQTVSWLFSGEIEHRDSLGSHAHVRPGELNLMTGGHGISHTEVSTPRTTVLHGVQLWVALPGEHRNAPRDFQHHVPEPVSV
DGAEIRVFLGSLAGSTSPVATFSPLLGAEIALAPGATVTLDVPAFEHGLLVDRGEVGMAGTPLRPADLGFLDAGSDTLTLVNAA
DTPARAVLIGGTPFDEEIVMWWNFIGRSHEDIVRARTDWQNASDRFGAVEGYPGDRLPAPALPNGALTPRGNPPRR

>WP_053134172.1 pirin family protein [Streptomyces ambofaciens]
MPAVTVENPLTLPRVSASADAVARPVLTVTTAPSGFEGEGFPVRRAFAGINYRHLDPFIMMDQMGEVEYAPGEPKGTPWHPHRG
FETVTYIIDGIFDHQDSNGGGGTITNGDTQWMTAGSGLLHIEAPPEQLVMSGGLFHGLQLWVNLPAKDKMMAPRYQDIRSGSVQ
LLTSPDGGALLRVIAGELDGHDGPGITHTPITMVHATLAPGAEVTLPWREDFNGLAYVMAGRGSVGAERRPIHLGQTAVFGAGG
SLTVRADDKQDAHTPDLEVVLLGGQPIREPMAHYGPFVMNTKDELMQAFEDFQKGRLGTVPAVHGMSAAGPEA

>WP_053140539.1 pirin family protein [Streptomyces ambofaciens]
MSNLDREAVPSLCGGRGFVVAEPVRELLSPRQVRLGESTEVRRLLPNLGRRMVGAWCFVDHYGPDDIADEPGMQVPPHPHMGLQ
TVSWLHEGEVLHRDSTGSLQTIRPRQLGLMTSGHAISHSEESPRSHARHLHGAQLWVALPDAHRHTDPHFEFHAELPRVTAPGL
TATVLLGSLDGTTSPGTTYTPLVGADLSLTAGTDVRLPLERDFEYAVLSMSGEAHVDGVPLVPGSMLYLGCGRGELPLRADSDA
DLMLLGGEPFEEELIMFWNWIGRSQEEIVQARRDWTEGTRFGEVKGYDGAPLAAPELPAVPLKPRGRAR
```

>WP_053142372.1 pirin family protein [Streptomyces ambofaciens]

MDVRRADERFPGGDPAAGIVSRHAFSFGPHYDPDNLRFGALLACNEERLAPGAGFDEHPHSHTEIVTWVVEGELTHRDTAGHET

VVRAGDVQRLSSAAGVRHVERNDGPAPLTFVQMWLAPLTPGGDPAYEIVHGIADSTPYAVPEAGAMFHVRRLAAGERTAVPDGA

YVYVHVVRGEVTLGGETLGAGDAARVTDADALDAVAVTRAEVLLWEMGGQ

fur query sequence:

>EFG08270.1 Putative ferric uptake regulatory protein [Streptomyces clavuligerus]

MPPAPCGPGARLRRLPTLGYVVSTDWKSDLRQRGYRLTPQRQLVLEAVDALDHATPDDILCEVRRTASGVNISTVYRTLELLEEL

GLVSHTHLGHGAPTYHLADRHHHIHLVCRDCTDVIETDVDFAADFTAKLRETFGFDTDLKHFAIFGRCQGCTSSRTDRADRTDQA

DRAERRDGPDGTDRVARPAAPSASGASRAPGASTSGS

Table B.14: Pirin homologues in *Streptomyces sp.* 3211.3 and *Streptomyces sp.* 3212.2

Gene ID	Gene Product Name	Genome Name	COG	COG Description
Ga0334680_112617	hypothetical protein	Streptomyces sp. 3211.3	COG1741	Redox-sensitive bicupin YhaK, pirin superfamily
Ga0334680_114081	hypothetical protein	Streptomyces sp. 3211.3	COG1741	Redox-sensitive bicupin YhaK, pirin superfamily
Ga0334680_116632	hypothetical protein	Streptomyces sp. 3211.3	COG1741	Redox-sensitive bicupin YhaK, pirin superfamily
Ga0181101_0882	hypothetical protein	Streptomyces sp. 3212.2	COG1741	Redox-sensitive bicupin YhaK, pirin superfamily
Ga0181101_1828	hypothetical protein	Streptomyces sp. 3212.2	COG1741	Redox-sensitive bicupin YhaK, pirin superfamily
Ga0181101_5334	hypothetical protein	Streptomyces sp. 3212.2	COG1741	Redox-sensitive bicupin YhaK, pirin superfamily
Ga0181101_7151	hypothetical protein	Streptomyces sp. 3212.2	COG1741	Redox-sensitive bicupin YhaK, pirin superfamily

Table B.15: Fur homologues in *Streptomyces sp.* 3211.3 and *Streptomyces sp.* 3212.2

Gene ID	Gene Product Name	Genome Name	COG	COG Description
Ga0334680_112690	Fur family zinc uptake regulator	Streptomyces sp. 3211.3	COG0735	Fe2+ or Zn2+ uptake regulation protein
Ga0334680_113611	Fur family nickel uptake regulator	Streptomyces sp. 3211.3	COG0735	Fe2+ or Zn2+ uptake regulation protein
Ga0334680_115082	Fur family ferric uptake transcriptional regulator	Streptomyces sp. 3211.3	COG0735	Fe2+ or Zn2+ uptake regulation protein
Ga0334680_11643	Fur family ferric uptake transcriptional regulator	Streptomyces sp. 3211.3	COG0735	Fe2+ or Zn2+ uptake regulation protein
Ga0334680_117451	Fur family ferric uptake transcriptional regulator	Streptomyces sp. 3211.3	COG0735	Fe2+ or Zn2+ uptake regulation protein
Ga0181101_5757	Fur family nickel uptake regulator	Streptomyces sp. 3212.2	COG0735	Fe2+ or Zn2+ uptake regulation protein
Ga0181101_6999	Fur family zinc uptake regulator	Streptomyces sp. 3212.2	COG0735	Fe2+ or Zn2+ uptake regulation protein
Ga0181101_8610	Fur family ferric uptake transcriptional regulator	Streptomyces sp. 3212.2	COG0735	Fe2+ or Zn2+ uptake regulation protein

B.21 Hierarchical clustering of primary metabolic genes with *pirin*-like homologues or *fur* homologues in 3211.3 and 3212.2

Figure B.34 ***(following page)*****:** Hierarchical clustering of primary metabolic genes and pirin-like homologues in 3212.2. Hierarchical clustering of primary metabolic genes and pirin-like homologues in 3212.2. Heatmap shows log2 fold change in gene expression of primary metabolic genes and putative pirin-like homologues from 3212.2 (filled black circles). A single pirin-like gene (open black circle) clusters with a subset of the primary metabolic genes.

Figure B.34: (continued)

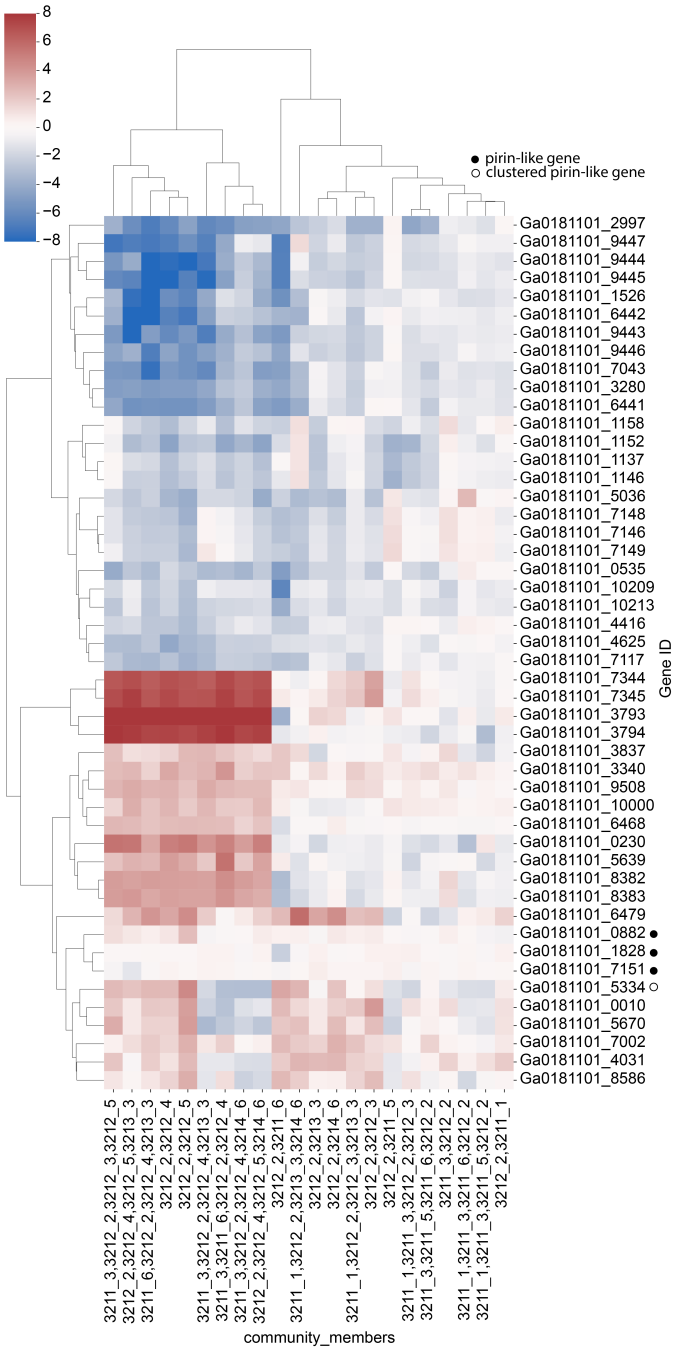


Figure B.35 ***(following page)*****:** Hierarchical clustering of primary metabolic genes and pirin-like homologues in 3211.3. Heatmap shows log2 fold change in gene expression of primary metabolic genes and putative pirin-like homologues from 3211.3 (filled black circles).

Figure B.35: (continued)

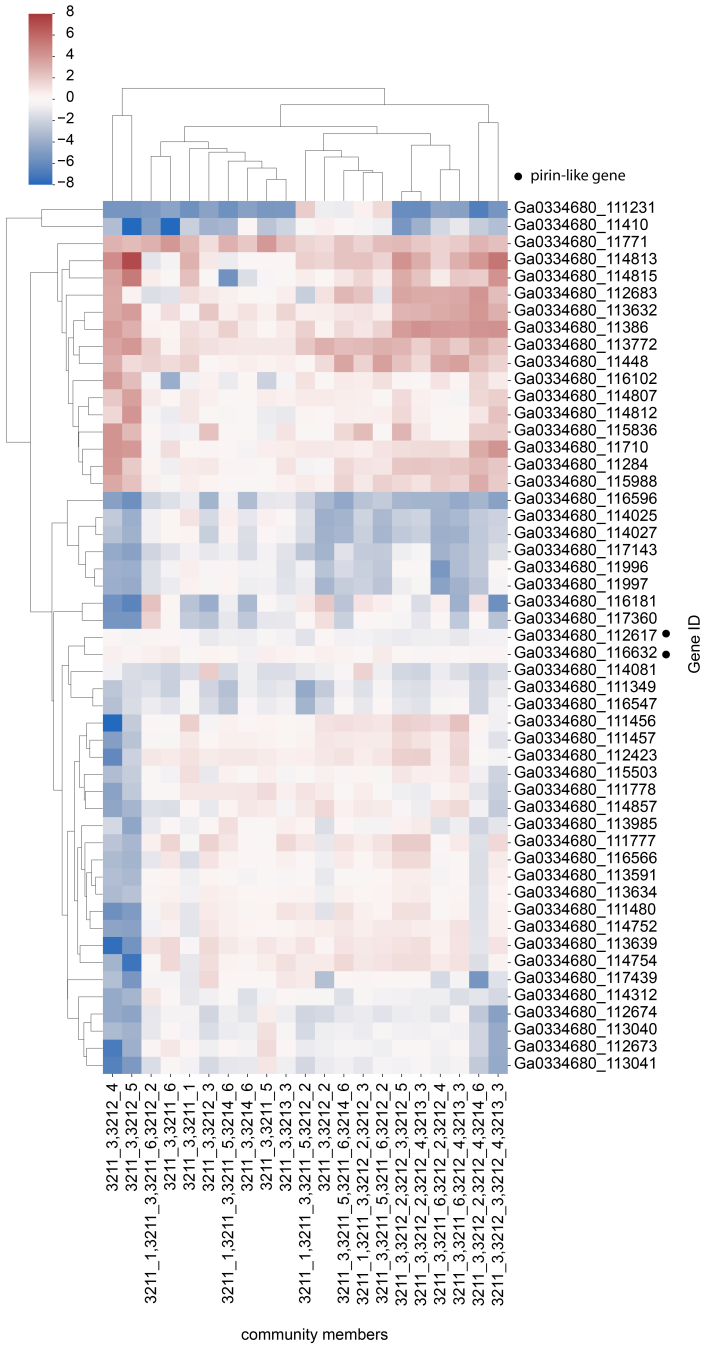


Figure B.36 *(following page)*: Hierarchical clustering of primary metabolic genes and fur homologues in 3211.3. Hierarchical clustering of primary metabolic genes and *fur* homologues in 3211.3. Heatmap shows log2 fold change in gene expression of primary metabolic genes and putative *fur* homologues from 3211.3 (filled black circles). A single *fur* gene (open black circle) clusters with a subset of the primary metabolic genes.

Figure B.36: (continued)

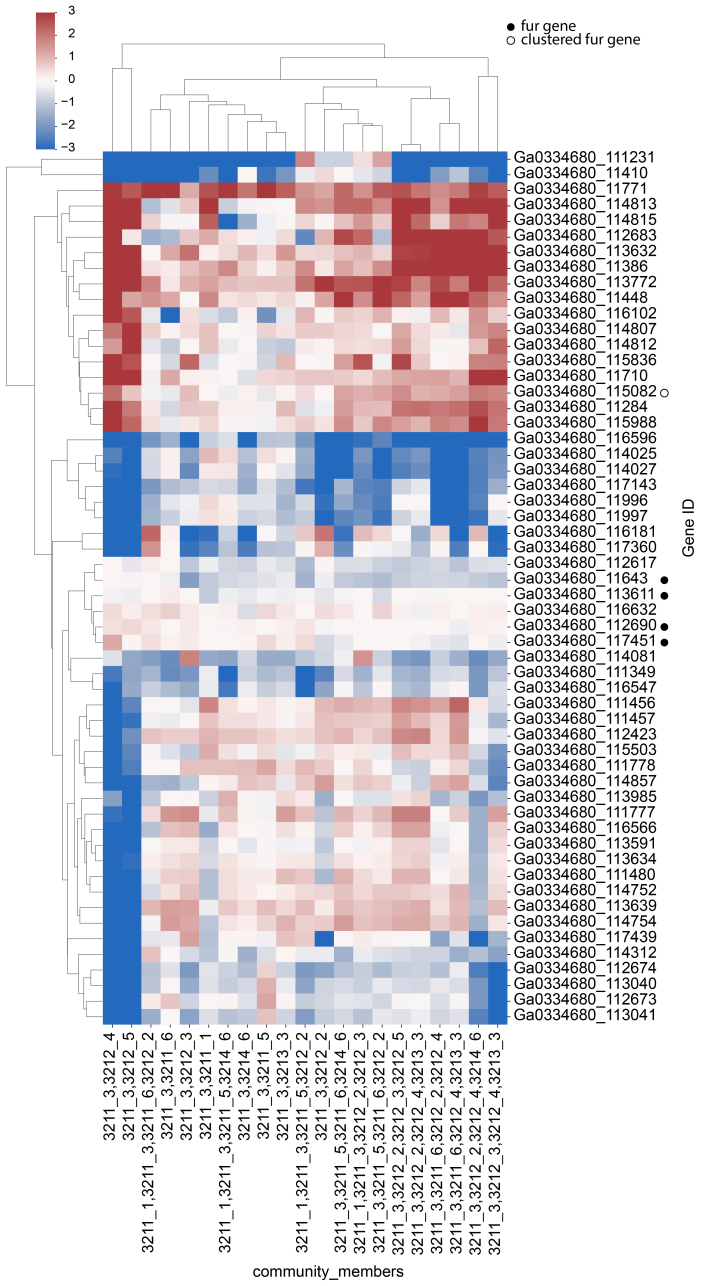
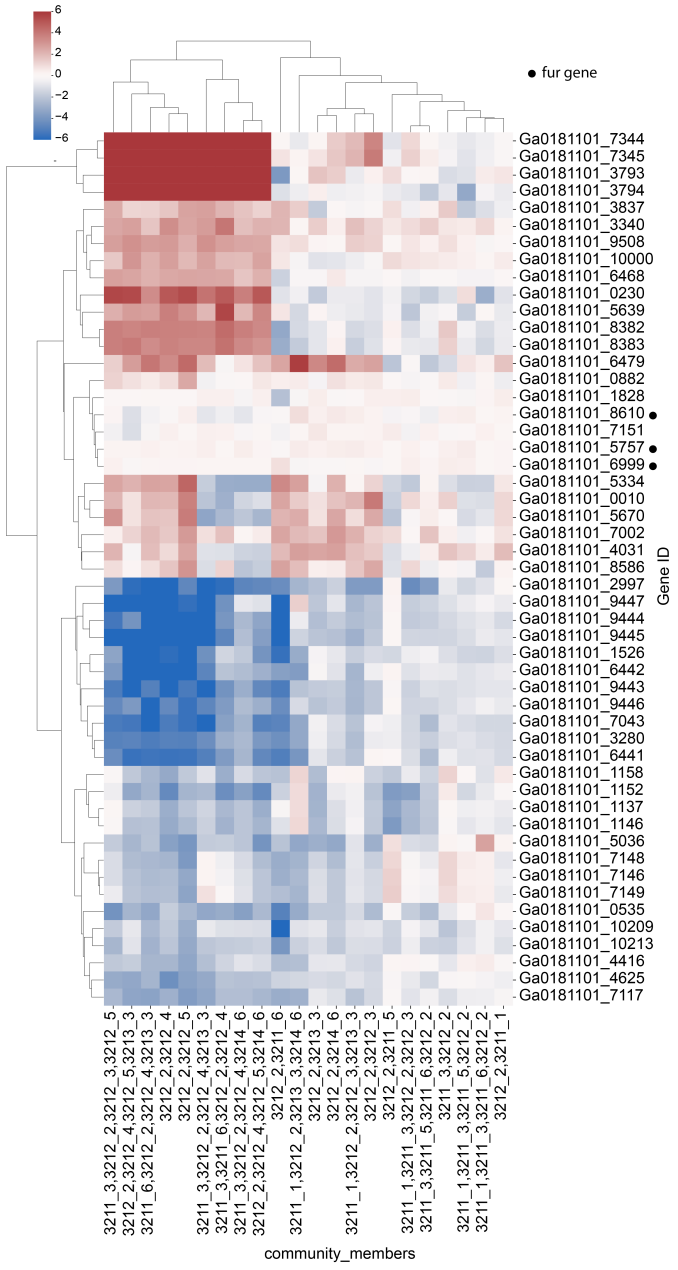


Figure B.37 *(following page)*: Hierarchical clustering of primary metabolic genes and fur homologues in 3212.2. Hierarchical clustering of primary metabolic genes and *fur* homologues in 3212.2. Heatmap shows log2 fold change in gene expression of primary metabolic genes and putative *fur* homologues from 3212.2 (filled black circles).

Figure B.37: (continued)



Appendix C

Supplemental materials for Simulation Modeling to Compare High-Throughput, Low-Iteration Optimization Strategies for Metabolic Engineering

C.1 Code availability

All code described in the text is available at https://github.com/smanskiLab/pathway_optimization