NON-CODING RNAS AND MRNA SECONDARY STRUCTURE IN *STREPTOMYCES*

NON-CODING RNAS AND THE ROLES OF MRNA SECONDARY STRUCTURES IN *STREPTOMYCES* BACTERIA

By MATTHEW JOHN MOODY, B.SC.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Doctor of Philosophy in Biology Degree

McMaster University DOCTOR OF PHILOSOPHY (2017) Hamilton, Ontario (Biology)

TITLE: Non-coding RNAs and the roles of mRNA secondary structures in *Streptomyces* bacteria

AUTHOR: Matthew John Moody (B.Sc)
SUPERVISOR: Dr. Marie A. Elliot

NUMBER OF PAGES: xvii, 232

Lay abstract

The flow of genetic information, from DNA to RNA to proteins, often portrays RNA as a mere intermediary molecule. An alternative, and perhaps more accurate, way to view RNA is that it is central to all cellular processes. Many RNAs are not translated into proteins and instead act as regulatory molecules, impacting the expression of other genes. In this work we found many examples of these regulatory RNAs in a group of bacteria known to produce many of the world's antibiotics. Understanding the roles these regulatory RNAs play in impacting gene expression will be important for the discovery of new molecules, such as antibiotics. In addition to distinct regulatory RNAs mentioned above, we found that RNA structures within the coding sequences of mRNAs that are translated into proteins have dramatic regulatory consequences. We describe the characterization of one such RNA structure in a gene involved in bacterial communication, and develop a bioinformatic tool to hunt for other such structures conserved throughout bacteria.

Abstract

Work over the past two decades has revealed that non-coding RNAs (ncRNAs) are prevalent in all kingdoms of life. Using RNA-seq we discovered hundreds of ncRNAs in the antibiotic-producing genus of bacteria, *Streptomyces*. These included *trans*-encoded small RNAS (sRNAs), *cis*-antisense RNAs, and a new type of antisense RNA we termed cutoRNAs (convergent untranslated overlapping RNAs) that arise when transcription termination does not occur in the intergenic region between two convergently arranged genes. Many of these ncRNAs feature prominently in the specialized metabolite biosynthetic clusters (*e.g.* antibiotics, anticancer agents, immunosuppressants). Hence, it is likely that understanding the functions of these RNAs will be important for new molecule discovery. We found that one highly expressed antisense RNA (ScbN) was expressed opposite the γ-butyrolactone synthase *scbA* in the model streptomycete *Streptomyces coelicolor*. However, ScbN had no detectible impact on the expression of *scbA*. Instead, the transcription terminator of *scbN,* which also forms a hairpin within the coding sequence of *scbA*, was found to reduce expression of *scbA* more than 10-fold. This led us to bioinformatically search for similar coding-sequence hairpins throughout all bacteria, leading to the discovery of many stable RNA structures with conserved locations throughout very divergent bacteria (*e.g. Streptomyces, Escherichia coli, Bacillus subtilis*).

Acknowledgments

Completing a PhD is truly a team effort, and I have been incredibly lucky to have been supported by many great people in both my academic and personal life.

First, I want to thank Dr. Marie Elliot for giving me the opportunity to work in her lab. Over the years she has been a fantastic mentor. Her knowledge, her attention to detail, the speed at which she responds to my sometimes rambling emails, the (literal) openness of her office door, her efficiency, her birthday cakes, her generosity, and her kindness are unrivaled. I truly consider myself lucky to have spent seven years under her supervision.

When James was born last April, I was less than an ideal grad student: I was exhausted and forgetful, making many experimental mistakes. Marie was so understanding during this exciting, but stressfull, time. Thank you so much for that.

As a scientific mentor, Marie is amazing. I would work on ideas for hours only to have her poke a giant hole in the middle of it in a matter of seconds! How she was able to do this (and do this simultaneously for the ~dozen projects going on in her lab at any given time) is beyond my comprehension. She let me spend months on absolutely crazy ideas, and even in the numerous cases where these ideas led to dead-ends, I learned so much from the process. In fact, I think I learned more about how bacteria work in these failed endeavors than I would have if all of my experiments worked.

My project was a frustrating one to work on, but it would have been unbearably frustrating without Marie's encouragement and guidance. And when it finally became time to wrap things up (see attached thesis) she was extremely supportive and encouraging. Marie, thank you so much for an absolutely amazing PhD experience.

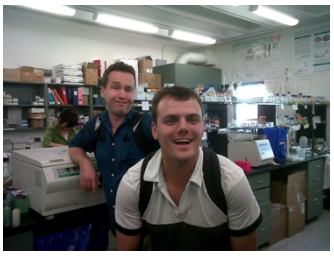
I would also like to thank the members of my supervisory committee, Dr. Brian Golding and Dr. Michael Surette, for providing critical feedback on my research. I remember being nervous prior to my first committee meeting after hearing horror stories from other graduate students of their meetings turning into intellectual attacks. I never experienced this. In fact, I always left those meetings excited about the new directions in which you both pointed me. I looked forward to subsequent meetings with the knowledge that both of you were there to help and advise. I'd further like to thank Dr. Golding for allowing me to access to his

computers, which I'm sure I slowed to a halt at times with my extraordinarily inefficient algorithms.

To my external examiner, Dr. Joe McCormick, thank you for providing valuable feedback on my thesis. And to Dr. Gerry Wright – thanks for being my internal examiner. And also thanks for letting me pretend to be a rock star each December! Dr. Jonathan Dushoff, thank you for chairing my defence.

Marie has assembled a great group of people in her lab who have made my time here amazing. Even during times when my experiments were not going as planned, I looked forward to getting into the lab. You were an undeniably remarkable group of people to work with, and you will all be missed. In particular, I would like to thank:

Dr. Henry Haiser – I remember starting in the lab and being absolutely amazed at your incredible focus and determination: you were *the* model grad-student when I started. And the second the lab door closed behind you at the end of the day, you were so much fun to chat with. You have an amazing combination of a dry, sarcastic, and cutting sense of humour, that I utterly love. While we only overlapped in the lab for about eight months, we've remained great friends. All the times we have hung out since together, in Boston or Toronto, we've picked up that friendship immediately without missing a beat.

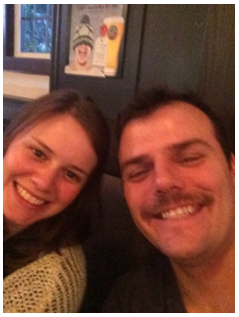


Dr. David Capstick, for his fantastic wit and for editing everything I wrote pre 2013 (you are such a talented writer). I think we've had entire 'group' conversations that consisted of just the two of us lobbing jokes back and forth to each other, jokes that went way over the rest of the group's heads. Thank you for "getting it".

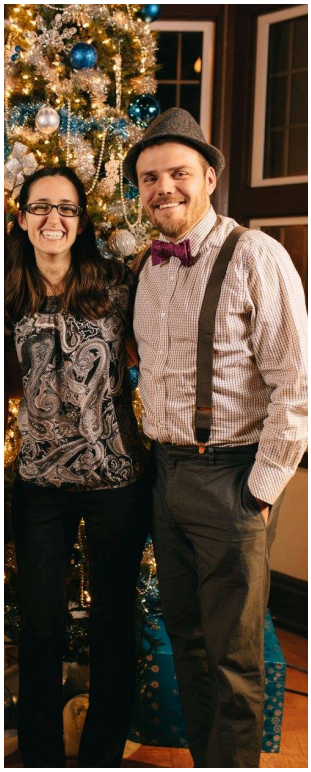
Andrew Duong, for being his affable self (and for letting me crash on his couch when isolating RNA had to occur at midnight).

Dr. Julia Swiercz, the only time I cried about a lab related event (edit: this isn't exactly true now post-defence) was on the train ride home after your last day – and that likely means something. There are a ton of things for which I owe you thanks: walking me through every protocol (literally holding my hand), listening to my stupid stories, ensuring I landed in soft soil instead of a hard rock when you flipped me into the

garden, making me laugh daily . . . but most importantly you genueinly are a wonderful person to spend time with. For about two years we spent every morning having coffee together, and because of this I always started my day off with a smile on my face.



Rachel Young, for that special slow dance we shared. I think I felt the earth stop spinning for a few minutes. Also, thank you for your help with the work in BMC Genomics paper!



Dr. Renee St. Onge, thank you for being the kind, but messy (*sarcasm*), lab bench-neighbour with a pirate-like mouth. Pardon *your* French! I always appreciated how you'd call me *Moo moo*, even at conferences when I'm talking to somebody important.

Danielle Sexton- You left. You left for the UK without saying bye while I was writing my thesis. And for this I will never forgive you. Edit: You came back! You came back and we had a great time the night of my defence, you are so great!

David Crisante, thank you for (unintentionaly?) making me laugh. For example, by saying things like "vanity? I'm not van" and for "inventing" phrases for the world to use. Phrases such as: "I feel like".

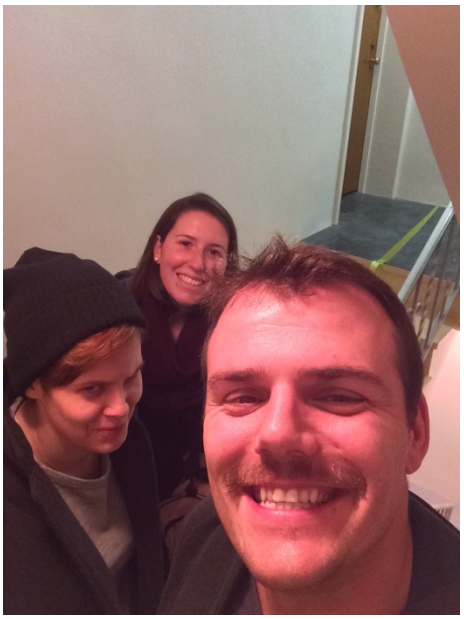
Dr. Emma Sherwood – I was very saddened when the Elliot Lab suffered its own Brexit, when our only British member exited in 2016. You've been missed. My favourite memories of you are all from Steph's cottage. The watermelon. You talking to the rope: *Not a knotty knotty?*

Chris Firby, I think I owe you some money because you predicted a while ago that Trump would be President? Nostradamous? Or fantastic bet-hedging strategy? I always enjoyed chatting with you when we shared the bus to Aldershot, and your unfailing optimism about the Blue Jays.

Fei Zhang, we didn't overlap for long, but I think I'll tell the story of you forgetting your ID for the bar and ordering a 50/50 apple/orange juice mix for a while.

Dr. Hindra – when you told me you were moving to Jupiter I thought you had lost your mind. Glad to hear it's just a city on Florida. Thanks for teaching me how to not contaminate everything when doing radioactive work (a skill I apparently did not pass on to the next generation).

Dr. Daphne Gao - I'll never forget your laugh, or your cough … in Andrew's face.

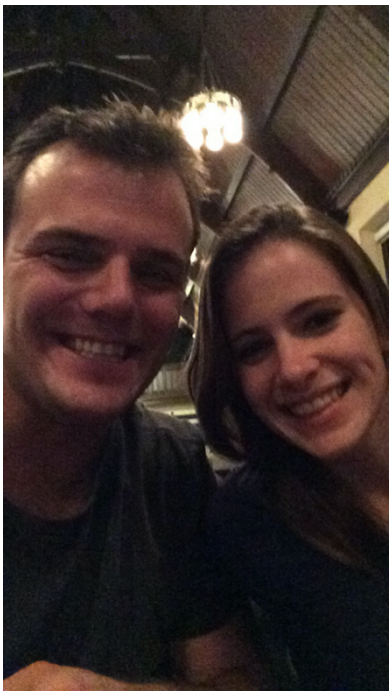


Savannah Colameco, you are best described as genuinely nice. So nice, in fact, that I cannot imagine you being mean. I love how even when you are trying to seem mad you cannot help but have a slight smile on your face. You are always a true pleasure to be around.

Suzanne Boursalie – My lowest moment in grad school was losing to you in Monoply. We need to have a rematch soon. One on one. No interference from lesser combatents… inferiors, underlings, clowns, fools! Just us.

Loryn, Christine, *new* Emma, Nicole, Dan, Alison, Dan, Jackie, Talha, Ahmed, Michael, Coral and Kristy – sorry. I ran out of time writing this and so I lumped most of the undergrads into one group. Sorry, I'm being such a typical grad student. I have memories of each of you that made my time in the lab memorable: from setting off the panic alarm with your hips, to (while being an undergrad in the lab) telling the post doc in the lab to finish your experiments because you want to go to the cottage, to not getting the *hot-tea* joke even after 10 minutes of explanations, to asking for help closing the freezer because "something was wrong with the door" (there was a giant autoclave glove covering the opening), to having a very precise solid-liquid ratio in your diet. A very special thanks to Nicole and Christine who worked on projects closely related to the work that follows.

Stephanie Jones – thank you for being your amazing self. I remember when you interviewed with the lab at the *old* Phoenix I arrived about 30 minutes late. I mainly came for the chicken fingers, and had no clue then that I was interviewing someonew who would end up being a truly great friend, my best friend, for the next few years - thank you so much. Some of my favourite moments were the dozens of times at PH1, PH2 and PH3, or us quite literally being the only two people in the entire city of Niagara Falls, and then you ordering the words most giant stack of pancakes the next morning and only eating the fruit off the top, playing squash "*boat*?", or you watching me sing karaoke all night and not complaining, or dancing like maniacs to the Null Hypothesis, or me destroying you at Wave Race, or you destroying me at Diddy Kong Racing, drawing things in the snow, everyone calling me "Mark" at Frisbee, or battling you in the hockey pool, hanging out at your cottage, hanging out at my cottage (and Rachèle's death grib while on the seado), you not getting too mad when I spilled a scoop of chocolate ice cream in your cutlery drawer didn't clean it up, not getting mad when I ate all of your pickles, our 200 meter Uber ride to campus from your old apartment, *Avatar*, and finally sorry for that time I almost let Goku out of your back porch and you almost . . . I'll stop there. I'll miss you so much, please let's remain friends. You are a brilliant scientist - keep up the great work, and I look forward to reading your subsequent *Nature* and *Science* and *Elife* (but not stupid *Cell*) papers for years to come. Just promise to never treat me as a lay-person!!



To the love of my life, Alison, I want to both apologize to and thank you.

First, sorry for the incredible 'bait and switch' I pulled on you: "Will you marry me? Yes? Ok, I'm quitting my job and going back to school for a significant portion of a decade! Muhahahah".

And of course thank you for being so supportive during these crazy few years. Getting a PhD is a team effort, and I think that makes you at least a half-doctor! A lot of people have a hard time emotionally during grad school, and I think my time would have been so much harder if I didn't have such an amazing wife to come home to at the end of every day.

About a year and a half ago we welcomed James into the world. James, I'm so excited to watch you grow and discover all there is to learn about the world. I think it's true what they say, that we are all born scientists. Last night I was giving you a bath and you were ever so slightly adjusting the angle of your spinning bath toy in the flow of the water to find the perect angel for maximum spin velocity. I might be reading too much into this as a highly biased parent, but I think you'll be able to read this thesis before you start school ;)

Thanks to Alison's parents, Ron and Mary E. I think I've spent at least a couple hours talking about science at Thanksgiving / Christmas dinners and you've always both shown so much interest and enthusiasm in my work. I've been very fortunate in marrying into such a loving and caring family.

I would also like to thank my family, for all their love and support. All four of my grandparents played enormous roles in my life. They taught me how to be kind, how to be funny, and how to work hard. I can attribute much of my success in life to having such an amazing upbringing. Both of my parents always encouraged me to follow my passions, was gave me every possible opportunity. My dad passed away exactly twelve years ago (September 9th, 2005). I really wish he was around to witness me getting my PhD, I like to think he would be proud (but very confused while listening to the defence). And thanks mom for being the best mom, and the glue that holds our family together. Thank you Allan for all your love and support over the years. And Taylor, oh Taylor, thanks for putting up with a goofy nerdy older brother! You've made me laugh throughout the years more than anyone above combined (although half of these times were due to your orthodontic headgear).

Finally, a prospective thank you to any researcher, in Marie's lab or elsewhere, who discovers what ScbN, my nemesis, actually does (if it does anything at all).

Dedication

This thesis is dedicated to two of the most important people in my life, my parents.

First, to my dad, Michael John Moody (1955 – 2005). You always were the largest personality in the room, a gentle giant with a booming laugh and the biggest smile. I've missed you every day.

And to my mom, Sharon Moody. You have supported and encouraged me for longer than I can remember. And since dad passed away, you have been an absolute rock that has kept our family so close.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

asRNA	*cis*-encoded antisense ribonucleic acid
DNA	deoxyribonucleic acid
DNA medium)	Difco nutrient agar (when in the context of growth
DNase	deoxyribonuclease
g	gram
GBL	gamma-butyrolactone or γ-butyrolactone
GC content	guanine / cytosine content
kb	kilobases
kV	kilovolts
Mb	megabases
mg	milligram
mL	milliliter
min	minutes or minimum (depending on context)
ncRNA	non-coding ribonucleic acid
nt	nucleotide
PCR	polymerase chain reaction
pH	potential of hydrogen
RBS	ribosome binding site
RNA	ribonucleic acid
RNase	ribonuclease
RNA-seq	ribonucleic acid sequencing
rRNA	ribosomal ribonucleic acid
rtPCR	reverse transcription polymerase chain reaction
s	seconds
sRNA	*trans*-encoded intergenic small ribonucleic acid
tRNA	transfer ribonucleic acid
UTR	untranslated region
v/v	volume per volume
w/v	weight per volume
XGluc	5-bromo-4-chloro-3-indolyl-β-D-glucoronide
µg	microgram
µL	microliter
~	approximately
Δ	change in
Δ	deletion (when in context of a gene)
ΔG	free energy of an RNA structure
2˚	secondary

Chapter 1: General Introduction

Streptomyces bacteria

Streptomyces is a genus of Gram-positive bacteria that primarily live in the soil and are well known for both their complex life-cycle and specialized metabolism. The *Streptomyces* life-cycle begins when a spore germinates upon encountering favourable growth conditions. The resulting germ tubes grow by hyphal tip extension and periodic branching to form a complex network of cells termed the vegetative mycelium. When nutrients become limiting, or in response to an as yet unknown signal, vegetative growth slows and the colony raises non-branching aerial hyphae up into the air. Chromosomes are segregated throughout individual aerial hyphae, and a rapid phase of synchronous cell division generates chains of unigenomic exospores (Flärdh and Buttner, 2009) (**Figure 1.1**). In addition to this classical model of *Streptomyces* development, recently a new mode of development was discovered. When interacting with fungi or when sensing a volatile signal, some *Streptomyces* can spread incredibly rapidly across surfaces using non-branching vegetative hyphae (Jones et al., 2017).

Streptomyces make many of the world's antibiotics and a number of other important chemicals such as anti-cancer and anti-parasitic agents, and therefore have a reputation for being one of nature's top chemists (Bérdy, 2012). These compounds are collectively referred to as 'specialized metabolites'. Specialized metabolites are made by clusters of contiguous genes. These specialized metabolite gene clusters encode most of the proteins / enzymes needed for the regulation, synthesis, and modification of these products. It is relatively easy to predict subsets of genes (*e.g.* polyketide and non-ribosomal peptide synthases) involved in specialized metabolism, and all *Streptomyces* species whose genomes have been sequenced to date encode far greater numbers of biosynthetic genes than would have been expected based on the metabolites they were known to produce. This has led to the suggestion that most specialized metabolites are 'cryptic', and are not detected under typical laboratory conditions (Nett et al., 2009). Hence the known metabolites produced by these species appear to represent a small subset of their biosynthetic potential.

It is currently not known why *Streptomyces* have such diverse chemical repertoires. It is possible that some metabolites are chemical weapons used to inhibit the growth of other microorganisms, while others may serve as signaling molecules (Bernier and Surette, 2013; Fajardo and Martínez, 2008; Goh et al., 2002; Keller and Surette, 2006; Yim et al., 2006, 2007). Regardless of their purpose in nature, fully understanding the regulation of specialized metabolism in *Streptomyces* bacteria will allow us to more effectively exploit their reservoir of new natural products.

Unlike most bacteria, *Streptomyces* is unusual in that its genome is contained on a very large (usually >8 Mb) linear chromosome. The large chromosome is subdivided into a central 'core' of ~4 Mb, with two ~2 Mb 'arms' on each side. Within the core region are housed highly conserved and housekeeping genes, while the arms contain *Streptomyces*- and species-specific genes. As such, the specialized metabolite clusters unique to individual *Streptomyces* species are often found on the arms of these linear chromosomes

(Kirby, 2011). The *Streptomyces* chromosome also has a very high GC-content (>70%). Since GC base-pairs are stronger than AU base-pairs in RNA, the high GC nature of genes in these species likely results in very strong RNA secondary structures which may have important regulatory consequences.

Regulation of specialized metabolite production in *Streptomyces*

Streptomyces produce about two thirds of the world's clinically-used antibiotics. As a result, there have been considerable efforts dedicated to studying the regulation of antibiotics and other specialized metabolites in various *Streptomyces* species. *Streptomyces coelicolor* has long served as the model system for understanding the regulation of specialized metabolism. These investigations have been greatly facilitated by the fact that *S. coelicolor* produces two pigmented antibiotics – a blue polyketide antibiotic called actinorhodin (Wright and Hopwood, 1976), and a red, cell-associated compound called undecylprodigiosin (Feitelson et al., 1985; Williamson et al., 2006). More recently, *S. coelicolor* has been shown to produce a third pigmented antibiotic, the yellow-pigmented coelimycin polyketide. Due to the ease in tracking these pigments visually, *S. coelicolor* has become the best-studied member of the streptomycetes in terms of the regulation of specialized metabolism.

As mentioned above, the synthesis of specialized metabolites is directed by clusters of genes that span contiguous tens to hundreds of thousands of base pairs of the chromosome (**Figure 1.2**). These gene clusters include those encoding pathway specific regulators, enzymes responsible for the synthesis of the core metabolite (*e.g.* polyketide or non-ribosomal peptide synthases), modifying and tailoring enzymes, and resistance determinants. In addition to pathway-specific regulators, global pleiotropic regulators located outside these clusters affect the expression of multiple biosynthetic clusters and often act by controlling the expression of pathway-specific regulators (Liu et al., 2013), as described below.

Regulation of model specialized biosynthetic clusters

Actinorhodin is a blue-pigmented molecule with antibiotic properties. The actinorhodin biosynthetic gene cluster spans ~21 kb and is made up of five operons. The associated polyketide synthases and tailoring enzymes are activated by the pathway specific regulator ActII-ORF4, itself encoded within the second operon (Malpartida and Hopwood, 1986). ActII-ORF4 has a DNA binding helix-turn-helix domain at its N-terminus, and a transcriptional-activation domain at its C-terminus. ActII-ORF4 is a founding member of the SARP (*Streptomyces* antibiotic regulatory protein) family of antibiotic regulators. SARPs are found in approximately one quarter of all known specialized metabolite clusters in *Streptomyces* (Liu et al., 2013), activating the expression of their associated clusters.

Actinorhodin production is highly regulated, but interestingly, this regulation is exerted primarily through the control of *actII-orf4* expression. In contrast, the undecylprodigiosin cluster (~32 kb, and 23 genes) is positively regulated by a cascade of pathway-specific regulators. The 'orphaned' two component response regulator RedZ

activates the expression of *redD,* which like *actII-ORF4* encodes a SARP-family regulator. RedD subsequently activates the expression of the rest of the *red* cluster.

actII-ORF4 and *redZ* are under the control of a number of different proteins which modulate their expression under various conditions. For example, at least eight transcription factors bind to the *actII-orf4* promoter (Davies, 2011; Gao et al., 2012; Park et al., 2009; Sheeler et al., 2005; Shu et al., 2009; Uguru et al., 2005; Wang et al., 2013). Many of these regulators (*e.g.* Crp, AfsQ1/Q2, AbsA2 AdpA) act on the *redZ* promoter as well (Gao et al., 2012; Sheeler et al., 2005; Shu et al., 2009; Wang et al., 2013; Williamson et al., 2006). Therefore these proteins are termed global antibiotic regulators, as they impact the expression of a number of different gene clusters.

Some of these global antibiotic regulators act in response to changes in nutrient levels. For example, the proteins DraR/K and AfsQ1/Q2 are important for responding to changes in nitrogen levels. Under conditions of excess nitrogen AfsQ1/Q2 is activated. AfsQ1/Q2 is a two-component system and is responsible for repressing genes responsible for nitrogen assimilation as well as activating multiple specialized metabolic clusters. AfsQ1 binds directly to both the pathways' specific regulator genes for actinorhodin (*actII-orf4*) and undecylprodigiosin (*redZ*), activating their expression (Wang et al., 2013). Similarly, the two component system DraR/K is activated under conditions of nitrogen excess. DraR subsequently activates *actII-orf4,* and directly negatively regulates the coelimycin cluster through its pathway specific regulator *cpkO*, while the negative effects on the undecylprodigiosin cluster by this regulator do not appear to be mediated directly through *redZ* (Yu et al., 2012).

N-acetylglucosamine is major component of bacterial cell walls. *Streptomyces* imports and phosphorylates N-acetylglucosamine, where it binds to the DNA-binding protein DasR. It is thought that phosphorylated N-acetylglucosamine acts as a signal that nutrients are being released by underlying vegetative hyphae during autolysis (Chater et al., 2010). N-acetylglucosamine binding to DasR relieves the repression of DarR target genes, which include N-acetylglucosamine uptake systems, creating a feedforward loop. In addition, this results in de-repression of the pathway specific regulators *actII-orf4* and *redZ* (Rigali et al., 2008).

In addition to responding to N-acetylglucosamine, specialized metabolism is also controlled indirectly by the phosphate starvation response system in *Streptomyces.* PhoR is a membrane bound sensor kinase which phosphorylates cytoplasmic PhoP under phosphate starvation conditions (Allenby et al., 2012). Phosphorylated PhoP has a number of targets involved in nutrient scavenging (Sola-Landa et al., 2005). Expression of actinorhodin and undecylprodigiosin biosynthetic gene clusters are also effected by the deletion of *phoP,* likely caused by at least two indirect mechanisms. First, PhoP activates the expression of the RNA polymerase sigma factor RpoZ, which is important for the transcription of the actinorhodin and undecylprodigiosin clusters (Santos-Beneit et al., 2011). Second, PhoP and AfsR compete for binding upstream of *afsS*, whose expression activates *actII-orf4* and *redZ.* Hence, PhoP binding to the *afsR* operator upstream of *afsS* negatively influences expression of both specialized metabolites (Martín et al., 2011; Santos-Beneit et al., 2009).

AbsA2 is a global repressor of many specialized metabolism clusters (Liu et al., 2013). It is a response regulator and is phosphorylated by its cognate transmembrane histidine kinase AbsA1 in response to an unknown signal. AbsA2 represses the pathway specific activators of the actinorhodin (*actII-orf4*), calcium-dependent antibiotic (*cdaR*), and undecylprodigiosin (*redZ*) (Sheeler et al., 2005; Uguru et al., 2005).

In addition to the biosynthetic and regulatory genes, export and resistance genes are also normally located within specialized metabolite clusters. Their timely expression is important to ensure that the potentially toxic metabolite (in the case of antibiotics) does not adversely affect or kill the producing colony. For the actinorhodin cluster, the TetR-family regulator ActR represses the divergently-expressed genes encoding actinorhodin efflux pumps (*actA* and *actB*) (**Figure 1.3AB**). As the actinorhodin cluster is activated and metabolic precursors are sequentially assembled into actinorhodin intermediates, ActR binds to these biosynthetic intermediates. This binding results in a conformational change that alters its ability to bind DNA, ultimately relieving its repression of *actAB* (**Figure 1.3C**). This elegant mechanism ensures that the efflux pumps are ready to export the final product prior to it being made (Tahlan et al., 2007). The derepression of *actAB* following binding of actinorhodin biosynthetic intermediates is short-lived, and it takes binding of the final actinorhodin product to get sustained de-repression of *actAB* (**Figure 1.3D**), ensuring that colony-wide resistance to actinorhodin occurs (Xu et al., 2012). How ActR de-represses *actAB* differently when bound to actinorhodin biosynthetic intermediates versus the final product is not currently known.

GBLs and the regulation of the coelimycin biosynthetic cluster

In addition to the multitude of protein regulators that impact actinorhodin and undecylprodigiosin biosynthesis, these clusters, as well as the coelimycin cluster (discussed in more detail below), are also regulated by γ-butyrolactone (GBL) signaling molecules (Takano, 2006).

When considering GBL signaling, the best understood system is that of *Streptomyces griseus*. In *S. griseus*, GBL signaling controls both morphological development and specialized metabolism (Ohnishi et al., 2005) (**Figure 1.4**). In this species the γ-butyrolactone, called A-factor (autoregulatory factor), is synthesized by the gene product of *afsA* (A-factor synthase). A-factor is bound by a TetR-like DNA-binding protein called ArpA (A-factor receptor protein), which in the absence of A-factor, binds a target DNA sequence upstream of *adpA* (A-factor dependent protein). Upon binding to A-factor (once it accumulates to a threshold concentration), the DNA binding ability of ArpA is abolished, leading to the de-repression of *adpA*. AdpA then activates the expression of genes involved in A-factor synthesis, morphological development and regulation of specialized metabolite clusters, including that of the clinically important antibiotic streptomycin (Ohnishi et al., 2005).

Different *Streptomyces* species make different GBLs by varying the composition of the side chain, and GBL receptors can distinguish between these different molecules (Nishida et al., 2007). While most species encode one GBL synthase, many encode

multiple GBL receptors, suggesting that *Streptomyces* species have the ability to respond to GBLs produced by other relatives.

There are numerous differences between the GBL system of *S. griseus,* and that of *S. coelicolor*. In *S. griseus,* GBLs control both development and specialized metabolism, while in *S. coelicolor* their effect is restricted to regulating the latter. Also, in *S. griseus* the GBL receptor ArpA has a single target, which in turn regulates numerous downstream genes (Ohnishi et al., 1999). Conversely, in *S. coelicolor,* the GBL receptor ScbR directly controls dozens of target genes (Li et al., 2015).

The GBL-encoding system in *S. coelicolor* is located adjacent to the coelimycin biosynthetic cluster (**Figure 1.5**). Coelimycin and a related yellow compound (yCpk) are cryptic polyketides that were discovered when a regulatory gene (*scbR2*) within the cluster was deleted (Gottelt et al., 2010). In addition to being negatively regulated by ScbR2, expression of the coelimycin pathway-specific regulator (CpkO) is controlled by ScbR. ScbR is a GBL-receptor protein with homology to ScbR2, and it represses expression of *cpkO* in the absence of GBLs (Takano et al., 2005). Encoded divergently from *scbR* is *scbA,* which encodes the GBL synthase. In *S. coeliocolor,* this GBL is known as SCB1. When ScbA levels rise, more SCB1 is produced (Takano et al., 2001). These molecules then bind to ScbR, relieving repression of *cpkO* (in cells lacking ScbR2) and driving coelimycin production. ScbR2 binds to both actinorhodin and undecylprodigiosin, and the exogenously produced jadomycin antibiotic (Wang et al., 2011; Xu et al., 2010). However, the amount of metabolites used in these *in vitro* assays exceed the amounts of cells are likely to experience in nature. Therefore, it is possible that *in vivo,* ScbR2 binds to a currently unknown metabolite (Liu et al., 2013).

In addition to the quorum-sensing systems of *S. griseus* and *S. coelicolor,* other butyrolactone systems have been studied in the genus. For example, five GBLs (VB-A, VB-B, VB-C, VB-D, and VB-E) are produced by *Streptomyces virginiae* and they regulate production of the specialized metabolite virginiamycin (Kinoshita et al., 1997; Nakano et al., 1998). The GBL IM-2 regulates showdomycin and minimycin production in *Streptomyces lavendulae* (Kitani et al., 1999)**.** The GBL Factor I impacts both anthracyclin and morphological development in *Streptomyces viridochromogenes* (Grafe et al., 1982). In all of the above cases, it is thought that the GBL binds to and changes the activity of a GBL-receptor protein in similar to SCB1/ScbR and A-factor/ArpA in *S. coelicolor* and *S. griseus*, respectively (McCormick and Flärdh, 2011; Nishida et al., 2007). In addition to GBL-based quorum sensing, *S. coelicolor* also uses diffusible methylenomycin furans to control methylenomycin production. These furans are chemically distinct from GBLs, and their mechanism of action is not yet understood (Davis et al., 2009; McCormick and Flärdh, 2011).

Non-coding RNA regulators

Most RNAs transcribed in eukaryotes are not translated into proteins. For a long time, it was thought that these transcripts were not functional and they were dismissed simply as ‘junk’. However, functions for more and more regulatory non-coding RNAs have been determined and it has been argued that organismal complexity appears to be

correlated with the numbers of non-coding genes, not with protein coding genes, suggesting that regulatory non-coding RNAs may serve important roles (Pennisi, 2012). Bacterial genomes are more densely packed with protein-coding genes than their eukaryotic counterparts, and therefore it was thought that they would not have comparable numbers of non-coding DNA. However, recent advances in RNA-sequencing have revealed that the transcriptomes of bacteria are incredibly complex, and hundreds or thousands of non-coding transcripts have been identified in a typical species (Wade and Grainger, 2014). There are two main types of non-coding RNAs (ncRNAs) that are important for this thesis, and they are differentiated mainly by their genomic context: *trans*-encoded intergenic small RNAs (sRNAs) and *cis*-antisense RNAs (asRNAs) (Gottesman and Storz, 2010).

trans-encoded intergenic small RNAs (sRNAs)

sRNAs are typically expressed in intergenic regions and are, as their name implies, small (~50-400 nucleotides in length). Their functions are diverse but typically they act by interacting post-transcriptionally with mRNAs or proteins, and can have either negative or positive regulatory outcomes (Storz et al., 2011). Often, the interactions with other RNAs are facilitated by the RNA chaperone Hfq (Vogel and Luisi, 2011) (**Figure 1.6A**).

Individual sRNAs can have multiple targets and can regulate each target in a different manner. A good example to illustrate the diversity by which sRNAs can act is the sRNA RyhB. RyhB is activated under conditions of low-iron and differentially affects its many mRNA targets. RyhB negatively regulates the expression of many mRNAs by binding near the ribosome-binding site, sequestering it away from ribosomes and in doing so, reducing or eliminating translation initiation. These interactions require Hfq (Aiba, 2007; Massé and Gottesman, 2002; Massé et al., 2003). RyhB does, however, have a positive effect on the expression of genes involved in the synthesis of iron-scavenging siderophores. The *shiA* mRNA encodes a shikimate importer, where shikimate is an important precursor for siderophore synthesis. The 5′ UTR of the *shiA* mRNA is naturally structured, particularly near the ribosome binding site (RBS), resulting in little translation of this mRNA. In low iron conditions, RyhB binds to the *shiA* 5′ UTR, relaxing the RNA secondary structure and increasing ribosome accessibility to the transcript (Prévost et al., 2007).

In addition to acting at the translation initiation level, sRNAs can also act by modulating the stability of mRNAs. Many sRNAs, including RyhB, bind to some target RNAs in complex with Hfq, and directly recruit ribonucleases (*e.g.* RNase E). This reduces the RNA half-lives of these targeted transcripts (Morita and Aiba, 2011; Morita et al., 2005). Regulation at the level of translation initiation and RNA stability are not mutually exclusive, as blocking translation initiation is often accompanied by decreased target RNA half-lives. This is likely due to translating ribosomes protecting the mRNA from ribonucleases (RNases), with the absence of ribosome protection leading to increased accessibility for RNases (Deneke et al., 2013). In the RhyB example mentioned above translation blockage is sufficient for sRNA-mediated repression of expression, but

it is thought that the added step of mRNA degradation adds another level of regulatory insurance (Prévost et al., 2007).

As outlined above, the RyhB sRNA acts in a variety of ways, including modifying accessibility to RNases, and positively or negatively regulating translatability of different mRNAs. While this certainly makes it an outstanding regulatory molecule, this level of regulatory complexity exerted by individual sRNAs may be more common than was originally thought. A recent report showed that a single quorum-sensing sRNA (Qrr3) in *Vibrio harveyi* regulates multiple mRNA targets using four different mechanisms. Qrr3 promotes the catalytic degradation of *luxR* mRNA (*i.e.* the sRNA promotes degradation of the mRNA by ribonucleaeses but not itself degraded along with this target), while it is co-degraded with the *luxM* mRNA. Qrr3 further blocks translation of the *luxO* mRNA by binding near the RBS, while it activates translation of *aphA* mRNA by relieving intrinsic inhibitory secondary structure near the RBS (Feng et al., 2015).

sRNAs also can directly target proteins. Usually this is accomplished by RNA mimicry, where the sRNA mimics the normal RNA or DNA target of the protein (Marzi and Romby, 2012). A classic example of this is the ubiquitous 6S sRNA. The 6S sRNA is induced upon entry into stationary phase and it structurally mimics the open promoter complex recognized by housekeeping RNA polymerase holoenzyme. It functions to sequester the housekeeping RNA polymerase homoenzyme, and in doing so, reduces transcription of certain genes not required during stationary phase (Cavanagh and Wassarman, 2014).

cis-antisense RNAs

While *trans*-encoded sRNAs are located in intergenic regions and have multiple targets, asRNAs are located on the opposite strand from protein encoding genes and are generally thought to target their cognate sense mRNAs (Brantl, 2007; Thomason and Storz, 2010). As they share perfect complementarity with their associated sense mRNA, asRNAs can act in similar ways to sRNAs. The mechanism of action of asRNAs is often assumed by their genetic context. For example, an asRNA is expressed opposite a RBS, it is assumed that it negatively regulates translation of its target mRNA; if expressed opposite a coding region, it is assumed that they promote degradation by RNases (**Figure 1.6B**). However, due to the challenges associated with modifying antisense RNA expression without also impacting their associated sense gene, proof of any regulatory mechanism is usually lacking.

Often asRNAs are expressed opposite the 5′ UTR of a protein-encoding gene. In these cases, translation modulation may occur by either blocking or promoting ribosome binding to the RBS. For example, in cyanobacteria, an antisense RNA is expressed opposite the RBS of *furA*. In a mutant unable to produce the antisense RNA, FurA levels are increased, consistent with the antisense RNA blocking *furA* translation (Hernández et al., 2006).

Antisense RNAs can also alter the activity of RNases targeting mRNA transcripts. RNase III is a double strand-specific ribonuclease that is a prime candidate for being involved in antisense RNA-mediated gene regulation. In *Staphylococcus aureus,* RNA-sequencing (RNA-seq) in wild-type and Δ*rnaseIII* strains revealed that ~75% of mRNAs

are associated with antisense RNAs that are only detectable in the mutant strain (Lasa et al., 2011). The function of such pervasive antisense transcription is currently unknown (Lasa et al., 2012). In addition to RNase III, RNase E has also been proposed as being important for antisense RNA regulation of mRNAs. For example, in *Prochlorococcus,* it has been suggested that an antisense RNA protects mRNAs from RNase E cleavage (Stazic et al., 2011). However, these experiments were performed *in vitro* and it is unclear how such protection would occur *in vivo*.

asRNA control of translation and transcript stability, as described above, are similar to those mediated by *trans*-encoded sRNAs. However, due to their location opposite the protein-encoding gene, asRNAs also may act at the transcriptional level, and influence the likelihood of RNA polymerase collisions, and overall RNA polymerase processivity (Shearwin et al., 2005).

Blurring lines of noncoding RNA regulators

As we learn more about the functioning of regulatory ncRNA, the lines are blurring between what used to be considered distinct categories of ncRNAs. For example, there have been recent reports of tRNA-associated transcripts acting as 'sponges' in binding sRNAs (Lalaouna et al., 2015). sRNA transcripts having also been found to have dual roles as both ncRNAs and as mRNAs for small proteins. Also, asRNAs can serve as both *cis* regulators and as alternative promoters for adjacent divergently transcribed genes (Sesto et al., 2012). Since only a few asRNAs have been characterized in any detail, it is likely that many other surprises await discovery.

Aims of this thesis

Previous work by Swiercz et al. (2008) sought to identify non-coding RNAs in *S. coelicolor* using a combined bioinformatics and cloning approach (Swiercz et al., 2008). Shortly after this work was published, RNA-seq emerged as a powerful tool for studying transcriptomes in bacteria (Pinto et al., 2011; Sharma and Vogel, 2009; van Vliet, 2010). We used RNA-seq to identify hundreds of new non-coding RNAs in three *Streptomyces* species (Moody et al., 2013), greatly expanding the number of known genes in these species. Many of these non-coding RNAs are located within specialized metabolite biosynthetic clusters, suggesting a possible role in regulating the production of these potentially important metabolites. This work is described in Chapter 3.

One of the most interesting non-coding RNAs discovered in the above work was an asRNA (ScbN) expressed opposite the GBL synthase *scbA*. Chapter 4 outlines efforts to characterize this antisense asRNA. Surprisingly, the expression of ScbN had little effect on *scbA*, and instead a hairpin structure within *scbA,* corresponding to the reverse complement of the *scbN* terminator, had profound effects on *scbA* expression. This work led to an investigation into hairpins found within the coding regions of other mRNAs. Chapter 5 details the development of a bioinformatics tool that finds conserved coding-region hairpins throughout bacteria, and highlights a number of potentially interesting hairpins for follow-up investigation.

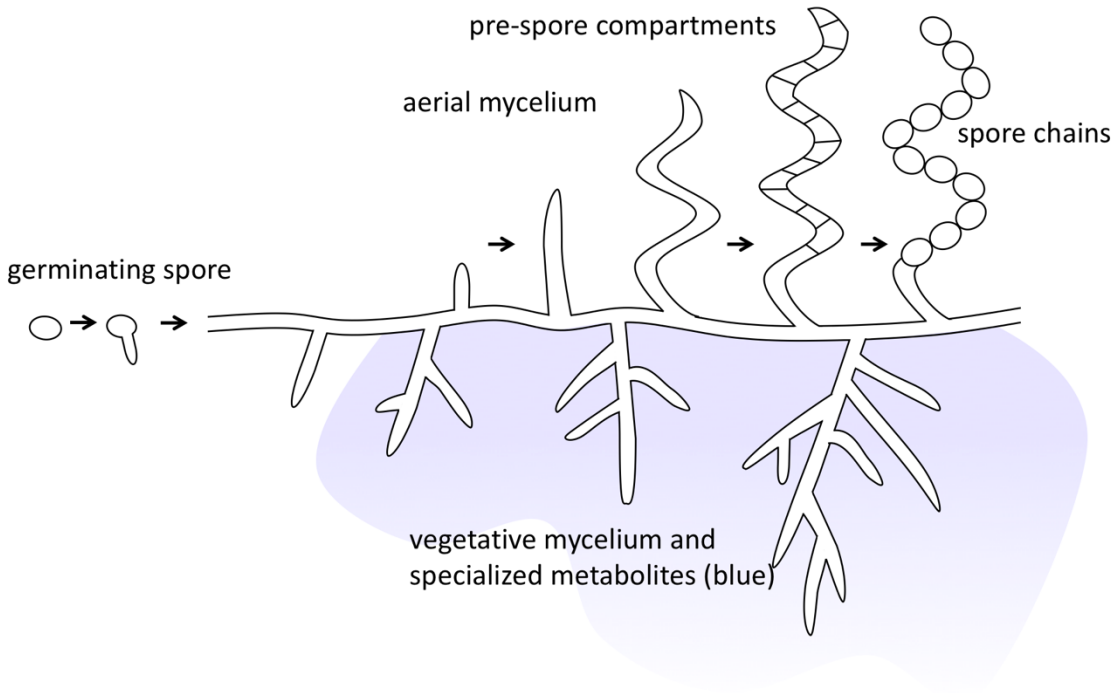


Figure 1.1: The *Streptomyces* life-cycle. Upon encountering favourable growth conditions, a germ tube emerges from a spore. Tip extension and periodic branching create a complex mesh of vegetative mycelium. Aerial mycelia are raised out of the growth substrate and the underlying vegetative mycelium produce many specialized metabolites (highlighted in blue). The chromosomes in the multi-genomic aerial hyphae are separated into pre-spore compartments, which mature into spore chains. The life-cycle begins again from a unigenomic exospore. Modified from Swiercz and Elliot (2012).

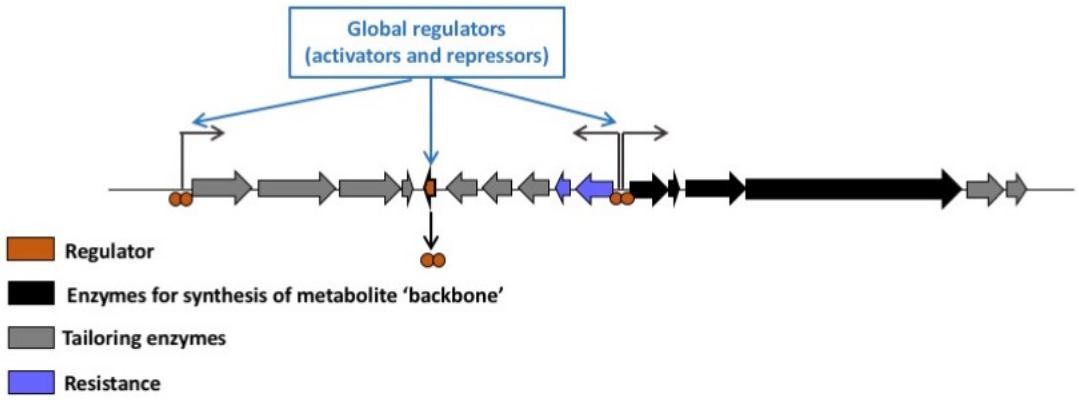


Figure 1.2: Regulation of a typical specialized metabolite cluster. The pathway-specific regulators, enzymes responsible for synthesis of the backbone of the metabolite, tailoring enzymes that modify the metabolite and resistance determinants are all typically encoded by a single cluster of genes. The pathway specific regulators usually control the expression of operons encoding the synthases and tailoring enzymes. Expression of these pathway specific regulatory genes is normally controlled by multiple global regulators, which are expressed from genome locations outside the biosynthetic clusters and often control more than one specialized metabolite cluster.

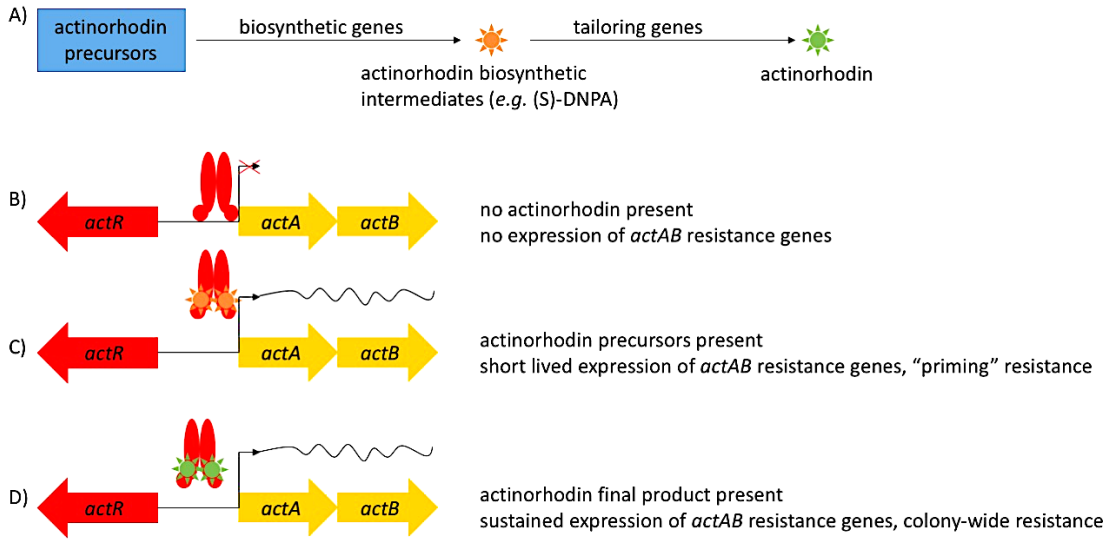


Figure 1.3. Actinorhodin biosynthetic intermediates and final product trigger resistance to actinorhodin. A) Actinorhodin precursors are converted into intermediate products by biosynthetic gene products encoded within the actinorhodin gene cluster, and these intermediates are further converted into the final actinorhodin product by tailoring enzymes. **B)** In the absence of actinorhodin (and intermediates), ActR (red) binds to the *actAB* promoter region and represses expression. **C)** Biosynthetic intermediates of actinorhodin bind to ActR and change its conformation such that it no longer binds DNA. This results in short-lived depression of *actAB,* and priming of the actinorhodin export / resistance system. **D)** Binding of the final actinorhodin product to ActR results in sustained derepression and colony-wide expression of *actAB*.

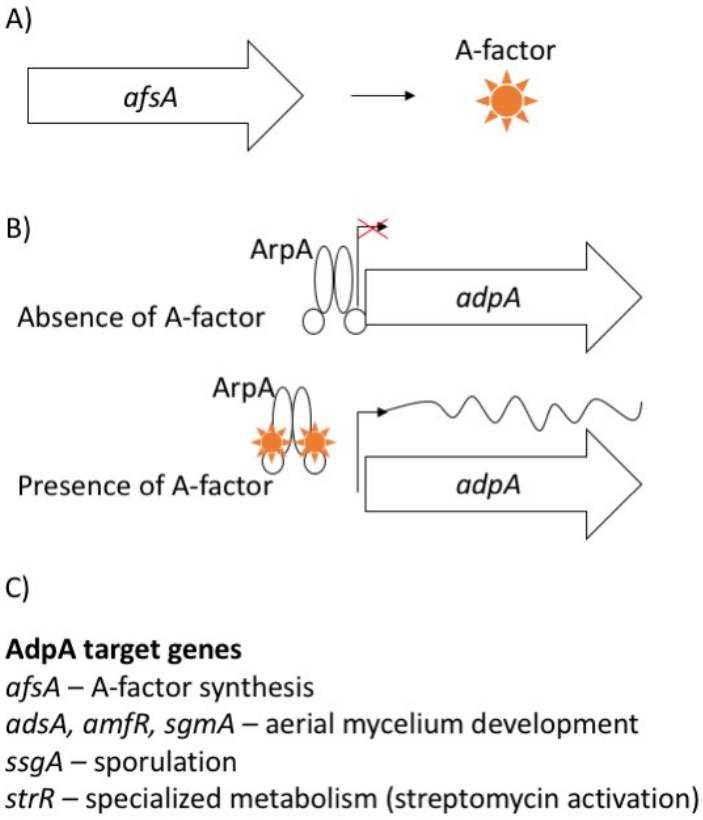


Figure 1.4. The GBL system of *Streptomyces griseus*. A) *afsA* encodes an enzyme that synthesizes the GBL A-factor. **B)** In the absence of A-factor, ArpA binds and represses *adpA*. A-factor binds to ArpA, changing its conformation such that it does not bind to DNA, relieving the repression of *adpA*. **C)** AdpA is a transcriptional activator that has many targets involved in morphological development and specialized metabolism.

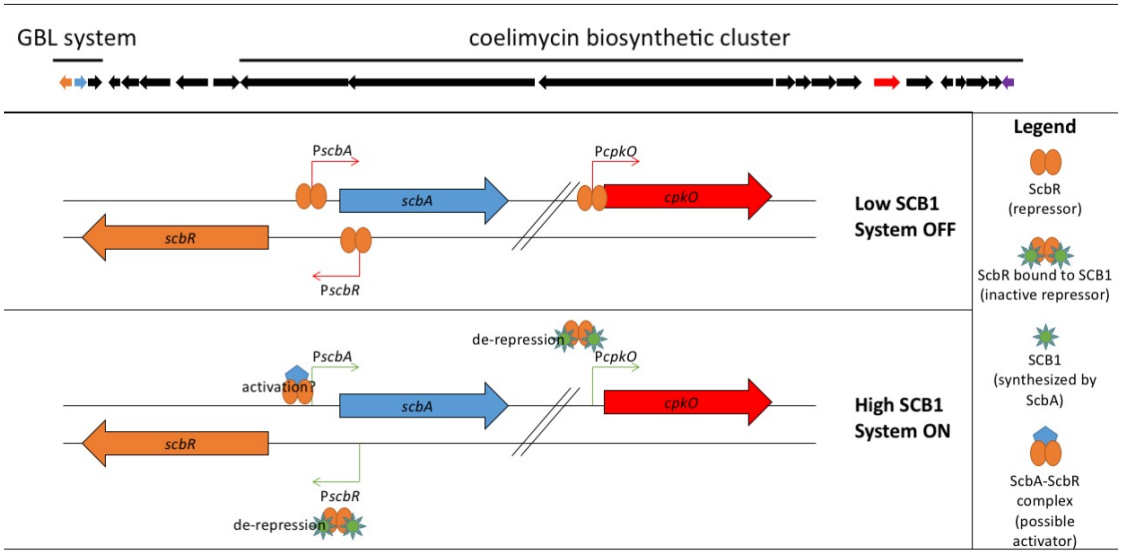


Figure 1.5. Overview of the *Streptomyces coelicolor* quorum-sensing system. The GBL genes (*scbA, scbR*) are located to the left of the coelimycin cluster (shown in orange and blue). The pathway specific activator of the coelimycin cluster is CpkO (gene is shown in red near the middle of the cluster). At low SCB1 levels, ScbR binds and represses transcription from its own promoter, and that of *scbA* and *cpkO*. At high SCB1 levels, a non-repressive SCB1:ScbR complex is formed resulting in de-repression of the *scbR* and *cpkO* promoters. Transcription of *scbA* requires both *scbA* and *scbR*. The mechanism of *scbA* activation is not yet clear; however, some groups have proposed that a ScbA:ScbR complex could act as an activator (D'Alia et al., 2011; Mehra et al., 2008). At the right-end of the cluster is shown *scbR2* (purple). ScbR2 acts in a similar manner to ScbR. However, instead, its transcriptional repression is relieved by binding to actinorhodin, undecylprodigiosin, and jadomycin (Wang et al., 2011; Xu et al., 2010), instead of SCB1.

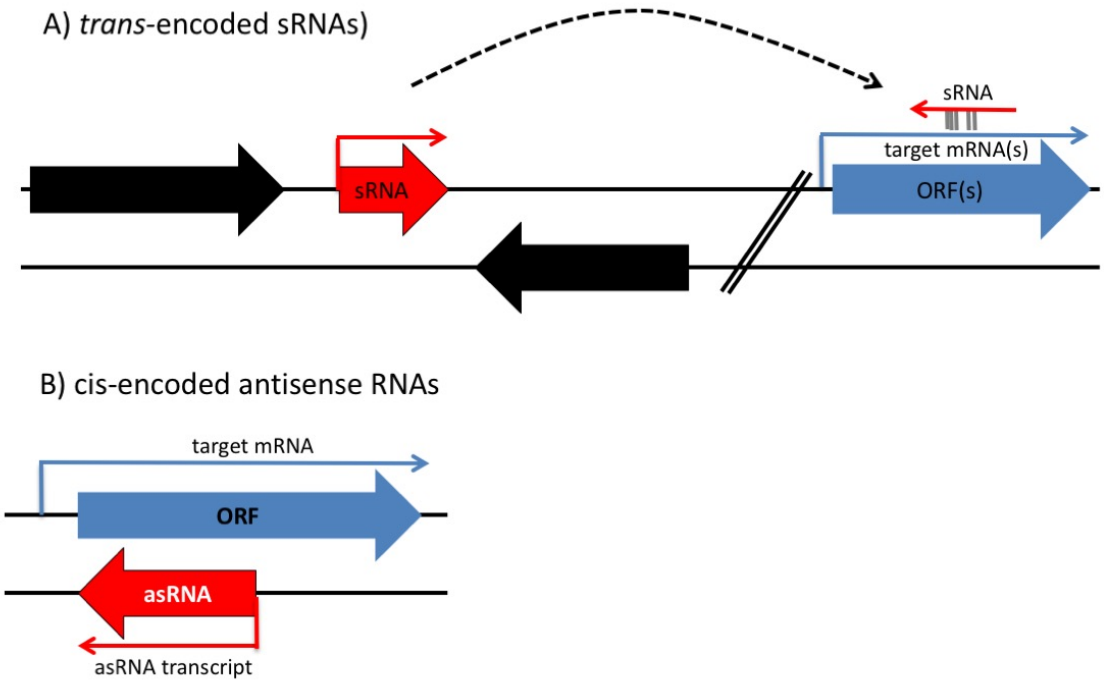


Figure 1.6: Schematic illustration of *trans*-encoded sRNAs and *cis*-encoded antisense RNAs. Genes are depicted as thick arrows, with protein-coding genes shown in blue and labeled as 'ORF's (open reading frame), non-coding RNA genes and transcripts shown in red, and black depicting genes of either type. RNA transcripts are shown above their corresponding gene, with transcription initiating at the vertical line, and terminating at the small arrowhead. **A)** sRNA genes are most commonly found in the intergenic region between genes, and typically target (by imperfect complementary base-pairing) one or more mRNAs expressed from disparate chromosomal locations. **B)** Antisense RNAs (asRNAs) are expressed from a promoter on the strand opposite a protein-coding gene, and are thought to regulate that gene.

Chapter 2: General Materials and Methods

Culturing and microbiological techniques

Growth media

All media for growing bacteria was sterilized in an autoclave at 121˚C for 45 min. Some *E. coli* media prepared by other lab members may have been autoclaved for 20 or 35 min. Prior to use in media, concentrated stock antibiotic solutions were prepared in distilled water and filter-sterilized through a 0.2 µm Pall membrane and stored at -20˚C. Exceptions to this included chloramphenicol, which was prepared in ethanol and not filter-sterilized, naladixic acid, which was prepared in 0.2 M NaOH, and hygromycin B, which was purchased as a solution from Invitrogen. Unless indicated otherwise, antibiotics were used at the final concentrations listed in **Table 2.1**.

E. coli agar plates were prepared with ~25 mL medium per plate, while ~33 mL was used per *Streptomyces* plates to ensure that the media did not dry out during longer incubations. For liquid growth of *S. coelicolor*, either small metal springs or glass beads were added to the bottoms of universals or flasks to help ensure that the mycelia grew in a dispersed manner.

Streptomyces were cultured in different media types for different purposes. For much of the *scbN* work described in Chapter 4, *S. coelicolor* was grown in liquid supplemented minimal medium (SMM) or on supplemented minimal medium solid (SMMS) (Kieser et al., 2000; Takano et al., 1992, 2001).

A 1:1 mixture of YEME (yeast extract-malt extract) and TSB (tryptone soya broth) (Kieser et al., 2000) was used for growing overnight cultures of *Streptomyces* prior to sub-culturing into SMM (*S. coelicolor* does not germinate well in SMM) and for growing *Streptomyces* strains for isolating genomic DNA. These two media types were prepared separately and then combined prior to use. To germinate spores prior to performing conjugations, 2×YT medium was used (Kieser et al., 2000).

Soy flour mannitol (SFM) agar (Floriano and Bibb, 1996; Kieser et al., 2000) was used for propagating *Streptomyces* exconjugants and for preparing *S. coelicolor* spore stocks. Difco nutrient agar (DNA) (Kieser et al., 2000) was used for rapid, non-differentiating growth of *Streptomyces* (*e.g.* to count colony forming units).

E. coli strains were grown on LB (Luria Bertani) or DNA solid medium, or in liquid LB or super optimal broth (SOB) (Kieser et al., 2000). DNA and SOB both lack NaCl, and thus were used to select for hygromycin resistant *E. coli* strains (where hygromycin is salt sensitive). SOB was also used when growing strains prior to electroporation, as salt can cause electrical discharge (arcing) during the electroporation protocol.

MYM (maltose, yeast extract, malt extract) (Stuttard, 1982) is the medium of choice for growing *Streptomyces venezuelae*, and was used here to compare the gene expression profiles of *S. venezuelae, S. coelicolor* and *Streptomyces avermitilis*. MYM was prepared by dissolving 4 g maltose, 4 g yeast extract, 10 g malt extract in a mixture of 500 mL tap water and 500 mL dH_2O. After autoclaving, 2 mL of trace element solution was added to the medium. The trace element solution consisted of 40 mg $ZnCl_2$, 200 mg

$FeCl_3 \cdot 6H_2O$, 10 mg $CaCl_2 \cdot 2H_2O$, 10 mg $MnCl_2 \cdot 4H_2O$, 10 mg $Na_2B_4O_7 \cdot 10H_2O$ and $(NH_4)_6Mo_7O_{24} \cdot 4H_2O$.

.

Strains and plasmids

To make spore stock suspensions of *Streptomyces* strains, individual colonies were broken up in ~50-100 µL of sterile 20% glycerol and spread on top of a cellophane disc placed on top of SFM agar. After growth to sporulation (~4 to 7 days for *S. coelicolor*) biomass was scraped off of the cellophane discs with a sterile spatula and placed into ~10 mL of sterilized water in a small universal vial. The biomass was then sonicated in a water bath until a near-homogenous suspension was achieved, after which it was passed through a sterile cotton filter. The filtered suspension was then pelleted by centrifugation, and the resulting spore pellet was re-suspended in sterile 20% glycerol and stored at -20˚C.

To make stocks of *E. coli* strains, an equal volume of overnight culture and 40% glycerol was mixed, and the resulting stock was stored at -80˚C. Strains and plasmids used in this work are summarized in **Table 2.2 and Table 2.3**. Primers and PCR products used to create and check the above strains are listed in **Tables 2.4** and **2.5**. Cloning details of the strains, constructs, and PCRs listed are described in detail where appropriate in subsequent chapters.

Genetic and molecular biology techniques

DNA amplification by polymerase chain reaction

DNA amplification was achieved using one of two polymerase chain reaction (PCR) protocols, depending on the circumstances. When the error-rate of the polymerase was not critical (*e.g.* when checking for presence of an insert in a plasmid), *Taq* DNA polymerase was used (Norgen Biotek Corporation or BioShop), as per the manufacturer's instructions. When checking for plasmid inserts in *E. coli*, or checking for DNA integration into the *Streptomyces* chromosome, these *Taq* PCR amplifications were often performed using *E. coli* or *Streptomyces* colonies as template. When performing colony PCR on *E. coli,* a small portion of the colony was transferred directly into the PCR and the initial denaturing step was performed for 10 min at 95˚C. When performing colony PCR using sporulating *Streptomyces* colonies, a small amount (~ one cubic mm) of a colony was first placed in 20 µL dH_2O and boiled for 10-15 min. An aliquot (1-2 µL) of this cell lysate was then added as template to the PCR. Cycles for *Taq* PCRs were performed with an initial denaturation step at 95˚C for 5-10 min. This was then followed by 30-37 cycles of: denaturation at 95˚C for 30 s, annealing at 40-72˚C for 30 s (depending on the oligonucleotide primers being used), and extension at 72˚C for 30 s to 5 min (depending on the length of the desired amplicon). Reactions were completed with a final extension step of 10 min at 72˚C.

When performing PCRs that required error-free products (*e.g.* when amplifying a DNA fragment for cloning), a high fidelity DNA polymerase [iProof^{TM} High-Fidelity

DNA polymerase (BioRad) or Phusion® High-Fidelity DNA Polymerase (New England BioLabs)]. When using the Phusion enzyme, the specialized "high GC buffer" was used when amplifying *Streptomyces* DNA. Phusion amplifications involved an initial denaturation step at 98˚C for 2 min. Subsequently, 30-37 cycles of the following were performed: denaturation at 98˚C for 10-30 s, annealing at 40-72˚C (depending on oligonucleotide primers being used) for 10-30 s, and extension at 72˚C for 15 s to 3 min (depending on the length of the desired amplified region). This all culminated in a final extension step of 5 min at 72˚C. Annealing temperatures were either calculated using an online tool (http://tmcalculator.neb.com), or optimized experimentally by performing temperature gradients. PCR products were analyzed on an agarose gel as described above.

Isolation of plasmid and cosmid DNA from *E. coli*

Plasmids were isolated using the Invitrogen PureLink™ Quick Plasmid Miniprep Kit (following the manufacturer's instructions), or when larger quantities of plasmid were required, using phenol-chloroform extraction. Cosmids were also isolated using a phenol-chloroform-based extraction protocol. Briefly, cell pellets from 40 mL overnight cultures were resuspended in 1 mL ice-cold solution I (50 mM Tris-HCl pH 8, 10 mM EDTA, 55 mM glucose) by vortexing and pipetting the mixture up and down repeatedly. Next, 2 mL of room-temperature solution II (200 mM NaOH, 1% SDS) were added and the tubes were mixed by gently inverting 10 times. The nucleic acids were precipitated by adding 1.5 mL of ice-cold solution III (3 M potassium acetate pH 5.5), mixed quickly by inverting, and centrifuged at >6,800 ×*g* for 10 min. Approximately 4 mL of the supernatant were then transferred to a fresh tube containing an equal volume of phenol:chloroform:isoamyl alcohol (50:50:1). The mixture was then vortexed vigorously for 30 s to 2 min, before being centrifuged at 12,000 ×*g* for 5 min. The nucleic acid-containing upper aqueous phase was then transferred to a fresh tube. To precipitate the nucleic acids, 2 volumes of ice cold ethanol and 0.1 volume of 5 M sodium acetate were added. The solutions were mixed by inversion, and incubated at -20˚C or on ice for ~30 min. The DNA was then pelleted by centrifugation at 12,000 ×*g* at 4˚C for 30 min. The pellet was then washed in ice-cold 70% ethanol to remove salts, and pelleted again. Residual ethanol was removed, and the pellet was air dried for ~5 min before being re-suspended in nuclease-free water. All plasmid/cosmid DNA was then stored at -20˚C. The volumes of the above protocol were scaled appropriately depending upon culture volume (*e.g.* when isolating from 5 mL cultures, smaller volumes of each solution was used). To remove RNAs, RNase A was added either directly in Solution I, or following re-suspension in water. In the latter case, the solution was incubated at 37˚C for 1 h before being subjected to a phenol:chloroform extraction and precipitation as described below.

Plasmid DNA sequence and integrity were confirmed by sequencing through MOBIX Labs at McMaster University.

Restriction digestion of DNA

Restriction enzymes were purchased either from New England BioLabs or ThermoFisher Scientific and used as recommended by the manufacturer. Digests (100-

200 ng) were typically performed in 50 µL reactions; however, for plasmid digests that required a subsequent dephosphorylation step, this volume was normally increased to 100-200 µL to ensure glycerol concentrations remained low following addition of the phosphatase enzyme. Following digestions, the DNA was typically recovered either by gel extraction (see below) or with an Invitrogen Quick PCR Purification Kit.

Dephosphorylation of DNA

When cloning DNA fragments into plasmids, use of a single restriction enzyme for plasmid digestion would yield a product with ligation-compatible ends. To minimize the opportunity for re-ligation, plasmids would be subjected to a dephosphorylation step following digestion. As calf intestine alkaline phosphatase (Roche) is compatible with the restriction digest buffers used here, calf intestine alkaline phosphatase was added to restriction digests and incubated at 37˚C for 1 h as per the manufacturer's instructions. Gel extraction (see below) or PCR purification kits were used to purify the digested and dephosphorylated DNA.

Agarose gel electrophoresis and gel purification

Agarose gels were prepared with TBE buffer (Tris base, boric acid, EDTA) with varying concentrations of agarose. Lower percentage agarose gels were used to analyze large (>2,000 bp) DNA fragments (*e.g.* 0.8% for analysis of plasmid digestions) and higher percentage gels were used to visualize smaller DNA fragments or total RNA (*e.g.* 3% for small inserts for cloning and 2% for total RNA). Immediately before pouring, 2.5 µL of ethidium bromide (BioShop) were added for every 50 mL of agarose. Gels were run at 100 V for ~30 min. DNA bands were visualized by exposure to UV light at 305 nm. UV exposure was minimized if fragments were to be isolated for subsequent cloning, by using a longer wave-length (365 nm) and by limiting UV exposure time. In these instances, DNA-containing regions were excised from the gel using a razor blade, and the DNA was then purified using a Monarch™ DNA Gel Extraction Kit (New England BioLabs) or MinElute® Gel Extraction kit (Qiagen), as per the manufacturer's instructions.

DNA ligations

DNA ligations were performed either using the Rapid DNA Ligation Kit (Roche) or T4 DNA Ligase (Invitrogen) as per the manufacturer's instructions with minor modifications. Typically, digested insert DNA was combined with digested vector DNA in a 5:1 molar ratio in a total reaction volume of 21 µL containing ~100 ng of total DNA. However, when using the Rapid DNA Ligation kit, this ratio was often increased to 10:1 or even 20:1 as these ratios tended to give better results. Also, the manufacturer's protocol for the Invitrogen T4 DNA ligase recommended 30-60 min ligation at room temperature. However, this reaction was normally performed overnight at 16˚C, as again, this tended to result in more successful ligations.

Preparing and transforming competent *E. coli* cells

When cloning or subcloning into plasmids, the resulting constructs were transformed into commercially competent DH5α cells. Plasmid-insert ligations (2 or 3 μL) were mixed with a 50 μL aliquot of these cells and left on ice for 30-120 min. The cells were subsequently heat shocked in a water bath at 37˚C for 20 s, after which they were placed immediately on ice for ~2 min. Next, 800 μL of LB were added to these tubes. The culture was then incubated at 37˚C, shaking at 200 rpm, for ~1 h to allow for expression of the antibiotic resistance genes. Finally, the cells were collected by centrifugation, resuspended in ~100 μL, and spread on LB or DNA agar plates containing appropriate antibiotics to select for the incoming plasmid. The cells were then incubated at 37˚C for 16-24 h until colonies were visible.

Plasmids and cosmids were moved into *E. coli* ET12567/pUZ8002 (for conjugation with *Streptomyces*) or BW25113/pIJ790 (for creating gene knockouts in cosmids) (**Table 2.2**) by electroporation. To create electro-competent *E. coli* cells, overnight cultures were sub-cultured into SOB containing kanamycin and chloramphenicol, and were grown to an OD_{600} of approximately 0.4. The cultures were then placed on ice for 1 h, after which the cells were pelleted by centrifugation. The cells were then washed three times in ice-cold 10% glycerol before being resuspended in the residual glycerol solution (~100 μL). Approximately 100 ng (0.5-2 μL) of plasmid/cosmid DNA were then added to 50 μL of electrocompetent cells. The DNA-cell suspension was mixed well by flicking the microcentrifuge tube before being transferred into an ice-cold electroporation cuvette (BioRad). Following electroporation at 2.5 kV, 800-1000 μL of ice-cold LB broth was added to the cell suspension, and the mixture was incubated shaking at 37˚C for ~1 h and then plated on solid agar media with appropriate antibiotics to select for the incoming plasmid/cosmid.

Streptomyces conjugation

Plasmid or cosmid conjugations from *E. coli* to *Streptomyces* were performed as described previously (Kieser et al., 2000). Exconjugants typically appeared 3-5 days following antibiotic overlay, at which point colonies were re-streaked onto SFM agar containing naladixic acid and the appropriate antibiotic for plasmid/cosmid selection. A single colony from this plate was then dispersed and spread onto SFM overlaid with cellophane discs to generate a spore stock, as described above.

Isolation of genomic DNA from _Streptomyces_

Genomic DNA was isolated from *S. coelicolor* strains grown for ~2 days in a YEME:TSB mixture, using a Bacterial Genomic DNA Isolation Kit (Norgen Biotek Corporation), as per the manufacturer's instructions.

RNA-isolation

RNA was isolated using a modified version of the guanidium thiocyanate protocol described previously (Chomczynski and Sacchi, 1987). Briefly, cells were lysed by vortexing with glass beads in a guanidium thiocyanate solution (4 M guanidium thiocyanate, 25 mM trisodium citrate dihydrate, 0.5% w/v sodium N-lauroylsarcosinate, 0.8% ß-mercaptoethanol) until homogeneous. The resulting suspension was subjected to

two phenol-chloroform extractions and one acid phenol-chloroform extraction. Total nucleic acids were precipitated overnight at −20°C in isopropanol, before being pelleted, washed with 70% ethanol and resuspended in nuclease free water. Contaminating DNA was removed using Turbo DNase (Ambion), and RNA purity and concentrations were determined using a Nanodrop spectrophotometer. RNA quality was assessed using an Agilent 2100 Bioanalyzer prior to RNA-seq, or by agarose gel electrophoresis for other applications.

Northern blotting

Northern blotting was performed as described previously (Haiser et al., 2008; Swiercz et al., 2008), only RNA was crosslinked to membranes using a 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide cross-linking solution at 55°C for 2 h (Pall and Hamilton, 2008). Membranes were stripped with high-stringency buffer (0.2% saline-sodium citrate, 0.1% sodium dodecyl sulfate) at 65°C. They were subsequently checked by exposure to a storage phosphor screen to ensure complete removal of all radiolabelled probe, and re-probed as necessary.

Table 2.1: Antibiotic concentrations used in growth media

	Concentration (µg/mL)	
Antibiotic	***Streptomyces***	***E. coli***
Apramycin	50	50
Kanamycin	50	50
Chloramphenicol	n/a	25
Ampicillin	n/a	100
Hygromycin B	50	50
Naladixic acid	n/a	25

Table 2.2: Strains used in this work

Strain name	Plasmids	Description	Reference
Streptomyces coelicolor			
M145	n/a	SCP1- SCP2-	(Bentley et al., 2002)
M1146	n/a	Δ*act* Δ*red* Δ*cpk* Δ*cda*	(Gomez-Escribano and Bibb, 2011)
M145 P*scbN-gusA*	pGus containing P*scbN*	*scbN* promoter reporter construct in wild-type background.	This work
M1146 P*scbN-gusA*	pGus containing P*scbN*	*scbN* promoter reporter construct in M1146 background.	This work
M145Δ*scbAR* P*scbN-gusA*	pGus containing P*scbN*	*scbN* promoter reporter construct in Δ*scbAR* background.	This work
M1146 P*scbN* (with terminator)	pGus containing P*scbN* (with terminator)	*scbN* promoter-driven reporter that ends after the RNA hairpin predicted to be the *scbN* terminator. Used to test *scbN* termination.	This work
M145 P*scbA-gusA*	pGus containing P*scbA*	*scbA* promoter reporter construct in wild-type background.	This work
M1146 P*scbA-gusA*	pGus containing P*scbA*	*scbA* promoter reporter construct in M1146 background.	This work
M145Δ*scbAR* P*scbA-gusA*	pGus containing P*scbA*	*scbA* promoter reporter construct in Δ*scbAR* background.	This work
M1146 P*scbR –gusA*	pGus containing P*scbR*	*scbR* promoter reporter construct in M1146 background.	This work
M145Δ*scbAR* P*ermE* scbA* (without *scbN*)-*gusA*	pGus containing P*ermE* scbA* (without *scbN*)	P*ermE** driven reporter of the *scbA* mRNA up to, but not including, the *scbN* promoter.	This work
M145Δ*scbAR* P*ermE* scbA* (with *scbN*)-*gusA*	pGus containing P*ermE* scbA* (with *scbN*)	P*ermE** driven reporter of the *scbA* mRNA up to, but not including, the *scbN* promoter.	This work
M145Δ*scbAR* P*ermE* scbA* (divergent)-*gusA*	pGus containing P*ermE* scbA* (divergent)	Monitoring effects of divergent transcription of *scbR* on *scbA* expression.	This work
M145Δ*scbAR* P*ermE* scbA* (divergent and wild-type hairpin)-igusA	pGus containing P*ermE* scbA* (divergent and wild-type hairpin)	Monitoring the combined effects of *scbR* divergent transcription and the *scbA* hairpin on *scbA* expression.	This work

M145Δ*scbAR* P*ermE* scbA* (divergent and disrupted hairpin)-*gusA*	pGus containing P*ermE* scbA* (divergent and disrupted hairpin)	Monitoring the combined effects of *scbR* divergent transcription and a disrupted version of the *scbA* hairpin on *scbA* expression.	This work
M145Δ*scbAR scbA* (wild-type hairpin)-*gusA*	pGus containing *scbA* (wild-type hairpin)	Monitoring the effects of the *scbA* hairpin on *scbA* expression.	This work
M145Δ*scbAR scbA* (disrupted hairpin)-*gusA*	pGus containing *scbA* (disrupted hairpin)	Monitoring the effects of a disrupted version of the *scbA* hairpin on *scbA* expression.	This work
Escherichia coli			
SE DH5α		Highly competent ('**s**ubcloning **e**fficiency') version of DH5α, used as a cloning host: [*fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80 Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*]	(Hanahan, 1983) and Invitrogen
ET12567		Conjugation and / or de-methylation strain: *dam*, *dcm*, *hsdS*, *cat*, *tet*; has plasmid pUZ8002 (below)	(MacNeil et al., 1993)
BW25113		*E. coli* host for manipulating *Streptomyces* cosmids: Δ*araBAD* Δ*rhaBAD*	(Datsenko and Wanner, 2000)

Table 2.3: Plasmids/cosmids used in this work

Plasmid /cosmid name	PCR product insert	Restriction sites used	Purpose / Description	Created
StAH10	n/a	n/a	Cosmid containing the *scbA, scbN, scbR* region. *bla* (AmpR), *kan*^{R}	(Bentley et al., 2002)
pUZ8002	n/a	n/a	Plasmid encoding conjugative apparatus to shuttle plasmids from *E. coli* into *Streptomyces*; *tra, neo,* RP4	(Paget et al., 1999)
pIJ790	n/a	n/a	Temperature-sensitive plasmid containing arabinose-inducible λ RED recombination system; used for making gene knockouts in *Streptomyces.* *cat* (chloramphenicol resistance)	(Gust et al., 2003)
pIJ797	n/a	n/a	A plasmid containing the hygromycin resistance gene and *oriT* (the 'hyg cassette'). Used as template for PCR reactions when generating gene knockouts;*hygR, oriT*	(Saleh et al., 2012)
pGus	n/a	n/a	Transcriptional reporter backbone with *gusA*. Negative control for reporter constructs. *aac(3)IV* (ApraR) ; ΦC31 *attP*]	(Myronovskyi et al., 2011)
pGus P*scbN*	P*scbN*	XbaI/KpnI	*scbN* promoter reporter construct.	This work
pGus P*scbN* (with terminator)	P*scbN* (with terminator)	XbaI/KpnI	*scbN* termination reporter construct.	This work
pGus P*scbA*	P*scbA*	XbaI/KpnI	*scbA* promoter reporter construct.	This work
pGus P*scbR*	P*scbR*	XbaI/KpnI	*scbR* promoter reporter construct.	This work.
pGus P*ermE**	P*ermE**	XbaI/KpnI	Strong promoter *ermE** reporter; often used as a positive control for reporter assays.	(St-Onge, 2016)
pGus P*ermE* scbA* (without *scbN*)	P*ermE**, *scbA* (without *scbN*)	XbaI/KpnI (P*ermE**), KpnI/SpeI (*scbA* mRNA without *scbN*)	*scbA* mRNA reporter construct lacking the *scbN* promoter on the opposite strand, with transcription initiating at the *ermE** promoter.	This work
pGus P*ermE* scbA* (with *scbN*)	P*ermE**, *scbA* (with *scbN*)	XbaI/KpnI (P*ermE**), KpnI/SpeI (*scbA*	*scbA* mRNA reporter construct containing the *scbN* promoter on the opposite strand, with *scbA* expression being driven by the *ermE** promoter.	This work

		mRNA with *scbN*)		
pGus P*ermE* scbA* (divergent)	P*ermE*, scbA* (divergent)	XbaI/KpnI (P*ermE**), KpnI/SpeI (*scbA* [divergent])	Monitoring effects of divergent transcription of *scbR* on *scbA,* with *scbA* expression being driven by the *ermE** promoter.	This work
pGus P*ermE* scbA* (divergent and wild-type hairpin)	P*ermE*, scbA* (divergent and wild-type hairpin)	XbaI/KpnI (P*ermE**), KpnI/SpeI (*scbA* [divergent and wild-type hairpin])	Monitoring the combined effects of *scbR* divergent transcription and the *scbA* hairpin on *scbA* expression (which is driven by the *ermE** promoter).	This work
pGus P*ermE* scbA* (divergent and disrupted hairpin)	P*ermE*, scbA* (divergent and disrupted hairpin)	XbaI/KpnI (P*ermE**), KpnI/SpeI (*scbA* [divergent and disrupted hairpin])	Monitoring the combined effects of *scbR* divergent transcription and a disrupted version of the *scbA* hairpin on *scbA* expression (which is driven by the *ermE** promoter).	This work
pGus *scbA* (wild-type hairpin)	P*ermE*, scbA* (wild-type hairpin)	XbaI/KpnI (P*ermE**), KpnI/SpeI (*scbA* [wild-type hairpin]	Monitoring the effects of the *scbA* hairpin on *scbA* expression (which is driven by the *ermE** promoter.	This work
pGus *scbA* (disrupted hairpin)	P*ermE*, scbA* (disrupted hairpin)	XbaI/KpnI (P*ermE**), KpnI/SpeI (*scbA* [distrupted hairpin])	Monitoring the effects of a disrupted version of the *scbA* hairpin on *scbA* expression (which is driven by the *ermE** promoter).	This work

Table 2.4: Oligonucleotides used in this study

Italics – extra sequence at 5` end to assist with digestion; Bold – restriction site; Underlined – sequence corresponding to target DNA			
Primer name	**Sequence**	**Restriction enzyme site**	**Primer combination(s) and purpose(s)**
16S fwd	AGAGTTTGATCCTGGCTCAG	n/a	DNA contamination check; transcription profiling for 16S rRNA
16S rev	CGAACCTCGCAGATGCCTG	n/a	DNA contamination check; RT and transcription profiling for 16S rRNA
sco4606 in fwd	GTCCCCTTTCTGATCGGTCT	n/a	RT (as_nuoL2M2) and transcription profiling for as_nuoL2M2 and nuoL2/sco4606
sco4606 in rev	ATCGTCGAGTAGGCCAGGAC	n/a	RT (nuoL2/sco4606) and transcription profiling for as_nuoL2/M2 and nuoL2/sco4606
sco4567 in fwd	GATCGCGTACGTCAAGGAGT	n/a	RT (as_nuoEF) and transcription profiling for as_nuoEF and nuoF/sco4567
sco4567 in rev	GTGAGGGTGAGATCGAGTCC	n/a	RT (nuoF/sco4567) and transcription profiling for as_nuoEF and nuoF/sco4567
sco6762 in fwd	CTCGCCTTCTCCTTCCAGC	n/a	RT (as_hopC) and transcription profiling for as_hopC and hopC/sco6762
sco6762 in rev	CGATCCGGTCGAGGAACC	n/a	RT (hopC/sco6762) and transcription profiling for as_hopC and hopC/sco6762
sco3578 5′ in fwd	ATGATGCTCGACATGAAGCG	n/a	Transcription profiling for 5′ end of sco3578
sco3578 5′ in rev	GAGCGATTGGTAGAAGGGGT	n/a	RT and transcription profiling for 5′ end of sco3578
sco3578 3′ in fwd	CGGGAAAGCTGGACTTCGTA	n/a	RT (3′ end wblA/sco3579) and transcription profiling for 3′ end of sco3578 and wblA/sco3579
sco3578 3′ in rev	CGTACGTGTCGGGAGAGTTG	n/a	RT (3′ end sco3578) and transcription profiling for 3′ end of wblA/sco3579 and sco3578
sco3579 5′ in fwd	CTGCTCCCTGAACGAACAGT	n/a	RT and transcription profiling for 5′ end of wlbA/sco3579
sco3579 5′ in rev	GCTGGGTAACCGACTGGAG	n/a	Transcription profiling for 5′ end of wblA/sco3579

		n/a	
Conserved sRNA1	GCCCGGTGAAGGTTGAGAAGAC GATCACGA	n/a	Northern probe to detect scr2634, sar5413, and svr2416
Conserved sRNA 2	GGGGGAGCCGAGTCGGGCAGTT CGGGA	n/a	Northern probe to detect scr5583, sar2652, and svr5279
Conserved sRNA 3	AACGAAGCCCCCGATCACGGGG GAACAA	n/a	Northern probe to detect scr1434, sar6912, and svr1031
sar2765	ACACGCCAGACTCTGCACTCTC CCG	n/a	Northern probe to detect sar2765.
sar3980	TCGCCTCCATTGCCGCGTTCAG ACA	n/a	Northern probe to detect sar3980.
scr3716	TGCCTGGAGGCCTCGAACCGTT CCA	n/a	Northern probe to detect sar3716.
scr3931	TGGAAAGCGAACGGGTGGCCCG GTT	n/a	Northern probe to detect sar3931
svr5535	GCCGCTCCCGGCACCGGGGCCC TCG	n/a	Northern probe to detect svr5535.
sven_2734_sr1	CCAGGGTGCCCGACACGCCGTC CGC	n/a	Northern probe to detect a stable degradation product within the coding region of sven2734.
5S_rRNA	CCCTGCAGTACCATCGGC	n/a	Northern probe to detect 5S rRNA.
PscbN_up	*ATATCC***TCTAGA**GAAGAGGGTC GGGTGACTG	XbaI	Forward primer for cloning *scbN* promoter.
PscbN_dn	*ATATCC***GGTACC**CCACGACGCG CTTCACC	KpnI	Reverse primer for cloning *scbN* promoter.
PscbN&scbN_dn	*ATATCC***GGTACC**ATCGAACAAA CCGCATTGCC	KpnI	Reverse primer to monitor *scbN* termination.
PscbA_fwd	*ATATCC***TCTAGA**TAGAGGGCTC CCTTGGTCAC	XbaI	Forward primer for cloning *scbN* promoter.
PscbA_rev	*ATATCC***GGTACC**TGGCATCGGA CGCAGAATTG	KpnI	Reverse primer for cloning *scbN* promoter.
P*scbR*_dn	*ATATCC***GGTACC**GTCTCCCCCG GGAAGGATAG	KpnI	Reverse primer for cloning *scbR* promoter.
scbA_mRNA_L	*ATATCC***GGTACC**AGAACAGCTC GGCATCACAT	KpnI	Forward primer stating at the transcriptional start site of the *scbA* mRNA.

scbA_mRNA_R3	*ATATCC***ACTAGT**CGCGTCGTGGCAGTCC	SpeI	Reverse primer at the end of the *scbA:scbN* interaction sequence, avoiding the P*scbN* promoter.
scbA_mRNA_R4	*ATATCC***ACTAGT**GAACTCGCCGACAGGACC	SpeI	Reverse primer within the *scbA* mRNA capturing the P*scbN* promoter.
scbA_mRNA_R1	*ATATCC***ACTAGT**CCTTCGGTATCCAGCTGACC	SpeI	*scbA* reverse primer upstream of hairpin.
scbA_mRNA_R1.1_SpeI	*CC***ACTAGT**TCAGCAGCGGATCGTGTC	SpeI	*scbA* reverse primer downstream of hairpin.
scbA_mRNA_L2_KpnI	*CC***GGTACC**CCGGTCAGCTGGATACCGAA	KpnI	*scbA* forward primer downstream of scbR promoter
scbAR_KO_L	tcagtccttcccggtcggtgccagttcggccgcgaggcgATTCCGGGGATCCGTCGACC	N/ap	Forward knockout primer beginning at the stop codon of *scbR*. (lower-case corresponds to *scbR* sequence, upper-case binds to hygromycin cassette)
scbAR_KO_R	tcagccggagaacgcggggccggacagcgtggtgaggaaTGTAGGCTGGAGCTGCTTC	N/ap	Reverse knockout primer beginning at the stop codon of *scbA*. (lower-case corresponds to *scbR* sequence, upper-case binds to hygromycin cassette)
Secondary structure disruption G-Block	Modifications to wild-type *scbA* sequence are described in main text under **Creation of hairpin disruption reporter strains** section.	N / ap	Used as template for hairpin disruption reporter constructs.
pGus_F	AAGCTTGCTCAATCAATCACC	N/ap	Used for checking for inserts in pGus multiple cloning site.
pGus_R	CGTCCAGCTTCTTGATCTCG	N/ap	Used for checking for inserts in pGus multiple cloning site.
gusAR2	TCGATACCGCAGTTCTCC	N/ap	Used for sequencing inserts in pGus.
Cassette_checkR	TGCTTCGGGGTCATTATAGC	N/ap	Used in confirmation gene deletions. Binds within deletion cassete.

Table 2.5: PCR products generated in this work

PCR product name	Primers used	Template	Purpose	Created
P*scbN*	PscbN_up, PscbN_dn	StAH10	*scbN* promoter reporter construct.	This work
P*scbN* (with terminator)	PscbN_up, PscbN&scbN_dn	StAH10	*scbN* termination reporter construct.	This work
P*scbA*	PscbA_fwd, PscbA_rev	StAH10	*scbA* promoter reporter construct	This work
P*scbR*	PscbN_up, PscbR_R	StAH10	scbR promoter reporter construct.	This work
scbA (without *scbN*)	scbA_mRNA_L scbA_mRNA_R3	StAH10	Promoterless fragment of the *scbA* mRNA up to just upstream of the *scbN* promoter. Combined with P*ermE** during cloning.	This work
scbA (with *scbN*)	scbA_mRNA_L scbA_mRNA_R4	StAH10	Promoterless fragment of the *scbA* mRNA including the *scbN* promoter. Combined with P*ermE** during cloning.	This work
scbA (divergent)	scbA_mRNA_L scbA_mRNA R1	StAH10	Monitoring effects of divergent transcription of *scbR* on *scbA*.	
scbA (divergent and wild-type hairpin)	scbA_mRNA_L scbA_mRNA_R1.1	StAH10	Monitoring the combined effects of *scbR* divergent transcription and the *scbA* hairpin on *scbA*.	
scbA (divergent and disrupted hairpin)	scbA_mRNA_L scbA_mRNA_R1.1	2˚ disruption G block	Monitoring the combined effects of *scbR* divergent transcription and a disrupted version of the *scbA* hairpin on *scbA*.	
scbA (wild-type hairpin)	scbA_mRNA_L2 scbA_mRNA_R1.1	StAH10	Monitoring the effects of the *scbA* hairpin on *scbA*.	
scbA (disrupted hairpin)	scbA_mRNA_L2 scbA_mRNA_R1.1	2˚ disruption G block	Monitoring the effects of a disrupted version of the *scbA* hairpin on *scbA*.	
scbAR KO	scbAR_KO_L, scbAR_KO_R	StAH10	Δ*scbAR* strain creation.	This work

Chapter 3: Comparative analysis of non-coding RNAs in the antibiotic-producing *Streptomyces* bacteria

The following chapter was previously published in *BMC Genomics* in May 2013 (reference below). Figures and Tables have been renumbered for the sake of continuity.

Moody, M.J., Young, R.A., Jones, S.E., and Elliot, M.A. (2013) Comparative analysis of non-coding RNAs in the antibiotic-producing Streptomyces bacteria. *BMC Genomics* ***14****:558*

I performed the majority of the experiments and data analysis, and wrote the majority of the manuscript, with the following exceptions. Rachel Young performed RT-PCR experiments shown in Figures 3.3 and 3.4, and wrote the associated captions. Stephanie Jones populated parts of Table 3.4.

Abstract

Background: ncRNAs are key regulatory elements that control a wide range of cellular processes in all bacteria in which they have been studied. Taking advantage of recent technological innovations, we set out to fully explore the ncRNA potential of the multicellular, antibiotic-producing *Streptomyces* bacteria.

Results: Using a comparative RNA-seq analysis of three divergent model streptomycetes (*S. coelicolor, S. avermitilis* and *S. venezuelae*), we discovered hundreds of novel asRNAs and intergenic sRNAs. We identified a ubiquitous antisense RNA species that arose from the overlapping transcription of convergently-oriented genes; we termed these RNA species 'cutoRNAs', for **c**onvergent **u**n**t**ranslated **o**verlapping RNAs. Conservation between different classes of ncRNAs varied greatly, with sRNAs being more conserved than antisense RNAs. Many species-specific ncRNAs, including many distinct cutoRNA pairs, were located within antibiotic biosynthetic clusters, including the actinorhodin, undecylprodigiosin, and coelimycin clusters of *S. coelicolor*, the chloramphenicol cluster of *S. venezuelae*, and the avermectin cluster of *S. avermitilis.*

Conclusions: These findings indicate that ncRNAs, including a novel class of antisense RNA, may exert a previously unrecognized level of regulatory control over antibiotic production in these bacteria. Collectively, this work has dramatically expanded the ncRNA repertoire of three *Streptomyces* species and has established a critical foundation from which to investigate ncRNA function in this medically and industrially important bacterial genus.

Background

Over the last fifteen years, there has been a growing appreciation for the multifaceted roles played by regulatory RNAs in organisms ranging from bacteria to mammals. In bacteria, regulatory ncRNAs come in many forms, and can impact protein function, transcription initiation, mRNA stability and translation initiation/elongation (Waters and Storz, 2009). As described previously, ncRNA transcripts can be broadly

divided into asRNAs and *trans*-encoded sRNAs (Gottesman and Storz, 2010). A notable difference between asRNAs and sRNAs is that asRNAs share complete complementarity with their mRNA targets, whereas the trans-encoded sRNAs have much shorter complementary regions, and different sequences within a sRNA may bind different mRNA targets. ncRNA-mediated regulation has been implicated in a multitude of cellular processes, including stress responses (Hobbs et al., 2010), quorum sensing (Bejerano-Sagie and Xavier, 2007) and pathogenicity (Toledo-Arana et al., 2007).

The ncRNA potential of bacteria has been explored most thoroughly in *E. coli* (Argaman et al., 2001; Hu et al., 2006; Lybecker et al., 2014; Raghavan et al., 2011; Wassarman et al., 2001; Zhang et al., 2003), but in recent years, technological advances in the form of tiling microarrays (Toledo-Arana et al., 2009) and RNA-seq (Attaiech et al., 2016; Cohen et al., 2016; Irnov et al., 2010; Sharma et al., 2010; Sittka et al., 2009) have begun to reveal the extent - and the complexity - of ncRNAs in a wide range of bacteria.

The non-coding RNA capacity of *Streptomyces* bacteria is expected to be extensive. The streptomycetes are predominantly soil-dwelling bacteria, and as such must have the means of coping with diverse environmental stresses. They also have a large chromosome (>8 Mb), and a complex life cycle that involves progression through distinct developmental and metabolic stages - processes that are subject to multi-level regulation. Most streptomycetes grow vegetatively in liquid culture, although several species including *S. venezuelae*, sporulate under these conditions. Along with their morphological complexity, the streptomycetes are best known for their ability to produce a vast array of specialized metabolites having medical and agricultural importance, including the majority of naturally synthesized antibiotics. Specialized metabolism is coordinately regulated with development, initiating during the transition from vegetative to aerial growth (or vegetative to 'mycelial fragmentation', for those species that sporulate in liquid culture); in liquid culture, specialized metabolism initiates during entry into stationary phase (Flärdh and Buttner, 2009) for the majority of (non-sporulating) streptomycetes.

We were interested in exploring the ncRNA potential of *Streptomyces* bacteria throughout the course of their developmental and metabolic cycles. A series of initial investigations had confirmed the existence of ncRNAs in these bacteria (D'Alia et al., 2010; Pánek et al., 2008; Swiercz et al., 2008; Tezuka et al., 2009), and this ncRNA repertoire was expanded considerably by an early RNA sequencing study undertaken by Suess and colleagues (Vockenhuber et al., 2011), who identified many asRNAs and sRNAs in the model species *S. coelicolor*. This pioneering study focused on RNA expression at a single time point during *S. coelicolor* growth in liquid culture. To gain a more comprehensive view of the ncRNA potential of *Streptomyces* bacteria, we undertook a comparative genomics investigation into the transcriptomes of three evolutionarily divergent *Streptomyces* species (Ludwig et al., 2012) – *S. coelicolor*, *S. avermitilis*, and *S. venezuelae* - using RNA harvested at distinct metabolic and developmental stages. *S. coelicolor* and *S. venezuelae* represent classic and emerging model species, respectively, while *S. avermitilis* has been well studied in part due to its production of avermectin, a commercially important insecticidal and anti-parasitic

compound. We identified dozens of new conserved sRNAs and asRNAs, including a distinct group of asRNAs termed 'cutoRNAs' that resulted from overlap of the 3′ ends of convergently transcribed mRNAs (**Figure 3.1**). We also detected an abundance of unique ncRNAs, including many that featured prominently in specialized metabolic biosynthetic clusters.

Materials and Methods

Streptomyces growth and RNA isolation

Streptomyces strains were grown on cellophane disks on the surface of solid MYM agar medium (*S. coelicolor* M145 and *S. avermitilis* MA-4680), or shaken in flasks containing liquid MYM (*S. venezuelae* ATCC 10712) at 30°C. Cells were harvested at timepoints corresponding to vegetative growth, aerial hyphae formation (or mycelial fragmentation in the case of *S. venezuelae*) and sporulation. RNA was extracted and its quality was assessed as outlined in Chapter 2. PCR amplification of a 196-nucleotide region of the 16S rRNA gene was used to confirm the absence of DNA (lack of any product relative to a chromosomal DNA control) prior to RNA-seq library creation.

Library preparation and RNA sequencing

For each species, RNA samples from each of three timepoints were divided in two, with one half being subjected to rRNA depletion using the MICROBExpressTM Bacterial mRNA Enrichment Kit (Ambion) as per the manufacturer's instructions. Each of the rRNA-depleted samples was then combined to generate a species-specific 'pool,' such that each pool contained equal amounts of RNA from each time-point (vegetative growth, aerial growth/mycelial fragmentation, sporulation). Equivalent pools were created for the untreated (total) RNA samples. The two pools were destined to become two separate libraries: the rRNA-depleted pools were sequenced using a protocol optimized for full length transcripts, while the total RNA samples were enriched for sRNAs prior to sequencing.

All samples were treated with tobacco acid pyrophosphatase to create 5′ ends amenable for adapter ligation. To enrich for small RNAs, the total RNA samples were size selected (40 to 300 nucleotides) following polyacrylamide gel electrophoresis. RNA from both libraries was then fragmented in a buffered zinc solution and single stranded RNA adapters were ligated to the 5′ and 3′ ends to maintain strand specificity, prior to re-purifying on a polyacrylamide gel. Each pool was then reverse transcribed and PCR amplified to generate DNA colonies, which were then sequenced using an Illumina HiSeq 2000 sequencer. For the sRNA-enriched library, read lengths ranged from 24 to 94 nucleotides, while read lengths were as long as 150 nucleotides for the long transcript-enriched library.

Alignment of reads to genomes

Sequencing reads having low quality 3′ ends were trimmed using the program PrinSeq (Schmieder and Edwards, 2011). The quality trimmed reads were aligned to their respective genome sequence using Bowtie2 (Langmead et al., 2009) and then sorted, indexed and converted to BAM format using SAMtools (Version 0.1.18) (Li et al., 2009).

For the full length transcript libraries, 59,073,931 (~99.4% of total reads obtained during sequencing), 44,462,362 (~99.4%) and 94,358,187 reads (~99.4%) were mapped to the *S. coelicolor*, *S. avermitilis*, and *S. venezuelae* genomes respectively. For the sRNA-enriched libraries a total of 21,871,239 (~99.5%), 23,608,152 (~98.5%) and 21,880,716 (~97.6%) reads were mapped to the same respective genomes. The BAM files were visualized using Integrated Genomics Viewer (Version 2.0) (Robinson et al., 2011). BEDTools (Quinlan and Hall, 2010) was used to calculate the positive and negative-strand read depth at each nucleotide, and a Perl script was used to exploit the BEDTools output in calculating the average sense and antisense coverage per nucleotide of each annotated gene.

Non-coding RNA analysis

The genes with the highest levels of antisense coverage from the full length transcript libraries were chosen for analysis. These included genes with an antisense 'mean expression value' (MEV) (average read depth per nucleotide) greater than 10.0 in *S. coelicolor*, 3.5 in *S. avermitilis*, and 4.1 in *S. venezuelae*, where the different values were determined by normalizing the MEV to the number of non-ribosomal RNA sequences obtained for each species. Given that many known asRNAs overlap only a small fraction of their sense counterpart, we also used BEDTools to determine the highest peak antisense expression levels for each gene. We focused on those genes having a maximum expression level greater than 20.0 in *S. coelicolor*, 6.9 in *S. avermitilis*, and 8.2 in *S. venezuelae*. Again, cutoff values were chosen to reflect differences in numbers of non-ribosomal RNA sequences obtained for each species. Intergenic sRNAs were annotated manually using Integrated Genomics Viewer (Robinson et al., 2011).

Homologous sRNAs were identified using BLASTN (E-value less than 1e-06) and aligned using Clustal Omega (Goujon et al., 2010; Sievers et al., 2011). Selected alignments were computationally folded using CMFinder (Yao et al., 2006) on the WAR webserver (Torarinsson and Lindgreen, 2008). Compensatory mutations that maintained secondary structure were located manually and the VARNA software package (Darty et al., 2009) was used to illustrate RNA secondary structure. RNA secondary structure predictions were performed using the program Mfold (Zuker, 2003).

To evaluate the protein-coding potential of sRNAs, we focused on those sRNAs conserved in at least two of the three *Streptomyces* species investigated in this study. Frame Plot 2.3.2 was used to highlight potential open-reading frames, and amino acid sequences of similar lengths (≤10% difference) were aligned using Clustal Omega (Goujon et al., 2010; Sievers et al., 2011). Amino acid sequences with high similarity (≥65%) were deemed to have the potential to encode a conserved small peptide.

Analysis of stable degradation products

A custom Perl script was used to identify mRNAs with defined regions of high coverage compared with the average coverage for the entire gene in the short-read library. The top hits from this analysis were visualized using Integrated Genomics Viewer, and putative stable regions that were grossly overrepresented (more than 100-fold) compared with the rest of the mRNA were identified. These stable regions were classified according

to their location: 5′ end-associated; 3′ end-associated (or between genes in an operon); or internal to the mRNA coding sequence. A custom Perl script was also developed to analyze the nucleotide content for entire genes, stable regions, and sequences flanking the stable regions.

Semi-quantitative RT-PCR

Semi-quantitative RT-PCR was conducted using total RNA isolated from *S. coelicolor* at three distinct life-cycle stages. For each reverse transcription reaction, 3 μg of RNA was mixed with 10 nmol of each dNTP, and 2 pmol of each gene-specific primer (**Tables 2.4** and **2.5**). RNase-free water was added to give a total reaction volume of 12 μL. Following mixing, each sample was first incubated at 65°C for 10 min and then immediately chilled on ice for 5 min. Reverse transcription was performed using SuperScript® III reverse transcriptase (Invitrogen) according to the manufacture's instructions with a few modifications. Briefly, 4 μL of 5× First Strand Buffer, 2 μL of 0.1 M DTT and 1 μL of RNaseOUT^{TM} were added to each reaction. After incubation at 42°C for 2 min, 1 μL of SuperScript® III reverse transcriptase was added. Reverse transcription was performed at 42°C for 60 min and reactions were terminated by incubating at 70°C for 15 min.

The reverse transcription products (2 μL) were then used as template for PCR amplification. A standard PCR protocol using Taq DNA polymerase (Norgen) was used, with the primers indicated in (**Table 2.4** and **2.5**). Annealing temperatures were optimized for each primer combination, as were the number of amplifications cycles (to ensure that amplification remained within the linear range). PCR products were separated on 2-3% agarose gels. Negative controls containing nuclease free water in lieu of reverse transcriptase were included to ensure there was no residual genomic DNA present in the RNA samples. Primers targeting 16S rRNA (**Tables 2.4 and 2.5**) were used as positive controls for RNA quality. All reverse transcription/PCR amplifications were carried out in triplicate, using RNA isolated from three independent RNA time-courses.

Results and Discussion

To probe the ncRNA potential of *S. avermitilis*, *S. coelicolor* and *S. venezuelae*, we performed RNA-seq using species-specific RNA pools. Each species was grown on MYM medium, so as to effectively compare their RNA profiles, with the only difference being that *S. avermitilis* and *S. coelicolor* were grown on MYM agar, while *S. venezuelae*, which sporulates in liquid culture, was grown in liquid MYM. For each species, RNA was isolated from cells at major developmental stages (vegetative; aerial hyphae/fragmentation (in the case of *S. venezuelae*); spores). The RNA samples for each species were then pooled and used to generate two libraries for sequencing: one enriched for full-length transcripts, and a second enriched for shorter transcripts (*e.g.* sRNAs and stable RNA degradation products).

asRNAs are abundant in the streptomycetes, and are largely species-specific

Previous RNA-seq analyses in diverse bacterial species have revealed extensive asRNA expression (Albrecht et al., 2010; Sharma et al., 2010). Consistent with these

observations, we detected abundant asRNAs in all three *Streptomyces* species: 680, 592 and 536 asRNAs were identified in *S. coelicolor*, *S. venezuelae* and *S. avermitilis*, respectively. These asRNAs could be further subdivided into three categories, on the basis of their coverage profiles and their genomic context: (i) asRNAs expressed from a dedicated promoter within a protein-coding gene (referred to here as simply 'asRNAs'); (ii) asRNAs that arose from the overlap of 3′ untranslated regions (UTRs) from convergently oriented genes - an RNA species that we have termed 'cutoRNA' (see below); and (iii) asRNAs that resulted from divergent transcription, where promoters of divergently expressed genes overlapped (**Table 3.1**).

asRNAs expressed on the strand opposite that of a protein-encoding gene [class (i)], did not comprise a majority of the asRNAs identified here, with fewer than 100 identified in any of the three species. As has been observed for comparative analyses conducted in other bacteria (Raghavan et al., 2012), the majority of the 99 (*S. coelicolor*), 59 (*S. avermitilis*), and 79 (*S. venezuelae*) asRNAs identified were species specific (**Table 3.1**). We considered the possibility that this species specificity resulted from asRNA association with coding sequences confined to a single species. This turned out not to be the case: 129 broadly conserved genes (genes with homologues in all three species) were associated with asRNAs in at least one species, but only 11 (or 8.5%) of these genes exhibited antisense expression in all three species (**Table 3.2**). This level of asRNA conservation is slightly less than that reported for *E. coli* and *Salmonella*, where ~14% of antisense transcripts were conserved between species (Raghavan et al., 2012). Within bacteria, the regulatory impact of apparently unique asRNAs encoded opposite conserved open reading frames remains to be elucidated.

Of the 11 conserved asRNAs we identified, the most striking was found opposite the *nuo* gene cluster. The *nuo* genes direct the expression of NADH:quinone oxidoreductase, an enzyme complex found in archaea, bacteria, and within eukaryotic mitochondria and chloroplasts (Moparthi and Hägerhäll, 2011). This multi-protein complex, also known as complex I, is a key player in the respiratory transport chain (Efremov et al., 2010). Many bacteria encode a 14-subunit (NuoA-N) version of complex I; however, some groups have retained an ancestral 11-subunit form that lacks the 'N-module' subunits NuoE, NuoF and NuoG, while others have a 12-membered complex lacking only NuoE and NuoF (Weerakoon and Olson, 2008). It is within the N-module-encoding region that we identified one of the most highly expressed and conserved asRNAs. Transcription of the asRNA began within the coding region of *nuoF* and continued through the coding region of *nuoE* (**Figure 3.2A**), with the asRNA extending for up to 1,600 nucleotides in *S. coelicolor* and *S. avermitilis*; a shorter asRNA was observed in *S. venezuelae*. An intriguing possibility is that the asRNA provides a checkpoint in complex I assembly, down-regulating the expression of N-module-encoding genes until the rest of the complex has been synthesized/assembled. In order for such regulation to occur, both sense and antisense transcripts would need to be coordinately expressed. To test this, we conducted semi-quantitative RT-PCR experiments, and found that both sense and antisense transcripts were expressed at the same time (**Figure 3.3A**), supporting a possible regulatory role for this asRNA.

Streptomyces species also possess an additional copy of many of the complex I genes (*nuoA2*, *B2*, *D2*, and *H2* to *N2*) encoded from a disparate chromosomal location. Like the standard *nuo* gene cluster, these genes are organized contiguously (with the exception of *nuoD2*) and our data suggest that they are expressed as a single operon. While this second cluster lacked the N-module-encoding genes, it was associated with a second conserved asRNA extending from *nuoM2* to *nuoL2*. (**Figure 3.3B**). Both *nuoM2* and *nuoL2* encode antiporter-like proteins (Mathiesen and Hägerhäll, 2002). In the cyanobacterium *Synechocystis*, different antiporter subunits can be incorporated into complex I for different tasks related to photosynthesis (Battchikova et al., 2011; Moparthi and Hägerhäll, 2011; Peng et al., 2011). The presence of additional *nuoL* and *nuoM* genes in *Streptomyces* genomes means there is the potential for analogous differential incorporation of these gene products into complex I, and this incorporation could be controlled by conserved asRNA activity. As for the *nuoEF*-associated asRNA, semi-quantitative RT-PCR revealed similar expression patterns for both *nuoL2* and its cognate asRNA, although the latter appeared to be expressed at lower levels relative to the mRNA (**Figure 3.3B**). Intriguingly, an asRNA has been reported opposite the *nuoM* homologue in rat mitochondria (Lung et al., 2006), raising the possibility that this asRNA arose before the evolution of eukaryotes, over 2-billion years ago.

In addition to the conserved asRNAs associated with the *nuo* gene clusters, we also identified conserved asRNAs associated with the genes targeted by the lysogenic phage φBT1 (Gregory et al., 2003) and φC31 (Kuhstoss and Rao, 1991), although for the φC31-targeted genes, only the *S. venezuelae*-associated sequence met the relatively stringent cutoff we used in assigning asRNA designations. φBT1 integrates into the coding sequence of an integral membrane protein-encoding gene (*sco4848*/*sav3412*/*sven4521*), while φC31 targets the coding sequence of a conserved pirin-like protein (*sco3798*/*sav4392*/*sven3566*). An asRNA encompassed the φBT1 integration site in all three *Streptomyces* species (**Figure 3.2B**), while for the φC31-associated genes, the asRNA was found immediately adjacent to the phage integration site (**Figure 3.2B**). There are a number of intriguing functional possibilities that could be ascribed to these asRNAs. They may simply act to control their associated protein coding genes, or they may contribute to a novel phage resistance mechanism, perhaps minimizing phage integration by sequestering these regions into transcriptionally active complexes. Alternatively, phage integration at these sites may be the result of positive selective pressure, as φC31, and presumably φBT1, integrate in an antisense orientation such that the integrase promoter is separated from its coding sequence (Gregory et al., 2003; Kuhstoss and Rao, 1991). As integrase activity is required for phage excision, a productive infection could only be achieved with the assistance of an endogenous (bacterial) promoter. For φBT1, such a promoter could obviously be provided by the asRNA (**Figure 3.2B**); for φC31, the asRNA promoter may well lie upstream of the integration site, but asRNA transcript levels were more abundant downstream of this region (**Figure 3.2B**).

'cutoRNAs' are a common and well-conserved phenomenon in *Streptomyces* species

In addition to the class (i) asRNAs, we also identified a second major class of asRNAs in all three *Streptomyces* species, termed 'cutoRNAs,' for **c**onvergent **u**n**t**ranslated **o**verlapping RNAs. These RNAs arose from the expression of convergent genes, whereby the transcription of one or both genes extended beyond its respective coding sequence into the downstream coding regions (**Table 3.1**). Whilst we identified only 11 conserved asRNAs, there were 19 cutoRNA pairs conserved in *S. avermitilis*, *S. coelicolor* and *S. venezuelae* (**Table 3.3**).

We examined the genetic organization of these 19 gene pairs in other streptomycetes, and found this organization to be highly conserved. For example, in *Streptomyces scabies* and *S. griseus*, a convergent configuration was observed for 19/19 (*S. scabies*) and 18/19 (*S. griseus*) gene pairs. We extended our analyses to include more diverse actinobacteria, but found many of the genes involved were *Streptomyces*-specific; only the *wblA-sco3578* gene pair was conserved and convergently arranged in the more distantly related *Frankia alni*, *Thermobifida fusca*, and *Mycobacterium tuberculosis*. In *M. tuberculosis*, 'antisense RNAs' to both genes have been previously reported (Arnvig et al., 2011), suggesting broad cutoRNA conservation across the actinobacteria for this gene pair.

Given the extent of its conservation, we sought to further investigate the expression of the *wblA* and *sco3578* cutoRNA. *wblA* encodes a transcription factor that impacts both antibiotic production and aerial morphogenesis in *S. coelicolor* (Fowler-Goldsworthy et al., 2011), while *sco3578* encodes a putative ion-transporting ATPase. Our RNA-seq data revealed that the 3′ UTR of *wblA* covered the entire coding region of the downstream ATPase-encoding gene in both *S. coelicolor* (**Figure 3.2C**) and *S. avermitilis* (**Figure 3.2C**), extending more than 1.2 kb beyond the *wblA* translation stop site. In *S. venezuelae*, *wblA* transcripts extended ~500 nucleotides beyond the *wblA* coding sequence, well into the downstream coding sequence (**Figure 3.2C**). While the ATPase-encoding gene was expressed at much lower levels than *wblA*, its 3′ UTR still extended into *wblA*. Semi-quantitative RT-PCR analyses were conducted to follow the expression of these genes. We found each gene and its corresponding 3′ UTR, was expressed throughout development (**Figure 3.4**). This suggested that, as for the asRNAs examined here, there is the potential for base pairing of these convergent transcripts, with possible downstream regulatory implications.

Outside of the *wblA*-associated cutoRNA, *M. tuberculosis* has previously been shown to have abundant asRNAs arising from the transcriptional read-through of convergently transcribed genes (Arnvig et al., 2011). A similar phenomenon has also been noted in the more distantly-related (Gram-positive) bacterium *Bacillus subtilis* (Nicolas et al., 2012), suggesting that cutoRNAs may be widespread in bacteria. Studies in *B. subtilis* have also revealed intriguing correlations between flexible transcription termination and growth conditions (Nicolas et al., 2012). In archaea, transcription of many genes (>30%) is controlled by multiple terminators, giving rise to different isoforms of mRNAs that may have regulatory consequences (Dar et al., 2016). It will be interesting to see whether cutoRNA occurrence in the streptomycetes is similarly impacted by different growth conditions.

There are a number of different scenarios by which cutoRNAs could function in the cell. Simultaneous expression of cutoRNA gene pairs could lead to altered stability of one or both transcripts. This is supported by an analysis of recently published data comparing gene expression in wild-type and RNase III deficient strains of *S. coelicolor* (Gatewood et al., 2012) (where RNase III specifically cleaves double stranded RNA), which revealed that one gene in each of seven different cutoRNA pairs was significantly impacted by the loss of RNase III (*SCO1150, SCO4283, SCO4749, SCO5106, SCO5146, SCO6716, SCO6729*; **Table 3.1**). Alternatively, cutoRNAs could serve to 'tether' the convergently expressed mRNAs such that their protein products are produced in close proximity. This would imply a functional correlation between the convergent genes and their resulting products. Currently, there is no experimental evidence supporting related functions for any of the conserved cutoRNA gene pairs, as the majority of these genes have not been characterized. It is worth noting, however, that cutoRNAs were abundant in the species-specific specialized metabolic gene clusters, where they were shared between genes with obvious functional relationships (**Table 3.1** and below).

In *E. coli*, cutoRNA-like transcription is thought to be deleterious, and it has been proposed that the Rho transcription termination factor acts to prevent such asRNA expression (Peters et al., 2012). Rho activity can be inhibited by the antibiotic bicyclomycin, and studies in a close relative of *S. coelicolor*, *Streptomyces lividans*, have revealed that bicyclomycin has no effect on colony growth (Ingham et al., 1996), suggesting that the loss of Rho activity is not detrimental to the streptomycetes. This may imply that *Streptomyces* tolerate convergent transcription better than *E. coli*, or it may mean that they invoke other, as yet unknown, means of dealing with transcriptional conflicts caused by convergent transcription.

Of the remaining asRNAs identified, very few were the result of divergent expression from overlapping promoters: only five were observed in *S. coelicolor*, while none were detected in *S. avermitilis* or *S. venezuelae*. (**Table 3.1**). Instead, much of the antisense transcription we detected could not be readily categorized (**Table 3.1**). This was largely due to the lack of defined transcription start/stop sites and uneven transcript coverage, which made definitive classification challenging. It is conceivable that many of these transcripts were processed shortly after generation, possibly in conjunction with their corresponding sense transcripts, and consequently full-length asRNAs failed to accumulate. The idea that rapid processing masks the full extent of antisense transcription has been supported by findings in *S. aureus*, where full length asRNAs were detected only following RNase III depletion (Lasa et al., 2011). The number of genes with associated asRNAs in *Streptomyces* may therefore be much higher than reported here.

Expanding the *Streptomyces* sRNA landscape: conservation and organization of new sRNAs

To expand the existing library of sRNAs in *S. coelicolor*, and to begin to understand the distribution of sRNAs in different *Streptomyces* species, we endeavoured to mine our RNA-seq data for unannotated sRNA genes within the intergenic regions of *S. coelicolor*, *S. avermitilis* and *S. venezuelae* (**Figure 3.5A**). New sRNAs were given a designation that consisted of a species reference, followed by a number corresponding to

that of its right flanking protein-coding gene (*e.g. scr1434*/*sar6912*/*svr1031* for *S. coelicolor*, *S. avermitilis* and *S. venezuelae* sRNAs, respectively) (**Table 3.4**). We identified 90 sRNAs in *S. coelicolor*, of which 71 were novel, bringing the total number of confirmed sRNAs in *S. coelicolor* to 105. Interestingly, we detected greater numbers of sRNAs in *S. avermitilis* and *S. venezuelae*: 199 and 176, respectively, of which fewer than 20 in each species were homologous to previously identified sRNAs from *S. coelicolor*. We also observed 17 of 34 previously confirmed sRNAs from *S. coelicolor* (Pánek et al., 2008; Swiercz et al., 2008; Vockenhuber et al., 2011), along with another four that had been predicted but not experimentally validated (Pánek et al., 2008; Swiercz et al., 2008) (**Table 3.4**). An additional 12 previously confirmed/predicted sRNAs appeared, from our data, to be highly expressed 5′ UTRs and not independently encoded sRNAs. This did not, however, preclude these regions from having sRNA regulatory potential, as there are documented examples of functional sRNAs arising from UTR fragments (Chao et al., 2012; Gössringer and Hartmann, 2012; Loh et al., 2009).

Unlike the asRNAs, we found a significant number of intergenic sRNAs were conserved between the three species (**Figure 3.5A**; **Table 3.5**). Of the 92 sRNAs we identified in *S. coelicolor*, 28.7% were conserved at a sequence level (E-value less than 1e-06) in all three species, while 22.3% and 2.2% were shared with *S. avermitilis* or *S. venezuelae*, respectively. We considered the possibility that some of these conserved sRNA genes may - in addition, or alternatively – encode small proteins, as has been seen in *E. coli* (Hobbs et al., 2011). We scrutinized all conserved sRNA sequences for open reading frames that were also conserved between species, and found four of 58 with the potential to encode a conserved protein (**Table 3.5**). Further experimentation will be needed to assess the protein-coding capacity of these four genes.

Here, we directed our efforts towards the initial characterization of a number of highly expressed, non-protein-coding novel sRNAs. Using northern blotting, we probed the expression of three conserved sRNAs to verify our RNA-Seq data and to investigate their expression profiles. One of the most highly-expressed conserved sRNA had two equivalently expressed paralogues in *S. coelicolor* (*scr2634, scr0999*) (**Figure 3.5B**). In *S. avermitilis* and *S. venezuelae*, the equivalent sRNAs (*sar5413* and *svr2416*, respectively) were also highly expressed (**Figure 3.5B**). Structural predictions suggested that these sRNAs adopted near identical structures (**Figure 3.5B**), being largely unaffected by primary sequence differences. In each species, the sRNA was expressed from a site immediately downstream of *sodF* (within 18 nucleotides), where *sodF* encodes an iron/zinc superoxide dismutase involved in the defense against reactive oxygen species. While *sodF*-associated sRNAs have not been reported previously, sRNAs encoded within the 3′ regions of protein-coding genes are not unprecedented and have been described recently in *Salmonella* (Chao et al., 2012; Gössringer and Hartmann, 2012; Kawano et al., 2005). There is, however, evidence for control of *sodF*-like genes by small RNAs: expression of the *sodF* equivalent in *E. coli*, *sodB*, is controlled by the RyhB sRNA (Massé and Gottesman, 2002); we do not currently have any data supporting a regulatory connection between *sodF* and the associated downstream sRNA. Northern blot analysis revealed that this *sodF*-associated sRNA was expressed throughout development in all three *Streptomyces* species (**Figure 3.5B**). Subsequent to our publication of this

work it was discovered that this *sodF* sRNA is indeed processed from the 3′ UTR of the *sodF* mRNA and binds near the RBS of *sodN*, repressing its translation. Hence, repression of *sodF* results in lower amounts of the *sodF* sRNA and increased expression of *sodN*, explaining the inverse correlation in the expression of these two genes (Kim et al., 2014b).

We probed an additional conserved sRNA that was amongst the most highly expressed in all three species. *scr5583*, *sar2652*, and *svr5279* shared extensive sequence identity, and were predicted to have a structurally distinctive C-rich (67%) terminal loop (**Figure 3.5C**). Many well-characterized sRNAs, such as RNAIII in *S. aureus* (Geisinger et al., 2006), target mRNAs via C-rich loops; however, *Streptomyces* genomes are very GC-rich (>70%), so whether an equivalent phenomenon exists in these bacteria remains to be seen. Unexpectedly, northern blot analyses revealed that this sRNA was differentially expressed in each of the three *Streptomyces* species: it was expressed most highly during aerial hyphae formation and sporulation (later developmental stages) in *S. coelicolor* and *S. avermitilis*, whereas in *S. venezuelae*, it was most highly expressed during vegetative (early) growth (**Figure 3.5B**).

Finally, we examined the expression profiles of the highly expressed *scr1434*, *sar6912*, and *svr1031* sRNAs. Highest levels of each, as determined by northern blotting, were observed during aerial hyphae formation and sporulation (**Figure 3.5D**). This sRNA was predicted to form a very stable stem-loop structure, again, having a C-rich loop region (**Figure 3.5D**).

While many sRNAs were shared by all three *Streptomyces* species, there were notable species-specific differences as well. We focused our attention on select highly expressed unique sRNAs, and used northern blot analysis to assess their expression profiles (**Figure 3.6**). Within *S. avermitilis*, the 89 nucleotide *sar2765* was expressed exclusively during vegetative growth (**Figure 3.6**), while the equivalently sized *sar3980* (88 nucleotides) was expressed most highly during vegetative and aerial growth (**Figure 3.6A**). In *S. coelicolor*, *scr3716* (~128 nucleotides) was highly represented in our long transcript-enriched library and was not present in the sRNA-enriched library, unlike the majority of sRNAs identified in our study (this is in contrast to all classes of asRNA, which were almost exclusively detected in our long transcript library). *scr3716* was expressed at low levels during vegetative growth, with expression levels rising significantly during aerial development and sporulation (**Figure 3.6B**), in contrast to the smaller 70 nucleotide *scr3931*, which was expressed solely during vegetative growth (**Figure 3.6B**). In *S. venezuelae*, svr5535 was one of the shortest sRNAs identified in our study at only 41 nucleotides, and unlike many other sRNAs, it was expressed throughout development (**Figure 3.6C**). Apart from *svr5535*, which was predicted to form a single stem-loop structure, all other sRNAs were predicted to adopt two or three stem-loop configurations.

In considering species-specific versus conserved sRNAs, we explored whether any correlation could be drawn between conservation and genome position. *Streptomyces* chromosomes are linear, with a central ‘core’ region that is broadly conserved in all actinobacteria. This central core is flanked on either side by ‘arm’ regions whose sequences are more divergent. Comparative genomic analyses have suggested that the left

arm contains an actinomycete-specific region immediately adjacent to the core, while the equivalent position in the right arm is associated with *Streptomyces*-specific genes. The extreme ends of the chromosome arms contain predominantly species-specific genes (Kirby, 2011). We examined the position of each sRNA in *S. coelicolor* in relation to these different genetic bounds (**Table 3.6**). The majority of sRNAs (58 of 92) fell within the core region, with 50% of these conserved in at least one of the other two *Streptomyces* species. Of the 17 sRNAs located in the 'actinomycete-specific' region, a remarkable 82% were conserved, whereas somewhat surprisingly, only eight sRNAs were expressed from within the '*Streptomyces*-specific' region, and of these, only three were also found in *S. avermitilis* or *S. venezuelae*. In the divergent chromosomal ends, few sRNAs were identified, and all of these were unique to *S. coelicolor*.

In general, the 105 sRNAs identified here and elsewhere (Pánek et al., 2008; Swiercz et al., 2008; Vockenhuber et al., 2011), for *S. coelicolor* is comparable to the number of sRNAs detected in *E. coli* [currently estimated to be ~80 (Raghavan et al., 2011)]. This is fewer than might have been expected given the large *Streptomyces* genome (>8 Mb versus 4–5 Mb for *E. coli*), and the relatively large proportion of protein-encoding genes dedicated to regulation in *S. coelicolor* (12.3% of all protein-coding genes (Bentley et al., 2002)). It is likely, however, that sRNA saturation has not been reached in any *Streptomyces* species, given that there has yet to be an exhaustive search conducted using different growth and stress conditions, and that each investigation undertaken to date has identified unique sRNA subsets without considerable overlap. Also, it is no longer as straight forward to ascribe 'sRNA gene' designations to non-coding RNAs, as many are derived from mRNAs, tRNAs, and UTRs (DebRoy et al., 2014; Lalaouna et al., 2015; Vanderpool et al., 2011), and many sRNAs also encode small proteins (Hobbs et al., 2011).

ncRNAs feature prominently in many specialized metabolite clusters

Streptomyces species are renowned for their ability to produce a broad range of antibiotics, together with a host of other specialized metabolites having medical and agricultural utility. Our transcriptome analyses have revealed previously unrecognized complexity for some specialized metabolic clusters, largely in the form of asRNA expression.

asRNAs were abundant in the predicted specialized metabolic clusters for the three *Streptomyces* species examined here: 20% of *S. avermitilis*, 30% of *S. coelicolor* and 60% of *S. venezuelae* specialized metabolic clusters were associated with asRNAs of at least one type (**Table 3.1**). Given the lack of general asRNA conservation found both within the streptomycetes in this study, and in other bacteria (Toffano-Nioche et al., 2012), we were surprised to identify a strongly-expressed cis-asRNA within a hopanoid biosynthetic cluster in *S. coelicolor* and *S. avermitilis* (**Figure 3.7A**). Hopanoids are cholesterol-like pentacyclic molecules (Ourisson et al., 1987; Rohmer et al., 1979; Sáenz et al., 2012) found throughout bacteria (Sahm et al., 1993). Hopanoids recently have been shown to provide order to lipids in the outer membrane of *Methylobacterium extorquens* by directly interacting with glycolipids. In this species, disrupting hopanoid synthesis

worsens drug export, further suggesting a role in preventing membrane disorder (Sáenz et al., 2015).

In *S. coelicolor*, the 12 gene hopanoid biosynthetic cluster is most highly expressed during aerial development, and it has been proposed that hopanoids help promote water retention during aerial hyphae formation (Poralla et al., 2000). This may explain why the equivalent cluster in *S. venezuelae* (grown in liquid culture) was expressed at very low levels. The asRNA was transcribed opposite *hopC* (*sco6762*) (**Figure 3.7A**), a predicted phytoene dehydrogenase-encoding gene. Using semi-quantitative RT-PCR, we determined that both sense and antisense genes were expressed at the same time (**Figure 3.3**). The hopanoid cluster in *S. coelicolor* is thought to direct the synthesis of both hopane and the related aminotrihydroxybacteriohopane (Poralla et al., 2000). Little is known about the biosynthetic steps leading to the synthesis of these compounds, and nothing is known about the role of HopC. It is possible that *hopC* expression may be modulated by its cognate asRNA, which in turn could impact the production of these products.

Two well-characterized specialized metabolic clusters in *S. coelicolor* also encoded distinct asRNAs: the coelimycin P1 (cpk) biosynthetic cluster (**Figure 3.7B,C**) and the prodiginine (red) biosynthetic cluster (**Figure 3.7D**). The 16 gene coelimycin P1 biosynthetic cluster (*sco6273-6288*) (Gomez-Escribano et al., 2012; Gottelt et al., 2010; Pawlik et al., 2010) includes two genes with associated asRNAs: *cpkE*/*sco6277* (encoding a putative epoxide hydrolase) and *cpkH*/*sco6281* (encoding a putative FAD-binding protein). The *cpkE*-associated asRNA was expressed most highly in the centre of *cpkE*, while the *cpkH* antisense was expressed closer to the 3′ end of the coding sequence (**Figure 3.7B**). The roles of CpkE and CpkH in coelimycin P1 biosynthesis have yet to be elucidated. It is worth noting that *cpkE* is expressed as part of a larger operon (*cpkD-G*), and that the expression of this entire operon was increased by more than two-fold in an RNase III mutant strain (Gatewood et al., 2012), suggesting that the *cpkE* asRNA may function to destabilize its cognate polycistronic mRNA in an RNase III-dependent manner. In contrast, *cpkH* expression was not enhanced following the loss of RNase III, although transcript levels for both upstream (*cpkO*) and downstream flanking genes (*cpkI-K*) were increased (Gatewood et al., 2012), suggesting complex post-transcriptional dynamics in this area.

Regulators of *cpk* gene expression (ScbA and ScbR) also appeared to be subject to asRNA regulation (**Figure 3.7C**). ScbA directs the synthesis of the γ-butyrolactone quorum sensing molecule SCB1, which is sensed by ScbR - an SCB1 receptor and DNA binding transcription factor that represses *cpk* cluster expression in the absence of SCB1 (Takano et al., 2005). It has been previously reported that the promoters - and thus 5′ UTRs - of the divergently transcribed *scbR* and *scbA* genes overlap (Chatterjee et al., 2011; Mehra et al., 2008) (**Table 3.1**). In addition to the asRNAs resulting from this 5′ overlap, we also found there was a highly expressed asRNA within *scbA* (**Figure 3.7C**), adding an additional layer of regulation to an already transcriptionally intricate region.

The prodiginine cluster spans 22 genes and yields two major products: undecylprodigiosin and butyl-meta-cycloheptylprodigionine. RedG, a Rieske oxygenase-like enzyme, is thought to catalyze the conversion of undecylprodigiosin to butyl-meta-

cycloheptylprodigionine (Cerdeño et al., 2001; Withall et al., 2015), and it was opposite *redG* that a significant asRNA was detected (**Figure 3.7D**). The undecylprodigiosin:butyl-meta-cycloheptylprodigionine ratio is typically ~2:1 (Tsao et al., 1985), so the *redG*-specific asRNA could provide a means of selectively modulating *redG* expression without impacting that of the downstream *redF*, whose product acts earlier in the undecylprodigiosin biosynthetic pathway (Williamson et al., 2006).

In addition to expressing conventional asRNAs, specialized metabolic clusters were also a rich source of cutoRNAs, with eight (*S. coelicolor*), six (*S. avermitilis*) and three (*S. venezuelae*) cutoRNA pairs identified within these clusters (**Table 3.1**). These included a distinctive cutoRNA pair within the 22 gene actinorhodin (act) biosynthetic cluster of *S. coelicolor*. Actinorhodin is a blue-pigmented polyketide antibiotic whose synthesis is directed by one of the best-studied pathways in the streptomycetes. At the centre of this cluster are two convergently transcribed genes, *actVA6* and *actR*, whose coding sequences overlap, and whose transcripts extended the full length of their respective downstream genes (**Figure 3.8A**). *actVA6* encodes a monooxygenase that catalyzes an intermediate step in actinorhodin biosynthesis (Kendrew et al., 1997), while *actR* encodes a TetR-family repressor of the proposed actinorhodin resistance (ABC transporter) encoding genes *actAB* (Tahlan et al., 2007; Xu et al., 2012).

The intriguing genetic coupling of biosynthesis and transport-associated genes was also observed for the siderophore-producing coelichelin biosynthetic cluster (Lautru et al., 2005). Within this 11 gene cluster, the 3′ UTR of *sco0491* (*cchI*) extended into the coding region of *sco0490* (*cchJ*) (**Figure 3.8B**). Similar to the cutoRNA pair from the actinorhodin cluster, *sco0490* encodes a coelichelin biosynthetic enzyme, and *sco0491* encodes an ABC transporter that may participate in coelichelin export (Lautru et al., 2005).

This theme was further reiterated in the chloramphenicol biosynthetic cluster of *S. venezuelae*. A four gene region encompassing *sven0915*, *cmlF*, *cmlE*, and *cmlD* (the leftmost genes in the cluster) was transcribed such that expression of the *cmlE-cmlD* operon failed to terminate, and instead extended through *cmlF* (a major facilitator family transporter) encoded on the opposite strand, into *sven0915*, located approximately 1,600 base pairs downstream (**Figure 3.8C**). CmlE and CmlD are required for chorismic acid synthesis, where chorismic acid is a precursor for both aromatic amino acid and chloramphenicol biosynthesis (He et al., 2001); CmlF is a major facilitator transporter that may contribute to chloramphenicol resistance (although it is not the major resistance determinant).

Interestingly, while this cutoRNA phenomenon has not been reported previously, prior studies have identified short asRNA regions associated with antibiotic resistance genes that stemmed from overlapping divergent promoters (Bedford et al., 1991; Bibb et al., 1994; Janssen et al., 1989). It will be interesting to determine whether such overlap of 5′ and 3′ untranslated regions is important for the stability and/or function of the associated mRNAs and protein products.

Most specialized metabolic clusters exhibited transcriptional patterns that could be readily correlated with protein-coding genes or defined asRNAs; however, there were clusters in each species that exhibited unusual transcriptional complexity. The most

remarkable example of this was in the avermectin biosynthetic cluster of *S. avermitilis*. Avermectin is an important anti-parasitic agent that is used broadly in veterinary medicine (Omura and Crump, 2004). Its biosynthetic cluster spans ~81 kb and 19 annotated genes (Ikeda et al., 2003; Omura et al., 2001), and an 18 kb region at the left end of the cluster encompassed a multitude of ncRNAs. This region included genes encoding the pathway-specific activator AveR, two polyketide tailoring enzymes (AveF and AveG), and the first of four polyketide synthases (the type I polyketide synthase AveA1). Expression of the convergently-oriented *aveR* and *aveF* genes resulted in the generation of a cutoRNA (**Figure 3.8D**). There was also very high antisense RNA expression observed opposite *aveF* and *aveD*, which themselves appeared to be co-transcribed, despite being separated by >700 bp [this unusual operonic structure has been noted previously (Chen et al., 2001)].

mRNA-associated small RNAs are widespread in the streptomycetes

The sRNA-enriched library proved to be a rich source of not only intergenic sRNAs, but also mRNA-associated short RNAs. While the long transcript libraries yielded relatively even coverage throughout most mRNAs, the sRNA-enriched libraries were dominated by sequences from the 5′ and 3′ UTRs of coding sequences. These regions were, in some instances, represented 100–1000 fold more highly than their corresponding coding sequences. In addition to these stable 5′ and 3′ UTR-associated RNAs, we also detected short sequences (~30-90 nucleotides) within coding regions that were unusually highly represented relative to their flanking sequences. A typical example of both end- and internally-enriched RNAs was seen for *sven2374*, which showed significant over-representation of its 5′ end and an internal 62 nt region, relative to the rest of the coding sequence (**Figure 3.9A**). Stable secondary structures within 5′ UTRs have been previously described (Bricker and Belasco, 1999; Hambraeus et al., 2002; Unniraman et al., 2002) and can influence overall mRNA stability, whereas stable regions within 3′ UTRs may represent termination sequences (a class of sequence that has not been well-defined in the streptomycetes). In *E. coli*, 5′ and 3′ UTR fragments can accumulate to high levels, and in some cases appear to be expressed at times distinct from that of their corresponding mRNAs, suggesting that they may have the capacity to act as independent RNA molecules (Kawano et al., 2005). Similarly, in *Salmonella*, 3′ UTRs have significant sRNA-encoding potential (Miyakoshi et al., 2015a). The 3′ sRNAs arise either through an independent promoter located within the coding region of the mRNA (Chao et al., 2012) or as a stable mRNA degradation product (Miyakoshi et al., 2015b). In the latter of these two examples, upon degradation by RNase E, the polycistronic *gltIJKL* mRNA (encoding amino acid transporters) yields a small sRNA named SroC. The SroC sRNA subsequently degrades another sRNA GcvB, de-repressing the expression of multiple GcvB-repressed genes, many of which are other amino-acid transporters. Hence, it is thought that this arrangement links the degradation of one mRNA involved in amino-acid transport, with increased expression of alternative genes in the same pathway (Miyakoshi et al., 2015b).

To begin understanding how these stable RNA species could be generated, we selected 20 of the most highly represented sequences for further analysis (**Table 3.7**).

These sequences were confined to *S. venezuelae*, where this phenomenon was more predominant than in the other two species examined. Secondary structure predictions (Zuker, 2003) for the 20 highly represented sequences, suggested that all of these regions were highly structured, compared with sequences that were less abundant in the sRNA-enriched library (*e.g.* **Figure 3.9B**). We evaluated the GC content of these 20 sequences, comparing them with the nucleotide content of the entire coding sequence, and with a 15-nucleotide sequence window immediately upstream of the stable/structured sequences. We found the 20 structured sequences had a GC-content similar to that of the coding sequence as a whole. This is in contrast to the regions immediately preceding the structured sequences, which were significantly more AT-rich (an average of 32.3% versus 27.8%), and contained a higher proportion of poorly-used codons relative to the structured sequence immediately following. This suggested the potential for translational pausing, which when coupled with a highly structured downstream region, could promote stable RNA fragment accumulation. In *E. coli,* the major endoribonuclease RNase E cleaves in AU-rich regions near hairpin structures (Mackie and Genereaux, 1993). In an analysis of ~22,000 RNase E recognition sites in *Salmonella* it was found that the ribonuclease cleaves in unstructured regions upstream of a uracil. These regions themselves are often upstream of a stem-loop structure generating sRNA transcripts (Chao et al., 2017). It will be interesting to see whether *Streptomyces* RNase E (Hagège and Cohen, 1997) contributes to the accumulation of these RNA species. It is equally possible that these RNA products were expressed independently of the associated coding sequence, as the AT-rich upstream region would also be consistent with a promoter region. To determine whether these abundant intragenic RNAs existed in the cell as discrete RNA elements, we used northern blotting with a probe specific for the highly represented region within *sven2374*. We observed a stable product during the early vegetative growth phase of *S. venezuelae* (**Figure 3.9C**), suggesting that these RNAs do indeed accumulate; whether they have a functional role in the cell remains to be determined.

Conclusions

Using an RNA-seq approach to evaluate gene expression throughout the *Streptomyces* life cycle, we have identified hundreds of novel ncRNAs in three disparate *Streptomyces* species. These included novel sRNAs, asRNAs and a prominent new class of asRNA – the cutoRNAs – that result from overlapping convergent transcription. Comparative analysis of the ncRNAs revealed considerable differences between species and between ncRNA types: *S. coelicolor* and *S. avermitilis* shared far greater numbers of ncRNA elements than either did with *S. venezuelae*, and throughout all species, asRNAs were less well conserved than sRNAs. From a genome-scale perspective, sRNA conservation largely mirrored that of protein-coding genes: sRNAs expressed from the chromosome ends were species-specific, while chromosome core-localized sRNAs were more highly conserved. Notably, ncRNAs were common features in specialized metabolic biosynthetic clusters, and likely contribute to the regulatory control of these pathways. Uncovering the ncRNA capacity of the streptomycetes will facilitate the downstream

integration of these molecules into the regulatory networks governing growth, development and antibiotic production.

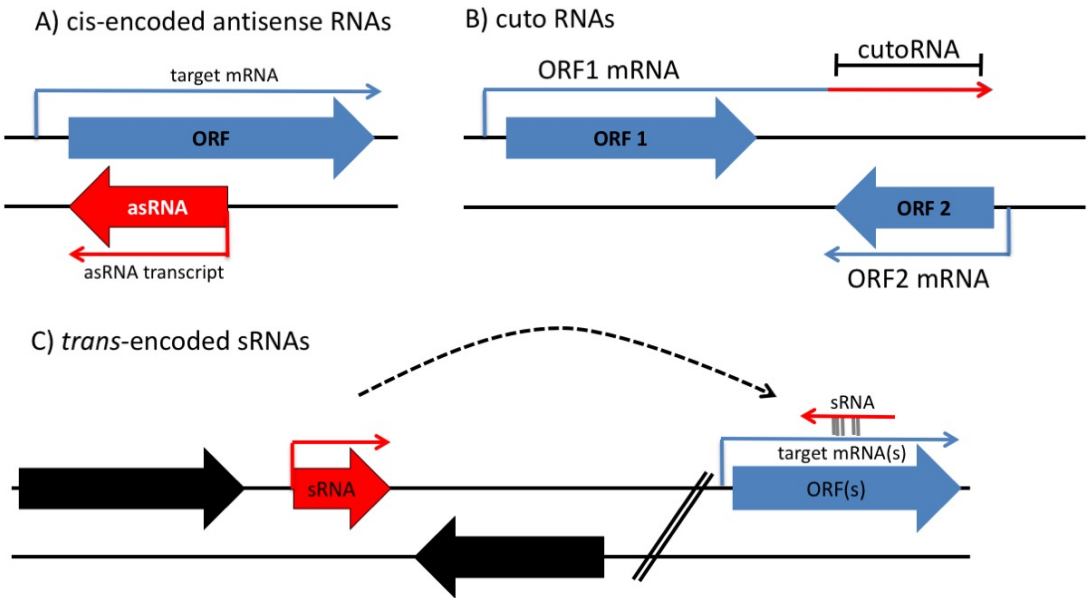


Figure 3.1: Schematic illustration of the different classes of non-coding RNAs identified. Genes are depicted as thick arrows, with protein-coding genes shown in blue and labeled as 'ORF's (open reading frame), ncRNAs shown in red, and black depicting genes of either type. RNA transcripts are shown above their corresponding gene, with transcription initiating at the vertical line, and terminating at the small arrowhead. **A)** Antisense RNAs (asRNAs) are expressed from a promoter on the strand opposite a protein-coding gene. **B)** cutoRNAs occur when a long 3′ UTR of an mRNA overlaps with a downstream, convergently transcribed gene. The region of overlap is indicated with a bracketed line. **C)** Small RNA (sRNA) genes are most commonly found in the intergenic region between genes, and typically target (by imperfect complementary base-pairing) one or more mRNAs expressed from disparate chromosomal locations.

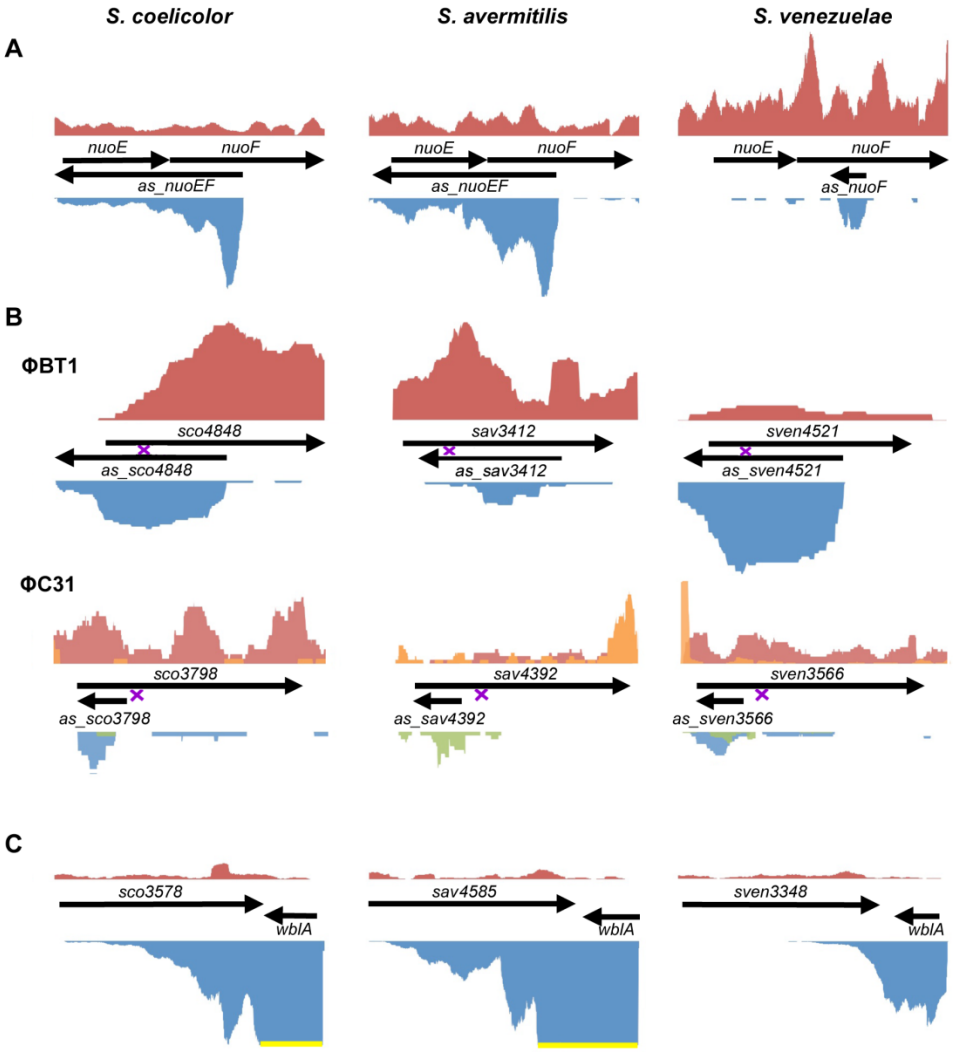


Figure 3.2: Expression profiles of select conserved cis-antisense RNAs and cutoRNAs. A) Expression profile of the antisense gene opposite *nuoE* and *nuoF*. **B)** Expression profile of a conserved asRNA opposite homologous genes in *S. coelicolor*, *S. avermitilis* and *S. venezuelae* which encompassed the site of ΦBT1 (Top) or ΦC31 (Bottom). The site of phage integration is marked with a purple 'x'. **C)** cutoRNA shared between *wblA* and a conserved downstream gene (*sco3578*/*sav4585*/*sven3348*). Red graphs (top) represent relative read coverage (long transcript library) for the positive strand (orange is the equivalent for the sRNA-enriched library), while blue graphs represent read coverage (long read library) for the negative strand (green represents the same in the sRNA-enriched library). As expression levels of different genes varied over several orders of magnitude, the y-axis for each gene set was scaled independently. When expression levels differed greatly for the positive and negative strand within a gene pair, the profile of the more highly expressed gene was cut off with a yellow line, to ensure that expression from the less highly expressed gene could be visualized.

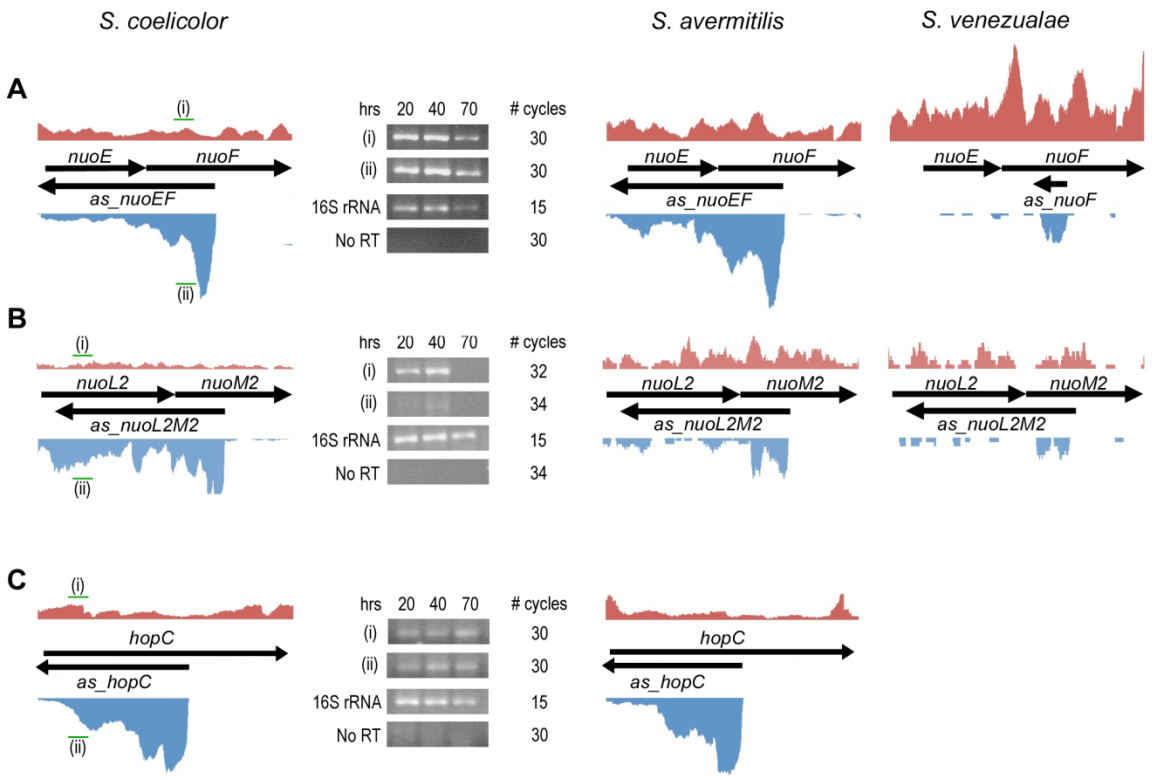


Figure 3.3: Expression profiles and semi-quantitative RT-PCR analysis of select asRNAs. A) Expression profiles (*S. coelicolor, S. avermitilis*, and *S. venezuelae*) and semi-quantitative RT-PCR analysis (*S. coelicolor*) of the asRNA expressed opposite *nuoE* and *nuoF*. **B)** Expression profiles (*S. coelicolor*, *S. avermitilis*, and *S. venezuelae*) and semi-quantitative RT-PCR analysis (*S. coelicolor*) of the asRNA expressed opposite *nuoL2* and *nuoM2*. **C)** Expression profiles (*S. coelicolor* and *S. avermitilis*) and semi-quantitative RT-PCR analysis (*S. coelicolor*) of the asRNA expressed opposite *hopC*. Relative sequence reads at each nucleotide position on the positive strand (top) are shown in red, while negative strand profiles are shown in blue. As expression levels of different genes varied over several orders of magnitude, the y-axis for each gene set was scaled independently. Semi-quantitative RT-PCR analyses examined the temporal expression of each sense and antisense gene (time of RNA extraction is indicated in hours). Reverse transcribed 16S rRNA was amplified as a control for RNA integrity and loading levels. Fragment (i) was reverse transcribed from transcripts originating from the positive strand, while (ii) was reverse transcribed from transcripts originating on the negative strand. The locations of these primers are indicated above or below the gene annotation.

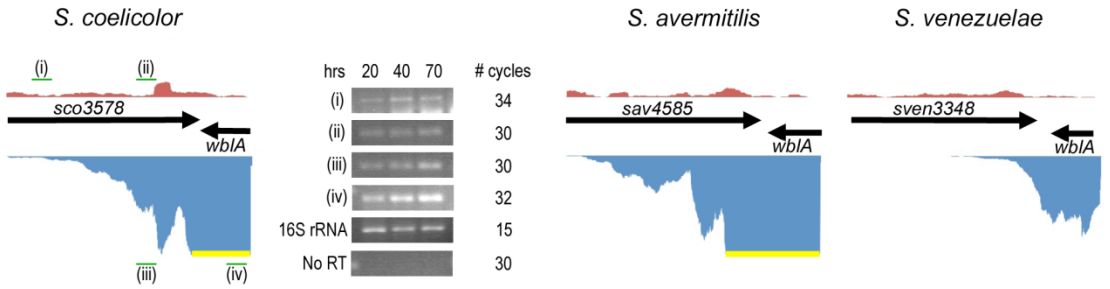


Figure 3.4: Expression profiles and semi-quantitative RT-PCR analysis of the *wblA* cutoRNA shared with its downstream gene in *S. coelicolor* (*sco3578*), *S. avermitilis* (*sav4585*) and *S. venezuelae* (*sven3348*). Relative sequence reads at each nucleotide position on the positive strand (top) are shown in red; negative strand profiles are shown in blue. When expression levels differed greatly for the positive and negative strand within a gene pair, the profile of the more highly expressed gene was cut off with a yellow line to ensure expression from the more lowly expressed gene could still be visualized. Semi-quantitative RT-PCR analyses over a developmental time course (time of RNA harvest is shown in hours), was conducted to assess cutoRNA expression. Reversed transcribed 16S rRNA was amplified to ensure RNA integrity and to serve as a loading control. Products (i) and (ii) were reverse transcribed from positive strand-derived transcripts, while (iii) and (iv) were reverse transcribed from transcripts originating from the negative strand. Primer locations are indicated above or below the gene annotation.

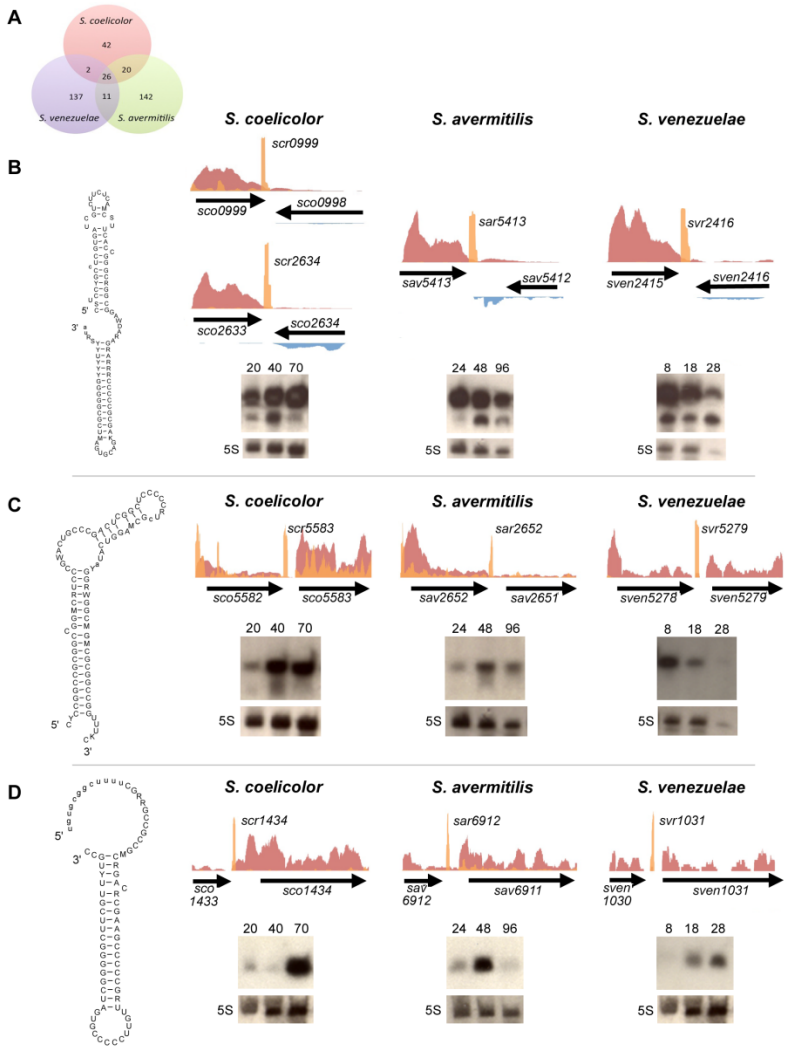


Figure 3.5: Comparing the structure and expression of conserved intergenic sRNAs. **A)** Venn diagram illustrating sRNA conservation in *S. coelicolor*, *S. avermitilis* and *S. venezuelae*. **B-D)** Structure, expression profiles and northern blot analyses of conserved sRNAs: **B)** *scr0999*, *scr2634*, *sar5413* and *svr2416*; **C)** *scr5583*, *sar2652* and *svr5279*; **D)** *scr1434*, *sar6912* and *svr1031*. For (B-D): predicted conserved secondary structures for each sRNA are shown to the far left. Non-standard bases are indicated as follows: M (A or C), R (A or G), W (A or U), S (G or C), Y (C or U), K (G or U), D (not C). Insertions are denoted in lower case. For the expression profiles: positive strand coverage is shown above the gene annotation in red (long transcript library) and orange (sRNA-enriched library); negative strand profiles (below annotation) are shown in blue (long transcript library) and green (sRNA-enriched library). Northern blots showing the temporal expression of each sRNA (time of RNA extraction is indicated in hours) are shown below each coverage graph. For each blot, 5S rRNA was also probed as a control for RNA integrity and abundance.

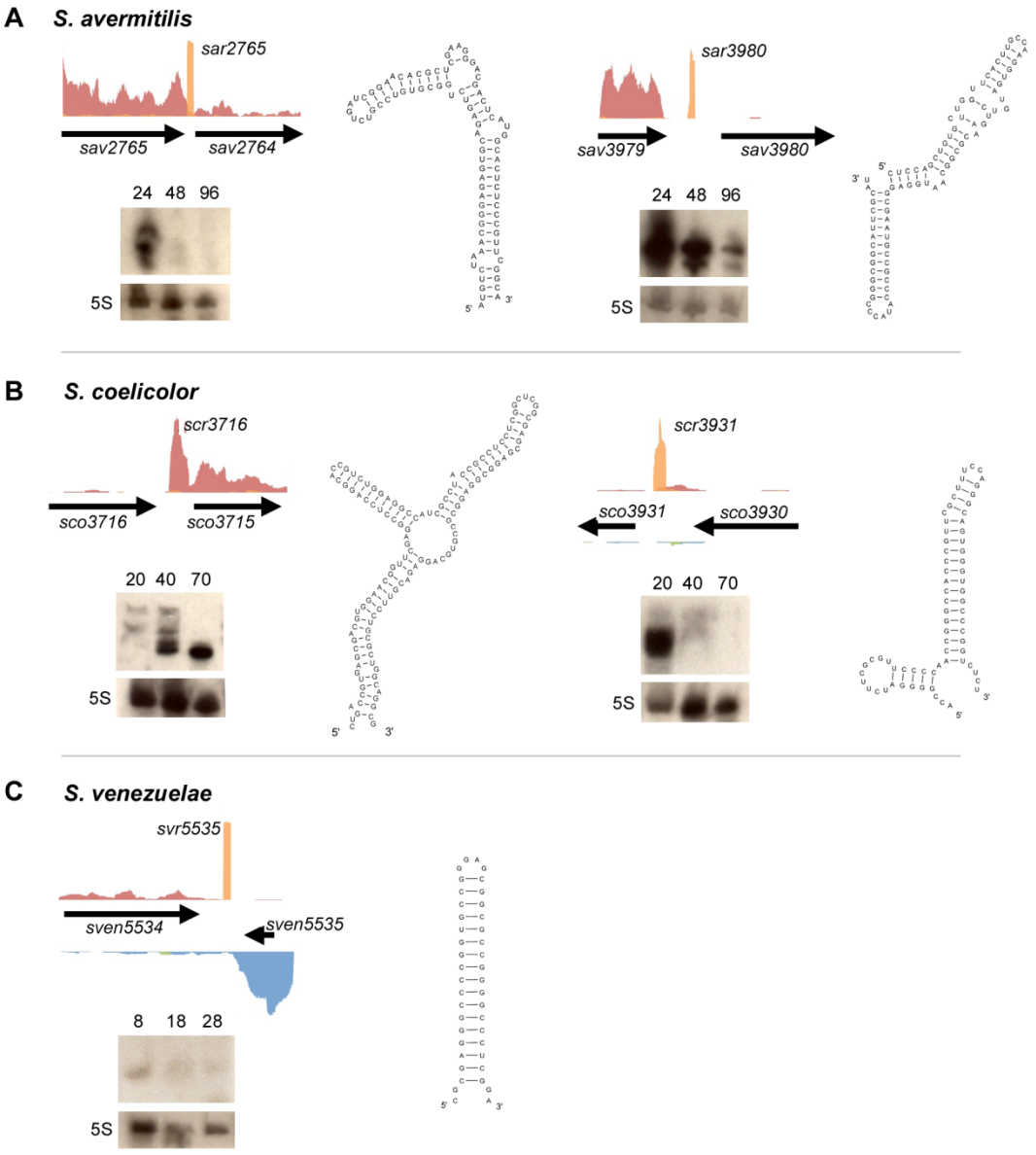


Figure 3.6: Structure and expression analyses of species-specific intergenic sRNAs. Expression profiles, northern analyses and structural predictions for: **A)** *S. avermitilis* sRNAs *sar2765* (left) and *sar3980* (right); **B)** *S. coelicolor* sRNAs *scr3716* (left) and *scr3931* (right); **C)** *S. venezuelae* sRNA *svr5535*. For each expression profile, relative sequence reads for genes encoded on the positive strand (top) are shown in red (long transcript library) and orange (sRNA-enriched library), while negative strand profiles are shown in blue (long transcript library). As expression levels for different genes varied greatly, the y-axes of each panel were scaled independently. Northern blots, shown below each coverage graph, revealed sRNA expression throughout development (time of RNA harvest is shown in hours). 5S rRNA was probed as a control for RNA quantity and integrity.

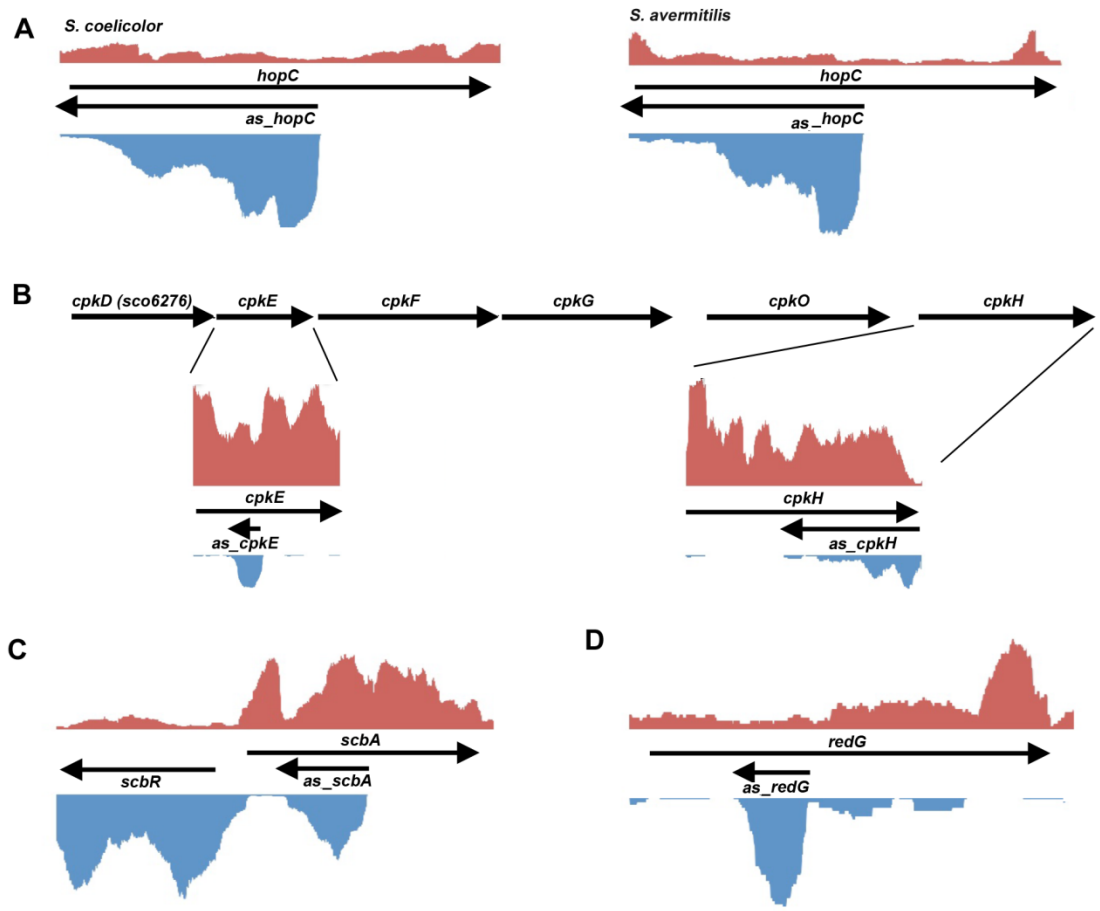


Figure 3.7: Expression profiles of asRNAs within secondary metabolite clusters. A) Expression profile of the asRNA expressed opposite *hopC* in *S. coelicolor* (left) and *S. avermitilis* (right). **B)** Expression of the two asRNAs expressed opposite *cpkE* and *cpkH* within the coelimycin P1 biosynthetic cluster of *S. coelicolor*. The relative position of these two genes within the cluster is shown above the coverage graphs. **C)** Expression levels for *scbA* and *scbR* within the coelimycin P1 biosynthetic cluster in *S. coelicolor*. Antisense RNAs resulted from the divergent transcription of these two genes (Chatterjee et al., 2011) and an independent antisense RNA was expressed within the coding region of *scbA*. **D)** Expression profile of the asRNA expressed opposite *redG* within the prodiginine biosynthetic cluster of *S. coelicolor*. For each of (A-D), relative sequence reads at each nucleotide position were shown in red (positive strand on the top), and blue (negative strand on the bottom). The y-axis of each gene set was scaled independently, as expression levels of different gene clusters varied.

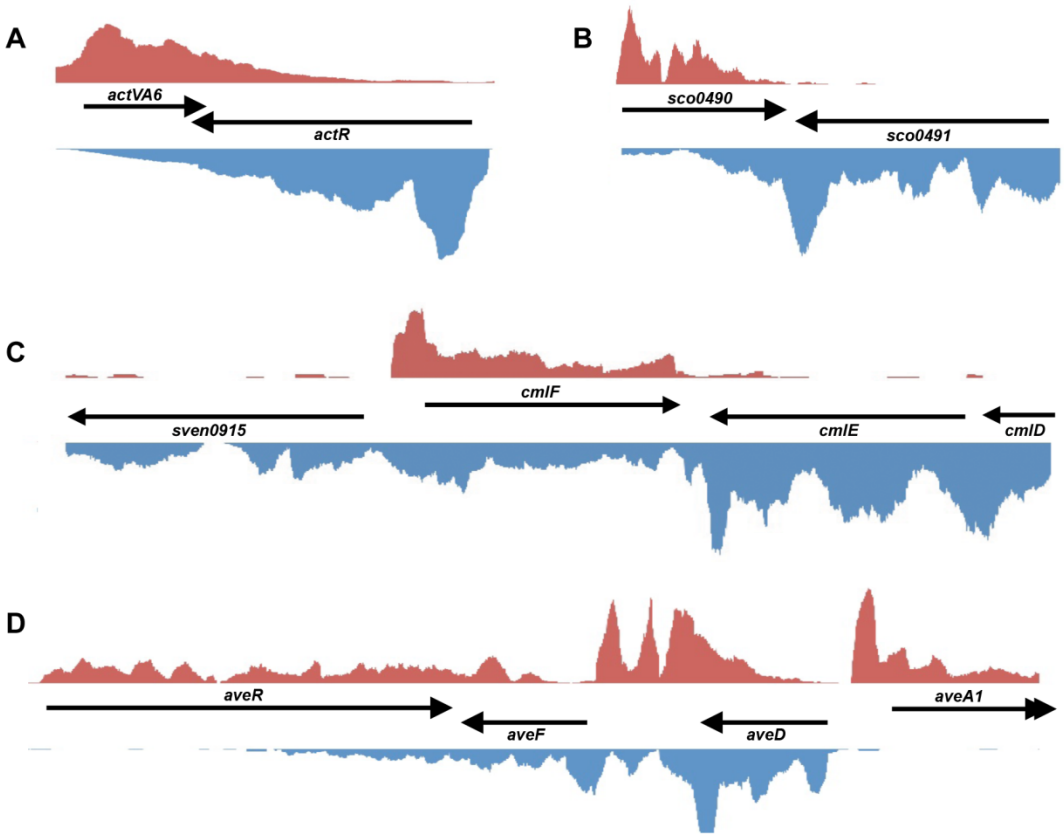


Figure 3.8: cutoRNA expression within specialized metabolic clusters. A) Expression levels for a cutoRNA pair shared by *actVA6* and *actR* within the actinorhodin cluster of *S. coelicolor*. **B)** Expression profiles for a cutoRNA pair (*sco0491* and *sco0490*) within the coelichelin siderophore biosynthetic cluster of *S. coelicolor*. **C)** Expression of a four gene region within the chloramphenicol biosynthetic cluster of *S. venezuelae*. Transcription of the *cmlDE* operon generates an asRNA/cutoRNA to *cmlF*. **D)** Expression profile of a transcriptionally complex region within the avermectin biosynthetic cluster of *S. avermitilis*. A cutoRNA pair exists between *aveR* and *aveF*. Multiple antisense RNAs were found opposite the *aveDF* operon. For (A-D), relative sequence reads at each nucleotide position were shown in red for the positive strand (top), and blue for the negative strand (bottom). Given the differing expression levels observed for each gene cluster, y-axes for each panel were scaled independently.

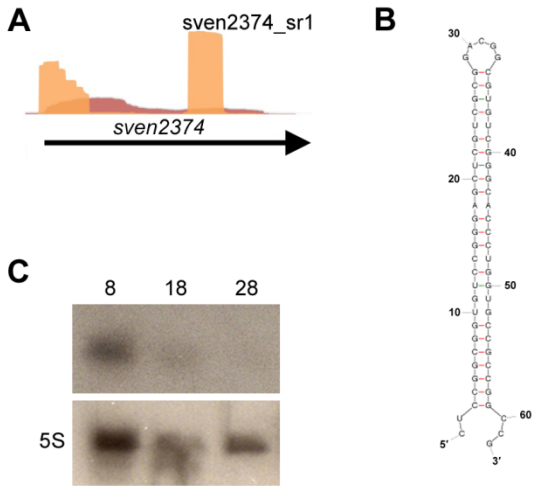


Figure 3.9: Expression profile, northern analysis and secondary structure prediction of an mRNA-associated small RNA in *S. venezuelae*. A) Expression profile of *sven2374*, which shows high read levels for both its 5′ end and an internal 62 nucleotide region. Relative sequence reads at each nucleotide position are shown in red (long transcript library) and orange (sRNA-enriched library). **B)** Predicted secondary structure of *sven2374_sr1* (sr1: stable region 1). **C)** Northern blot showing the temporal expression of *sven2374_sr1* (time of RNA extraction is indicated in hours). 5S rRNA was probed as a control for RNA integrity and RNA loading levels.

Table 3.1 Classification of antisense RNAs above threshold in *S. coelicolor*, *S. avermitilis*, and *S. venezuelae* (due to its large size, see Appendix 1)

Table 3.2: Homologous genes with conserved asRNAs

S. coelicolor	*S. avermitilis*	*S. venezuelae*	**Annotation of sense gene**
sco2364	*sav5807*	*sven2178*	Conserved hypothetical protein
sco2685	*sav5363*	*sven2472*	Putative ATP-binding protein
sco3318 - sco3317	*sav4741 - sav4740*	*sven3179 - sven3180*	Putative porphobilinogen deaminase (HemC) / uroporphyrinogen-III synthase (HemD)
sco3408	*sav4662*	*sven3260*	D-Ala-D-Ala carboxypeptidase
sco3671	*sav4484*	*sven3433*	DnaK - heat shock protein
sco4137	*sav4077*	*sven3895*	Pit-accessory protein (phosphate transport)
sco4440	*sav3782*	*sven3077*	Hypothetical protein
sco4566 - sco4567	*sav4841 - sav4842*	*sven4269 - sven4270*	NuoF / NuoE
sco4606 - sco4607	*sav4888 - sav4889*	*sven4315 - sven4316*	NuoL2 / NuoM2
sco4829	*sav3433*	*sven0126*	putative oxidoreductase
sco4848	*sav3412*	*sven4521*	putative integral membrane protein

Table 3.3 Conserved cutoRNA gene pairs (due to its large size, see Appendix 1)

Table 3.4 Intergenic sRNAs in S. coelicolor, S. avermitilis, and S. venezuelae (due to its large size, see Appendix 1)

Table 3.5: Conserved intergenic sRNAs in Streptomyces (due to its large size, see Appendix 1)

Table 3.6. Location of unique and conserved sRNAs in *S. coelicolor*

Chromosomal region*	**# sRNAs**	**# unique**	**# conserved**	**% conserved**
Left terminal region	2	2	0	0%
Left *Actinomycetales*-specific region	17	3	14	82.4%
Core region	58	29	29	50.0%
Right *Streptomyces*-specific region	8	5	3	37.5%
Right terminal region	7	7	0	0%

*: as per Kirby et al (2011)

Table 3.7: Locations and names of 19 highly expressed stable degradation products in *Streptomyces venezuelae*.

Gene	CDS Start	CDS Stop	Strand	Stable region designation*	Stable Start	Stable Stop	Length
sven0433	463756	464730	+	sven0433_sr1	463897	463972	76
sven0729	840508	841308	-	sven0729_sr1	840939	841016	78
sven0847	971829	972173	+	sven0847_sr1	972049	972100	52
sven1454	1631142	1632533	+	sven1454_sr1	1631298	1631375	78
sven1480	1656022	1656387	-	sven1480_sr1	1656329	1656370	42
sven2285	2465015	2466037	-	sven2285_sr1	2465419	2465480	62
sven2374	2564998	2566152	-	sven2374_sr1	2565389	2565477	89
sven2374	2564998	2566152	-	sven2374_sr2	2565153	2565214	62
sven2917	3176014	3176742	-	sven2917_sr1	3176510	3176563	54
sven3045	3331513	3334293	-	sven3045_sr1	3332336	3332420	85
sven3682	4003974	4004264	+	sven3682_sr1	4004048	4004088	41
sven3714	4031950	4032951	+	sven3714_sr1	4032618	4032681	64
sven3715	4032972	4036562	+	sven3715_sr1	4033149	4033208	60
sven3715	4032972	4036562	+	sven3715_sr2	4033257	4033295	39
sven3715	4032972	4036562	+	sven3715_sr3	4034482	4034531	50
sven4147	4488857	4491217	-	sven4147_sr1	4490785	4490858	74
sven4589	4944297	4945223	+	sven4589_sr1	4945085	4945170	86
sven4918	5296230	5296568	+	sven4918_sr1	5296312	5296357	46
sven5046	5429952	5430920	-	sven5046_sr1	5430157	5430227	71
sven5212	5633123	5634463	+	sven5212_sr1	5633165	5633207	43

* sr denotes ‘stable region’

Chapter 4: Characterization of the antisense RNA ScbN

I performed the majority of the experiments and data analysis for this chapter, with the following exceptions. Christine Pham made some of the strains in Figure 4.6, and Savannah Colameco performed *gusA* assays in Figure 4.3.

Introduction

Bacteria are classically viewed as being single-celled organisms; however, they frequently exist in multi-species communities, and have the ability to coordinate their behaviours with those of closely related neighbours. This inter-cell communication can be achieved using chemical signals that are secreted by individuals, and sensed and responded to by larger groups. Depending on the bacteria, these chemical signals can be either small molecules or small peptides, and the coordinating of group behaviors is collectively known as quorum-sensing. At certain signal thresholds, group behavior is induced. Group behaviors regulated by quorum-sensing include pathogenesis (Passador et al., 1993), antibiotic production (Hara and Beppu, 1982), bioluminescence (Nealson, 1977) and multicellular development (Ng and Bassler, 2009; Waters and Bassler, 2005).

Quorum-sensing is a general term that encompasses diverse strategies at both the molecular and gene regulatory network levels. Most quorum-sensing bacteria secrete the quorum-sensing signal at low levels when at low cell densities. This signal tends to be a small molecule in Gram-negative bacteria, and a small peptide in Gram-positive bacteria (Waters and Bassler, 2005). As cell density increases, so does the quorum-sensing signal, allowing it to be detected by cells in the vicinity. Most small molecule signals are able to passively diffuse across the membrane and are detected by transcription factor receptors, whose structures – and activities – are altered upon small molecule binding. In contrast, most small peptides are detected extracellularly by a two component system histidine kinase, which then activates its cognate cytoplasmic response regulator. The ensuing downstream regulatory networks in both cases are enormously diverse, although there are two basic principles that hold for most systems. First, once a threshold of quorum-sensing signal is reached, the genes required for synthesizing the signal are dramatically up-regulated, leading to a stable 'on' switch. Second, genes required for particular group behaviors are turned on (Ng and Bassler, 2009).

In contrast to most Gram-positive bacteria, the main quorum-sensing molecules produced by *Streptomyces* are not small peptides, but γ-butyrolactones (GBLs) (Takano, 2006). γ-butyrolactone signaling has been best studied in *S. griseus*, where it controls both morphological development and specialized metabolism (Ohnishi et al., 2005) (**Figure 1.3**).

While *S. griseus* is the best-studied *Streptomyces* species in terms of γ-butyrolactone-mediated regulation, *S. coelicolor* is the model system for understanding both development and specialized metabolism in the streptomycetes. *S. coelicolor* has homologs of many of the *S. griseus* γ-butyrolactone system genes, but the regulatory architecture associated with the *S. coelicolor* γ-butyrolactone system is different from that of *S. griseus* (Takano, 2006). Like *S. griseus, S. coelicolor* encodes a γ-butyrolactone

synthase (ScbA) and receptor (ScbR). However, unlike in *S. griseus* where these two genes are located hundreds of kilobases away from each other (Ohnishi et al., 2008), in *S. coelicolor, scbA* and *scbR* are found adjacent to each other in the chromosome. Furthermore, they are arranged divergently, such that their 5′ UTRs overlap.

S. coelicolor makes three GBLs, namely SCB1, SCB2 and SCB3, with SCB1 being the most abundant (Hsiao et al., 2009). Unlike *S. griseus* ArpA, which has a single known target (*adpA*) (Ohnishi et al., 2005), chromatin immunoprecipitation-sequencing experiments have shown that ScbR binds upstream of ~144 genes involved in primary and specialized metabolism (Li et al., 2015). The best studied targets of ScbR were discovered over a decade earlier and include *scbA, scbR*, and *cpkO* (Takano et al., 2001, 2005) (**Figure 4.1**).

In the absence of the GBL SCB1, the GBL-receptor protein ScbR binds to and represses *cpkO,* where *cpkO* encodes the coelimycin pathway-specific activator (Takano et al., 2005). As SCB1 levels rise, SCB1 binds to ScbR, leading to a structural change that abolishes the ability of ScbR to bind DNA. This leads to the de-repression of *cpkO,* and production of the antibiotic coelimycin. How ScbR regulates both its own expression and that of *scbA* are not entirely clear. In a Δ*scbA* strain, no GBLs are produced, and there is reduced expression of both *scbA* and *scbR*. This would suggest that in the absence of GBLs, ScbR represses both *scbA* and *scbR*. However, a Δ*scbR* strain shows increased expression from the *scbR* promoter, and concomitantly much lower expression of *scbA*. Taken together, these results suggest that ScbR represses itself, and that both ScbA and ScbR are required for *scbA* expression (Takano et al., 2001) (**Figure 4.1**). Some have suggested that ScbR acts as a repressor at some promoters and that an ScbR:ScbA complex activates the transcription of other genes (Takano et al., 2001). There is no experimental evidence that such a complex forms.

Interpretation of these data are further complicated by the fact that there are at least four other GBL-binding and/or GBL-like-binding proteins that may be capable of binding within the *scbAR* promoter regions: ScbR2 (Xu et al., 2010), SlbR (Yang et al., 2012), CprA (Li et al., 2015), CprB (Li et al., 2015). The evidence for these proteins binding to the *scbAR* promoters are quite strong for ScbR, ScbR2 and SlbR, being supported by multiple lines of *in vivo* and *in vitro* evidence (Li et al., 2015; Takano et al., 2001; Xu et al., 2010; Yang et al., 2012). The evidence for CprA and CprB binding to the *scbAR* is preliminary, and is based on indirect evidence and unpublished results (Li et al., 2015). The complex regulatory networks connecting all of these genes and their products (*i.e.* they regulate each other) make understanding the consequences of deleting a single gene challenging to decipher (Li et al., 2015). Recent reports have suggested that ScbR and ScbR2 can also form heterodimers, which adds further complexity to this regulatory system (Li et al., 2017).

As mentioned above, *scbA* and *scbR* are adjacent to each other and arranged in a divergent orientation such that their 5′ UTRs overlap (**Figure 4.1**). Mathematical modelling has suggested that this arrangement leads to repression of both genes, either due to collisions of convergently transcribing RNA polymerases, or post-transcriptionally through the interaction of the two 5′ UTRs (Chatterjee et al., 2011; Mehra et al., 2008).

An important feature of quorum-sensing is that there are two stable states – a phenomenon known as bistability. There is some debate as to how bistability may be achieved in the *S. coelicolor* system. One group (Mehra et al., 2008) has proposed that the hypothetical ScbA:ScbR complex contributes to the bistable (on and off) states. Another group (Chatterjee et al., 2011) has argued that there was no evidence for the ScbA:ScbR complex, and instead reasoned that bistability could be due to the divergent transcription and resulting asRNAs formed between the UTRs of *scbA* and *scbR*. It is worth noting that there is no experimental evidence that divergent transcription in this region has a repressive effect on either gene. There are, however, many instances where this sort of arrangement does cause repression in other species (Sesto et al., 2012).

RNA-sequencing experiments have revealed that asRNAs are encoded on the opposite strand of most protein-encoding genes in bacteria (Lasa et al., 2011). Yet, despite their prevalence, very few asRNAs have been characterized in any detail (Thomason and Storz, 2010). Most well-studied asRNAs are expressed near the translation initiation regions of mRNAs and modulate either RBS accessibility or RNase activity (Georg and Hess, 2011; Thomason and Storz, 2010). There have been no *in vivo* studies of asRNAs that initiate and terminate entirely opposite an open-reading-frame. How (and if) these asRNAs regulate their cognate mRNAs in the presence of translating ribosomes remains an open question.

We recently discovered hundreds of asRNAs in three phylogenetically diverse *Streptomyces* species (Moody et al., 2013) (Chapter 3). In the model streptomycete, *S. coelicolor,* we discovered an asRNA expressed opposite the coding region of the γ-butyrolactone synthase *scbA,* which we termed ScbN (*S*treptomyces *c*oelicolor butyrolactone non-coding RNA) (**Figure 4.1**).

We sought to explore the role of ScbN in controlling *scbA* expression. Surprisingly, we found that the asRNA itself had little effect on *scbA* expression. Instead, we discovered that the transcriptional terminator of *scbN* expression, which also forms a hairpin in the sense direction in the mRNA, has a profound impact on *scbA* transcription. Many *scbA* homologs in other *Streptomyces* species possess strong hairpins within their coding sequences. Interestingly, analysis of sequences revealed compensatory mutations that maintained RNA secondary structure for many of these hairpins, despite changes to the underlying coding sequences. This suggests that structured RNA within *scbA* has been selected for, and may therefore have an important functional role.

Materials and Methods

Creation of Δ*scbAR* deletion strain

To begin exploring the role of *scbN* within the context of *scbA* (and *scbN*) and *scbR*, the entire *scbARN* region was deleted using the ReDirect system (Gust et al., 2003) (**Table 2.2)**. The sequence extending from the stop codon of *scbR* to the stop codon of *scbA* was replaced by a hygromycin resistance cassette. This involved PCR amplifying (**Table 2.3**) the *oriT*-hygromycin resistance cassette from pIJ797 using primers scbAR_KO_L and scbAR_KO_R (**Tables 2.4 and 2.5**). Each of these primers have 39 nucleotide 5′ extensions corresponding to the sequence flanking the end of each of *scbR*

and *scbA,* such that the resulting PCR product can recombine and replace the intervening *scbR-scbA* coding sequences within the StAH10 cosmid, with the hygromycin resistance cassette.

StAH10 was first introduced into BW25113/pIJ790 by electroporation, and positive transformants were selected for using chloramphenicol (pIJ790) and ampicillin (StAH10) (**Table 2.3**). Next, the PCR fragment containing the hygromycin resistance cassette with flanking *scbR* and *scbA* sequences was electroporated into this strain and cosmid-containing colonies (where *scbA/scbR* had been replaced with the hygromycin resistance cassette) were selected using ampicillin (cosmid), and hygromycin (*scbAR* deletion). Successful replacement of *scbAR* with the hygromycin resistance cassette was confirmed by restriction enzyme digestion with BamHI, and by PCR using a primer that binds within the hygromycin cassette, together with scbAR_KO_L (**Table 2.4, Table 2.5**).

The confirmed StAH10 Δ*scbAR* cosmid was then transferred by electroporation into *E. coli* ET12567 carrying pUZ8002, and conjugated with wild-type *S. coelicolor* M145. Successful exconjugants were selected for using hygromycin, while nalidixic acid was used to select against the donor *E. coli* strain. To screen for double cross-over recombinants that had lost the cosmid but retained the hygromycin resistance cassette in place of *scbAR*, the exconjugants were replica plated onto SFM plates supplemented with naladixic acid and hygromycin, with and without kanamycin. Putative kanamycin sensitive colonies were re-streaked on SFM with hygromycin, with and without kanamycin, to confirm kanamycin sensitivity. Two sets of PCRs were then used to confirm the deletion of *scbAR.* The first utilized a primer upstream of *scbA* (PscbA_fwd) and one within the deletion cassette (cassette_checkR) (Table 2.4). This combination yielded a product when assessing the deletion strain, but not the wild-type. The second combination used the same primer upstream of *scbA* (PscbA_fwd), and a primer present within the deleted region (PscbN_up) (Table 2.4). This PCR yielded a product when testing wild-type strains, but not Δ*scbAR* strains.

Construction of β-glucuronidase reporter strains

Transcriptional β-glucuronidase reporters were used to assess the promoter activity and transcriptional processivity associated with *scbA, scbR,* and *scbN.* To follow *gusA* activity driven by the native promoters of these genes, the promoter regions were PCR amplified (**Table 2.5, Table 2.5**) using cosmid StAH10 as template. The resulting fragments were then digested with XbaI and KpnI and ligated into pGUS (digested with these same two enzymes and dephosphorylated) using the Rapid DNA Ligation Kit from Roche (see General Methods for more detailed information) (**Table 2.3**).

pGUS-*ermE** was previously created by cloning the constitutive *ermE** promoter into the XbaI/KpnI restriction sites of the pGUS vector (St-Onge, 2016). Additional sequences from *scbA* were then cloned downstream of *ermE** in the KpnI/SpeI sites of pGUS, in order assess the impact of different sequences on transcriptional readthrough (**Table 2.3, Table 2.4, Table 2.5**), in the absence of any confounding transcription initiation effects.

For all of the above, successful clones were identified by colony PCR using primers flanking the multiple cloning site of pGUS (pGusF and pGusR – **Tables 2.4** and **2.5**). Confirmed constructs were then isolated using PureLink® Quick Plasmid Miniprep Kit as described in Chapter 2 (*Isolation of plasmid and cosmid DNA from* E. coli) and their sequences were confirmed by sequencing using primer gusAR2 (**Table 2.4**) by MOBIX at McMaster University. The plasmids were moved into *E. coli* ET12567 carrying pIJ8002 (**Table 2.2, Table 2.3**) for conjugation into *S. coelicolor* as described in the General Methods section.

Creation of hairpin disruption reporter strains

The hairpin disruption reporter strains (**Table 2.2**) were created as above, except that they were amplified off of a G-block (synthesized DNA) (**Table 2.4, Table 2.5**) of the *scbA* region that contained three point mutations predicted to reduce RNA secondary structure, while maintaining the same amino acid sequence and general codon usage (we did not want to change the translatability of this region by introducing a rare codon where one was not there previously). Specifically, a region of the *scbA* gene was changed from 5′-GGT GAC CGG TTC TCG GTC-3′ to 5′-GG**A** GAC CGG TTC TC**C** GT**G**-3′ (altered bases bolded and underlined). The first change, from GGT to GGA, was a synonymous change, as both codons specify a glycine. The GGT codon is used at a frequency of 9.2/1000 codons throughout the *S. coelicolor* genome, while the GGA codon is used at a frequency of 7.2/1000. The second change converted a serine codon from TCG to TCC; these occur at a frequency of 20.2 and 13.7 per 1000 codons, respectively. Finally, the third change involved a valine, changing a GTC (46.9/1000) to GTG (35.0/1000). While all of these changes were to codons used slightly less often than the wild-type codon, in all cases the new codon choice was the closest option in terms of frequency of use. After cloning, the sequence changes were confirmed by sequencing.

Qualitative β-glucuronidase assays

Confirmed reporter strains were streaked from spore stocks onto DNA, MYM, or SMMS agar plates containing 200 µL of 40 mg/mL 5-bromo-4-chloro-3-indolyl-β-D-glucoronide (X-Gluc) per 100mL and incubated at 30˚C. Photographs were taken at specific time-points.

Quantitative β-glucuronidase assays

For spectrophotometric *gusA* assays, reporter strains were grown overnight in biological triplicate in 8 mL YEME/TSB liquid medium at 30˚C while shaking at 200 rpm, in glass universal vials containing ~8-12 4 mm glass beads to help ensure effective hyphal dispersal. The next day, 400 µL of each overnight culture were sub-cultured into 8 mL of SMM before being grown as above. At the indicated time-points (16-72 h), 1.2 mL of culture was removed, with 200 µL of culture being harvested for *gusA* assays, and 1 mL being used for dry weight measurement.

For the *gusA* assays, cells were pelleted by centrifugation before being resuspended in 1 mL of lysis buffer (50 mM phosphate buffer at pH 7, 0.1% Triton X-100, 0.27% β-mercaptoethanol, and 4 mg/mL lysozyme, where phosphate buffer was

prepared using 8 g/L $Na_2HPO_4 \cdot 7H_2O$, 2.35 g/L $NaH_2PO_4 \cdot H2O$, adjusted to pH 7) and placed in a water bath at 37˚C for 30 min. Following lysis, cell debris was removed by centrifugation and the cell lysate was transferred to a fresh microcentrifuge tube. Fifty microlitres of cell lysate were assayed in triplicate in a 96-well plate, using phosphate buffer as a blank; 100 µL of Z-buffer containing 20 µL of 4 mg/mL p-nitrophenyl-β-D-glucuronide. An OD_{420} was measured every minute for 1 h at 37˚C. The values were graphed, and the slope of the resulting line was measured to determine expression levels, before being normalized to dry weight or mg of protein.

To determine dry weight, cells were first collected by centrifugation in pre-weighed microcentrifuge tubes. The cells were then dried at 60˚C for ~2 days, after which the tubes were weighed again and dry-weight was calculated. When normalizing to protein levels, protein concentrations in the cell lysate were assessed using a Bradford assay (BioRad) using bovine serum albumin to generate a standard curve.

Results

Genomic location of ScbN

RNA-sequencing experiments revealed that ScbN was approximately 350 nucleotides in length, with transcription initiation and termination contained entirely within – but on the strand opposite – the coding region of the γ-butyrolactone synthase-encoding gene *scbA* (**Figure 4.1**) (Moody et al., 2013). We predicted that transcription termination of ScbN likely occurred at a strong hairpin sequence. To verify this, we created two reporter constructs, each containing the *scbN* promoter, with one including, and one lacking, the predicted *scbN* terminator (Myronovskyi et al., 2011), and monitored reporter activity. The P*scbN-gusA* construct captured the *scbN* promoter and all associated upstream regulatory regions, while the second construct *PscbN* (with terminator)-*gusA* included these same regions but also extended downstream of the predicted terminator (**Figure 4.2A**). We found that P*scbN-gusA* was expressed during growth on both SMMS and MYM media, while P*scbN* (with terminator)-*gusA* was off under both conditions (**Figure 4.2B**). Taken together, these results confirmed the location of *scbN* as predicted by our RNA-sequencing experiments.

Dynamics of ScbN and *scbA* expression

To determine whether *scbA* expression was inversely correlated with that of *scbN,* we fused the promoter of *scbA* (*PscbA*) (extending ~40 nucleotides into the coding region of *scbA*) to the *gusA* reporter gene (**Tables 2.2-2.5**). We monitored the expression of P*scbN* and P*scbA* reporter strains grown in various liquid media (MYM, DNB, R5, SMM, YEME/TSB). We found that the *scbN* promoter was most active in the first 16 h of growth, and gradually decreased over the next 28 h until it reached a low steady-state level. In contrast, *scbA* transcripts were stably expressed through the first 44 h of growth, after which they dropped to nearly undetectable levels. Interestingly, *scbA* levels were significantly lower than those of *scbN* throughout growth. (**Figure 4.3**).

In situ expression of ScbN does not impact P*scbA* reporter activity

The timing of expression of *scbA* and *scbN* meant it was possible that they could influence the expression of each other. Given the integrated nature of *scbA* and *scbN,* it wasn't possible to explore the effects of *scbN* on *scbA* by simply creating a gene knockout of *scbN,* as deleting this gene would also lead to mutation of *scbA*. Hence, we took a different approach to determining the consequences of ScbN expression on *scbA*. We designed a comprehensive suite of *gusA* reporter constructs (Myronovskyi et al., 2011), with each capturing a different region of the *scbA* mRNA **(Tables 2.2-2.5, Figure 4.4A)**. To focus on the effects of *scbN,* we placed each of these reporters under the control of the constitutive promoter P*ermE**, which allowed us to eliminate any effects caused by transcription initiation, and helped to overcome the low expression levels from the native *scbA* promoter (**Figure 4.3**). Downstream of P*ermE** were cloned regions of the *scbA* mRNA that either included or excluded *scbN* (**Figure 4.4A**). Both of these reporters contained the *scbA* mRNA up to and including the entire predicted *scbA:scbN* interaction region. The first reporter excluded the downstream *scbN* promoter (P*ermE* scbA* without *scbN*), while the second extended far enough downstream to capture P*scbN* (P*ermE* scbA* with *scbN*). To eliminate the possibility that endogenous ScbN could act in *trans* on these reporters, reporter assays were performed in an Δ*scbAR* background, which lacks *scbN*. Unexpectedly, we found that there was no significant different in activity between these two reporters at any time-point examined; however, we noted a drastic decrease in expression when the promoter only construct was compared with either of the two longer *scbA* mRNA-containing constructs (please see below) (**Figure 4.4B**).

The reverse complement of the *scbN* terminator is a hairpin within the coding region of *scbA*

While ScbN did not impact *scbA* expression in our reporter assays, we did observe a significant decrease in expression of both of these strains compared to the 'promoter only' (P*ermE**) control (**Figure 4.4B** – compare P*ermE** to both *scbA*-containing reporters).

We considered two models that could explain this decrease in expression. The first involved the overlapping 5′ UTRs for *scbA* and the divergently expressed *scbR* (Takano et al., 2005). This promoter organization has been proposed to reduce expression of the associated genes, either through collisions between RNA polymerases during transcription, or through RNase-targeting of the overlapping/base-pairing RNA regions (Chatterjee et al., 2011; Mehra et al., 2008) (**Figure 4.1A**). Several mathematical models have been proposed for this system (Chatterjee et al., 2011; Mehra et al., 2008), and in both, the authors assumed that the overlapping 5′ UTRs would have a negative effect on both genes. This was a reasonable assumption, given that this is the case for the vast majority of well-studied asRNAs (Georg and Hess, 2011; Sesto et al., 2012; Thomason and Storz, 2010); however, this has not been experimentally validated for the *scbAR* system. The second model involved a very stable hairpin located within the coding region of *scbA,* on the strand opposite of the predicted *scbN* terminator (**Figure 4.5A** and **Figure**

4.5B). Hairpins in other systems have been shown to adversely impact gene expression (Svoboda and Di Cara, 2006).

To test the relative impact of both the divergent *scbR* transcription, and the *scbA* hairpin, a series of reporter constructs were created. All reporters were driven by the constitutive promoter P*ermE**, downstream from which were cloned different regions of the *scbA* mRNA (*e.g.* region containing the P*scbR* divergent promoter, with and without the *scbA* hairpin). In addition, both wild-type DNA and a synthesized *scbA* gene with three point mutations disrupting the hairpin were used as templates in PCRs (**Figure 4.6AB**).

These reporter constructs were introduced into the Δ*scbAR(N)* strain to avoid any complications caused by endogenous ScbN expression. In this series of expression vectors, reporters containing the hairpin were expressed at much lower levels (~4-15×) than reporter only controls (**Figure 4.6C**). Interestingly, we found that the divergent *scbR* promoter resulted in a ~3-fold decrease in expression relative to promoter only controls; however, presence of this region resulted in slight stabilization in constructs that also contained the hairpin. Disruption of the hairpin resulted in a 1.5 fold increase in expression of *scbA* in the absence of divergent transcription but no change in the presence of divergent transcription (**Figure 4.6C**). While this increase in expression was statistically significant, it did not have as profound an effect as we had expected. It remains possible that the hairpin is targeted by an RNase/RNA binding protein site which is still able to recognize the modified sequence. In addition, future work could be directed at investigating a possible protective role played by the divergent transcription of *scbR*.

A conserved hairpin is found in many *scbA* homologs

To determine whether the *scbA*-associated hairpin was conserved in other *Streptomyces* species, a custom Perl script was used to assess the mRNA structure of 59 *scbA* homologues from other *Streptomyces* species. Coding sequences were examined using a sliding window of 30 nucleotides, and the free energy (ΔG) associated with the structure of each 30 nucleotide sequence was graphed. We focused our attention on sequences containing strong hairpins (ΔG< -20 kcal/mol). Surprisingly, we found that many (24 of 59) *scbA* homologues contained strong hairpins, although these were not typically found in the same location as in *scbA*. Indeed, only *Streptomyces sp. Ncost-T6T-1* had a hairpin at a similar location as *scbA* (**Figure 4.7A** – Group 1). There was little sequence similarity shared between the predicted hairpins for *S. coelicolor* and *Streptomyces sp. Ncost-T6T-1*, suggesting that the hairpin structure – and not its sequence - may be functionally important.

Interestingly, there was a second conserved hairpin identified in 12 disparate streptomycetes (**Figure 4.7A** – Group 2). Unlike the *scbA*-associated hairpin, the sequences of these were quite similar, and consequently, we were able to examine these for of the presence of compensatory mutations that maintained secondary structure. This analysis was restricted to *scbA* hairpins from five streptomycetes in this group (*Streptomyces alboviridis, Streptomyces fulvissimus, Streptomyces luridiscabiei,* and *Streptomyces toyocaensis, Streptomyces cyaneofuscatus*).

There was evidence of compensatory mutations at a position in the stem-loop (**Figure 4.7B**). Here, a GUG-encoding valine was found in *S. toyocaensis*, while a GCG-encoding alanine was observed in *S. cyaneofuscatus*, and a GCC-encoding alanine was found in the other three species (**Figure 4.7C**). In each case, there were point mutations found in equivalent positions on the other side of the stem that effectively compensated for these changes, thus maintaining RNA structure (**Figure 4.7C**). Again, this suggested that maintenance of RNA structure in these regions was under selection. At this point it is not clear if this second conserved hairpin serves a similar role as that found in *scbA,* and determining whether this is the case is a high priority.

Discussion and future directions

Here we report the initial characterization of ScbN, an antisense RNA whose transcription initiates and terminates on the strand opposite the coding region of the GBL synthase-encoding gene *scbA* (**Figure 4.1, 4.2**). The promoter of *scbN* is significantly more active than that of *scbA* (**Figure 4.3**) at most times, and under the majority of growth conditions; however, expression of *scbA* did not appear to be affected by *scbN* expression (**Figure 4.4**). Instead, we found that a dramatic decrease in the expression of reporter genes encompassing regions of the *scbA* mRNA compared to promoter alone (**Figure 4.4**) was caused by a hairpin that comprised the reverse complement of the *scbN* terminator (**Figure 4.5 and Figure 4.6)**. Mutations predicted to disrupt the hairpin of the hairpin resulted modest increases in *scbA* expression (**Figure 4.6**).

Probing the mechanism of action of the *scbA* hairpin

We do not know how the region of *scbA* containing the hairpin functions to reduce gene expression. As mentioned above, it could act at transcriptional, translational, or mRNA stability levels. Also, we do not yet know if there are conditions under which the repressive effects of the hairpin are abolished, leading to increased expression of ScbA. Interestingly, there are discrepancies between when the P*scbA* promoter is active and when SCB1 accumulates to high levels (Takano et al., 2001). Our results indicate that *scbA* transcription initiation levels – but not transcriptional readthrough - are high prior to any perceived burst in SCB1 production (Takano et al., 2001). It is tempting to speculate that the repressive effects of the hairpin are alleviated prior to this burst in production.

Mutations predicted to relax the secondary structure of the hairpin did not completely abolish the hairpin-mediated repression (**Figure 4.6**). However, codon rarity and RNA secondary structures are inversely correlated along most mRNAs (Gorochowski et al., 2015): regions of high RNA secondary structures tend to be encoded by well-used codons, while regions with less stable RNA secondary structures tend to contain rarer codons (Gorochowski et al., 2015). The authors propose that this levels the translation rate across mRNAs. The corollary of this is that regions near rare codons need to be relatively unstructured for rapid translation (Gorochowski et al., 2015). There is a very rare isoleucine codon immediately upstream of the *scbA* hairpin. Hence it is possible that the combination of this rare codon, and a hairpin structure having a modest ΔG value (the

ΔG value of the disrupted hairpin was still ~2-fold higher than an average 30 nucleotide region in *S. coelicolor* mRNAs), was sufficient to keep *scbA* expression levels low.

What is the mechanism of repression of the *scbA* hairpin? One way to begin to answer that question is to take advantage of a library of either chemically or transposon mutagenized *S. coelicolor,* into which we could introduce our P*ermE* scbA* (hairpin) *gusA* reporter construct. Differences in regulation of the *scbA* hairpin-containing region should cause changes in expression, which could be quickly visually screened using a colourimetric screen (Myronovskyi et al., 2011). It would be especially interesting if RNase(s), RNA helicases or potential RNA modifying genes caused changes in expression.

Possible biological functions of the *scbA* hairpin

What role does the *scbA* hairpin play in the regulatory network governing quorum sensing in *S. coelicolor*? Are there conditions under which hairpin-mediated transcription repression is relieved? There are many questions that remain to be addressed within this fascinatingly complex system.

It is worth noting that *scbA-scbR* do not fit the cannonical quorum-sensing model common to most bacteria, including the A-factor system in *S. griseus*. In *S. griseus,* there is low-level expression of *afsA* (the *scbA* homologue), which contributes to a gradual rise in GBL levels as cell density increases. Once a threshold level is reached, a bistable switch is triggered, and the quorum-sensing system is rapidly activated. In SMM liquid media *S. coelicolor,* no SCB1 (GBL) was detected until transition into stationary phase, where it then rapidly accumulates. Paradoxically, under these same conditions the *scbA* promoter is active during early growth stages (Takano et al., 2001). Therefore, it is possible that the *scbA* hairpin in this species acts to dampen any leaky expression of *scbA,* and is responsible for the differences seen between *S. griseus* and *S. coelicolor,* as in *S. griseus,* the *afsA* mRNA does not have a hairpin.

The expression of genes involved in quorum-sensing in *S. coelicolor* was largely media-dependent. In the media-type most commonly used to study this system (SMM) our reporter assays indicate that *scbA* promoter activity remains constant until stationary phase (where it is turned off), this means that there may exist some means of overcoming the negative effects of the *scbA* hairpin at appropriate time-points during the *S. coelicolor* life cycle. The role of RNA-modifications, a booming area of research in eukaryotes, is in its infancy in bacteria. Some RNA modifications have been shown to alter RNA structure, and correspondingly impact the ability of RNA-binding proteins to associate with their substrates (Liu et al., 2015), and thus temporally modified *scbA* mRNA could influence the activity of associating proteins. Finally, there are some examples of mRNA-derived sRNAs where mRNA fragments, many of which are highly structured, negatively impact expression of its originating mRNA in *trans* (Williamson et al., 2017). It is therefore possible that the *scbA* hairpin functions in trans on other RNAs or proteins in the cell.

Given that ScbN is highly transcribed and expressed immediately opposite the coding region of *scbA*, it was surprising that it did not appear to impact the expression of *scbA*, at least in our studies here. We cannot exclude the possibility that ScbN influences

scbA expression in a way that was not detectable in our reporter assays. For example, in eukaryotes, histone proteins and epigenetic modification of DNA can together influence chromatin compaction. This can in turn alter expression of nearby genes by making them more or less accessible to the transcriptional machinery. asRNAs can influence the recruitment of both histones and DNA methyltransferases (Daniel Holoch et al., 2015). Bacteria do not have histones, and instead chromosome compaction is modulated by nucleoid-associated proteins (Browning et al., 2010; Dorman, 2013; Peeters et al., 2015). It is currently not known whether bacterial asRNAs influence nucleoid-associated protein recruitment. Since our reporter constructs are integrated into the phi-C31 phage-integration site (Myronovskyi et al., 2011), and not into the native *scbAR* locus, it is possible that features from the native site (such as local nucleoid-associated proteins and the associated chromosome dynamics) are not recapitulated at this heterologous location. For example, it is known that in eukaryotes and archaea chromosome compaction impacts transcription and vice versa (Mondal et al., 2011; Peeters et al., 2015). Therefore, it is conceivable that high levels of *scbN* transcription could influence the overall accessibility of the *scbAR* region, permitting a rapid activation of these genes under appropriate conditions. The DNA at the phi-C31 integration site may not be as strongly affected as at the native *scbAR* locus, and thus ScbN expression at this site may have little impact.

It is also possible that ScbN targets a protein or an mRNA other than *scbA* in the cell. It is common for trans-encoded intergenic sRNAs to target proteins (Marzi and Romby, 2012) or multiple mRNAs (Storz et al., 2011). In contrast, *cis*-asRNAs (like ScbN) are generally expected to regulate their cognate sense gene, due to both proximity (*e.g.* opportunity for colliding RNA polymerases) and complete complementarity (Georg and Hess, 2011; Sesto et al., 2012). To date, there is only one reported example of a *cis*-asRNA acting in trans. In *Staphylococcus aureus,* the $SprA1_{AS}$ asRNA is expressed opposite the *sprA1* gene, and negatively regulates its expression. However, the 5′ end of the antisense RNA does not overlap with the coding region of the sense gene, and it is this region – and not the region of overlap - that is important for regulation of *sprA1*(Sayed et al., 2012). The region of the asRNA that did not overlap with the sense gene had complementarity to a region near the RBS of the *sprA1* mRNA, and its binding in *trans* negatively regulated translation. It is therefore possible that an asRNA that *does* overlap with its sense mRNA could also target other RNAs or proteins.

To investigate other possible protein or RNA targets of ScbN, affinity purification could be employed. *scbN* could be fused to an MS2 tag, where this RNA tag is specifically recognized and bound by the MS2 protein. The MS2 protein could then be used to pull down the *scbN* fusion, in association with any RNA/protein interacting partners (Rieder et al., 2012). Interacting proteins could be identified using mass-spectrometry, while any associated RNAs could be detected using RNA-sequencing or proteomic analyses. In addition, RNA-sequencing experiments could be used to compare the global gene expression of an Δ*scbAR* strain, with a strain complemented only with *scbN*. Differences in RNA expression might highlight potential target RNAs (McGettigan, 2013).

Finally, it is possible that ScbN has no defined function in the cell. This argument has been put forward by some groups based on the pervasiveness and lack of conservation

of many asRNAs, suggesting that these products are merely the result of spurious transcriptional noise and have no function (Lloréns-Rico et al., 2016; Raghavan et al., 2012). While this is possible, many sequences previously thought to be non-functional [*e.g.* junk DNA (Pennisi, 2012), and pseudogenes (Prieto-Godino et al., 2016)], are now understood to serve important functions. It is also possible that the mere act of asRNA transcription could have beneficial consequences (Wade and Grainger, 2014), even if the resulting RNA has no dedicated function. For example, active transcription detects and recruits complexes that repair DNA damage. As a consequence, poorly expressed genes tend to accumulate more mutations than more highly expressed genes (Savery, 2007). It is possible that a function of some asRNAs could be to reduce mutation rates at genes that are not commonly expressed (Wade and Grainger, 2014).

Non-coding RNAs involved in quorum-sensing in other species

The regulation of quorum-sensing is incredibly diverse across the bacterial kingdom (Ng and Bassler, 2009) making it difficult to highlight major themes common throughout bacteria. However, one common theme is the central role that non-coding RNAs play in these systems. At low AI-2 (the quorum-sensing signal in these species) concentration in *Vibrio,* phosphorylated LuxO activates the transcription of four *qrr* sRNAs (Lenz et al., 2004, 2005). The *qrr* sRNAs go on to repress the mRNA of the quorum-sensing master regulator *luxR,* and AI-2 synthase *luxM,* while activating the low-cell density regulator *aphA* (Feng et al., 2015; Shao and Bassler, 2012). Another example of a sRNA involved in regulating quorum-sensing is the RNAIII sRNA in *Staphylococcus aureus.* This sRNA is activated at high densities of the signaling peptide and acts as the main high-density regulator, activating virulence genes and repressing surface proteins, essential for evading the host immune system (Bronesky et al., 2016). While *trans*-acting sRNAs are common and well-studied in quorum-sensing, *cis*-acting antisense RNAs are not. To my knowledge there is only one characterized antisense RNA involved in quorum-sensing in bacteria, the as*rpaR* RNA from *Rhodopseudomonas palustris*. This antisense RNA is expressed opposite the signal receptor *rpaR* (not the synthase as in the *scbN* case) and acts like a canonical antisense RNA by repressing translation (Hirakawa et al., 2012).

How many antisense RNAs might not effect expression of their sense gene?

Thousands of antisense RNAs are known in bacteria; however, almost all characterized antisense RNAs are expressed and modulate accessibility to the RBS (Georg and Hess, 2011; Thomason and Storz, 2010). Other types of antisense RNAs have lagged behind in terms of characterization. My work in this Chapter, and other work from our lab (Hindra et al., 2014) have found that some asRNAs do not dramatically impact their corresponding sense gene expression.

Searching for other mRNAs throughout bacteria with conserved hairpins

Hairpins within coding sequences have been largely overlooked as potential regulatory elements. Instead, most groups have focused on the regulatory potential of structures within untranslated regions of mRNAs. However, mounting evidence is

suggesting that coding sequence-associated hairpins can have profound effects on gene expression (Song et al., 2012; Svoboda and Di Cara, 2006). In the next chapter of this thesis we describe the discovery of hundreds of conserved RNA structures throughout mRNA coding regions throughout the bacterial kingdom, and speculate about their importance. It is possible that many serve regulatory functions like that of the *scbA* hairpin, or may promote the evolution of asRNAs.

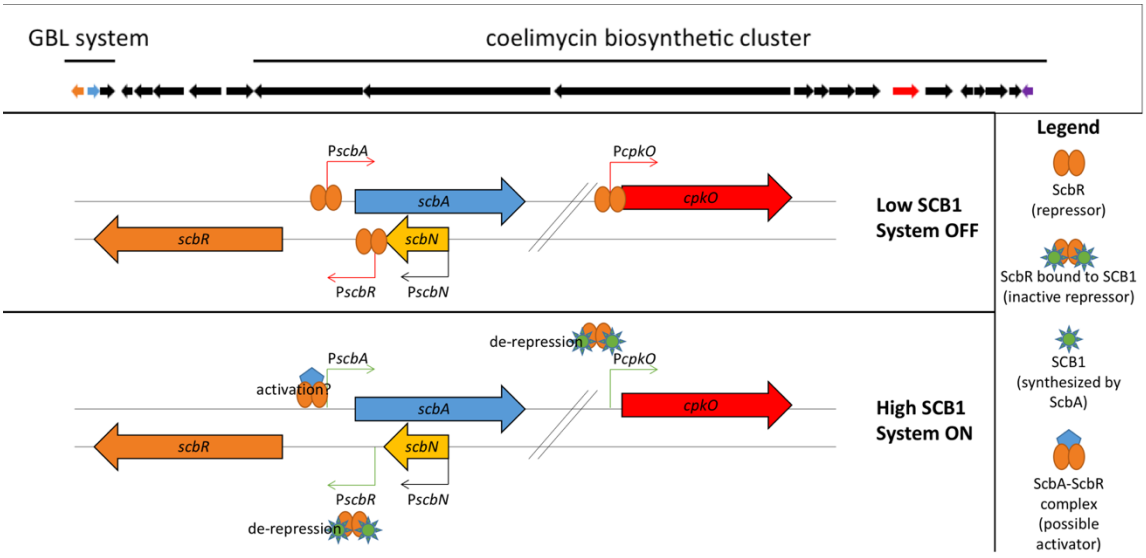


Figure 4.1. Overview of the *Streptomyces coelicolor* quorum-sensing system. The GBL genes (*scbA, scbR*) are located next to the coelimycin cluster. The pathway specific activator of the coelimycin cluster is CpkO. At low SCB1 levels, ScbR binds to and represses transcription from its own promoter, and that of *scbA* and *cpkO*. At higher SCB1 levels a non-repressive SCB1:ScbR complex is formed resulting in de-repression of the *scbR* and *cpkO* promoters. Transcription of *scbA* requires both *scbA* and *scbR*. The mechanism of activation of *scbA* is not yet clear; however, some groups have proposed that a ScbA:ScbR complex could act as an activator. The promoters of *scbA* and *scbR* are divergently arranged, such that there may be interactions at the transcriptional (RNA polymerase collisions) or post-transcriptional (antisense RNAs) level of these two mRNAs. The antisense RNA ScbN is expressed from the strand opposite the coding region of *scbA*. At the right-end of the cluster is shown *scbR2* (purple). ScbR2 acts in a similar manner to ScbR, except that instead of transcriptional repression is relieved by binding to actinorhodin, undecylprodigiosin, and jadomycin (Wang et al., 2011; Xu et al., 2010), instead of SCB1.

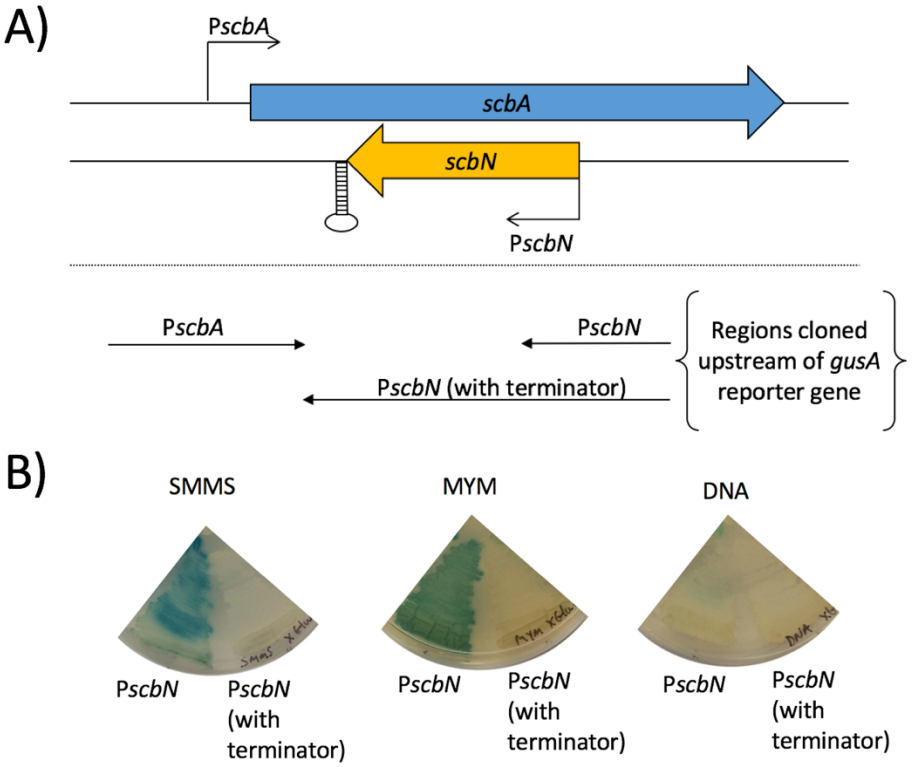


Figure 4.2: The *scbN* promoter and terminator are both located within the coding region of opposite *scbA*. A) Schematic diagram showing the location of the predicted *scbN* promoter and terminators relative to the sense gene *scbA*. Below the gene diagram are shown various regions (directionality shown by small arrows) of regions cloned upstream of *gusA*. The P*scbN* reporter captures the *scbN* promoter, while P*scbN* (with terminator) includes both the *scbN* promoter and its predicted terminator. **B)** Solid *gusA* assays on SMMS, MYM, and DNA media containing X-Gluc, which turns blue in the presence of the product of *gusA*.

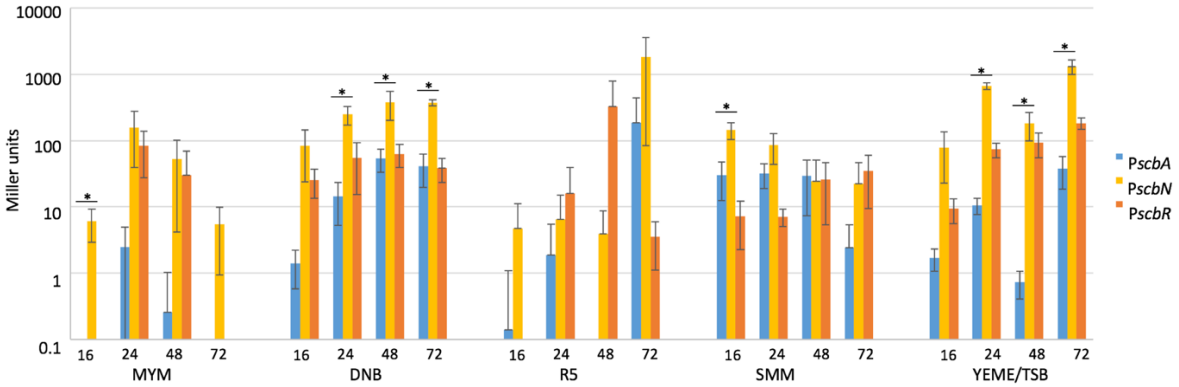


Figure 4.3: Dynamics of *scbA, scbN* and *scbR* expression. M1146 P*scbA*-, P*scbN*-, and P*scbR*-*gusA* fusions were grown in MYM, DNB, R5, SMM and YEME/TSB liquid media. Cells were harvested at 16, 24, 48 and 72 h and *gusA* assays were performed to measure gene expression. Expression patterns differed greatly depending on media type. In general, P*scbN* and P*scbR* were expressed at higher levels than P*scbA*. P*scbN* was more active than P*scbA* in many conditions (indicated by asterisks, $p < 0.05$), suggesting that ScbN has the potential to affect *scbA* expression.

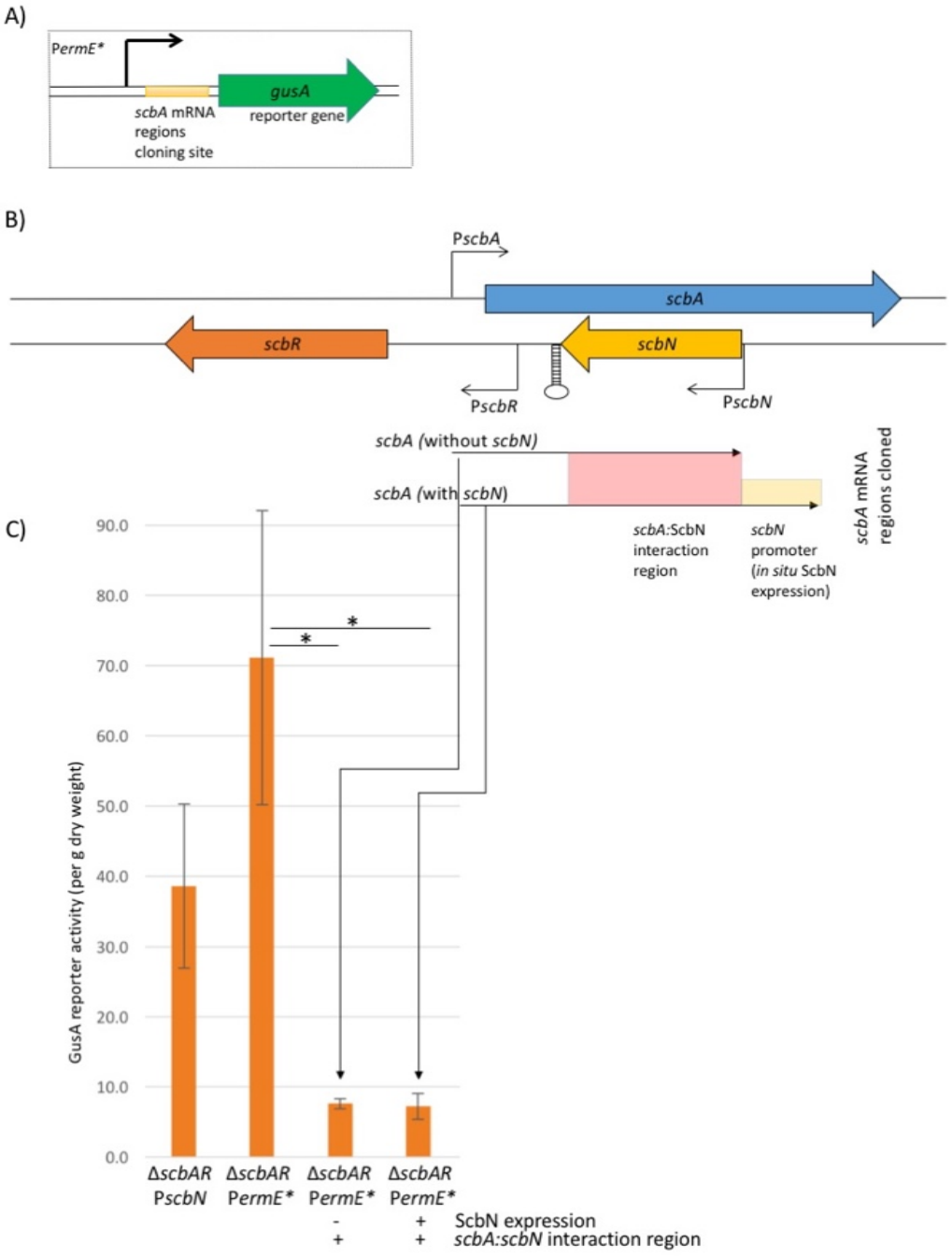


Figure 4.4: ScbN expression does not impact *scbA* expression. A) The P*ermE*-gusA* reporter system. Various regions are cloned between the constitutive promoter P*ermE** and the reporter gene *gusA.* **B)** Schematic of the *scbA, scbR* and *scbN* regions. Underneath are shown regions cloned into the P*ermE*gusA* reporter system shown in A). The region of hypothetical *scbA:scbN* interaction is highlighted in red, while the region responsible for *in situ* ScbN expression is highlighted in yellow. The two regions cloned

downstream of P*ermE** are indicated with horizontal arrows. *scbA* (without *scbN)* encompasses the *scbA* mRNA from the transcriptional start, but excludes the *scbN* promoter, while *scbA* (with *scbN*) extends further downstream to capture the *scbN* promoter on the opposite strand. **C)** GusA reporter assays showing expression of P*scbN*, P*ermE**** (positive control), and P*ermE**** driving the expression of the two regions of the *scbA* mRNA (with and without *in situ scbN* expression). There was no difference between the latter two constructs, while there was a large drop in expression comparing promoter alone to either one of the two *scbA* mRNA expression vectors ($p < 0.05$).

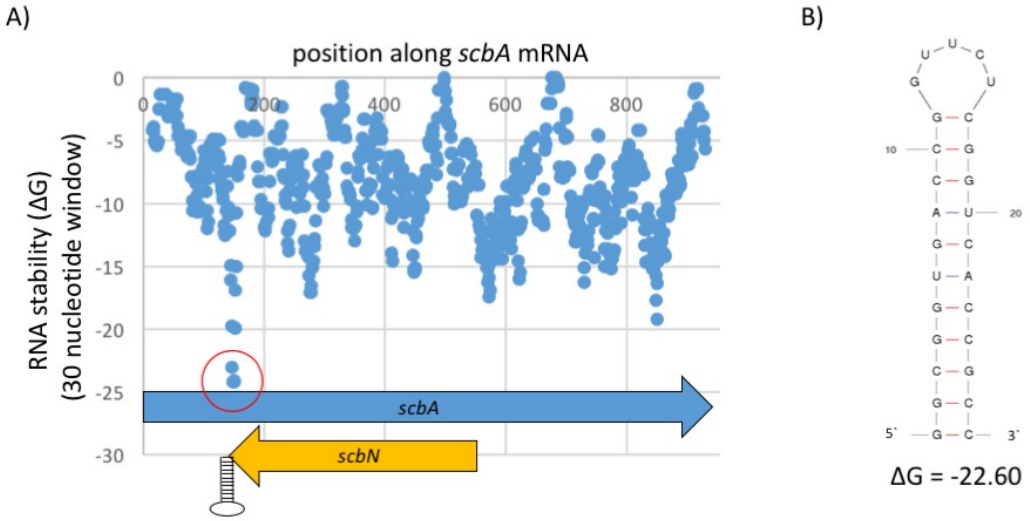


Figure 4.5: An mRNA hairpin reduces *scbA* expression. **A)** Schematic of the *scbA* mRNA and *scbN* showing calculated RNA stability (ΔG) for *scbA* mRNA, measured with 30 nucleotide windows across the entire coding region. The location of a strong hairpin, opposite the predicted *scbN* termination site, is circled in red. **B)** Predicted secondary structure of the *scbA* mRNA hairpin as predicted by RNA fold.

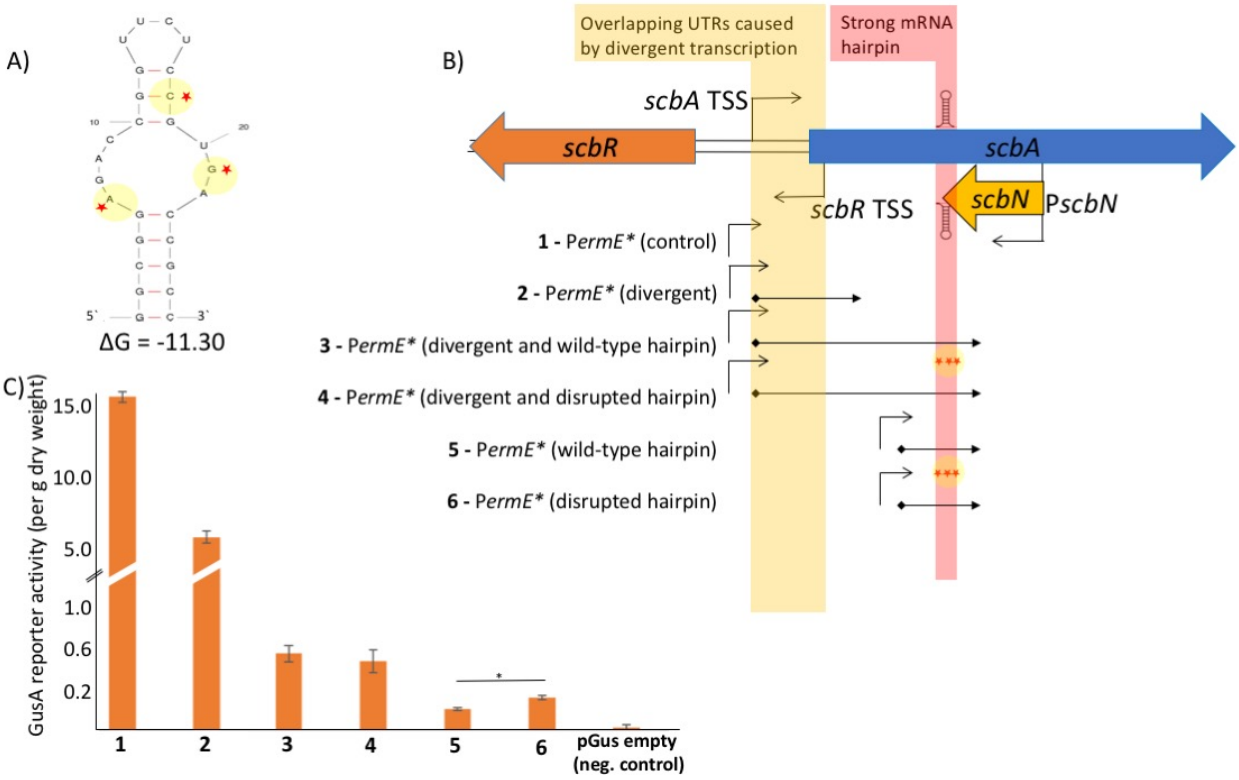


Figure 4.6: Disruption of the *scbA* hairpin has a small impact on *scbA* expression. A) Predicted secondary structure of the *scbA* mRNA hairpin following disruption as predicted by RNA Fold showing calculated ΔG value. Mutated bases are shown highlighted in yellow with red stars. **B)** Schematic diagram of the *scbA, scbN* and *scbR* region highlighting the regions important for divergent transcription (yellow highlight) and the strong hairpin (red highlight). Regions cloned into the *gusA* reporter constructs are shown below, and each of these is numbered and correspond to the numbers for each column in graph in (C). **C)** GusA reporter assays showing expression of various regions of the *scbA* mRNA. Divergent transcription results in a slight decrease in expression relative to promoter only control, but an increase in expression relative to constructs containing the hairpin. Disruption of the hairpin has a minor (~1.5 fold, $p < 0.05$) effect in the absence of divergent transcription.

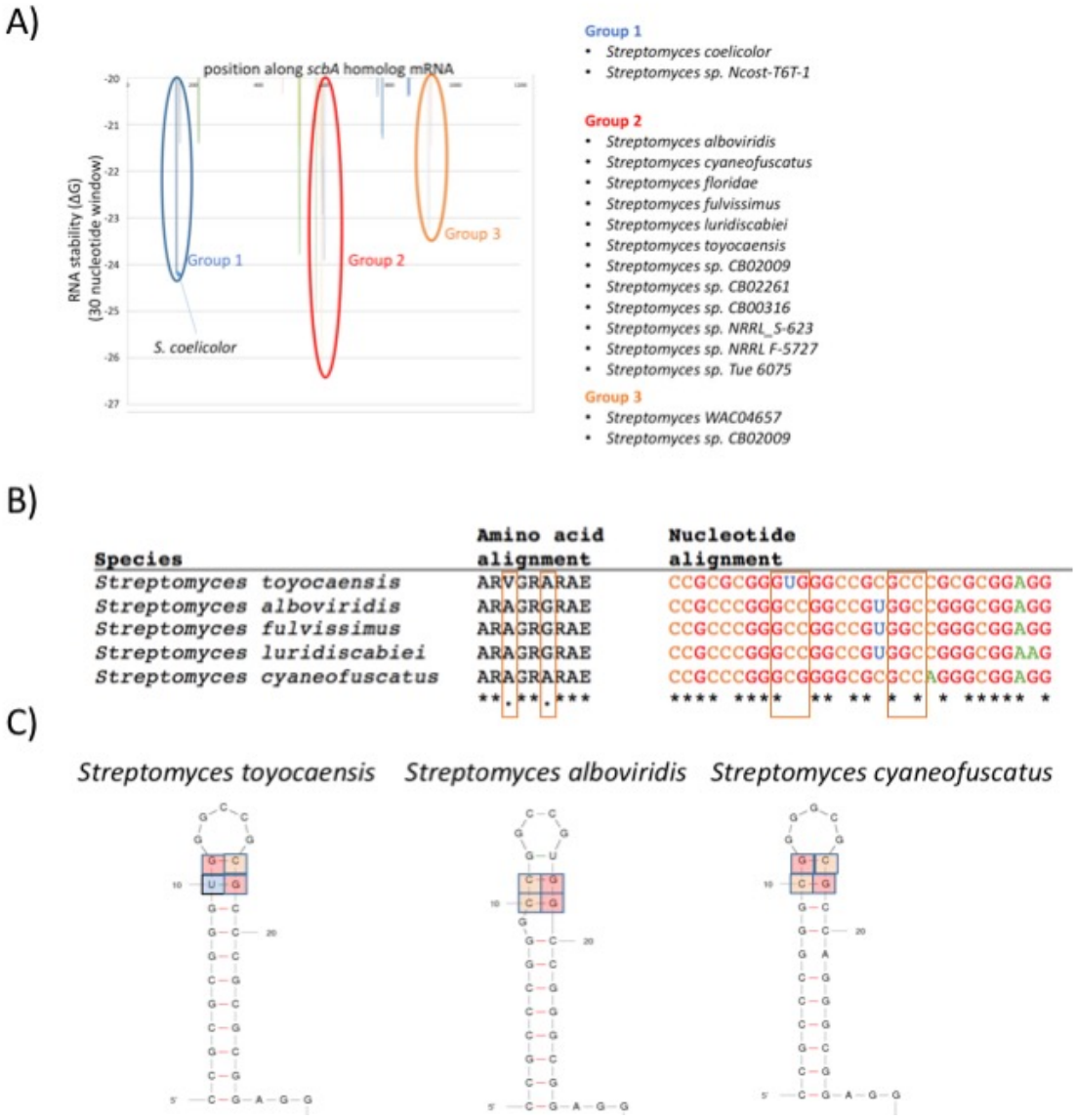


Figure 4.7. Hairpins at conserved locations show evidence for compensatory mutations in other *Streptomyces* species: A) Schematic of the *scbA* mRNA for various *scbA* homologs. The y-axis scale has been formatted to show only values less than -20 kcal/mol (*i.e.* very stable hairpins). Three groups of species have hairpins at similar locations. **B)** Detailed analysis of the most common hairpin – Group 2. Nucleotide and amino acid alignments with variable amino acids and corresponding codons boxed in orange. **C)** Compensatory mutations exist at corresponding positions in the opposite side of the stem loop that maintain RNA secondary structure at non-synonymous codons.

Chapter 5: The role of stable local structures in mRNAs

Introduction

Bacterial messenger RNAs encompass a coding region that is translated, with this sequence often being flanked by two untranslated regions – the 5′ and 3′ UTRs. While the coding sequences of mRNAs have historically garnered the most attention, the two untranslated regions can house many possible regulatory elements. For example, riboswiches can be found in the 5′ UTRs of some mRNAs, where they bind to small cellular ligands and affect the expression (through a variety of mechanisms) of the downstream coding sequence (Mandal and Breaker, 2004). Regions in UTRs can also be targeted by other RNA molecules, with many asRNAs and sRNAs binding at or near the ribosome binding sites, altering the translatability of the target mRNA (Thomason and Storz, 2010; Waters and Storz, 2009). Recently the 3′ UTRs have been found to be a major target for the regulatory RNA chaperone Hfq (Chao et al., 2012).

The coding regions of mRNAs, perhaps unsurprisingly, are typically not considered much beyond their coding sequence capabilities, with their sequences being constrained by the relative conservation and importance of the amino acids they encode. Consequently, most searches for RNA structures with regulatory potential have not focused on these regions. It is important to note, however, that given the redundancy of the genetic code (*i.e.* almost all amino acids can be encoded by more than one codon), there is the potential for selection acting on coding sequences beyond their amino acid specification. For example, codon usage itself can impact mRNA translation, with codons corresponding to rarer tRNAs being translated more slowly than those decoded by more abundant codons (Chaney and Clark, 2015; Kolmsee and Hengge, 2011; Plotkin and Kudla, 2011; Quax et al., 2015). In *Streptomyces* this is exemplified by the *bldA* tRNA, whose corresponding TTA codon is present in many genes, but none whose product is essential. Consequently, the *bldA* tRNA is dispensible for viability. There are, however, several TTA-containing genes whose products are important for development and antibiotic production. For example, there is a TTA codon in the important morphological developmental gene *adpA*, and as such, strains lacking the *bldA* tRNA fail to raise aerial hyphae. BldA activity also impacts specialized metabolism, with TTA codons found in the genes for the pathway specific activators *actII-orf4* (actinorhodin) and *redZ* (undecylprodigiosin) (Chater and Chandra, 2008; Leskiw et al., 1993; Takano et al., 2003).

In addition to codon usage directly influencing translation, the redundancy in the genetic code also permits some limited flexibility in sequence – and correspondingly in the RNA secondary structures that can form within mRNAs (Mathews et al., 2010; Petrillo et al., 2006; Svoboda and Di Cara, 2006). RNA secondary structures can have significant impacts on transcription and translation (Ciampi, 2006; Gorochowski et al., 2015; Mitra et al., 2009; Peters et al., 2011; Santangelo and Artsimovitch, 2011). For instance, they can be targeted for binding by regulatory RNA binding proteins (Babitzke and Yanofsky, 1993; Figueroa-Bossi et al., 2014; Jutras et al., 2013; Phadtare and Severinov, 2010) and RNases (Deutscher, 2015; Mackie and Genereaux, 1993). They can

further promote RNA localization (Hartswood et al., 2012; Nevo-Dinur et al., 2011, 2012), and can affect ribosomal frame-shifting (Kim et al., 2014a). With the exception of the impact of secondary structures involved in translation and RNase cleavage sites, there has been little study of structured RNAs within coding sequences, and to date, most investigations have focused on the 3′ and (especially) the 5′ UTRs. My previous work in Chapter 3 (Moody et al., 2013) found that short stable regions (30-80 nts) within coding regions of genes can be greatly over-represented compared to the rest of the mRNA. It is therefore possible that these short regions can accumulate and act as sRNA-like molecules on other cellular targets.

Recent advances in sequencing technologies have allowed for genome-wide probing of RNA secondary structures. These new techniques exploit the unique susceptibility of different RNA structures to cleavage or modification by RNases or chemicals, and can be coupled with massively parallel RNA-sequencing. For example, a technique termed PARS (parallel analysis of RNA structure) involves treating separate pools of cellular RNAs with RNase S1 and V1. These nucleases target single and double stranded RNAs, respectively, and when coupled with RNA-seq, can yield genome-wide information on the single/double stranded nature of any RNA at single-nucleotide resolution (Kertesz et al., 2010). More recently, a technique termed SHAPE-seq uses the chemical 1-methyl-7-nitroisatoic anhydride to selectively acetylate the 2′ hydroxyl of flexible nucleotides (*i.e.* those nucleotides not base-paired). This 2′-acetylation blocks reverse transcriptase, and thus, by combining randomly primed reverse transcription with deep sequencing, global RNA structures can be determined (Aviran et al., 2011; Lucks et al., 2011). While powerful, these methods do, however, have their limitations. They can only be used to define the structure of moderately to highly expressed mRNAs, and those structures must form under the conditions in which the experiment was performed. On their own, they also provide no insight into any structural conservation, in the absence of equivalent analyses performed in multiple species.

The work described in Chapter 4 revealed a hairpin within the *scbA* gene that dramatically reduced its expression. Bioinformatic searches for similar hairpins at this location in *scbA* homologs in other species suggested that while hairpins were common, they were rarely found at analogous locations. Interestingly, it has been reported that GBL synthases and receptors are evolving at an unusually rapid pace (Nishida et al., 2007), and thus the lack of direct conservation is perhaps not surprising. We wondered whether there might exist hairpins within genes that were conserved across many bacteria, and this question served as the driving force behind the work described below. This chapter describes the development of a Perl script designed specifically for identifying coding sequence-associated hairpin structures, and assessing their degree of conservation in multiple bacterial species. I describe some well-conserved hairpins that will be of interest for future study.

Materials and Methods

Coding to find conserved hairpins within mRNA coding sequences

A Perl script, dubbed LocoRNAStrut (Location conservation of RNA structures), was created to search for local regions capable of forming stable secondary structures (**Appendix 2**). Two sets of FASTA nucleotide files were used as input for the script. The first FASTA file contained the query sequences, which could represent anything from a single gene, to all of the genes from an organism of interest (*i.e.* downloaded coding sequences in Genbank). The second FASTA file comprised the nucleotide coding sequences from organisms within the group in which you want to identify conserved hairpins found in the query species. This second group of FASTA files were saved separately, in a folder named './others/'. To streamline the process, and minimize redundancy, a script - 'uniquify-fasta.pl' (Ben Furman, unpublished) - was used to remove any duplicate FASTA headers (**Figure 5.1).**

Once all duplicate FASTA headers had been removed, the newly created program translated all sequences in the './others/' folder, and used the resulting amino acid sequences to build a BLAST database (Altschul et al., 1990). The nucleotide sequences from the query species were then also translated and BLAST was used to identify potential homologues within the 'other' sequences. Significant homologues (E-value < 10^{-9}) were aligned using MUSCLE (Edgar, 2004) and these amino acid alignments were used to generate a corresponding *aligned* nucleotide sequence (**Figure 5.1**).

For each set of query species/potential homologs from the 'other' species, the nucleotide sequences were then divided into short overlapping windows. The default state for this portion of the program was set to a 30 nucleotide window, with a one nucleotide slide size, such that the window would 'slide' by one nucleotide with each iteration. These short sequences were then computationally folded by RNAFold, and the minimum free energy (ΔG) was recorded. GC content varied considerably between species of interest (*e.g. Streptomyces* genomes were >70% GC while others like many Firmicutes were <30%), and GC content is a large determinant of ΔG. Therefore, minimum ΔG cutoffs were determined for each species by taking the average ΔG value across mRNAs and multiplying by an investigator-determined factor (default was set to 2.5). This would result in the identification of thermodynamically stable regions if they were 2.5 times more stable than the average 30 nucleotide region across all genes (**Figure 5.1**).

Next, the position of stable regions of interest were compared between the query species and the other species. Here, the interest was in conserved hairpins, and as a result, hairpins positioned within 10 nucleotides one of each other in different species were flagged for follow up (**Figure 5.1**). A sample output of the program is shown in **Figure 5.2.**

Results and Discussion

The Perl script successfully identifies known hairpins

To test LoCoRNAStrut, we ran the program using *E. coli* as query species, against twelve other species: *B. subtilis, Streptococcus pyogenes, S. coelicolor, S. griseus, S. scabies, S. avermitilis, Haemophilus influenza, Helicobacter pylori, Salmonella enterica, Pseudomonas aeruginosa,* and *Yersinia pestis*. The script identified widely conserved hairpins in important genes such as *dnaE, clpP*, and *ftsZ* (**Table 5.1**).

As a control, to ensure that the script successfully identified known secondary structures of regulatory importance, we confirmed that LoCoRNAStrut correctly identified a hairpin known to impact expression of the RNA chaperone Hfq. Hfq is required in many bacteria to facilitate the interaction between non-coding RNAs and their mRNA targets (Sobrero and Valverde, 2012; Vogel and Luisi, 2011). In *E. coli,* Hfq binds to two RNA sites within its own mRNA, autoregulating its own expression. Both Hfq binding sites are located within the 5′ UTR; however, a hairpin sequence downstream of the translation start site (centered at +41) is critical for Hfq-mediated repression (Vecerek et al., 2005). LoCoRNAStrut found that this hairpin was conserved in *H. influenza, Y. pestis*, and *S. enterica.* To see if this hairpin was commonly found in other related bacteria, the search was manually expanded to assess the conservation throughout the Enterobacterea family, and confirmed hairpin presence in *Shigella flexneri, Ctirobacter rodentium, Kosakonia radicincitans, Klebsiella pneumonia, Cedecea neteri, Pluralibacter gergoviae* and *Serratia marcescens* (**Figure 5.3**).

These results suggest that LoCoRNAStrut is capable of finding hairpins of regulatory importance within coding regions of mRNAs. Having established its utility, LoCoRNAStrut was then used to assess the hairpin potential of *S. coelicolor* mRNAs.

Select hairpins with locational conservation between *Streptomyces* and other bacteria

The script was run using all *S. coelicolor* mRNAs as query, which were compared with all mRNAs in *S. venezuelae*, *M. tuberculosis,* and the more distantly related Gram positive bacterium *S. pyogenes* (**Table 5.2**). A total of 2,569 potential hairpins were identified in *S. coelicolor*. Of these, 193 (~8%) were conserved in *S. venezuelae*, 46 (~2%) in *M. tuberculosis,* and 91 (~4%) in *S. pyogenes*. It was surprising that more hairpins were conserved between *Streptomyces* and *Streptococcus*, than between *Streptomyces* and the more closely related *Mycobacterium*; however, this might have been due low GC content of *Streptococcus* resulting in lower ΔG cutoffs and more potential hairpins to analyze.

Hairpins at conserved locations in at least one other species (**Table 5.3**) were further analyzed in additional bacteria, including: *S. avermitilis, S. scabies, B. subtilis, E. coli, S. enterica, H. influenza*, *H. pylori*, *P. aeruginosa, Y. pestis,* and the human mitochondrion. Conserved hairpins (both structure and location) in the greatest numbers of species were collated and summarized in **Table 5.4**. Below are highlighted hairpins that show remarkable positional conservation across many groups of bacteria.

***rpsJ* hairpin –** *rpsJ* encodes NusE (also known as the 30S ribosomal subunit protein S10), which is involved both in antitermination and in coupling transcription and translation. In bacteria, the coupling of transcription and translation means that efficient translation can involve ribosomes enhancing transcriptional processivity by pushing the transcribing RNA polymerases forward (Proshkin et al., 2010). Conditions that slow translation (*e.g.* amino acid depletion or RNA secondary structure) can cause a separation between the ribosome and the RNA polymerase resulting in early transcriptional termination. This occurs due to the action of Rho, in conjunction with NusG: Rho binds

to naked mRNAs (unprotected by translating ribosomes), and NusG associates with both Rho and RNA polymerase to promote transcription termination. The interaction between Rho and NusG can be abolished through competition with NusE, which can also complex with NusG (Burmann et al., 2010; Santangelo and Artsimovitch, 2011).

Across diverse bacteria, *rpsJ* is the first gene in a large operon that encodes many ribosomal proteins (Cortes and Cox, 2015; Zurawski and Zurawski, 1985). Hence it isn't surprising that it would be under tight control. In fact, in *E. coli,* ribosomal proteins (*e.g.* ribosomal protein L4) are known to bind to hairpins within the 5′ UTR of *rpsJ* in an autoregulatory manner (Freedman et al., 1987; Johnsen et al., 1982). In addition to the hairpins found in the 5′ UTR of *rpsJ,* we found a second highly conserved hairpin centered at ~105 nucleotides into the coding region in *S. coelicolor, S. venezuelae, M. tuberculosis, S. pyogenes, E. faecalis*, *S. aureus*, and *B. subtilis* (**Figure 5.4**). It would be interesting if this hairpin also plays a role in autogenous regulation of the cluster, being a binding site of one of the many ribosomal proteins encoded in the *rpsJ* operon, or acting in an analogous fashion to the *hfq* example discussed above.

***purA / purR* hairpins** – *purA* (*SCO3629*) encodes adenylosuccinate synthetase, an enzyme required for purine biosynthesis that is widely distributed throughout bacteria (Wolfe and Smith, 1988). A conserved hairpin was located near position ~908 in *S. coelicolor, S. venezuelae, M. tuberculosis, Y. pestis, B. subtilis,* and *H. pylori* (**Figure 5.4A**). *purR* (*SCO1186*) encodes a regulatory protein that in *E. coli* represses *purA* at the transcriptional level (He and Zalkin, 1994). While not as conserved as the *purA* hairpin, we found a hairpin in *S. coelicolor* that had sequence complementarity within the loop of the stem loop compared to the loop within *purA* (**Figure 5.4B**). The presence of a complementary loops in two genes that have a known regulatory connection was intriguing as it is possible that in addition to PurR regulation of *purA,* there may be a regulatory connection at the mRNA level as well. This could be either through interactions between full length mRNAs, or the degradation product of one mRNA acting as an sRNA on the second mRNA.

***prfA* hairpin** – Protein release factor 1 is responsible for releasing ribosomes at UAA and UAG stop codons (Elliott and Wang, 1991). It achieves this by mimicking tRNAs that pair with UAA and UAG stop codons and promoting ribosome release (Ito et al., 1996; Laurberg et al., 2008). A hairpin was found at a conserved location (~480 nucleotides into the coding sequence) in many diverse bacteria, including *Streptomyces* species, *M. tuberculosis, Y. pestis, H. pylori, B. subtilis, S. pyogenes,* and *E. coli* (**Figure 5.6**). Very little is known about the regulation of *prfA*; however, giving the conservation of the location of this hairpin across both Gram negative and positive bacteria, it would be interesting to determine if it has a role in regulating this critical gene.

Conserved hairpins are not associated with highly abundant mRNA-associated sRNAs

In Chapter 3 we found many short (30-90 nucleotide) mRNA-derived transcripts were highly abundant, and these were presumed to be highly stable regions. It was not clear whether these transcripts were short degradation products or if they were transcribed

from their own promoter (Moody et al., 2013). We compared the most highly expressed of these mRNA-associated sRNAs in *S. venezuelae* discovered by RNA-seq (Moody et al., 2013), with our bioinformatically generated list of conserved hairpins. Twelve of 17 of these genes had homologs in *S. coelicolor*; however, none of these mRNAs had hairpins that met our stability criteria here. Many of these short stable regions were >50 nucleotides long, while the conserved hairpin-finding script was run with a window size of 30 nucleotides. It will be interesting to alter the window size of the hairpin-finding script, to determine whether there are any correlations in the two data sets.

Conserved hairpins are not located opposite conserved antisense RNAs

In Chapter 4 we found that the *scbN* terminator was capable of forming a hairpin in the sense direction, and that this had a major impact on RNA polymerase processivity through *scbA*. We cross-referenced the lists of the antisense RNAs that were conserved between *S. coelicolor, S. venezuelae* and *S. avermitilis* (**Table 3.2**), with our list of hairpins having conserved locations. There was no overlap between these two RNA sets; however, this might be due to the limited number (13) of antisense RNAs conserved between these species (Moody et al., 2013).

Future characterization of hairpins of interest

Having identified a number of well-conserved hairpins that were distributed widely across major groups of bacteria (*e.g. purA, rpsJ, prfA*), the next step will be to establish whether these hairpins serve a functional purpose.

It seems most likely that these structures would exert a regulatory effect at transcriptional, post-transcriptional, or translational levels. Transcriptional and translational fusions could be used to monitor the effects of the hairpin on transcription and translation respectively. This could be performed by comparing the activity of reporters of the gene that include or exclude the hairpin. Alternatively, mutant variants could be assessed, whereby the hairpin structure is disrupted (but the amino acid sequence is unaffected), analogous to experiments carried out in Chapter 4 to analyze the *scbA* hairpin.

It is also possible that some hairpins may adopt conformations that are recognized and bound by proteins or metabolites. A bioinformatics survey using RFAM (Griffiths-Jones et al., 2003; Nawrocki et al., 2015) could reveal any consensus motifs. As well there is a growing list of software that bioinformatically predicts RNA-binding proteins and their targets: SVMProt (Han et al., 2004), APRICOT (Sharan et al., 2017), RBPMap (Paz et al., 2014). There also exist protocols (e.g. aptamer tagging of sRNAs combined with affinity chromatography) for experimentally identifying RNA binding proteins that associate with RNAs of interest (Rieder et al., 2012). These techniques could be performed either in an unbiased manner using whole cell extracts, or in a more directed way – for example, with *rpsJ,* any binding may be by any of a number of ribosomal proteins that can act as post-transcriptional regulators (Cortes and Cox, 2015; Zurawski and Zurawski, 1985).

Enhancing the predictive power of LoDoRNAStrut through program optimization

A major bottleneck in analyzing the LoCoRNAStrutoutput is the time it takes the script to run. Run time of the script using *S. coelicolor* as the query sequence against four other species exceeded one week. A few minor changes (some of which were attempted but caused small glitches) would be expected to lead to more efficient analysis.

One stage that could be significantly streamlined would be the homologue detection stage. Currently, BLAST is used to find potential homologues, after which a filtering algorithm is used to identify the best hit for each species. BLAST has built in options for ensuring that only one hit is reported for each species; however, this option was incompatible with the output format required for the rest of the program. Utilizing specialized homologue finding scripts (*e.g.* Fong et al., 2008) would likely speed up this process and be more accurate (see below).

The hairpin-identification stage is also one where additional optimization would be beneficial. Here, hairpins were searched for within every mRNA query sequence, across the entire mRNA length, for a given focus species. The resulting hits were then compared to see if hairpins were found at similar locations in other species. An alternative approach could involve identifying hairpins in the query species, as was done here. But then instead of searching entire mRNAs in comparator species, the search could be focused in discrete regions, guided by the position of the hairpin in the query species. This modification could decrease the search space by at least ~1/40th (where ~1/10th of the length of each homologue would be analyzed in four comparator species). We would expect that optimizing the homologue identification and hairpin conservation aspects of the program could reduce program run-time to less than one day, which would allow a much broader homolog search against dozens or hundreds of other species to be undertaken.

In the existing program, RNA secondary structure was assessed by folding short regions within the mRNA coding sequences, and the ΔG values of folding energies were calculated. Next, ΔG values below a certain threshold were used to determine ΔG cutoffs (2.5 times the average ΔG), below which the region was flagged as a potential RNA structure of interest (**Figure 5.1**). This approach worked well for species with moderate to high GC content. For example, *S. coelicolor* has a high GC content and an average ΔG values of ~ -7 kcal/mol, and therefore cutoffs were set at ~-17 kcal/mol. Some organisms with low GC content (*e.g. Campylobacter jejuni*), have average ΔG values close to zero, and therefore assigning cutoffs based on geometric scaling would not be appropriate. This resulted in false positives, as manual inspection of ΔG trend for individual species did not show a hairpin at the predicted location for low GC species. Such examples had to be manually filtered. A better method for calculating thresholds (perhaps using melting temperature instead of ΔG values) might be more appropriate, as melting temperature would never approach 0˚K, and therefore geometric scaling to find cutoffs would not suffer the above-mentioned pitfall.

It is worth noting that some RNA regions were conserved between *S. coelicolor* and more distantly related Gram positive bacteria, but not between *S. coelicolor* and *S. avermitilis* or *S. scabies.* This could be due to a hairpin being slightly over the ΔG cutoff in *S. coelicolor* and being present in lower GC organisms, while in higher GC *Streptomyces* species, the hairpin could be just under the GC cutoff due to very high

overall ΔG values in these species. One possible way of addressing this issue would be to have less-absolute cutoff values for ΔG, and to simply assign 'scores' for each potential region. For example, if in the query species, a hairpin was found to have a ΔG value ~2.5 times more stable than average, hairpins in other species with values at 2.0, and 2.4 and 3.0 would all be recorded, but more those with lower values would be given lower scores. A combination of percent conservation and conservation scores could be used to pull out the most widely distributed hairpins.

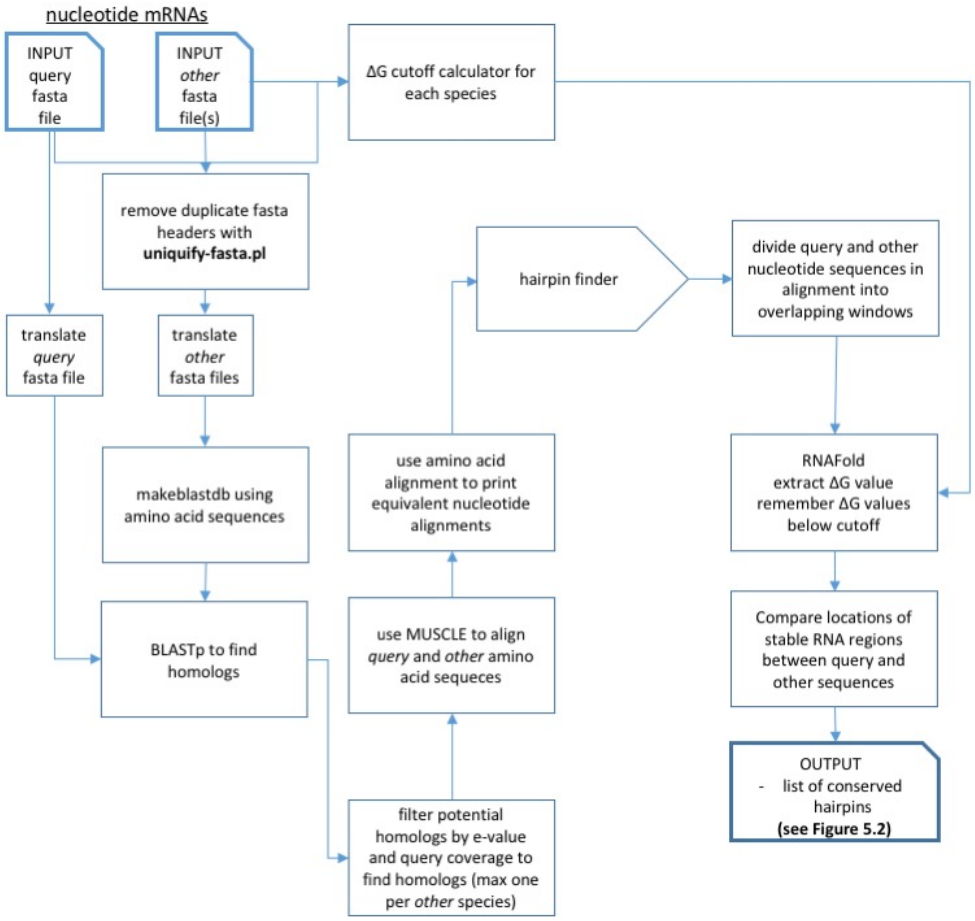
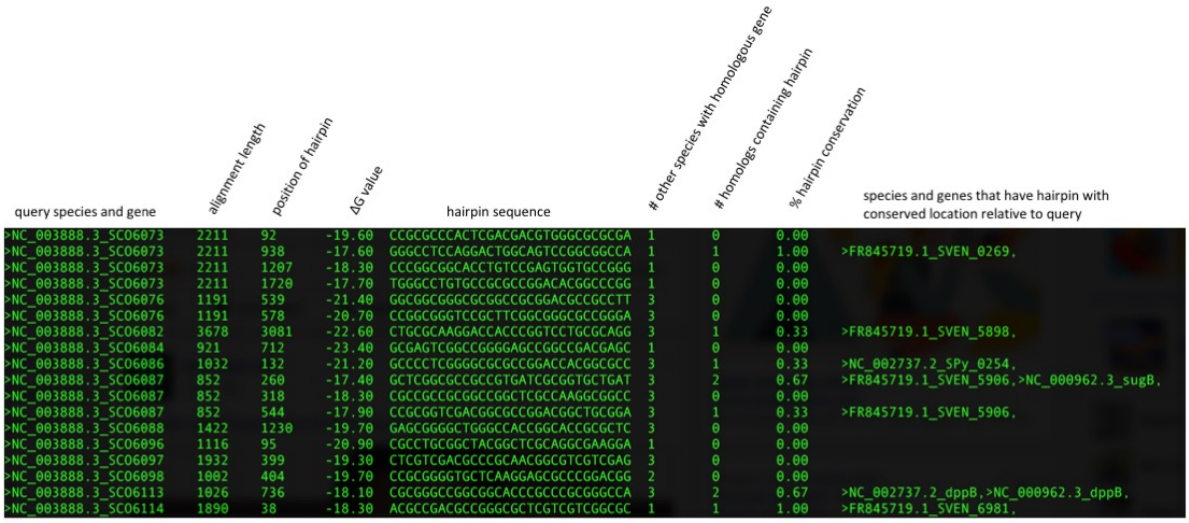


Figure 5.1: Flowchart of LoCoRNAStrut to find hairpins with conserved locations. Nucleotide FASTA input files for both query and *other* species are input into the script. Duplicate headers were removed using uniquify-fasta.pl, and then both the query and *other* nucleotide sequences were translated into equivalent amino acid FASTA files. The *other* amino acid FASTA file was used to generate a BLAST database. BLASTp was run using each query amino acid sequence against the BLAST database. Potential homologs were determined using e-values and query coverage, and was limited to one hit per species. Homologs were then aligned using MUSCLE, and amino acid alignments were used to print equivalent nucleotide alignments. The hairpin finder section of the script split each nucleotide sequence for each alignment into short windows (default 30 nucleotides) with a slide size of default 1 nucleotide. Each short window was then folded using RNAFold (Gruber et al., 2008) and ΔG values below ΔG cutoffs (calculated individually for each species) were extracted. Locations of stable regions were compared between the input and *other* species and a report was generated showing hairpins with conserved locations between species.



query species and gene	alignment length	position of hairpin	ΔG value	hairpin sequence	# other species with homologous gene	# homologs containing hairpin	% hairpin conservation	species and genes that have hairpin with conserved location relative to query
>NC_003888.3_SCO6073	2211	92	-19.60	CCGCGCCCACTCGACGACGTGGGCGCGCGA	1	0	0.00	
>NC_003888.3_SCO6073	2211	938	-17.60	GGGCCTCCAGGACTGGCAGTCCGGCGGCCA	1	1	1.00	>FR845719.1_SVEN_0269,
>NC_003888.3_SCO6073	2211	1207	-18.30	CCCGGCGGCACCTGTCCGAGTGGTGCCGGG	1	0	0.00	
>NC_003888.3_SCO6073	2211	1720	-17.70	TGGGCCTGTGCCGCGCCGGACACGGCCCGG	1	0	0.00	
>NC_003888.3_SCO6076	1191	539	-21.40	GGCGGCGGGGCGCGGCCGCGGACGCCGCCTT	3	0	0.00	
>NC_003888.3_SCO6076	1191	578	-20.70	CCGGCGGGTCCGCTTCGGCGGGCGCCGGGA	3	0	0.00	
>NC_003888.3_SCO6082	3678	3081	-22.60	CTGCGCAAGGACCACCCGGTCCTGCGCAGG	3	1	0.33	>FR845719.1_SVEN_5898,
>NC_003888.3_SCO6084	921	712	-23.40	GCGAGTCGGCCGGGGAGCCGGCCGACGAGC	1	0	0.00	
>NC_003888.3_SCO6086	1032	132	-21.20	GCCCCTCGGGGCGCGCCGGACCACGGCGCC	3	1	0.33	>NC_002737.2_SPy_0254,
>NC_003888.3_SCO6087	852	260	-17.40	GCTCGGCGCCGCCGTGATCGCGGTGCTGAT	3	2	0.67	>FR845719.1_SVEN_5906,>NC_000962.3_sugB,
>NC_003888.3_SCO6087	852	318	-18.30	CGCCGCCGCGGCCGGCTCGCCAAGGCGGCC	3	0	0.00	
>NC_003888.3_SCO6087	852	544	-17.90	CCGCGGTCGACGGCGCCGGACGGCTGCGGA	3	1	0.33	>FR845719.1_SVEN_5906,
>NC_003888.3_SCO6088	1422	1230	-19.70	GAGCGGGGCTGGGCCACCGGCACCGCGCTC	3	0	0.00	
>NC_003888.3_SCO6096	1116	95	-20.90	CGCCTGCGGCTACGGCTCGCAGGCGAAGGA	1	0	0.00	
>NC_003888.3_SCO6097	1932	399	-19.30	CTCGTCGACGCCCGCAACGGCGTCGTCGAG	3	0	0.00	
>NC_003888.3_SCO6098	1002	404	-19.70	CCGCGGGGTGCTCAAGGAGCGCCCGGACGG	2	0	0.00	
>NC_003888.3_SCO6113	1026	736	-18.10	CGCGGGCCGGCGGCACCCGCCCGCGGGCCA	3	2	0.67	>NC_002737.2_dppB,>NC_000962.3_dppB,
>NC_003888.3_SCO6114	1890	38	-18.30	ACGCCGACGCCGGGCGCTCGTCGTCGGCGC	1	1	1.00	>FR845719.1_SVEN_6981,

Figure 5.2: Sample output of Perl script for finding hairpins with conserved locations. The first column reported the FASTA header of the query species, indicating species and gene number. The next two columns reported the position of the center of the window of the potential hairpin relative to the translational start site (3rd column), compared to the alignment length (2nd column). The fourth and fifth columns reported the ΔG values and the nucleotide sequences of the hairpin in the query species, respectively. The next three columns indicated the conservation of the gene and hairpin location. Column six reported how many species had a homologue of the query gene, while column seven reported how many of these homologs had a conserved hairpin (0: no hairpin; 1: hairpin). The eighth column reported the fraction of genes with a conserved hairpin location out of the number of homologs searched. The final column showed fasta header information (species and gene names) for species/genes that had a hairpin at a conserved location relative to the query species/gene.

Figure 5.3: LoCoRNAStrut identifies a known hairpin of regulatory importance within the *hfq* gene that is widely distributed in bacteria. Aligned nucleotide sequences of the *hfq* mRNA from *E. coli, S. flexneri, C. rodentium, K. radicincitans, K. pneumonia, R. ornithinolytica, C. neteri, P. gergoviae and S. marcescens* were folded using Mfold (30 nucleotide window, slide size 1 nucleotide). The output for each species was normalized to the average ΔG value for 30 nucleotide windows in all mRNAs. A hairpin was identified near position ~105 in the alignment, with a ΔG value ~3.0 times the average ΔG value for each species at +41. Average value is plotted in blue, and error bars (standard deviation) are shown in grey.

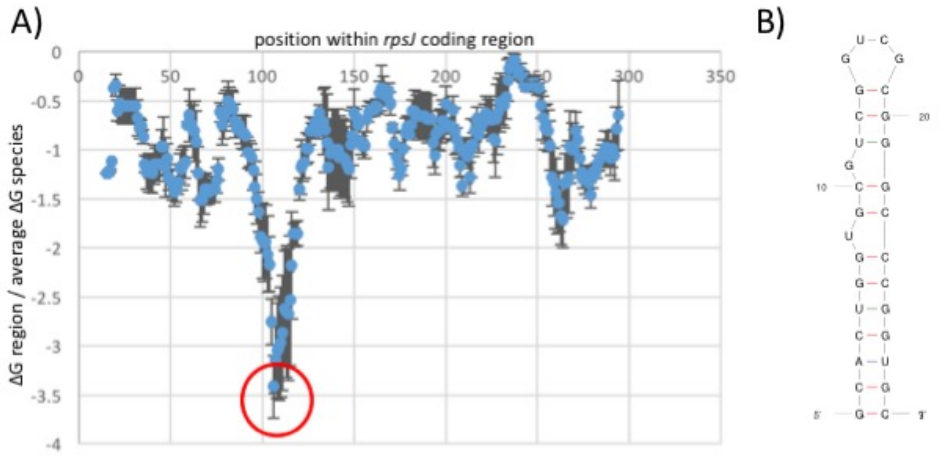


Figure 5.4: A hairpin with conserved location within the *rpsJ* gene. A) Aligned nucleotide sequences of the *rpsJ* mRNA from *S. coelicolor, S. venezuelae, M. tuberculosis, S. pyogenes, E. faecalis, S. aureus,* and *B. subtilis* were folded using Mfold (30 nucleotide window, slide size 1 nucleotide). The output for each species was normalized to the average ΔG value for 30 nucleotide windows in all mRNAs. A hairpin was identified near position ~105 in the alignment, with a ΔG value ~3.5 times the average ΔG value for each species (red circle). Average value is plotted in blue, and error bars (standard deviation) are shown in grey. **B)** Representative hairpin from *S. coelicolor,* as predicted by Mfold (Zuker, 2003).

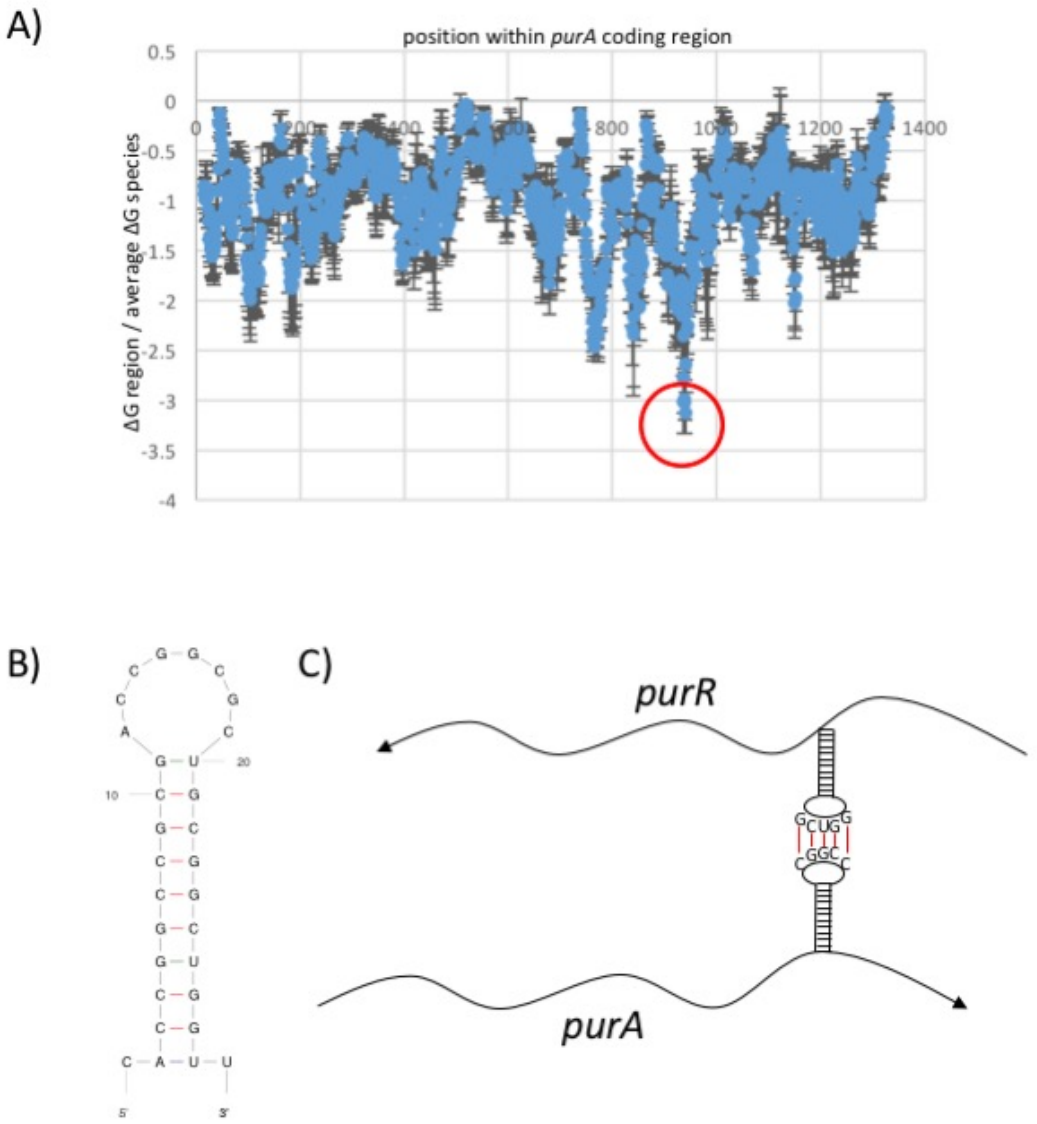


Figure 5.5: A conserved hairpin within the *purA* gene. A) Aligned nucleotide sequences of the *purA* mRNA from *S. coelicolor, S. venezuelae, M. tuberculosis, Y. pestis, B. subtilis,* and *H. pylori* were folded using Mfold (30 nucleotide window, slide size 1 nucleotide). The output for each species was normalized to the average ΔG value for 30 nucleotide windows in all mRNAs. A hairpin was identified near position ~950 in the alignment, having a ΔG value ~3.25 times the average ΔG value for each species (red circle). **B)** Representative *purA* hairpin from *S. coelicolor,* as predicted by Mfold (Zuker, 2003). Average value is plotted in blue, and error bars (standard deviation) are shown in grey.**C)** PurR is a transcriptional repressor of *purA*. The *purR* mRNA also has a strong stem-loop (although not as widely conserved as *purA*) with the potential to form a duplex with the *purA* sequence. mRNAs from *S. coelicolor* are shown.

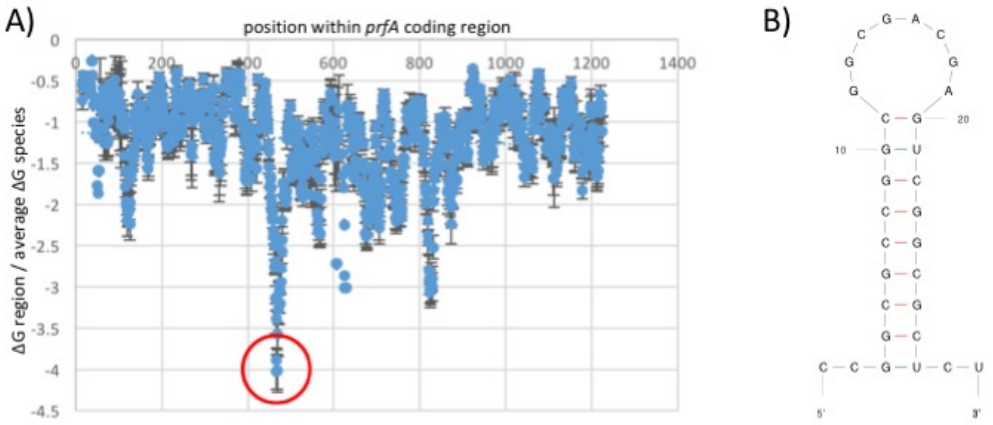


Figure 5.6: A hairpin with conserved location within the *prfA* gene. A) Aligned nucleotide sequences of the *prfA* mRNA from *Streptomyces* species, *M. tuberculosis, Y. pestis, H. pylori, B. subtilis, S. pyogenes,* and *E. coli* were folded using Mfold (30 nucleotide window, slide size 1 nucleotide). The output for each species was normalized to an average ΔG value for 30 nucleotide windows in all mRNAs. A hairpin was identified near position ~480 in the alignment with a ΔG value ~4 times the average ΔG value for each species (red circle). Average value is plotted in blue, and error bars (standard deviation) are shown in grey. B) Representative hairpin from *S. coelicolor,* as predicted by Mfold (Zuker, 2003)

Table 5.1: The most widely distributed hairpins using *Escherichia coli* as the query species (see Appendix 1)

Table 5.2: Summary of all conserved hairpins with *S. coelicolor* as query species and *S. venezuelae, M. tuberculosis,* and *Streptococcus pyogenes* as other species

Number of hairpins found	2569
Conserved in *S. venezuelae*	193
Conserved in *M. tuberculosis*	46
Conserved in *Streptococcus pyogenes*	91
Conserved in *S. venezuelae* and *Mycobacterium*	6
Conserved in *S. venezuelae* and *Streptococcus progenes*	11
Conserved in all three 'other' species	1

Table 5.3: Summary of hairpins conserved in at least two other species

		Homologs with hairpins at same location		
Query gene	**ΔG**	***S. venezuelae***	***Mycobacterium tuberculosis***	***Streptococcus pyogenes***
SCO4655 (rpoC)	-19.5	*SVEN_4346 (rpoC)*	*rpoC*	*rpoC*
SCO3045	-19.8	*SVEN_2785*	*Rv_3484*	
SCO4383	-18.2	*SVEN_4199*	*fadD3*	
SCO5117	-16.2	*SVEN_4765*	*dppA*	
rpsJ	-16.2	*SVEN_4378*	*rpsJ*	
SCO2789	-18.6	*SVEN_2576*		*glmS*
SCO2949	-20.2	*SVEN_2716*		*SPy_0763*
SCO3404	-22	*SVEN_3256*		*ftsH*
SCO3629 (purA)	-18.1		*purA*	*purA*
SCO5111	-17.9	*SVEN_4758*		*Spy_1527*
SCO5188	-22.5	*SVEN_4840*		*pcrA*
SCO5226	-20.8	*SVEN_4877*		*nrdE.1*
SCO5523	-20	*SVEN_5206*		*bcaT*
SCO5661	-23.5	*SVEN_5326*		*SPy_0839*
SCO6087	-17.4	*SVEN_5906*	*sugB*	
SCO6113	-18.1		*dppB*	*dppB*
SCO6587	-20.9		*Rv0223c*	*gapN*
SCO7012	-17.9	*SVEN_7354*		*SPy_0254*
SCO7281	-19.6	*SVEN_2403*		*clpP*
thrS	-19.2	*SVEN_1133*		*thrS*

Table 5.4: Hairpins showing the most widely distributed locational conservation

Query gene	ΔG	*S. avermitilis*	*S. griseus*	*S. scabies*	*B. subtilis*	*S. pyogenes*	*E. coli*	*H. pylori*	*S. enterica*	*P. aeruginosa*	*Y. pestis*
prfA	-18	Y			Y	Y	Y	Y			Y
SCO5111	-17.9				Y	Y	Y	Y			Y
SCO3890 (trxB)	-19					Y	Y	Y	Y		
SCO5188	-22.5				Y	Y		Y			Y
SCO4797	-20.9			Y	Y			Y			
SCO5117	-16.2	Y	Y	Y							Y
SCO7197	-16.9					Y		Y	Y		
SCO7197	-16.3				Y	Y		Y			
ileS	-20									Y	Y
rplM	-20.5	Y	Y	Y							
SCO1409	-19.8	Y	Y	Y							
SCO1621	-19.6				Y	Y				Y	
SCO1865	-21.7		Y	Y							Y
SCO2301	-22		Y	Y		Y					
SCO3071	-22.7	Y	Y	Y							
SCO3629	-18.1				Y			Y			Y
SCO3960	-23.5	Y			Y		Y				
SCO4655	-19.5					Y	Y		Y		
SCO4989	-19.8	Y			Y						
SCO5399	-22.3		Y	Y		Y					
SCO5566	-21.9				Y			Y			
SCO5920	-20.6		Y					Y	Y		
SCO6039	-27.1	Y		Y	Y						
SCO6517	-21.1	Y		Y		Y					
ulaA	-20.4					Y	Y		Y		

Chapter 6: Summary and Future Directions

Summary of research

Chapter 3 describes the discovery of hundreds of asRNAs, *trans*-encoded sRNAs and a new type of RNA (the 'cutoRNA') resulting from long 3′ UTRs. Many of these new ncRNAs were associated with specialized metabolite clusters, and it is likely that some will prove to be important for regulating the expression of these clusters. In addition, many mRNAs had short (30-90 nucleotide), internal transcripts that were highly abundant, and may represent highly stable degradation products.

The work in Chapter 4 originally started with the goal of characterizing one of the interesting asRNAs identified in our Chapter 3 work. This antisense RNA, named ScbN, was expressed opposite the coding region of *scbA,* a γ-butyrolactone synthase. Surprisingly, we found that ScbN had little or no impact on the expression of *scbA*, at least under the conditions examined here. However, the terminator of ScbN, which also forms a structured RNA hairpin in the sense direction in the *scbA* mRNA, caused a drastic decrease in expression of *scbA*.

The work described in Chapter 5 emerged from this discovery, and led to the development of a bioinformatics pipeline, focused on identifying regions of strong RNA secondary structures at conserved locations across many bacterial species. The resulting program yielded hundreds of genes with structured RNA regions in locations conserved not only in the Actinobacteria, but also in genes shared between species that diverged billions of years ago (*Streptomyces, Bacillus,* and *E. coli*). Determining whether these conserved structures have regulatory or functional relevance will be the focus of future investigations.

Future directions

Determining the functions of ncRNAs

RNA-sequencing has greatly facilitated the discovery of ncRNAs in all domains of life; however, characterization of these regulators has not kept pace with their discovery. This is especially true for species like *Streptomyces*, which lacks an obvious Hfq-like chaperone. In addition to the efforts described in this thesis aimed at characterizing ScbN, some minor projects have also been started which should help get a handle on functions of some ncRNAs in *Streptomyces*.

A collaborative project with Stephanie Jones focused on cloning 26 conserved sRNA promoters upstream of a reporter gene, and introducing these reporters into *S. venezuelae.* As many ncRNA regulators are important for regulating genes under stress conditions, these strains could be used to monitor expression under a range of different conditions. Preliminary testing revealed that some sRNAs were expressed specifically upon exposure to certain antibiotics and other stresses, which might allow us to predict potential cellular targets. For example, we found that svr1924 was induced upon exposure to tetracycline, but not other ribosome-targeting antibiotics such as erythromycin.

It is becoming apparent that most genes in bacteria have associated asRNAs (Lasa et al., 2011; Raghavan et al., 2012; Wade and Grainger, 2014); however, very few have

been characterized. asRNAs have classically been assumed to regulate their sense gene (Georg and Hess, 2011; Thomason and Storz, 2010). However, the work described in Chapter 4, and by others in the lab (*e.g.* Hindra et al., 2014) have suggested that even highly expressed antisense RNAs might not have great effects on the transcription, translation, or stability of their sense gene counterparts. Given the perfect complementarity between mRNAs/asRNAs, how is this possible? And if they are not regulating the sense gene, what – if anything - are they doing? In the case of *scbA/*ScbN, comparative RNA-seq between a Δ*scbAN* strain and strains complemented with *scbAN* and *scbN* only, will be critical in determining if ScbN has an impact on the transcriptome, other than the obvious candidate *scbA*. Alternatively, to find targets impacted at the translational level, iTRAQ (Manteca et al., 2010; Wiese et al., 2007) or MUDPIT (Washburn et al., 2001) could be used to find differences in the proteome between wild-type, deletion, complementation, and overexpression strains (Vockenhuber et al., 2015).

In addition to monitoring changes in transcriptome / proteome in strains with altered ncRNA expression, direct targets of ncRNAs could also be identified using other strategies. For example, MS2 RNA aptamers can be added to RNAs of interest (e.g. sRNAs, asRNAs), and MS2-binding protein can be used to pull out interacting proteins or RNAs which can be identified by mass-spectrometry or RNA-seq respectively (Lalaouna and Masse, 2015; Said et al., 2009).

Understanding the roles of mRNA secondary structures

The work outlined in Chapter 5 highlighted a number of mRNA-associated hairpins positioned at conserved sites throughout bacteria. In some instances, this conservation had been maintained over billions of years of independent evolution, and they share little primary sequence conservation (despite being present in coding sequences). Why are these structures conserved? Are they important for the regulation of their associated genes/operons? If so, what is their function? Pursuing functional characterization of these hairpins, particularly those that are conserved in *B. subtilis* and *E .coli*, may be more rapidly conducted in these model species (instead of *Streptomyces*), given the rapid growth, and abundant genetic tools and relevant mutant libraries that are available for these organisms. Reporters similar to those designed in Chapter 4 for the *scbN* hairpin could be introduced into a range of genetic backgrounds and differences in expression could be monitored. In addition, comprehensives suites of mutations disrupting the mutation could be systematically tested.

How do antisense RNAs evolve?

Given the connection between the *scbA* hairpin and the *scbN* terminator, we wondered whether mRNA hairpins arise from antisense RNA terminators, or if antisense RNAs preferentially arise in association with existing mRNA hairpins? If the mRNA hairpin arose first, one can imagine an asRNA opposite this gene employing existing hairpin structures as terminators, and arising as a result of downstream mutations that lead to promoter formation. Alternatively, asRNA terminators might have evolved for the purpose of terminating transcription in the antisense direction, and a byproduct of some of these is the ability to form strongly structured sequences in the sense direction as well.

Once these sense RNA structures arise, they could evolve to acquire a regulatory function distinct from that of the asRNA (as we expect for the *scbA* hairpin, ScbN case). As RNA discovery under numerous conditions becomes more saturated across many species, analyzing the distribution of hairpins and comparing these with known antisense RNA locations might allow us to answer this question.

References

Aiba, H. (2007). Mechanism of RNA silencing by Hfq-binding small RNAs. Curr. Opin. Microbiol. *10*, 134–139.

Albrecht, M., Sharma, C.M., Reinhardt, R., Vogel, J., and Rudel, T. (2010). Deep sequencing-based discovery of the *Chlamydia trachomatis* transcriptome. Nucleic Acids Res. *38*, 868–877.

Allenby, N.E.E., Laing, E., Bucca, G., Kierzek, A.M., and Smith, C.P. (2012). Diverse control of metabolism and other cellular processes in *Streptomyces coelicolor* by the PhoP transcription factor: genome-wide identification of in vivo targets. Nucleic Acids Res. *40*, 9543–9556.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic Local Alignment Search Tool. J. Mol. Biol. 403–410.

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E.G., Margalit, H., and Altuvia, S. (2001). Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. Curr Biol *11*, 941–950.

Arnvig, K.B., Comas, I., Thomson, N.R., Houghton, J., Boshoff, H.I., Croucher, N.J., Rose, G., Perkins, T.T., Parkhill, J., Dougan, G., et al. (2011). Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of *Mycobacterium tuberculosis*. PLoS Pathog. *7*, e1002342.

Attaiech, L., Boughammoura, A., Brochier-Armanet, C., Allatif, O., Peillard-Fiorente, F., Edwards, R.A., Omar, A.R., MacMillan, A.M., Glover, M., and Charpentier, X. (2016). Silencing of natural transformation by an RNA chaperone and a multitarget small RNA. Proc. Natl. Acad. Sci. U. S. A. *113*, 1–6.

Aviran, S., Trapnell, C., Lucks, J.B., Mortimer, S.A., Luo, S., Schroth, G.P., Doudna, J.A., Arkin, A.P., and Pachter, L. (2011). Modeling and automation of sequencing-based characterization of RNA structure. Proc. Natl. Acad. Sci. U. S. A. *108*, 11069–11074.

Babitzke, P., and Yanofsky, C. (1993). Reconstitution of *Bacillus subtilis trp* attenuation in vitro with TRAP, the trp RNA-binding attenuation protein. Proc. Natl. Acad. Sci. U. S. A. *90*, 133–137.

Battchikova, N., Eisenhut, M., and Aro, E.-M. (2011). Cyanobacterial NDH-1 complexes: novel insights and remaining puzzles. Biochim. Biophys. Acta *1807*, 935–944.

Bedford, D.J., Lewis, C.G., and Buttner, M.J. (1991). Characterization of a gene conferring bialaphos resistance in *Streptomyces coelicolor* A3(2). Gene *104*, 39–45.

Bejerano-Sagie, M., and Xavier, K.B. (2007). The role of small RNAs in quorum sensing. Curr. Opin. Microbiol. *10*, 189–198.

Bentley, S.D., Chater, K.F., Cerdeño-Tárraga, A.-M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., et al. (2002). Complete

genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature *417*, 141–147.

Bérdy, J. (2012). Thoughts and facts about antibiotics: Where we are now and where we are heading. J. Antibiot. (Tokyo). *65*, 385–395.

Bernier, S.P., and Surette, M.G. (2013). Concentration-dependent activity of antibiotics in natural environments. Front. Microbiol. *4*, 20.

Bibb, M.J., White, J., Ward, J.M., and Janssen, G.R. (1994). The mRNA for the 23S rRNA methylase encoded by the *ermE* gene of *Saccharopolyspora erythraea* is translated in the absence of a conventional ribosome-binding site. Mol. Microbiol. *14*, 533–545.

Brantl, S. (2007). Regulatory mechanisms employed by *cis*-encoded antisense RNAs. Curr. Opin. Microbiol. *10*, 102–109.

Bricker, A.L., and Belasco, J.G. (1999). Importance of a 5` stem-loop for longevity of *papA* mRNA in *Escherichia coli*. J. Bacteriol. *181*, 3587–3590.

Bronesky, D., Wu, Z., Marzi, S., Walter, P., Geissmann, T., Moreau, K., Vandenesch, F., Caldelari, I., and Romby, P. (2016). *Staphylococcus aureus* RNAIII and its regulon link quorum sensing, stress sesponses, metabolic adaptation, and regulation of virulence gene expression. Annu. Rev. Microbiol. *70*, 299–316.

Browning, D.F., Grainger, D.C., and Busby, S.J. (2010). Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression. Curr. Opin. Microbiol. *13*, 773–780.

Burmann, B., Schweimer, K., Luo, X., Wahl, M., Stitt, B., Gottesman, M., and Rosch, P. (2010). A NusE:NusG complex links transcription and translation. Science *13*, 773–380.

Cavanagh, A.T., and Wassarman, K.M. (2014). 6S RNA, a global regulator of transcription in *Escherichia coli, Bacillus subtilis*, and beyond. Annu. Rev. Microbiol. *68*, 45–60.

Cerdeño, A.M., Bibb, M.J., and Challis, G.L. (2001). Analysis of the prodiginine biosynthesis gene cluster of *Streptomyces coelicolor* A3(2): new mechanisms for chain initiation and termination in modular multienzymes. Chem. Biol. *8*, 817–829.

Chaney, J.L., and Clark, P.L. (2015). Roles for synonymous codon usage in protein biogenesis. Annu. Rev. Biophys. *44*, 143–166.

Chao, Y., Papenfort, K., Reinhardt, R., Sharma, C.M., and Vogel, J. (2012). An atlas of Hfq-bound transcripts reveals 3' UTRs as a genomic reservoir of regulatory small RNAs. EMBO J. *31*, 4005–4019.

Chao, Y., Li, L., Girodat, D., Forstner, K.U., Said, N., Corcoran, C., Smiga, M., Papenfort, K., Reinhardt, R., Wieden, H.J., et al. (2017). *in vivo* cleavage map illuminates the central cole of RNase E in coding and non-coding RNA pathways. Mol. Cell *65*, 39–51.

Chater, K.F., and Chandra, G. (2008). The use of the rare UUA codon to define

"expression space" for genes involved in secondary metabolism, development and environmental adaptation in *Streptomyces*. J. Microbiol. *46*, 1–11.

Chater, K.F., Biró, S., Lee, K.J., Palmer, T., and Schrempf, H. (2010). The complex extracellular biology of *Streptomyces*. FEMS Microbiol. Rev. *34*, 171–198.

Chatterjee, A., Drews, L., Mehra, S., Takano, E., Kaznessis, Y.N., and Hu, W.-S. (2011). Convergent transcription in the butyrolactone regulon in *Streptomyces coelicolor* confers a bistable genetic switch for antibiotic biosynthesis. PLoS One *6*, e21974.

Chen, Z., Song, Y., Wen, Y., and Li, J. (2001). Effect of gene disruption of *aveD* on avermectins production in *Streptomyces avermitilis*. Wei Sheng Wu Xue Bao *41*, 440–446.

Chomczynski, P., and Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem. *162*, 156–159.

Ciampi, M.S. (2006). Rho-dependent terminators and transcription termination. Microbiology *152*, 2515–2528.

Cohen, O., Doron, S., Wurtzel, O., Dar, D., Edelheit, S., Karunker, I., Mick, E., and Sorek, R. (2016). Comparative transcriptomics across the prokaryotic tree of life. Nucleic Acids Res. *44*, W46–W53.

Cortes, T., and Cox, R.A. (2015). Transcription and translation of the *rpsJ, rplN* and rRNA operons of the tubercle bacillus. Microbiology *161*, 719–728.

D'Alia, D., Nieselt, K., Steigele, S., Müller, J., Verburg, I., and Takano, E. (2010). Noncoding RNA of glutamine synthetase I modulates antibiotic production in *Streptomyces coelicolor* A3(2). J. Bacteriol. *192*, 1160–1164.

D'Alia, D., Eggle, D., Nieselt, K., Hu, W.-S., Breitling, R., and Takano, E. (2011). Deletion of the signalling molecule synthase *scbA* has pleiotropic effects on secondary metabolite biosynthesis, morphological differentiation and primary metabolism in *Streptomyces coelicolor* A3(2). Microb. Biotechnol. *4*, 239–251.

Daniel Holoch, Danesh Moazed, Holoch, D., and Moazed, D. (2015). RNA-mediated epigenetic regulation of gene expression. Nat. Rev. Genet. *16*, 71–84.

Dar, D., Prasse, D., Schmitz, R.A., and Sorek, R. (2016). Widespread formation of alternative 3′ UTR isoforms via transcription termination in archaea. Nat. Microbiol. *1*, 16143.

Darty, K., Denise, A., and Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics *25*, 1974–1975.

Datsenko, K.A., and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc. Natl. Acad. Sci. U. S. A. *97*, 6640–6645.

Davies, J. (2011). How to discover new antibiotics: harvesting the parvome. Curr. Opin. Chem. Biol. *15*, 5–10.

Davis, J.B., Bailey, J.D., and Sello, J.K. (2009). Biomimetic synthesis of a new class of bacterial signaling molecules. Org. Lett. *11*, 2984–2987.

DebRoy, S., Gebbie, M., Ramesh, A., Goodson, J.R., Cruz, M.R., van Hoof, A., Winkler, W.C., and Garsin, D.A. (2014). Riboswitches. A riboswitch-containing sRNA controls gene expression by sequestration of a response regulator. Science *345*, 937–940.

Deneke, C., Lipowsky, R., and Valleriani, A. (2013). Effect of ribosome shielding on mRNA stability. Phys. Biol. *10*, 46008.

Deutscher, M.P. (2015). How bacterial cells keep ribonucleases under control. FEMS Microbiol. Rev. *39*, 350–361.

Dorman, C.J. (2013). Genome architecture and global gene regulation in bacteria: making progress towards a unified model? Nat. Rev. Microbiol. *11*, 349–355.

Edgar, R.C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Efremov, R.G., Baradaran, R., and Sazanov, L.A. (2010). The architecture of respiratory complex I. Nature *465*, 441–445.

Elliott, T., and Wang, X. (1991). *Salmonella typhimurium prfA* mutants defective in release factor 1. J Bacteriol *173*, 4144–4154.

Fajardo, A., and Martínez, J.L. (2008). Antibiotics as signals that trigger specific bacterial responses. Curr. Opin. Microbiol. *11*, 161–167.

Feitelson, J.S., Malpartida, F., and Hopwood, D.A. (1985). Genetic and biochemical characterization of the red gene cluster of *Streptomyces coelicolor* A3(2). J. Gen. Microbiol. *131*, 2431–2441.

Feng, L., Rutherford, S.T., Papenfort, K., Bagert, J.D., van Kessel, J.C., Tirrell, D.A., Wingreen, N.S., and Bassler, B.L. (2015). A *qrr* noncoding RNA deploys four different regulatory mechanisms to optimize quorum-sensing dynamics. Cell *160*, 228–240.

Figueroa-Bossi, N., Schwartz, A., Guillemardet, B., D'Heygère, F., Bossi, L., and Boudvillain, M. (2014). RNA remodeling by bacterial global regulator CsrA promotes Rho-dependent transcription termination. Genes Dev. *28*, 1239–1251.

Flärdh, K., and Buttner, M.J. (2009). *Streptomyces* morphogenetics: dissecting differentiation in a filamentous bacterium. Nat. Rev. Microbiol. *7*, 36–49.

Floriano, B., and Bibb, M. (1996). *afsR* is a pleiotropic but conditionally required regulatory gene for antibiotic production in *Streptomyces coelicolor* A3(2). Mol Microbiol *21*, 385–396.

Fong, C., Rohmer, L., Radey, M., Wasnick, M., and Brittnacher, M.J. (2008). PSAT: A web tool to compare genomic neighborhoods of multiple prokaryotic genomes. BMC Bioinformatics *9*, 170.

Fowler-Goldsworthy, K., Gust, B., Mouz, S., Chandra, G., Findlay, K.C., and Chater,

K.F. (2011). The actinobacteria-specific gene *wblA* controls major developmental transitions in *Streptomyces coelicolor* A3(2). Microbiology *157*, 1312–1328.

Freedman, L.P., Zengel, J.M., Archer, R.H., and Lindahl, L. (1987). Autogenous control of the S10 ribosomal protein operon of *Escherichia coli*: genetic dissection of transcriptional and posttranscriptional regulation. Proc. Natl. Acad. Sci. U. S. A. *84*, 6516–6520.

Gao, C., Hindra, Mulder, D., Yin, C., and Elliot, M.A. (2012). Crp is a global regulator of antibiotic production in *Streptomyces*. MBio *3*, e00407-12.

Gatewood, M.L., Bralley, P., Weil, M.R., and Jones, G.H. (2012). RNA-Seq and RNA immunoprecipitation analyses of the transcriptome of *Streptomyces coelicolor* identify substrates for RNase III. J. Bacteriol. *194*, 2228–2237.

Geisinger, E., Adhikari, R.P., Jin, R., Ross, H.F., and Novick, R.P. (2006). Inhibition of *rot* translation by RNAIII, a key feature of *agr* function. Mol. Microbiol. *61*, 1038–1048.

Georg, J., and Hess, W.R. (2011). *cis*-antisense RNA, another level of gene regulation in bacteria. Microbiol. Mol. Biol. Rev. *75*, 286–300.

Goh, E.-B., Yim, G., Tsui, W., McClure, J., Surette, M.G., and Davies, J. (2002). Transcriptional modulation of bacterial gene expression by subinhibitory concentrations of antibiotics. Proc. Natl. Acad. Sci. U. S. A. *99*, 17025–17030.

Gomez-Escribano, J.P., and Bibb, M.J. (2011). Engineering *Streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. Microb. Biotechnol. *4*, 207–215.

Gomez-Escribano, J.P., Song, L., Fox, D.J., Yeo, V., Bibb, M.J., and Challis, G.L. (2012). Structure and biosynthesis of the unusual polyketide alkaloid coelimycin P1, a metabolic product of the *cpk* gene cluster of *Streptomyces coelicolor* M145. Chem. Sci. *3*, 2716.

Gorochowski, T.E., Ignatova, Z., Bovenberg, R.A.L., and Roubos, J.A. (2015). Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. Nucleic Acids Res. *43*, 3022–3032.

Gössringer, M., and Hartmann, R.K. (2012). 3'-UTRs as a source of regulatory RNAs in bacteria. EMBO J. *31*, 3958–3960.

Gottelt, M., Kol, S., Gomez-Escribano, J.P., Bibb, M., and Takano, E. (2010). Deletion of a regulatory gene within the *cpk* gene cluster reveals novel antibacterial activity in *Streptomyces coelicolor* A3(2). Microbiology *156*, 2343–2353.

Gottesman, S., and Storz, G. (2010). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. Cold Spring Harb. Perspect. Biol. *3*, a003798.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., and Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res. *38*, W695-9.

Grafe, U., Shcade, W., Eritt, I., and Fleck, W.F. (1982). A new inducer of anthracycline biosynthesis from *Streptomyces viridochromogenes*. J. Antibiot. (Tokyo). *35*, 1722.

Gregory, M., Till, R., and Smith, M.C.M. (2003). Integration site for Streptomyces phage phi-BT1 and development of site-specific integrating vectors. J. Bacteriol. *185*, 5320–5323.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. (2003). Rfam: An RNA family database. Nucleic Acids Res. *31*, 439–441.

Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R., and Hofacker, I.L. (2008). The Vienna RNA websuite. Nucleic Acids Res. *36*, W70-4.

Gust, B., Challis, G.L., Fowler, K., Kieser, T., and Chater, K.F. (2003). PCR-targeted *Streptomyces* gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. Proc. Natl. Acad. Sci. U. S. A. *100*, 1541–1546.

Hagège, J.M., and Cohen, S.N. (1997). A developmentally regulated *Streptomyces* endoribonuclease resembles ribonuclease E of *Escherichia coli*. Mol. Microbiol. *25*, 1077–1090.

Haiser, H.J., Karginov, F. V, Hannon, G.J., and Elliot, M.A. (2008). Developmentally regulated cleavage of tRNAs in the bacterium *Streptomyces coelicolor*. Nucleic Acids Res. *36*, 732–741.

Hambraeus, G., Karhumaa, K., and Rutberg, B. (2002). A 5' stem-loop and ribosome binding but not translation are important for the stability of *Bacillus subtilis aprE* leader mRNA. Microbiology *148*, 1795–1803.

Han, L.Y., Cai, C.Z., Lo, S.L., Chung, M.C.M., and Chen, Y.Z. (2004). Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. RNA *10*, 355–368.

Hanahan, D. (1983). Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. *166*, 557–580.

Hara, O., and Beppu, T. (1982). Mutants blocked in streptomyces production in *Streptomyces griseus* - the role of A-factor. J. Antibiot. (Tokyo). *35*, 349–358.

Hartswood, E., Brodie, J., Vendra, G., Davis, I., and Finnegan, D.J. (2012). RNA:RNA interaction can enhance RNA localization in *Drosophila* oocytes. RNA *18*, 729–737.

He, B., and Zalkin, H. (1994). Regulation of *Escherichia coli purA* by purine repressor, one component of a dual control mechanism. J. Bacteriol. *176*, 1009–1013.

He, J., Magarvey, N., Piraee, M., and Vining, L.C. (2001). The gene cluster for chloramphenicol biosynthesis in *Streptomyces venezuelae* ISP5230 includes novel shikimate pathway homologues and a monomodular non-ribosomal peptide synthetase gene. Microbiology *147*, 2817–2829.

Hernández, J.A., Muro-Pastor, A.M., Flores, E., Bes, M.T., Peleato, M.L., and Fillat, M.F. (2006). Identification of a *furA cis* antisense RNA in the cyanobacterium *Anabaena*

sp. PCC 7120. J. Mol. Biol. *355*, 325–334.

Hindra, Moody, M.J., Jones, S.E., and Elliot, M.A. (2014). Complex intra-operonic dynamics mediated by a small RNA in *Streptomyces coelicolor*. PLoS One *9*, e85856.

Hirakawa, H., Harwood, C.S., Pechter, K.B., Schaefer, A.L., and Greenberg, E.P. (2012). Antisense RNA that affects *Rhodopseudomonas palustris* quorum-sensing signal receptor expression. Proc. Natl. Acad. Sci. U. S. A. *109*, 12141–12146.

Hobbs, E.C., Astarita, J.L., and Storz, G. (2010). Small RNAs and small proteins involved in resistance to cell envelope stress and acid shock in *Escherichia coli*: analysis of a bar-coded mutant collection. J. Bacteriol. *192*, 59–67.

Hobbs, E.C., Fontaine, F., Yin, X., and Storz, G. (2011). An expanding universe of small proteins. Curr. Opin. Microbiol. *14*, 167–173.

Hsiao, N.-H., Nakayama, S., Merlo, M.E., de Vries, M., Bunet, R., Kitani, S., Nihira, T., and Takano, E. (2009). Analysis of two additional signaling molecules in *Streptomyces coelicolor* and the development of a butyrolactone-specific reporter system. Chem. Biol. *16*, 951–960.

Hu, Z., Zhang, A., Storz, G., Gottesman, S., and Leppla, S.H. (2006). An antibody-based microarray assay for small RNA detection. Nucleic Acids Res. *34*, e52.

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

Ingham, C.J., Hunter, I.S., and Smith, M.C. (1996). Isolation and sequencing of the rho gene from *Streptomyces lividans* ZX7 and characterization of the RNA-dependent NTPase activity of the overexpressed protein. J. Biol. Chem. *271*, 21803–21807.

Irnov, I., Sharma, C.M., Vogel, J., and Winkler, W.C. (2010). Identification of regulatory RNAs in *Bacillus subtilis*. Nucleic Acids Res. *38*, 6637–6651.

Ito, K., Ebihara, K., Uno, M., and Nakamura, Y. (1996). Conserved motifs in prokaryotic and eukaryotic polypeptide release factors: tRNA-protein mimicry hypothesis. Proc. Natl. Acad. Sci. U. S. A. *93*, 5443–5448.

Janssen, G.R., Ward, J.M., and Bibb, M.J. (1989). Unusual transcriptional and translational features of the aminoglycoside phosphotransferase gene (*aph*) from *Streptomyces fradiae*. Genes Dev. *3*, 415–429.

Johnsen, M., Christensen, T., Dennis, P.P., and Fiil, N.P. (1982). Autogenous control: ribosomal protein L10-L12 complex binds to the leader sequence of its mRNA. EMBO J. *1*, 999–1004.

Jones, S.E., Ho, L., Rees, C.A., Hill, J.E., Nodwell, J.R., and Elliot, M.A. (2017). *Streptomyces* exploration is triggered by fungal interactions and volatile signals. Elife *6*, e21738.

Jutras, B.L., Jones, G.S., Verma, A., Brown, N.A., Antonicello, A.D., Chenail, A.M., and

Stevenson, B. (2013). Posttranscriptional self-regulation by the Lyme disease bacterium's BpuR DNA/RNA-binding protein. J. Bacteriol. *195*, 4915–4923.

Kawano, M., Reynolds, A.A., Miranda-Rios, J., and Storz, G. (2005). Detection of 5'- and 3'-UTR-derived small RNAs and *cis*-encoded antisense RNAs in *Escherichia coli*. Nucleic Acids Res. *33*, 1040–1050.

Keller, L., and Surette, M.G. (2006). Communication in bacteria: an ecological and evolutionary perspective. Nat. Rev. Microbiol. *4*, 249–258.

Kendrew, S.G., Hopwood, D.A., and Marsh, E.N. (1997). Identification of a monooxygenase from *Streptomyces coelicolor* A3(2) involved in biosynthesis of actinorhodin: purification and characterization of the recombinant enzyme. J. Bacteriol. *179*, 4305–4310.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. Nature *467*, 103–107.

Kieser, T., Bibb, M.J., Buttner, M.J., Chater, K.F., and Hopwood, D.A. (2000). Practical Streptomyces genetics. John Innes Centre, Ltd.

Kim, H.-K., Liu, F., Fei, J., Bustamante, C., Gonzalez, R.L., and Tinoco, I. (2014a). A frameshifting stimulatory stem loop destabilizes the hybrid state and impedes ribosomal translocation. Proc. Natl. Acad. Sci. U. S. A. *111*, 5338–5343.

Kim, H.M., Shin, J.-H., Cho, Y.-B., and Roe, J.-H. (2014b). Inverse regulation of Fe- and Ni-containing SOD genes by a Fur family regulator Nur through small RNA processed from 3'UTR of the *sodF* mRNA. Nucleic Acids Res. *42*, 2003–2014.

Kinoshita, H., Ipposhi, H., Okamoto, S., Nakano, H., Nihira, T., and Yamada, Y. (1997). Butyrolactone autoregulator receptor protein (BarA) as a transcriptional regulator in *Streptomyces virginiae*. J. Bacteriol. *179*, 6986–6993.

Kirby, R. (2011). Chromosome diversity and similarity within the Actinomycetales. FEMS Microbiol. Lett. *319*, 1–10.

Kitani, S., Kinoshita, H., Nihira, T., and Yamada, Y. (1999). *In vitro* analysis of the butyrolactone autoregulator receptor protein (FarA) of Streptomyces lavendulae FRI-5 reveals that FarA acts as a DNA-binding transcriptional regulator that controls its own synthesis. J Bacteriol. *181*, 5081–5084.

Kolmsee, T., and Hengge, R. (2011). Rare codons play a positive role in the expression of the stationary phase sigma factor RpoS in *Escherichia coli*. RNA Biol. *8*, 913–921.

Kuhstoss, S., and Rao, R.N. (1991). Analysis of the integration function of the streptomycete bacteriophage phi C31. J. Mol. Biol. *222*, 897–908.

Lalaouna, D., and Masse, E. (2015). Identification of sRNA interacting with a transcript of interest using MS2-affinity purification coupled with RNA sequencing (MAPS) technology. Genomics Data *5*, 136–138.

Lalaouna, D., Carrier, M.-C., Semsey, S., Brouard, J.-S., Wang, J., Wade, J.T., and Massé, E. (2015). A 3′ external transcribed spacer in a tRNA transcript acts as a sponge for small RNAs to prevent transcriptional noise. Mol. Cell *58*, 393–405.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lasa, I., Toledo-Arana, A., Dobin, A., Villanueva, M., de Los Mozos, I.R., Vergara-Irigaray, M., Segura, V., Fagegaltier, D., Penadés, J.R., Valle, J., et al. (2011). Genome-wide antisense transcription drives mRNA processing in bacteria. Proc. Natl. Acad. Sci. U. S. A. *108*, 20172–20177.

Lasa, I., Toledo-Arana, A., and Gingeras, T.R. (2012). An effort to make sense of antisense transcription in bacteria. RNA Biol. *9*, 1039–1044.

Laurberg, M., Asahara, H., Korostelev, A., Zhu, J., Trakhanov, S., and Noller, H.F. (2008). Structural basis for translation termination on the 70S ribosome. Nature *454*, 852–857.

Lautru, S., Deeth, R.J., Bailey, L.M., and Challis, G.L. (2005). Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. Nat. Chem. Biol. *1*, 265–269.

Lenz, D.H., Mok, K.C., Lilley, B.N., Kulkarni, R. V, Wingreen, N.S., and Bassler, B.L. (2004). The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in *Vibrio harveyi* and *Vibrio cholerae*. Cell *118*, 69–82.

Lenz, D.H., Miller, M.B., Zhu, J., Kulkarnl, R. V., and Bassler, B.L. (2005). CsrA and three redundant small RNAs regulate quorum sensing in *Vibrio cholerae*. Mol. Microbiol. *58*, 1186–1202.

Leskiw, B.K., Mah, R., Lawlor, E.J., and Chater, K.F. (1993). Accumulation of *bldA*-specified tRNA is temporally regulated in *Streptomyces coelicolor* A3(2). J. Bacteriol. *175*, 1995–2005.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1000 Genome Project Data Processing (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, X., Wang, J., Li, S., Ji, J., Wang, W., and Yang, K. (2015). ScbR- and ScbR2-mediated signal transduction networks coordinate complex physiological responses in Streptomyces coelicolor. Sci. Rep. *5*, 14831.

Li, X., Wang, J., Shi, M., Wang, W., Corre, C., and Yang, K. (2017). Evidence for the formation of ScbR/ScbR2 heterodimers and identification of one of the regulatory targets in Streptomyces coelicolor. Appl. Microbiol. Biotechnol. *101*, 5333–5340.

Liu, G., Chater, K.F., Chandra, G., Niu, G., and Tan, H. (2013). Molecular regulation of antibiotic biosynthesis in *Streptomyces*. Microbiol. Mol. Biol. Rev. *77*, 112–143.

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., and Pan, T. (2015). N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. Nature *518*, 560–564.

Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.-H., Bork, P., Glass, J.I., Serrano, L., and Lluch-Senar, M. (2016). Bacterial antisense RNAs are mainly the product of transcriptional noise. Sci. Adv. *2*, 1–10.

Loh, E., Dussurget, O., Gripenland, J., Vaitkevicius, K., Tiensuu, T., Mandin, P., Repoila, F., Buchrieser, C., Cossart, P., and Johansson, J. (2009). A *trans*-acting riboswitch controls expression of the virulence regulator PrfA in *Listeria monocytogenes*. Cell *139*, 770–779.

Lucks, J.B., Mortimer, S.A., Trapnell, C., Luo, S., Aviran, S., Schroth, G.P., Pachter, L., Doudna, J.A., and Arkin, A.P. (2011). Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc. Natl. Acad. Sci. U. S. A. *108*, 11063–11068.

Ludwig, W., Euzéby, J., Schumann, P., Busse, H.-J., Trujillo, M.E., Kämpfer, P., and Whitman, W.B. (2012). Road map of the phylum Actinobacteria. Bergey's Manual® Syst. Bacteriol. Vol. Five Actinobacteria, Part A 1–28.

Lung, B., Zemann, A., Madej, M.J., Schuelke, M., Techritz, S., Ruf, S., Bock, R., and Hüttenhofer, A. (2006). Identification of small non-coding RNAs from mitochondria and chloroplasts. Nucleic Acids Res. *34*, 3842–3852.

Lybecker, M., Zimmermann, B., Bilusic, I., Tukhtubaeva, N., and Schroeder, R. (2014). The double-stranded transcriptome of *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. *111*, 3134–3139.

Mackie, G.A., and Genereaux, J.L. (1993). The role of RNA structure in determining RNase E-dependent cleavage sites in the mRNA for ribosomal protein S20 in vitro. J. Mol. Biol. *234*, 998–1012.

MacNeil, T., Gewain, K.M., and MacNeil, D.J. (1993). Deletion analysis of the avermectin biosynthetic genes of *Streptomyces avermitilis* by gene cluster displacement. J. Bacteriol. *175*, 2552–2563.

Malpartida, F., and Hopwood, D.A. (1986). Physical and genetic characterisation of the gene cluster for the antibiotic actinorhodin in *Streptomyces coelicolor* A3(2). Mol Gen Genet *205*, 66–73.

Mandal, M., and Breaker, R.R. (2004). Gene regulation by riboswitches. Nat. Rev. Mol. Cell Biol. *5*, 451–463.

Manteca, A., Jung, H.R., Schwämmle, V., Jensen, O.N., and Sanchez, J. (2010). Quantitative proteome analysis of *Streptomyces coelicolor* nonsporulating liquid cultures demonstrates a complex differentiation process comparable to that occurring in sporulating solid cultures. J. Proteome Res. *9*, 4801–4811.

Martín, J.F., Sola-Landa, A., Santos-Beneit, F., Fernández-Martínez, L.T., Prieto, C., and

Rodríguez-García, A. (2011). Cross-talk of global nutritional regulators in the control of primary and secondary metabolism in Streptomyces. Microb. Biotechnol. *4*, 165–174.

Marzi, S., and Romby, P. (2012). RNA mimicry, a decoy for regulatory proteins. Mol. Microbiol. *83*, 1–6.

Massé, E., and Gottesman, S. (2002). A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. *99*, 4620–4625.

Massé, E., Escorcia, F.E., and Gottesman, S. (2003). Coupled degradation of a small regulatory RNA and its mRNA targets in *Escherichia coli*. Genes Dev. *17*, 2374–2383.

Mathews, D.H., Moss, W.N., and Turner, D.H. (2010). Folding and finding RNA secondary structure. Cold Spring Harb. Perspect. Biol. *2*, a003665.

Mathiesen, C., and Hägerhäll, C. (2002). Transmembrane topology of the NuoL, M and N subunits of NADH:quinone oxidoreductase and their homologues among membrane-bound hydrogenases and bona fide antiporters. Biochim. Biophys. Acta *1556*, 121–132.

McCormick, J.R., and Flärdh, K. (2011). Signals and regulators that govern *Streptomyces* development. FEMS Microbiol. Rev. *36*, 206–231.

McGettigan, P.A. (2013). Transcriptomics in the RNA-seq era. Curr. Opin. Chem. Biol. *17*, 4–11.

Mehra, S., Charaniya, S., Takano, E., and Hu, W.-S. (2008). A bistable gene switch for antibiotic biosynthesis: the butyrolactone regulon in *Streptomyces coelicolor*. PLoS One *3*, e2724.

Mitra, A., Angamuthu, K., Jayashree, H.V., and Nagaraja, V. (2009). Occurrence, divergence and evolution of intrinsic terminators across eubacteria. Genomics *94*, 110–116.

Miyakoshi, M., Chao, Y., and Vogel, J. (2015a). Regulatory small RNAs from the 3' regions of bacterial mRNAs. Curr. Opin. Microbiol. *24*, 132–139.

Miyakoshi, M., Chao, Y., and Vogel, J. (2015b). Cross talk between ABC transporter mRNAs via a target mRNA-derived sponge of the GcvB small RNA. EMBO J. *34*, 1478–1492.

Mondal, J., Bratton, B.P., Li, Y., Yethiraj, A., and Weisshaar, J.C. (2011). Entropy-based mechanism of ribosome-nucleoid segregation in *E. coli* cells. Biophys. J. *100*, 2605–2613.

Moody, M.J., Young, R.A., Jones, S.E., and Elliot, M.A. (2013). Comparative analysis of non-coding RNAs in the antibiotic-producing *Streptomyces* bacteria. BMC Genomics *14*, 558.

Moparthi, V.K., and Hägerhäll, C. (2011). The evolution of respiratory chain complex I from a smaller last common ancestor consisting of 11 protein subunits. J. Mol. Evol. *72*, 484–497.

Morita, T., and Aiba, H. (2011). RNase E action at a distance: degradation of target mRNAs mediated by an Hfq-binding small RNA in bacteria. Genes Dev. *25*, 294–298.

Morita, T., Maki, K., and Aiba, H. (2005). RNase E-based ribonucleoprotein complexes: mechanical basis of mRNA destabilization mediated by bacterial noncoding RNAs. Genes Dev. *19*, 2176–2186.

Myronovskyi, M., Welle, E., Fedorenko, V., and Luzhetskyy, A. (2011). Beta-glucuronidase as a sensitive and versatile reporter in actinomycetes. Appl. Environ. Microbiol. *77*, 5370–5383.

Nakano, H., Takehara, E., Nihira, T., and Yamada, Y. (1998). Gene replacement analysis of the *Streptomyces virginiae barA* gene encoding the butyrolactone autoregulator receptor reveals that BarA acts as a repressor in virginiamycin biosynthesis. J. Bacteriol. *180*, 3317–3322.

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J., et al. (2015). Rfam 12.0: Updates to the RNA families database. Nucleic Acids Res. *43*, D130–D137.

Nealson, K.H. (1977). Autoinduction of bacterial luciferase. Occurrence, mechanism and significance. Arch. Microbiol. *112*, 73–79.

Nett, M., Ikeda, H., and Moore, B.S. (2009). Genomic basis for natural product biosynthetic diversity in the actinomycetes. Nat. Prod. Rep. *26*, 1362–1384.

Nevo-Dinur, K., Nussbaum-Shochat, A., Ben-Yehuda, S., and Amster-Choder, O. (2011). Translation-independent localization of mRNA in *E. coli*. Science *331*, 1081–1084.

Nevo-Dinur, K., Govindarajan, S., and Amster-Choder, O. (2012). Subcellular localization of RNA and proteins in prokaryotes. Trends Genet. *7*, 314–322.

Ng, W.-L., and Bassler, B.L. (2009). Bacterial quorum-sensing network architectures. Annu. Rev. Genet. *43*, 197–222.

Nicolas, P., Mäder, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., Bidnenko, E., Marchadier, E., Hoebeke, M., Aymerich, S., et al. (2012). Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. Science *335*, 1103–1106.

Nishida, H., Ohnishi, Y., Beppu, T., and Horinouchi, S. (2007). Evolution of gamma-butyrolactone synthases and receptors in *Streptomyces*. Environ. Microbiol. *9*, 1986–1994.

Ohnishi, Y., Kameyama, S., Onaka, H., and Horinouchi, S. (1999). The A-factor regulatory cascade leading to streptomycin biosynthesis in *Streptomyces griseus*: Identification of a target gene of the A-factor receptor. Mol. Microbiol. *34*, 102–111.

Ohnishi, Y., Yamazaki, H., Kato, J.-Y., Tomono, A., and Horinouchi, S. (2005). AdpA, a central transcriptional regulator in the A-factor regulatory cascade that leads to morphological development and secondary metabolism in *Streptomyces griseus*. Biosci.

Biotechnol. Biochem. *69*, 431–439.

Ohnishi, Y., Ishikawa, J., Hara, H., Suzuki, H., Ikenoya, M., Ikeda, H., Yamashita, A., Hattori, M., and Horinouchi, S. (2008). Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. J. Bacteriol. *190*, 4050–4060.

Omura, S., and Crump, A. (2004). The life and times of ivermectin - a success story. Nat. Rev. Microbiol. *2*, 984–989.

Omura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M., Takahashi, Y., Horikawa, H., Nakazawa, H., Osonoe, T., et al. (2001). Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. Proc. Natl. Acad. Sci. U. S. A. *98*, 12215–12220.

Ourisson, G., Rohmer, M., and Poralla, K. (1987). Prokaryotic hopanoids and other polyterpenoid sterol surrogates. Annu. Rev. Microbiol. *41*, 301–333.

Paget, M.S.B., Chamberlin, L., Atrih, A., Foster, S.J., and Buttner, M.J. (1999). Evidence that the extracytoplasmic function sigma factor σ^E is required for normal cell wall structure in *Streptomyces coelicolor* A3(2). J. Bacteriol. *181*, 204–211.

Pall, G.S., and Hamilton, A.J. (2008). Improved northern blot method for enhanced detection of small RNA. Nat. Protoc. *3*, 1077–1084.

Pánek, J., Bobek, J., Mikulík, K., Basler, M., and Vohradský, J. (2008). Biocomputational prediction of small non-coding RNAs in *Streptomyces*. BMC Genomics *9*, 217.

Park, S.S., Yang, Y.H., Song, E., Kim, E.J., Kim, W.S., Sohng, J.K., Lee, H.C., Liou, K.K., and Kim, B.G. (2009). Mass spectrometric screening of transcriptional regulators involved in antibiotic biosynthesis in *Streptomyces coelicolor* A3(2). J. Ind. Microbiol. Biotechnol. *36*, 1073–1083.

Passador, L., Cook, J.M., Gambello, M.J., Rust, L., and Iglewski, B.H. (1993). Expression of *Pseudomonas aeruginosa* virulence genes requires cell-to-cell communication. Science *260*, 1127–1130.

Pawlik, K., Kotowska, M., and Kolesiński, P. (2010). *Streptomyces coelicolor* A3(2) produces a new yellow pigment associated with the polyketide synthase Cpk. J. Mol. Microbiol. Biotechnol. *19*, 147–151.

Paz, I., Kosti, I., Ares, M., Cline, M., and Mandel-Gutfreund, Y. (2014). RBPmap: A web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res. *42*, W361-367.

Peeters, E., Driessen, R.P.C., Werner, F., and Dame, R.T. (2015). The interplay between nucleoid organization and transcription in archaeal genomes. Nat. Rev. Microbiol. *13*, 333–341.

Peng, L., Yamamoto, H., and Shikanai, T. (2011). Structure and biogenesis of the chloroplast NAD(P)H dehydrogenase complex. Biochim. Biophys. Acta - Bioenerg. *1807*, 945–953.

Pennisi, E. (2012). Genomics. ENCODE project writes eulogy for junk DNA. Science *337*, 1159, 1161.

Peters, J.M., Vangeloff, A.D., and Landick, R. (2011). Bacterial transcription terminators: the RNA 3'-end chronicles. J. Mol. Biol. *412*, 793–813.

Peters, J.M., Mooney, R.A., Grass, J.A., Jessen, E.D., Tran, F., and Landick, R. (2012). Rho and NusG suppress pervasive antisense transcription in *Escherichia coli*. Genes Dev. *26*, 2621–2633.

Petrillo, M., Silvestro, G., Di Nocera, P.P., Boccia, A., and Paolella, G. (2006). Stem-loop structures in prokaryotic genomes. BMC Genomics *7*, 170.

Phadtare, S., and Severinov, K. (2010). RNA remodeling and gene regulation by cold shock proteins. RNA Biol. *7*, 788–795.

Pinto, A.C., Melo-Barbosa, H.P., Miyoshi, A., Silva, A., and Azevedo, V. (2011). Application of RNA-seq to reveal the transcript profile in bacteria. Genet. Mol. Res. *10*, 1707–1718.

Plotkin, J.B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. *12*, 32–42.

Poralla, K., Muth, G., and Härtner, T. (2000). Hopanoids are formed during transition from substrate to aerial hyphae in *Streptomyces coelicolor* A3(2). FEMS Microbiol. Lett. *189*, 93–95.

Prévost, K., Salvail, H., Desnoyers, G., Jacques, J.-F., Phaneuf, E., and Massé, E. (2007). The small RNA RyhB activates the translation of *shiA* mRNA encoding a permease of shikimate, a compound involved in siderophore synthesis. Mol. Microbiol. *64*, 1260–1273.

Prieto-Godino, L.L., Rytz, R., Bargeton, B., Abuin, L., Arguello, J.R., Peraro, M.D., and Benton, R. (2016). Olfactory receptor pseudo-pseudogenes. Nature *539*, 93–97.

Proshkin, S., Rahmouni, A.R., Mironov, A., and Nudler, E. (2010). Cooperation between translating ribosomes and RNA polymerase in transcription elongation. Science *328*, 504–508.

Quax, T.E.F., Claassens, N.J., Söll, D., and van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. Mol. Cell *59*, 149–161.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Raghavan, R., Groisman, E.A., and Ochman, H. (2011). Genome-wide detection of novel regulatory RNAs in *E. coli*. Genome Res. *21*, 1487–1497.

Raghavan, R., Sloan, D.B., and Ochman, H. (2012). Antisense transcription is pervasive but rarely conserved in enteric bacteria. MBio *3*, e00156-12.

Rieder, R., Reinhardt, R., Sharma, C., and Vogel, J. (2012). Experimental tools to identify

RNA-protein interactions in *Helicobacter pylori*. RNA Biol. *9*, 520–531.

Rigali, S., Titgemeyer, F., Barends, S., Mulder, S., Thomae, A.W., Hopwood, D.A., and van Wezel, G.P. (2008). Feast or famine: the global regulator DasR links nutrient stress to antibiotic production by *Streptomyces*. EMBO Rep. *9*, 670–675.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

Rohmer, M., Bouvier, P., and Ourisson, G. (1979). Molecular evolution of biomembranes: structural equivalents and phylogenetic precursors of sterols. Proc. Natl. Acad. Sci. U. S. A. *76*, 847–851.

Sáenz, J.P., Sezgin, E., Schwille, P., and Simons, K. (2012). Functional convergence of hopanoids and sterols in membrane ordering. Proc. Natl. Acad. Sci. U. S. A. *109*, 14236–14240.

Sáenz, J.P., Grosser, D., Bradley, A.S., Lagny, T.J., Lavrynenko, O., Broda, M., and Simons, K. (2015). Hopanoids as functional analogues of cholesterol in bacterial membranes. Proc. Natl. Acad. Sci. *112*, 11971–11976.

Sahm, H., Rohmer, M., Bringer-Meyer, S., Sprenger, G.A., and Welle, R. (1993). Biochemistry and physiology of hopanoids in bacteria. Adv. Microb. Physiol. *35*, 247–273.

Said, N., Rieder, R., Hurwitz, R., Deckert, J., Urlaub, H., and Vogel, J. (2009). In vivo expression and purification of aptamer-tagged small RNA regulators. Nucleic Acids Res. *37*, e133.

Saleh, O., Bonitz, T., Flinspach, K., Kulik, A., Burkard, N., Mühlenweg, A., Vente, A., Polnick, S., Lämmerhofer, M., Gust, B., et al. (2012). Activation of a silent phenazine biosynthetic gene cluster reveals a novel natural product and a new resistance mechanism against phenazines. Medchemcomm *3*, 1009–1019.

Santangelo, T.J., and Artsimovitch, I. (2011). Termination and antitermination: RNA polymerase runs a stop sign. Nat. Rev. Microbiol. *9*, 319–329.

Santos-Beneit, F., Rodríguez-García, A., Sola-Landa, A., and Martín, J.F. (2009). Cross-talk between two global regulators in Streptomyces: PhoP and AfsR interact in the control of *afsS, pstS* and *phoRP* transcription. Mol. Microbiol. *72*, 53–68.

Santos-Beneit, F., Barriuso-Iglesias, M., Fernández-Martínez, L.T., Martínez-Castro, M., Sola-Landa, A., Rodríguez-García, A., and Martín, J.F. (2011). The RNA polymerase omega factor RpoZ is regulated by PhoP and has an important role in antibiotic biosynthesis and morphological differentiation in *Streptomyces coelicolor*. Appl. Environ. Microbiol. *77*, 7586–7594.

Savery, N.J. (2007). The molecular mechanism of transcription-coupled DNA repair. Trends Microbiol. *15*, 326–333.

Sayed, N., Jousselin, A., and Felden, B. (2012). A *cis*-antisense RNA acts in trans in

Staphylococcus aureus to control translation of a human cytolytic peptide. Nat. Struct. Mol. Biol. *19*, 105–112.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

Sesto, N., Wurtzel, O., Archambaud, C., Sorek, R., and Cossart, P. (2012). The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. Nat. Rev. Microbiol. *11*, 75–82.

Shao, Y., and Bassler, B.L. (2012). Quorum-sensing non-coding small RNAs use unique pairing regions to differentially control mRNA targets. Mol. Microbiol. *83*, 599–611.

Sharan, M., Förstner, K.U., Eulalio, A., and Vogel, J. (2017). APRICOT: an integrated computational pipeline for the sequence-based identification and characterization of RNA-binding proteins. Nucleic Acids Res. *45*, e96.

Sharma, C.M., and Vogel, J. (2009). Experimental approaches for the discovery and characterization of regulatory small RNA. Curr. Opin. Microbiol. *12*, 536–546.

Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., et al. (2010). The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature *464*, 250–255.

Shearwin, K.E., Callen, B.P., and Egan, J.B. (2005). Transcriptional interference - A crash course. Trends Genet. *21*, 339–345.

Sheeler, N.L., MacMillan, S. V., and Nodwell, J.R. (2005). Biochemical activities of the AbsA two-component system of *Streptomyces coelicolor*. J. Bacteriol. *187*, 687–696.

Shu, D., Chen, L., Wang, W., Yu, Z., Ren, C., Zhang, W., Yang, S., Lu, Y., and Jiang, W. (2009). AfsQ1-Q2-SigQ is a pleiotropic but conditionally required signal transduction system for both secondary metabolism and morphological development in *Streptomyces coelicolor*. Appl. Microbiol. Biotechnol. *81*, 1149–1160.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

Sittka, A., Sharma, C.M., Rolle, K., and Vogel, J. (2009). Deep sequencing of *Salmonella* RNA associated with heterologous Hfq proteins in vivo reveals small RNAs as a major target class and identifies RNA processing phenotypes. RNA Biol. *6*, 266–275.

Sobrero, P., and Valverde, C. (2012). The bacterial protein Hfq: much more than a mere RNA-binding factor. Crit. Rev. Microbiol. *4*, 276–299.

Sola-Landa, A., Rodríguez-García, A., Franco-Domínguez, E., and Martín, J.F. (2005). Binding of PhoP to promoters of phosphate-regulated genes in *Streptomyces coelicolor*: Identification of PHO boxes. Mol. Microbiol. *56*, 1373–1385.

Song, Y., Liu, Y., Ward, C.B., Mueller, S., Futcher, B., Skiena, S., Paul, A.V., and

Wimmer, E. (2012). Identification of two functionally redundant RNA elements in the coding sequence of poliovirus using computer-generated design. Proc. Natl. Acad. Sci. U. S. A. *109*, 14301–14307.

St-Onge, R. (2016). Regulation of a muralytic resuscitation-promoting factor in Streptomyces coelicolor. PhD Thesis, McMaster Univ.

Stazic, D., Lindell, D., and Steglich, C. (2011). Antisense RNA protects mRNA from RNase E degradation by RNA-RNA duplex formation during phage infection. Nucleic Acids Res. *39*, 4890–4899.

Storz, G., Vogel, J., and Wassarman, K.M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. Mol. Cell *43*, 880–891.

Stuttard, C. (1982). Temperate phages of *Streptomyces venezuelae*: lysogeny and host specificity shown by phages SV1 and SV2. Microbiology *128*, 115–121.

Svoboda, P., and Di Cara, A. (2006). Hairpin RNA: A secondary structure of primary importance. Cell. Mol. Life Sci. *63*, 901–918.

Swiercz, J.P., and Elliot, M.A. (2012). Chapter 4: *Streptomyces* sporulation. In Bacterial Spores: Current Research and Applications, pp. 39–56.

Swiercz, J.P., Hindra, Bobek, J., Bobek, J., Haiser, H.J., Di Berardo, C., Tjaden, B., and Elliot, M.A. (2008). Small non-coding RNAs in *Streptomyces coelicolor*. Nucleic Acids Res. *36*, 7240–7251.

Tahlan, K., Ahn, S.K., Sing, A., Bodnaruk, T.D., Willems, A.R., Davidson, A.R., and Nodwell, J.R. (2007). Initiation of actinorhodin export in *Streptomyces coelicolor*. Mol. Microbiol. *63*, 951–961.

Takano, E. (2006). Gamma-butyrolactones: *Streptomyces* signalling molecules regulating antibiotic production and differentiation. Curr. Opin. Microbiol. *9*, 287–294.

Takano, E., Gramajo, H.C., Strauch, E., Andres, N., White, J., and Bibb, M.J. (1992). Transcriptional regulation of the *redD* transcriptional activator gene accounts for growth-phase-dependent production of the antibiotic undecylprodigiosin in *Streptomyces coelicolor* A3(2). Mol. Microbiol. *6*, 2797–2804.

Takano, E., Chakraburtty, R., Nihira, T., Yamada, Y., and Bibb, M.J. (2001). A complex role for the gamma-butyrolactone SCB1 in regulating antibiotic production in *Streptomyces coelicolor* A3(2). Mol. Microbiol. *41*, 1015–1028.

Takano, E., Tao, M., Long, F., Bibb, M.J., Wang, L., Li, W., Buttner, M.J., Bibb, M.J., Deng, Z.X., and Chater, K.F. (2003). A rare leucine codon in *adpA* is implicated in the morphological defect of *bldA* mutants of *Streptomyces coelicolor*. Mol. Microbiol. *50*, 475–486.

Takano, E., Kinoshita, H., Mersinias, V., Bucca, G., Hotchkiss, G., Nihira, T., Smith, C.P., Bibb, M., Wohlleben, W., and Chater, K. (2005). A bacterial hormone (the SCB1) directly controls the expression of a pathway-specific regulatory gene in the cryptic type I

polyketide biosynthetic gene cluster of *Streptomyces coelicolor*. Mol. Microbiol. *56*, 465–479.

Tezuka, T., Hara, H., Ohnishi, Y., and Horinouchi, S. (2009). Identification and gene disruption of small noncoding RNAs in *Streptomyces griseus*. J. Bacteriol. *191*, 4896–4904.

Thomason, M.K., and Storz, G. (2010). Bacterial antisense RNAs: how many are there, and what are they doing? Annu. Rev. Genet. *44*, 167–188.

Toffano-Nioche, C., Nguyen, A.N., Kuchly, C., Ott, A., Gautheret, D., Bouloc, P., and Jacq, A. (2012). Transcriptomic profiling of the oyster pathogen *Vibrio splendidus* opens a window on the evolutionary dynamics of the small RNA repertoire in the *Vibrio* genus. RNA *18*, 2201–2219.

Toledo-Arana, A., Repoila, F., and Cossart, P. (2007). Small noncoding RNAs controlling pathogenesis. Curr. Opin. Microbiol. *10*, 182–188.

Toledo-Arana, A., Dussurget, O., Nikitas, G., Sesto, N., Guet-Revillet, H., Balestrino, D., Loh, E., Gripenland, J., Tiensuu, T., Vaitkevicius, K., et al. (2009). The *Listeria* transcriptional landscape from saprophytism to virulence. Nature *459*, 950–956.

Torarinsson, E., and Lindgreen, S. (2008). WAR: Webserver for aligning structural RNAs. Nucleic Acids Res. *36*, W79-84.

Tsao, S.W., Rudd, B.A., He, X.G., Chang, C.J., and Floss, H.G. (1985). Identification of a red pigment from *Streptomyces coelicolor* A3(2) as a mixture of prodigiosin derivatives. J. Antibiot. (Tokyo). *38*, 128–131.

Uguru, G.C., Stephens, K.E., Stead, J.A., Towle, J.E., Baumberg, S., and McDowall, K.J. (2005). Transcriptional activation of the pathway-specific regulator of the actinorhodin biosynthetic genes in *Streptomyces coelicolor*. Mol. Microbiol. *58*, 131–150.

Unniraman, S., Chatterji, M., and Nagaraja, V. (2002). A hairpin near the 5' end stabilises the DNA gyrase mRNA in *Mycobacterium smegmatis*. Nucleic Acids Res. *30*, 5376–5381.

Vanderpool, C.K., Balasubramanian, D., and Lloyd, C.R. (2011). Dual-function RNA regulators in bacteria. Biochimie *93*, 1943–1949.

Vecerek, B., Moll, I., and Bläsi, U. (2005). Translational autocontrol of the *Escherichia coli hfq* RNA chaperone gene. RNA *11*, 976–984.

van Vliet, A.H.M. (2010). Next generation sequencing of microbial transcriptomes: challenges and opportunities. FEMS Microbiol. Lett. *302*, 1–7.

Vockenhuber, M.-P., Sharma, C.M., Statt, M.G., Schmidt, D., Xu, Z., Dietrich, S., Liesegang, H., Mathews, D.H., and Suess, B. (2011). Deep sequencing-based identification of small non-coding RNAs in *Streptomyces coelicolor*. RNA Biol. *8*, 468–477.

Vockenhuber, M.P., Heueis, N., and Suess, B. (2015). Identification of *metE* as a second

target of the sRNA *scr5239* in <i>Streptomyces coelicolor<i>. PLoS One *10*.

Vogel, J., and Luisi, B.F. (2011). Hfq and its constellation of RNA. Nat. Rev. Microbiol. *9*, 578–589.

Wade, J.T., and Grainger, D.C. (2014). Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat. Rev. Microbiol. *12*, 647–653.

Wang, J., Wang, W., Wang, L., Zhang, G., Fan, K., Tan, H., and Yang, K. (2011). A novel role of pseudo γ-butyrolactone receptors in controlling γ-butyrolactone biosynthesis in *Streptomyces*. Mol. Microbiol. *82*, 236–250.

Wang, R., Mast, Y., Wang, J., Zhang, W., Zhao, G., Wohlleben, W., Lu, Y., and Jiang, W. (2013). Identification of two-component system AfsQ1/Q2 regulon and its cross-regulation with GlnR in *Streptomyces coelicolor*. Mol. Microbiol. *87*, 30–48.

Washburn, M.P., Wolters, D., and Yates III, J.R. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. *19*, 242–247.

Wassarman, K.M., Repoila, F., Rosenow, C., Storz, G., and Gottesman, S. (2001). Identification of novel small RNAs using comparative genomics and microarrays. Genes Dev. *15*, 1637–1651.

Waters, C.M., and Bassler, B.L. (2005). Quorum sensing: cell-to-cell communication in bacteria. Annu. Rev. Cell Dev. Biol. *21*, 319–346.

Waters, L.S., and Storz, G. (2009). Regulatory RNAs in bacteria. Cell *136*, 615–628.

Weerakoon, D.R., and Olson, J.W. (2008). The *Campylobacter jejuni* NADH:ubiquinone oxidoreductase (complex I) utilizes flavodoxin rather than NADH. J. Bacteriol. *190*, 915–925.

Wiese, S., Reidegeld, K.A., Meyer, H.E., and Warscheid, B. (2007). Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. Proteomics *7*, 340–350.

Williamson, L., Saponaro, M., Boeing, S., East, P., Mitter, R., Kantidakis, T., Kelly, G.P., Lobley, A., Walker, J., Spencer-Dene, B., et al. (2017). UV irradiation induces a non-coding RNA that functionally opposes the protein encoded by the same gene. Cell *168*, 843–855.

Williamson, N.R., Fineran, P.C., Leeper, F.J., and Salmond, G.P.C. (2006). The biosynthesis and regulation of bacterial prodiginines. Nat. Rev. Microbiol. *4*, 887–899.

Withall, D.M., Haynes, S.W., and Challis, G.L. (2015). Stereochemistry and mechanism of undecylprodigiosin oxidative carbocyclization to streptorubin B by the rieske oxygenase RedG. J. Am. Chem. Soc. *137*, 7889–7897.

Wolfe, S. a, and Smith, J.M. (1988). Nucleotide sequence and analysis of the *purA* gene encoding adenylosuccinate synthetase of *Escherichia coli* K12. J. Biol. Chem. *263*, 19147–19153.

Wright, L.F., and Hopwood, D.A. (1976). Actinorhodin is a chromosomally-determined antibiotic in *Streptomyces coelicolor* A3(2). J. Gen. Microbiol. *96*, 289–297.

Xu, G., Wang, J., Wang, L., Tian, X., Yang, H., Fan, K., Yang, K., and Tan, H. (2010). Pseudo gamma-butyrolactone receptors respond to antibiotic signals to coordinate antibiotic biosynthesis. J. Biol. Chem. *285*, 27440–27448.

Xu, Y., Willems, A., Au-Yeung, C., Tahlan, K., and Nodwell, J.R. (2012). A two-step mechanism for the activation of actinorhodin export and resistance in *Streptomyces coelicolor*. MBio *3*, e00191-12.

Yang, Y.-H., Song, E., Kim, J.-N., Lee, B.-R., Kim, E.-J., Park, S.-H., Kim, W.-S., Park, H.-Y., Jeon, J.-M., Rajesh, T., et al. (2012). Characterization of a new ScbR-like γ-butyrolactone binding regulator (SlbR) in *Streptomyces coelicolor*. Appl. Microbiol. Biotechnol. *96*, 113–121.

Yao, Z., Weinberg, Z., and Ruzzo, W.L. (2006). CMfinder--a covariance model based RNA motif finding algorithm. Bioinformatics *22*, 445–452.

Yim, G., Wang, H.H., and Davies, J. (2006). The truth about antibiotics. Int. J. Med. Microbiol. *296*, 163–170.

Yim, G., Wang, H.H., and Davies, J. (2007). Antibiotics as signalling molecules. Philos. Trans. R. Soc. London Ser. B, Biol. Sci. *362*, 1195–1200.

Yu, Z., Zhu, H., Dang, F., Zhang, W., Qin, Z., Yang, S., Tan, H., Lu, Y., and Jiang, W. (2012). Differential regulation of antibiotic biosynthesis by DraR-K, a novel two-component system in *Streptomyces coelicolor*. Mol. Microbiol. *85*, 535–556.

Zhang, A., Wassarman, K.M., Rosenow, C., Tjaden, B.C., Storz, G., and Gottesman, S. (2003). Global analysis of small RNA and mRNA targets of Hfq. Mol. Microbiol. *50*, 1111–1124.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415.

Zurawski, G., and Zurawski, S.M. (1985). Structure of the *Escherichia coli* S10 ribosomal protein operon. Nucleic Acids Res. *13*, 4521–4526.

Appendix 1: Large Tables from Chapter 3 and Chapter 5

Table 3.1: Classification of antisense RNAs above threshold in *S. coelicolor*, *S. avermitilis*, and *S. venezuelae*

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco0058	-	-	-	Y	-
sco0102	Y	-	-	-	-
sco0103	Y	-	-	-	-
sco0166	-	Y	-	-	-
sco0195	Y	-	-	-	-
sco0224	-	Y	-	-	-
sco0297	Y	-	-	-	-
sco0298	Y	-	-	-	-
sco0324	-	Y	-	-	-
sco0402	-	Y	-	-	-
sco0412	-	Y	-	-	-
sco0460	-	Y	-	-	-
sco0477	-	Y	-	-	-
sco0490	-	Y	-	-	Coelichelin
sco0491	-	Y	-	-	Coelichelin
sco0500[d]	-	-	-	Y	-
sco0562	-	-	-	Y	-
sco0591	-	-	-	Y	-
sco0592	-	Y	-	-	-
sco0681	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco0682	-	Y	-	-	-
sco0695	-	-	-	Y	-
sco0729	-	-	-	Y	-
sco0732	-	Y	-	-	-
sco0733	-	-	-	Y	-
sco0735	-	Y	-	-	-
sco0742	-	Y	-	-	-
sco0763	-	Y	-	-	-
sco0798	-	Y	-	-	-
sco0884	-	Y	-	-	-
sco0908	-	Y	-	-	-
sco0909	-	Y	-	-	-
sco0927	-	Y	-	-	-
sco0954	Y	-	-	-	-
sco0973	-	Y	-	-	-
sco0974	-	Y	-	-	-
sco0975	-	Y	-	-	-
sco0976	-	Y	-	-	-
sco0988	-	-	-	Y	-
sco0989	-	Y	-	-	-
sco0998	-	-	-	Y	-
sco1019	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco1025	-	-	-	Y	-
sco1039	-	Y	-	-	-
sco1069	-	-	-	Y	-
sco1071	Y	-	-	-	-
sco1075	-	Y	-	-	-
sco1086	-	Y	-	-	-
sco1104	-	Y	-	-	-
sco1117	-	-	-	Y	-
sco1140	-	-	-	Y	-
sco1142	-	-	-	Y	-
sco1150[d]	-	Y	-	-	-
sco1153	-	Y	-	-	-
sco1154	-	-	-	Y	-
sco1157	-	Y	-	-	-
sco1165	-	Y	-	-	-
sco1172	Y	-	-	-	-
sco1193	-	Y	-	-	-
sco1211	-	-	-	Y	-
sco1224	-	-	-	Y	-
sco1240	Y	-	-	-	-
sco1242	-	-	-	Y	-
sco1243	-	-	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco1248	-	Y	-	-	-
sco1261	-	Y	-	-	-
sco1293	-	-	-	Y	-
sco1294	-	Y	-	-	-
sco1297	Y	-	-	-	-
sco1315	-	Y	-	-	-
sco1320	-	Y	-	-	-
sco1334	-	-	-	Y	-
sco1335	-	-	-	Y	-
sco1347	-	-	-	-	-
sco1358	-	Y	-	-	-
sco1394	-	-	-	Y	-
sco1404	-	Y	-	-	-
sco1427	-	Y	-	-	-
sco1428	-	Y	-	-	-
sco1451	-	Y	-	-	-
sco1489	-	-	-	Y	-
sco1504	-	-	-	Y	-
sco1530	Y	-	-	-	-
sco1535	Y	-	-	-	-
sco1542	-	Y	-	-	-
sco1544	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco1545	-	Y	-	-	-
sco1552	-	-	-	Y	-
sco1562	-	Y	-	-	-
sco1588	-	-	-	Y	-
sco1589	-	Y	-	-	-
sco1590	-	-	-	Y	-
sco1602	-	-	-	Y	-
sco1610	-	Y	-	-	-
sco1622	Y	-	-	-	-
sco1624	Y	-	-	-	-
sco1625	-	Y	-	-	-
sco1650	-	Y	-	-	-
sco1651	Y	-	-	-	-
sco1669	Y	-	-	-	-
sco1671	-	Y	-	-	-
sco1673	Y	-	-	-	-
sco1674	-	Y	-	-	-
sco1690	-	-	-	Y	-
sco1698	-	Y	-	-	-
sco1730	-	Y	-	-	-
sco1750	-	Y	-	-	-
sco1753	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco1757	Y	-	-	-	-
sco1764	-	-	-	Y	-
sco1784	-	Y	-	-	-
sco1795	-	-	Y	-	-
sco1796	-	-	-	Y	-
sco1800	-	-	-	Y	-
sco1808	-	Y	-	-	-
sco1809	-	Y	-	-	-
sco1813	-	-	-	Y	-
sco1821	-	-	-	-	-
sco1831	-	Y	-	-	-
sco1832	-	Y	-	-	-
sco1837	-	Y	-	-	-
sco1841	-	Y	-	-	-
sco1842	-	Y	-	-	-
sco1859	Y	-	-	-	-
sco1903	-	Y	-	-	-
sco1904	-	Y	-	-	-
sco1908	-	Y	-	-	-
sco1909	-	Y	-	-	-
sco1929	Y	-	-	-	-
sco1931	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco1961	-	-	-	Y	-
sco1966	-	-	-	Y	-
sco1975	-	Y	-	-	-
sco1977	-	-	-	Y	-
sco1981	-	Y	-	-	-
sco1988	-	-	-	Y	-
sco1992	-	Y	-	-	-
sco2000	-	-	-	Y	-
sco2009	-	-	-	Y	-
sco2012	-	-	-	Y	-
sco2022	-	Y	-	-	-
sco2027	-	-	-	Y	-
sco2068	-	Y	-	-	-
sco2077	-	-	-	Y	-
sco2100	-	-	-	Y	-
sco2101	-	Y	-	-	-
sco2113	-	Y	-	-	-
sco2114	-	Y	-	-	-
sco2124	-	Y	-	-	-
sco2125	-	Y	-	-	-
sco2153	Y	-	-	-	-
sco2157	Y	-	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco2169	-	-	-	Y	-
sco2171	-	Y	-	-	-
sco2190	-	-	-	Y	-
sco2204	-	Y	-	-	-
sco2219	-	-	-	Y	-
sco2224	-	-	-	Y	-
sco2234[a]	Y	-	-	-	-
sco2237	-	Y	-	-	-
sco2238	-	Y	-	-	-
sco2244	-	-	-	Y	-
sco2251	-	Y	-	-	-
sco2259	-	Y	-	-	-
sco2260	-	Y	-	-	-
sco2268	-	Y	-	-	-
sco2269	-	-	-	Y	-
sco2274	-	-	-	Y	-
sco2279	-	Y	-	-	-
sco2293	-	-	-	Y	-
sco2299	-	Y	-	-	-
sco2311	Y	-	-	-	-
sco2320	-	-	-	Y	-
sco2331	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco2338	-	Y	-	-	-
sco2346	-	-	-	Y	-
sco2351	-	Y	-	-	-
sco2357	-	Y	-	-	-
sco2358	-	Y	-	-	-
sco2364	Y	Y	-	-	-
sco2372	-	Y	-	-	-
sco2376	-	-	-	-	-
sco2383	Y	-	-	-	-
sco2384	-	Y	-	-	-
sco2385	-	-	-	Y	-
sco2438	-	Y	-	-	-
sco2439	-	Y	-	-	-
sco2465	-	-	-	Y	-
sco2484	-	-	-	Y	-
sco2497	-	-	-	Y	-
sco2509	-	Y	-	-	-
sco2526	-	Y	-	-	-
sco2527	-	Y	-	-	-
sco2540	-	-	-	-	-
sco2557	-	Y	-	-	-
sco2581	-	-	Y	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco2583	-	Y	-	-	-
sco2595	Y	-	-	-	-
sco2618	Y	-	-	-	-
sco2626	-	Y	-	-	-
sco2634	-	-	-	Y	-
sco2654	-	Y	-	-	-
sco2670	-	Y	-	-	-
sco2674	-	Y	-	-	-
sco2685	Y	-	-	-	-
sco2726	-	-	-	Y	-
sco2772	-	Y	-	-	-
sco2786	-	Y	-	-	-
sco2789	-	Y	-	-	-
sco2790	-	Y	-	-	-
sco2806	Y	-	-	-	-
sco2810	Y	-	-	-	-
sco2823	Y	-	-	-	-
sco2838	-	Y	-	-	-
sco2884	-	Y	-	-	-
sco2885	-	Y	-	-	-
sco2888	Y	-	-	-	-
sco2893	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco2897	-	-	-	Y	-
sco2902	-	Y	-	-	-
sco2906	-	-	-	Y	-
sco2910	-	-	-	Y	-
sco2917	-	-	-	Y	-
sco2918	-	Y	-	-	-
sco2919	-	Y	-	-	-
sco2921	-	Y	-	-	-
sco2922	-	Y	-	-	-
sco2925	-	Y	-	-	-
sco2926	-	Y	-	-	-
sco2938	-	-	-	Y	-
sco2943	-	-	-	Y	-
sco2949	-	-	-	Y	-
sco2975	-	Y	-	-	-
sco2985	-	Y	-	-	-
sco2989	-	Y	-	-	-
sco3007[c]	-	Y	-	-	-
sco3015	Y	-	-	-	-
sco3017	-	-	-	Y	-
sco3023	-	-	-	Y	-
sco3024	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco3039	-	-	-	Y	-
sco3043	-	-	-	Y	-
sco3047	-	Y	-	-	-
sco3049	-	Y	-	-	-
sco3050	-	Y	-	-	-
sco3052	-	-	-	Y	-
sco3053	-	Y	-	-	-
sco3057	-	-	-	Y	-
sco3076	-	Y	-	-	-
sco3080	-	Y	-	-	-
sco3087	-	-	-	Y	-
sco3106	-	-	-	Y	-
sco3107	-	Y	-	-	-
sco3118	-	-	-	Y	-
sco3133	-	Y	-	-	-
sco3158	-	Y	-	-	-
sco3173	-	-	-	Y	-
sco3175	Y	-	-	-	-
sco3176	-	-	-	Y	-
sco3191	-	Y	-	-	-
sco3195	-	Y	-	-	-
sco3196	-	Y	-	-	-

Sense gene[abcd]	**cis-antisense RNA**	**cutoRNA**	**Divergent transcription**	**Unclassified***	**Biosynthetic clusters**[x]
sco3202	Y	-	-	-	-
sco3203	-	-	-	Y	-
sco3227	-	Y	-	-	Calcium dependent antibiotic
sco3275	-	-	-	Y	-
sco3283	-	-	-	Y	-
sco3287[ab]	-	-	-	Y	-
sco3289	-	-	-	Y	-
sco3290	Y	-	-	-	-
sco3291	Y	-	-	-	-
sco3298	-	Y	-	-	-
sco3310	-	-	-	Y	-
sco3317[a]	Y	-	-	-	-
sco3318	Y	-	-	-	-
sco3320	Y	-	-	-	-
sco3321[a]	-	-	-	Y	-
sco3324	-	Y	-	-	-
sco3325	-	Y	-	-	-
sco3344	-	Y	-	-	-
sco3349	-	-	-	-	-
sco3350	-	-	-	Y	-
sco3352	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco3360	-	Y	-	-	-
sco3364	-	Y	-	-	-
sco3372	-	-	-	Y	-
sco3374	-	Y	-	-	-
sco3394	-	Y	-	-	-
sco3402	-	Y	-	-	-
sco3403	-	Y	-	-	-
sco3408	Y	-	-	-	-
sco3410	-	-	-	Y	-
sco3413	-	Y	-	-	-
sco3423	-	Y	-	-	-
sco3424	-	Y	-	-	-
sco3432	-	Y	-	-	-
sco3433	Y	-	-	-	-
sco3481	Y	-	-	-	-
sco3482	Y	-	-	-	-
sco3483	Y	-	-	-	-
sco3487	Y	-	-	-	-
sco3497	-	-	-	Y	-
sco3503	-	-	-	Y	-
sco3505	-	-	-	Y	-
sco3535	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco3536	-	Y	-	-	-
sco3550	-	-	Y	-	-
sco3559	-	-	-	Y	-
sco3560	Y	-	-	-	-
sco3567	-	-	-	Y	-
sco3572	-	Y	-	-	-
sco3578	-	Y	-	-	-
sco3579	-	-	-	Y	-
sco3602	-	-	-	-	-
sco3606	-	Y	-	-	-
sco3620	Y	-	-	-	-
sco3621	-	-	-	Y	-
sco3624	-	Y	-	-	-
sco3636	-	-	-	Y	-
sco3638	-	-	-	Y	-
sco3665	-	-	-	Y	-
sco3671	Y	-	-	-	-
sco3693	-	Y	-	-	-
sco3712	-	Y	-	-	-
sco3713	-	Y	-	-	-
sco3734	-	Y	-	-	-
sco3769	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco3793	-	Y	-	-	-
sco3794	-	Y	-	-	-
sco3825	-	Y	-	-	-
sco3832	-	Y	-	-	-
sco3836	-	Y	-	-	-
sco3855	-	Y	-	-	-
sco3902	-	Y	-	-	-
sco3910	-	-	-	Y	-
sco3920	-	-	-	Y	-
sco3927	-	Y	-	-	-
sco3939	-	Y	-	-	-
sco3941	Y	-	-	-	-
sco3942	-	-	-	Y	-
sco3949	-	Y	-	-	-
sco3960	-	-	-	Y	-
sco3973	-	-	-	Y	-
sco3976	-	-	-	-	-
sco3989	Y	-	-	-	-
sco3991	-	-	-	Y	-
sco4008	-	-	-	Y	-
sco4014	-	Y	-	-	-
sco4025	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco4034	-	-	-	Y	-
sco4038	Y	-	-	-	-
sco4041	-	-	-	Y	-
sco4042	-	Y	-	-	-
sco4056	-	Y	-	-	-
sco4057	-	-	-	Y	-
sco4079	-	-	-	Y	-
sco4080	-	Y	-	-	-
sco4087	-	Y	-	-	-
sco4088	-	Y	-	-	-
sco4090	-	-	-	Y	-
sco4118	Y	-	-	-	-
sco4134	-	Y	-	-	-
sco4135	-	Y	-	-	-
sco4136	-	-	-	Y	-
sco4137	Y	-	-	-	-
sco4164	Y	-	-	-	-
sco4172	-	Y	-	-	-
sco4182	-	Y	-	-	-
sco4183	-	-	-	Y	-
sco4186	-	Y	-	-	-
sco4188	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco4201	Y	-	-	-	-
sco4206	-	-	-	Y	-
sco4223	-	-	-	Y	-
sco4232	-	-	-	Y	-
sco4259	-	-	-	Y	-
sco4261[a]	Y	-	-	-	-
sco4262	Y	-	-	-	-
sco4264	-	-	-	Y	-
sco4269	-	-	-	Y	-
sco4279	-	Y	-	-	-
sco4283[d]	-	Y	-	-	-
sco4290	-	Y	-	-	-
sco4298	-	Y	-	-	-
sco4299	-	Y	-	-	-
sco4310	-	Y	-	-	-
sco4316	-	-	-	Y	-
sco4317	-	-	-	Y	-
sco4318	-	Y	-	-	-
sco4323	-	Y	-	-	-
sco4324	-	-	-	Y	-
sco4325	-	Y	-	-	-
sco4326	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco4334	-	Y	-	-	-
sco4358	-	-	-	Y	-
sco4365	-	Y	-	-	-
sco4371	Y	-	-	-	-
sco4388	-	-	-	Y	-
sco4401	-	Y	-	-	-
sco4407	-	Y	-	-	-
sco4413	-	-	-	Y	-
sco4424	-	Y	-	-	-
sco4425[c]	-	Y	-	-	-
sco4426	-	-	-	Y	-
sco4430	-	Y	-	-	-
sco4431	-	Y	-	-	-
sco4440	Y	-	-	-	-
sco4443	-	Y	-	-	-
sco4447	-	-	-	Y	-
sco4474	-	-	-	Y	-
sco4488	-	Y	-	-	-
sco4489	-	-	-	Y	-
sco4511	-	Y	-	-	-
sco4524	Y	-	-	-	-
sco4536	Y	-	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco4537	Y	-	-	-	-
sco4557	-	Y	-	-	-
sco4560	-	-	-	Y	-
sco4564	Y	-	-	-	-
sco4565	Y	-	-	-	-
sco4566[a]	Y	-	-	-	-
sco4567[a]	Y	-	-	-	-
sco4579	-	-	-	Y	-
sco4588	-	Y	-	-	-
sco4589	-	Y	-	-	-
sco4591	-	Y	-	-	-
sco4602	Y	-	-	-	-
sco4603	Y	-	-	-	-
sco4604	Y	-	-	-	-
sco4605	Y	-	-	-	-
sco4606	Y	-	-	-	-
sco4607	Y	-	-	-	-
sco4613	-	Y	-	-	-
sco4614	-	Y	-	-	-
sco4632	-	-	-	Y	-
sco4657	-	-	-	Y	-
sco4662	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco4680	-	Y	-	-	-
sco4692[a]	-	-	-	Y	-
sco4693[a]	Y	-	-	-	-
sco4699[a]	-	-	-	Y	-
sco4702	-	-	-	Y	-
sco4721	-	-	-	Y	-
sco4731	-	Y	-	-	-
sco4737	-	-	-	Y	-
sco4749[d]	-	Y	-	-	-
sco4761	Y	-	-	-	-
sco4772	-	-	-	-	-
sco4773	-	Y	-	-	-
sco4789	-	Y	-	-	-
sco4792	-	Y	-	-	-
sco4793	-	Y	-	-	-
sco4794	-	-	-	Y	-
sco4795	-	-	-	Y	-
sco4801	-	Y	-	-	-
sco4811	-	-	-	Y	-
sco4825	Y	-	-	-	-
sco4829	Y	-	-	-	-
sco4840	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco4841	-	Y	-	-	-
sco4846	-	Y	-	-	-
sco4848	Y	-	-	-	-
sco4872	-	-	-	Y	-
sco4909	-	-	-	Y	-
sco4912	-	Y	-	-	-
sco4921	-	-	-	Y	-
sco4928	-	Y	-	-	-
sco4929	-	Y	-	-	-
sco4933	-	-	-	Y	-
sco4953	-	Y	-	-	-
sco4957	-	-	-	Y	-
sco4960	-	-	-	Y	-
sco4965	Y	-	-	-	-
sco4969	-	-	-	Y	-
sco4975	-	-	-	Y	-
sco4995	-	Y	-	-	-
sco5049	-	Y	-	-	-
sco5054	-	Y	-	-	-
sco5060	-	Y	-	-	-
sco5062	-	-	-	Y	-
sco5063	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco5073	Y	-	-	-	Actinorhodin
sco5080	Y	-	-	-	Actinorhodin
sco5081	-	Y	-	-	Actinorhodin
sco5082	-	Y	-	-	Actinorhodin
sco5099	Y	-	-	-	-
sco5106[d]	-	Y	-	-	-
sco5121	Y	-	-	Y	-
sco5123	-	Y	-	-	-
sco5124	-	Y	-	-	-
sco5125	-	Y	-	-	-
sco5134	-	-	-	Y	-
sco5138	-	-	-	-	-
sco5143	-	-	-	Y	-
sco5146[d]	-	Y	-	-	-
sco5151	-	Y	-	-	-
sco5152	-	Y	-	-	-
sco5155	-	Y	-	-	-
sco5164	-	Y	-	-	-
sco5167	-	Y	-	-	-
sco5168	-	Y	-	-	-
sco5177	-	-	-	Y	-
sco5178	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco5182	-	-	-	Y	-
sco5183	-	-	-	Y	-
sco5197	-	Y	-	-	-
sco5198	-	Y	-	-	-
sco5200	-	-	Y	-	-
sco5202	-	Y	-	-	-
sco5203	-	-	Y	Y	-
sco5209	-	-	-	Y	-
sco5223	-	Y	-	-	Albaflavenone
sco5224	-	-	-	Y	-
sco5241	Y	-	-	-	-
sco5242	-	Y	-	-	-
sco5243	-	-	-	Y	-
sco5255	-	Y	-	-	-
sco5256	-	Y	-	-	-
sco5294	-	Y	-	-	-
sco5329	Y	-	-	-	-
sco5330	-	Y	-	-	-
sco5333	-	Y	-	-	-
sco5334	-	-	-	Y	-
sco5356	-	-	-	Y	-
sco5380	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco5389	-	-	-	Y	-
sco5390	-	Y	-	-	-
sco5392	-	-	-	Y	-
sco5398	Y	-	-	-	-
sco5410	-	-	-	Y	-
sco5411	-	-	-	Y	-
sco5418	-	-	-	Y	-
sco5422	-	-	-	Y	-
sco5427	-	Y	-	-	-
sco5445	-	-	-	Y	-
sco5465	-	Y	-	-	-
sco5466	-	Y	-	-	-
sco5488	-	Y	-	-	-
sco5489	-	Y	-	-	-
sco5506	-	-	-	Y	-
sco5507	-	Y	-	-	-
sco5516	-	-	-	Y	-
sco5518	-	-	-	Y	-
sco5543	Y	-	-	-	-
sco5545	Y	-	-	-	-
sco5558	Y	-	-	-	-
sco5561	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco5573	-	Y	-	-	-
sco5576	-	-	-	Y	-
sco5581	-	Y	-	-	-
sco5630	-	-	-	Y	-
sco5641	-	-	-	Y	-
sco5650	-	-	-	Y	-
sco5661	Y	-	-	-	-
sco5664	-	Y	-	-	-
sco5706	-	-	-	Y	-
sco5709	-	-	-	Y	-
sco5722	-	Y	-	-	-
sco5727	-	Y	-	-	-
sco5760	-	Y	-	-	-
sco5781	-	Y	-	-	-
sco5782	-	-	-	Y	-
sco5788	-	-	-	Y	-
sco5802	-	-	-	Y	-
sco5810	-	-	-	Y	-
sco5814	-	Y	-	-	-
sco5821	-	-	-	Y	-
sco5828	-	Y	-	-	-
sco5835	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco5841	-	-	-	Y	-
sco5842	Y	-	-	-	-
sco5856	-	Y	-	-	-
sco5866	-	Y	-	-	-
sco5897	Y	-	-	-	Prodiginine
sco5986	-	Y	-	-	-
sco5988	-	Y	-	-	-
sco5989	-	-	-	Y	-
sco6002	-	Y	-	-	-
sco6003	-	-	-	Y	-
sco6029	Y	-	-	-	-
sco6039	-	-	-	Y	-
sco6042	-	Y	-	-	-
sco6043	-	Y	-	-	-
sco6066	-	-	-	Y	-
sco6077	-	Y	-	-	-
sco6085	-	Y	-	-	-
sco6093	-	-	-	Y	-
sco6106	-	-	-	Y	-
sco6127	-	Y	-	-	-
sco6156	-	-	-	Y	-
sco6196	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco6222	-	Y	-	-	-
sco6246	-	-	-	-	-
sco6265	-	-	-	Y	-
sco6266	Y	-	-	-	SCB1
sco6268	-	Y	-	-	-
sco6271	-	-	-	Y	-
sco6273	-	-	-	-	yCPK
sco6274	-	-	-	Y	yCPK
sco6277[d]	Y	-	-	-	yCPK
sco6281	Y	-	-	-	yCPK
sco6284	-	-	-	Y	yCPK
sco6285	-	-	-	Y	yCPK
sco6311	-	-	-	Y	-
sco6312	-	-	-	Y	-
sco6373	Y	-	-	-	-
sco6375	Y	-	-	-	-
sco6377	-	-	-	Y	-
sco6390	-	-	-	Y	-
sco6393	Y	-	-	-	-
sco6394	Y	-	-	-	-
sco6396	-	-	-	Y	-
sco6422	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco6438	-	Y	-	-	Dipeptide
sco6455	-	Y	-	-	-
sco6477	-	-	-	Y	-
sco6492	-	Y	-	-	-
sco6598	Y	-	-	-	-
sco6624[c]	Y	-	-	-	-
sco6636	-	Y	-	-	-
sco6646	-	Y	-	-	-
sco6654	-	-	-	Y	-
sco6704	-	-	-	Y	-
sco6715	Y	-	-	-	-
sco6716[d]	-	Y	-	-	-
sco6717	-	Y	-	-	-
sco6721[a]	-	-	-	Y	-
sco6723	-	-	-	Y	-
sco6725	-	Y	-	-	-
sco6729[d]	-	Y	-	-	-
sco6733	-	-	-	Y	-
sco6735	-	Y	-	-	-
sco6736	-	-	-	-	-
sco6740	-	Y	-	-	-
sco6742	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco6749	-	Y	-	-	-
sco6760	-	-	-	Y	Hopene
sco6761	Y	-	-	-	Hopene
sco6762	Y	-	-	-	Hopene
sco6770	-	-	-	Y	Hopene
sco6771	-	Y	-	-	
sco6772	-	Y	-	-	-
sco6775	Y	-	-	-	-
sco6777	-	Y	-	-	-
sco6828	-	-	-	Y	-
sco6861	-	-	-	Y	-
sco6947	-	-	-	Y	-
sco6949	-	-	-	Y	-
sco6998	-	Y	-	-	-
sco7009	-	Y	-	-	-
sco7022	Y	-	-	-	-
sco7047	-	-	-	Y	-
sco7101	-	-	-	Y	-
sco7173	-	-	-	Y	-
sco7219	-	-	-	Y	-
sco7234	-	-	-	Y	-
sco7249	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sco7250	-	Y	-	-	-
sco7252	-	-	-	-	-
sco7253	-	Y	-	-	-
sco7258	-	Y	-	-	-
sco7268	-	Y	-	-	-
sco7300	-	Y	-	-	-
sco7343	-	Y	-	-	-
sco7455	-	-	-	Y	-
sco7456	-	-	-	Y	-
sco7480	-	Y	-	-	-
sco7603	-	-	-	Y	-
sco7605	-	Y	-	-	-
sco7652	-	Y	-	-	-
sco7653	-	-	-	Y	-
sco7716	-	Y	-	-	-
sco7717	Y	-	-	-	-
sco7718	-	-	-	Y	-
sco7757	-	-	-	Y	-
sco7835	-	-	-	Y	-
sav0047	-	-	-	Y	-
sav0112	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav0164	-	-	-	Y	-
sav0193	-	-	-	Y	-
sav0200	-	-	-	Y	-
sav0220	-	-	-	Y	-
sav0234	-	-	-	Y	-
sav0243	-	-	-	Y	-
sav0244	-	-	-	Y	-
sav0247	-	-	-	Y	-
sav0259	-	-	-	Y	Microcin
sav0263	-	-	-	Y	-
sav0272	-	-	-	Y	-
sav0274	-	-	-	Y	-
sav0277	-	-	-	Y	-
sav0334	-	-	-	Y	-
sav0410	Y	Y	-	-	-
sav0435	-	-	-	Y	-
sav0514	-	-	-	Y	-
sav0541	-	-	-	Y	-
sav0573	-	-	-	Y	-
sav0675	-	-	-	Y	-
sav0749	-	-	-	Y	-
sav0779	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav0783	-	-	-	Y	-
sav0784	-	-	-	Y	-
sav0824	-	-	-	Y	-
sav0825	-	-	-	Y	-
sav0886	-	-	-	Y	-
sav0896	-	-	-	Y	-
sav0897	-	-	-	Y	-
sav0935	-	Y	-	-	Avermectin
sav0936	Y	Y	-	-	Avermectin
sav0937	Y	-	-	-	Avermectin
sav0938	-	-	-	Y	Avermectin
sav0939	-	-	-	Y	Avermectin
sav0942	-	-	-	Y	Avermectin
sav0945	-	Y	-	-	Avermectin
sav0946	-	Y	-	-	Avermectin
sav0948	-	-	-	Y	Avermectin
sav0949	-	Y	-	-	Avermectin
sav0962	-	-	-	Y	-
sav0980	-	-	-	Y	-
sav1016	-	-	-	Y	-
sav1043	-	-	-	Y	-
sav1047	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav1048	Y	Y	-	-	-
sav1060	-	-	-	Y	-
sav1083	-	-	-	Y	-
sav1090	Y	-	-	-	-
sav1092	-	Y	-	-	-
sav1100	-	-	-	Y	-
sav1104	-	-	-	Y	-
sav1111	-	-	-	Y	-
sav1118	-	-	-	Y	-
sav1138	-	Y	-	-	-
sav1161	-	-	-	Y	-
sav1162	-	Y	-	-	-
sav1164	-	-	-	Y	-
sav1187	-	-	-	Y	-
sav1235	-	Y	-	-	-
sav1236	-	Y	-	-	-
sav1237	-	-	-	Y	-
sav1238	-	Y	-	-	-
sav1247	-	-	-	Y	-
sav1300	-	-	-	Y	-
sav1308	-	-	-	Y	-
sav1332	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav1410	-	-	-	Y	-
sav1423	-	-	-	Y	-
sav1463	-	-	-	Y	-
sav1486	-	Y	-	-	-
sav1502	-	-	-	Y	-
sav1526	-	-	-	Y	-
sav1572	-	-	-	Y	-
sav1574	-	-	-	Y	-
sav1578	-	-	-	Y	-
sav1589	-	Y	-	-	-
sav1635	-	-	-	Y	-
sav1642	-	Y	-	-	-
sav1644	-	-	-	Y	-
sav1652	Y	-	-	-	Hopene
sav1663	-	Y	-	-	-
sav1664	-	-	-	Y	-
sav1676	Y	-	-	-	-
sav1684	-	-	-	Y	-
sav1692	-	-	-	Y	-
sav1704	-	-	-	Y	-
sav1705	-	-	-	Y	-
sav1792	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav1799	-	-	-	Y	-
sav1837	-	Y	-	-	-
sav1895	-	Y	-	-	-
sav1939	-	-	-	Y	-
sav1956	-	-	-	Y	-
sav1959	-	-	-	Y	-
sav2001	-	-	-	Y	-
sav2034	-	-	-	Y	-
sav2048	-	-	-	Y	-
sav2061	-	-	-	Y	-
sav2075	-	-	-	Y	-
sav2085	-	-	-	Y	-
sav2089	-	-	-	Y	-
sav2116	-	-	-	Y	-
sav2160	-	-	-	Y	-
sav2169	-	-	-	Y	-
sav2203	-	-	-	Y	-
sav2220	-	-	-	Y	-
sav2225	-	-	-	Y	-
sav2227	-	-	-	Y	-
sav2232	-	-	-	Y	-
sav2261	-	-	-	Y	-

Sense gene[abcd]	**cis-antisense RNA**	**cutoRNA**	**Divergent transcription**	**Unclassified***	**Biosynthetic clusters**[x]
sav2268	-	-	-	Y	2-alkyl-4-hydroxymethyl furan-3-carboxylic acids
sav2272	-	Y	-	-	-
sav2325	-	Y	-	-	-
sav2326	-	-	-	Y	-
sav2363	-	-	-	Y	-
sav2382	-	-	-	Y	Aromatic polyketide
sav2406	-	-	-	Y	-
sav2410	-	-	-	Y	-
sav2417	Y	-	-	-	-
sav2424	-	-	-	Y	-
sav2435	-	-	-	Y	-
sav2446	-	-	-	Y	-
sav2451	-	-	-	Y	-
sav2456	-	-	-	Y	-
sav2464	-	-	-	Y	-
sav2490	-	-	-	Y	-
sav2539	-	-	-	Y	-
sav2551	Y	-	-	-	-
sav2565	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav2574	-	-	-	Y	-
sav2628	-	-	-	Y	-
sav2653	-	-	-	Y	-
sav2677	-	-	-	Y	-
sav2678	-	-	-	Y	-
sav2693	Y	-	-	-	-
sav2719	-	-	-	Y	-
sav2738	-	-	-	Y	-
sav2770	-	-	-	Y	-
sav2797	-	-	-	Y	-
sav2799	-	Y	-	-	-
sav2821	-	Y	-	-	-
sav2822	-	-	-	Y	-
sav2829	-	-	-	Y	-
sav2831	-	-	-	Y	-
sav2854	-	-	-	Y	-
sav2865	-	Y	-	-	-
sav2866	-	Y	-	-	-
sav2875	-	-	-	Y	-
sav2906	-	-	-	Y	-
sav2918	Y	-	-	-	-
sav2933	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav2957	-	-	-	Y	-
sav2970	-	-	-	Y	-
sav2971	-	-	-	Y	-
sav2985	Y	-	-	-	-
sav2986	-	Y	-	-	-
sav2987	-	-	-	Y	-
sav2988	-	Y	-	-	-
sav3014	-	-	-	Y	-
sav3050	-	Y	-	-	-
sav3051	-	-	-	Y	-
sav3074	-	-	-	Y	-
sav3082	-	-	-	Y	-
sav3097	-	-	-	Y	-
sav3119	Y	-	-	-	-
sav3123	-	-	-	Y	-
sav3142	-	Y	-	-	-
sav3143	-	-	-	Y	-
sav3147	-	-	-	Y	-
sav3148	-	-	-	Y	-
sav3149	-	-	-	Y	-
sav3164	Y	-	-	-	Non-ribosomal peptide
sav3165	Y	-	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav3167	-	-	-	Y	-
sav3168	-	Y	-	-	-
sav3198	Y	-	-	-	Non-ribosomal peptide
sav3208	-	-	-	Y	-
sav3214	Y	-	-	-	-
sav3215	Y	-	-	-	-
sav3265	-	-	-	Y	-
sav3291	-	-	-	Y	-
sav3297	-	Y	-	-	-
sav3299	Y	-	-	-	-
sav3306	-	Y	-	-	-
sav3310	-	Y	-	-	-
sav3315	-	-	-	Y	-
sav3328	-	-	-	Y	-
sav3335	-	-	-	Y	-
sav3336	-	-	-	Y	-
sav3338	-	-	-	Y	-
sav3355	-	-	-	Y	-
sav3364	-	-	-	Y	-
sav3407	-	-	-	Y	-
sav3429	-	-	-	Y	-
sav3447	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav3459	-	Y	-	-	-
sav3462	-	Y	-	-	-
sav3465	-	-	-	Y	-
sav3469	-	-	-	Y	-
sav3478	-	-	-	Y	-
sav3492	-	-	-	Y	-
sav3493	Y	-	-	-	-
sav3494	-	-	-	Y	-
sav3502	-	-	-	Y	-
sav3503	-	-	-	Y	-
sav3513	-	-	-	Y	-
sav3547	-	-	-	Y	-
sav3559	-	-	-	Y	-
sav3571	-	Y	-	-	-
sav3580	-	-	-	Y	-
sav3596	-	-	-	Y	-
sav3597	-	-	-	Y	-
sav3626	-	Y	-	-	-
sav3673	-	-	-	Y	-
sav3682	-	-	-	Y	-
sav3687	-	-	-	Y	-
sav3695	Y	-	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav3702	-	Y	-	-	-
sav3703	-	Y	-	-	-
sav3711	-	-	-	Y	-
sav3714	-	-	-	Y	-
sav3730	-	-	-	Y	-
sav3775	-	-	-	Y	-
sav3783	Y	-	-	-	-
sav3790	-	-	-	Y	-
sav3794	-	Y	-	-	-
sav3806	-	-	-	Y	-
sav3841	-	-	-	Y	-
sav3848	-	-	-	Y	-
sav3853	-	-	-	Y	-
sav3860	-	-	-	Y	-
sav3889	-	-	-	Y	-
sav3897	-	Y	-	-	-
sav3906	-	Y	-	-	-
sav3910	-	-	-	Y	-
sav3913	-	-	-	Y	-
sav3931	-	-	-	Y	-
sav3936	-	-	-	Y	-
sav3942	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav3965	Y	-	-	-	-
sav3970	-	-	-	Y	-
sav3982	-	-	-	Y	-
sav3983	-	-	-	Y	-
sav4024	-	-	-	Y	-
sav4048	-	-	-	Y	-
sav4054	-	-	-	Y	-
sav4059	-	-	-	Y	-
sav4077	Y	-	-	-	-
sav4078	-	Y	-	-	-
sav4081	-	-	-	Y	-
sav4111	-	-	-	Y	-
sav4133	-	Y	-	-	-
sav4134	-	-	-	Y	-
sav4177	-	-	-	Y	-
sav4181	-	-	-	Y	-
sav4185	-	-	-	Y	-
sav4187	-	-	-	Y	-
sav4193	-	Y	-	-	-
sav4235	Y	-	-	-	-
sav4266	-	Y	-	-	-
sav4275	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav4277	-	-	-	Y	-
sav4326	-	-	-	Y	-
sav4328	Y	-	-	-	-
sav4349	-	-	-	Y	-
sav4357	-	-	-	Y	-
sav4361	-	-	-	Y	-
sav4367	Y	-	-	-	-
sav4369	-	-	-	Y	-
sav4387	-	-	-	Y	-
sav4397	-	-	-	Y	-
sav4398	-	-	-	Y	-
sav4401	-	-	-	Y	-
sav4421	-	-	-	Y	-
sav4428	-	-	-	Y	-
sav4429	-	-	-	Y	-
sav4435	-	Y	-	-	-
sav4437	-	-	-	Y	-
sav4489	-	-	-	Y	-
sav4510	-	Y	-	-	-
sav4516	-	-	-	Y	-
sav4548	-	-	-	Y	-
sav4554	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav4563	-	Y	-	-	-
sav4585	-	Y	-	-	-
sav4590	-	-	-	Y	-
sav4595	-	-	-	Y	-
sav4601	-	-	-	Y	-
sav4613	-	-	-	Y	-
sav4626	-	-	-	Y	-
sav4625	-	-	-	Y	-
sav4662	Y	-	-	-	-
sav4667	-	-	-	Y	-
sav4668	-	-	-	Y	-
sav4679	-	-	-	Y	-
sav4696	-	-	-	Y	-
sav4712	-	Y	-	-	-
sav4733	-	-	-	Y	-
sav4737	-	-	-	Y	-
sav4738	Y	-	-	-	-
sav4739	-	-	-	Y	-
sav4740	Y	-	-	-	-
sav4741	Y	-	-	-	-
sav4743	-	-	-	Y	-
sav4745	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav4754	-	-	-	Y	-
sav4757	-	-	-	Y	-
sav4762	-	-	-	Y	-
sav4782	-	-	-	Y	-
sav4783	-	-	-	Y	-
sav4792	-	-	-	Y	-
sav4794	-	-	-	Y	-
sav4795	-	-	-	Y	-
sav4819	-	-	-	Y	-
sav4823	-	-	-	Y	-
sav4835	-	Y	-	-	-
sav4840	Y	-	-	-	-
sav4841	Y	-	-	-	-
sav4842	Y	-	-	-	-
sav4843	Y	-	-	-	-
sav4852	-	-	-	Y	-
sav4886	Y	-	-	-	-
sav4889	Y	-	-	-	-
sav4894	-	-	-	Y	-
sav4895	-	-	-	Y	-
sav4905	-	-	-	Y	-
sav4955	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav4960	-	-	-	Y	-
sav4971	-	-	-	Y	-
sav5004	-	-	-	Y	-
sav5024	-	-	-	Y	-
sav5025	Y	-	-	-	-
sav5053	-	-	-	Y	-
sav5054	-	-	-	Y	-
sav5059	-	-	-	Y	-
sav5061	-	-	-	Y	-
sav5068	-	-	-	Y	-
sav5069	-	-	-	-	-
sav5085	-	-	-	Y	-
sav5117	-	-	-	Y	-
sav5164	-	Y	-	-	-
sav5180	-	-	-	Y	-
sav5188	-	Y	-	-	-
sav5189	-	-	-	Y	-
sav5190	-	-	-	Y	-
sav5231	-	-	-	Y	-
sav5240	Y	-	-	-	-
sav5245	-	-	-	Y	-
sav5246	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav5268	-	Y	-	-	-
sav5293	-	-	-	Y	-
sav5304	-	-	-	Y	-
sav5305	-	-	-	Y	-
sav5363	Y	-	-	-	-
sav5369	-	Y	-	-	-
sav5371	-	-	-	Y	-
sav5379	-	-	-	Y	-
sav5407	Y	-	-	-	-
sav5412	-	-	-	Y	-
sav5439	-	-	-	Y	-
sav5444	Y	-	-	-	-
sav5475	-	-	-	Y	-
sav5483	-	-	-	Y	-
sav5485	-	-	-	Y	-
sav5489	-	-	-	Y	-
sav5494	-	Y	-	-	-
sav5525	-	-	-	Y	-
sav5548	-	Y	-	-	-
sav5587	-	-	-	Y	-
sav5592	-	-	-	Y	-
sav5596	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav5624	-	-	-	Y	-
sav5638	-	Y	-	-	-
sav5650	-	Y	-	-	-
sav5679	-	-	-	Y	-
sav5690	-	-	-	Y	-
sav5695	-	-	-	Y	-
sav5699	-	Y	-	-	-
sav5700	-	Y	-	-	-
sav5701	-	-	-	Y	-
sav5708	-	-	-	Y	-
sav5709	-	-	-	Y	-
sav5711	-	Y	-	-	-
sav5741	Y	-	-	-	-
sav5756	-	-	-	Y	-
sav5764	-	-	-	Y	-
sav5780	-	-	-	Y	-
sav5781	-	-	-	Y	-
sav5782	-	-	-	Y	-
sav5794	-	Y	-	-	-
sav5807	Y	-	-	-	-
sav5810	-	-	-	Y	-
sav5823	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav5836	-	Y	-	-	-
sav5837	-	Y	-	-	-
sav5862	-	-	-	Y	-
sav5872	-	-	-	Y	-
sav5882	-	-	-	Y	-
sav5913	-	-	-	Y	-
sav5929	-	Y	-	-	-
sav5933	-	Y	-	-	-
sav5936	-	-	-	Y	-
sav5937	-	-	-	Y	-
sav5985	-	-	-	Y	-
sav5988	-	-	-	Y	-
sav6012	-	-	-	Y	-
sav6015	-	Y	-	-	-
sav6034	-	-	-	Y	-
sav6048	-	-	-	Y	-
sav6050	-	-	-	Y	-
sav6057	-	-	-	Y	-
sav6059	-	-	-	Y	-
sav6061	Y	-	-	-	-
sav6088	-	-	-	Y	-
sav6090	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav6130	-	-	-	Y	-
sav6134	-	-	-	Y	-
sav6138	-	-	-	Y	-
sav6139	-	-	-	Y	-
sav6171	Y	-	-	-	-
sav6228	-	-	-	Y	-
sav6233	Y	-	-	-	-
sav6237	-	-	-	Y	-
sav6241	-	Y	-	-	-
sav6256	-	-	-	Y	-
sav6261	-	-	-	Y	-
sav6287	-	-	-	Y	-
sav6303	-	-	-	Y	-
sav6320	-	-	-	Y	-
sav6391	-	-	-	Y	-
sav6399	Y	-	-	-	-
sav6410	-	-	-	Y	-
sav6423	-	Y	-	-	-
sav6429	-	-	-	Y	-
sav6434	-	-	-	Y	-
sav6456	-	-	-	Y	-
sav6464	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav6483	-	Y	-	-	-
sav6498	-	-	-	Y	-
sav6533	-	-	-	Y	-
sav6542	-	Y	-	-	-
sav6543	-	Y	-	-	-
sav6556	-	-	-	Y	-
sav6561	-	-	-	Y	-
sav6567	-	-	-	Y	-
sav6585	-	-	-	Y	-
sav6597	-	-	-	Y	-
sav6602	-	-	-	Y	-
sav6606	-	-	-	Y	-
sav6608	-	-	-	Y	-
sav6636	-	-	-	Y	-
sav6637	Y	-	-	-	-
sav6648	-	-	-	Y	-
sav6674	Y	-	-	-	-
sav6675	-	Y	-	-	-
sav6696	-	-	-	Y	-
sav6707	-	-	-	Y	-
sav6711	-	-	-	Y	-
sav6712	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav6731	-	-	-	Y	-
sav6787	-	-	-	Y	-
sav6805	-	Y	-	-	-
sav6816	-	-	-	Y	-
sav6823	-	Y	-	-	-
sav6888	-	-	-	Y	-
sav6893	-	Y	-	-	-
sav6894	-	Y	-	-	-
sav6897	-	-	-	Y	-
sav6956	-	Y	-	-	-
sav6959	-	-	-	Y	-
sav6961	-	-	-	Y	-
sav6991	-	-	-	Y	-
sav6993	-	-	-	Y	-
sav7004	-	-	-	Y	-
sav7039	-	-	-	Y	-
sav7058	-	-	-	Y	-
sav7059	-	-	-	Y	-
sav7075	-	-	-	Y	-
sav7087	-	-	-	Y	-
sav7088	-	Y	-	-	-
sav7090	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav7094	-	Y	-	-	-
sav7096	-	-	-	Y	-
sav7098	-	Y	-	-	-
sav7118	-	-	-	Y	-
sav7129	-	-	-	Y	-
sav7165	-	-	-	Y	Non-ribosomal peptide
sav7188	-	-	-	Y	-
sav7218	-	-	-	Y	-
sav7247	-	-	-	Y	-
sav7248	-	Y	-	-	-
sav7249	-	-	-	Y	-
sav7250	-	Y	-	-	-
sav7251	-	Y	-	-	-
sav7289	-	-	-	Y	-
sav7299	-	-	-	Y	-
sav7311	-	-	-	Y	-
sav7313	-	-	-	Y	-
sav7327	-	-	-	Y	-
sav7337	-	-	-	Y	-
sav7343	-	-	-	Y	-
sav7346	-	Y	-	-	-
sav7388	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sav7389	-	Y	-	-	-
sav7419	-	-	-	Y	-
sav7429	-	-	-	Y	-
sav7442	Y	-	-	-	-
sav7448	-	-	-	Y	-
sav7449	Y	-	-	-	-
sav7477	-	-	-	Y	-
sav7481	-	-	-	Y	-
sav7486	-	-	-	Y	-
sav7487	-	Y	-	-	-
sav7498	-	-	-	Y	-
sav7513	-	-	-	Y	-
sav7519	-	-	-	Y	-
sav7536	-	-	-	Y	-
sav7544	-	-	-	Y	-
sven0002	-	-	-	Y	-
sven0003	-	-	-	Y	-
sven0013	-	-	-	Y	-
sven0053	-	-	-	Y	-
sven0066	-	-	-	Y	-
sven0181	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven0203	-	-	-	Y	-
sven0204	-	-	-	Y	-
sven0216	-	-	-	Y	-
sven0227	-	-	-	Y	Ectoine
sven0253	-	-	-	Y	-
sven0265	-	Y	-	-	Terpene
sven0287	-	-	-	Y	-
sven0301	-	-	-	Y	-
sven0335	-	-	-	Y	-
sven0338	-	Y	-	-	-
sven0347	Y	-	-	-	-
sven0363	-	-	-	Y	-
sven0365	-	Y	-	-	-
sven0367	-	-	-	Y	-
sven0369	-	-	-	Y	-
sven0375	-	-	-	Y	-
sven0422	-	-	-	Y	-
sven0432	-	-	-	Y	-
sven0434	-	-	-	Y	-
sven0452	Y	-	-	-	-
sven0480	-	-	-	Y	Polyketide / non-ribosomal peptide hybrid

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven0501	-	-	-	Y	Polyketide / non-ribosomal peptide hybrid
sven0516	-	-	-	Y	Polyketide / non-ribosomal peptide hybrid
sven0519	-	-	-	Y	Polyketide / non-ribosomal peptide hybrid
sven0552	-	Y	-	-	Terpene
sven0579	-	-	-	Y	-
sven0591	-	Y	-	-	-
sven0604	-	-	-	Y	-
sven0605	Y	-	-	-	-
sven0611	-	-	-	Y	-
sven0645	-	-	-	Y	Lantipeptide
sven0658	-	-	-	Y	Lantipeptide
sven0659	Y	-	-	-	Lantipeptide
sven0666	-	-	-	Y	-
sven0689	-	-	-	Y	-
sven0693	-	-	-	Y	-
sven0712	-	-	-	Y	-
sven0719	-	-	-	Y	-
sven0724	-	-	-	Y	-
sven0735	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven0738	-	-	-	Y	-
sven0741	-	-	-	Y	-
sven0747	-	-	-	Y	-
sven0767	-	Y	-	-	Indole
sven0787	-	-	-	Y	-
sven0826	-	-	-	Y	-
sven0827	-	-	-	Y	-
sven0842	-	-	-	Y	-
sven0843	-	-	-	Y	-
sven0856	-	-	-	Y	-
sven0867	-	-	-	Y	-
sven0868	-	-	-	Y	-
sven0879	-	-	-	Y	-
sven0891	-	-	-	Y	-
sven0898	-	-	-	Y	-
sven0904	Y	-	-	-	-
sven0912	-	-	-	Y	-
sven0913	-	-	-	Y	-
sven0916	-	-	-	Y	-
sven0936	-	-	-	Y	-
sven0954	-	-	-	Y	-
sven0964	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven1046	-	-	-	Y	-
sven1050	-	-	-	Y	-
sven1064	-	-	-	Y	-
sven1065	-	-	-	Y	-
sven1075	-	-	-	Y	-
sven1079	Y	-	-	-	-
sven1094	-	-	-	Y	-
sven1098	Y	-	-	-	-
sven1105	-	-	-	Y	-
sven1116	Y	-	-	-	-
sven1128	-	-	-	Y	-
sven1130	Y	Y	-	-	-
sven1131	-	Y	-	-	-
sven1132	-	-	-	Y	-
sven1158	Y	-	-	-	-
sven1160	-	Y	-	-	-
sven1167	-	-	-	Y	-
sven1178	-	-	-	Y	-
sven1184	-	-	-	Y	-
sven1187	-	-	-	Y	-
sven1197	-	Y	-	-	-
sven1202	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven1230	-	-	-	Y	-
sven1237	Y	-	-	-	-
sven1239	Y	-	-	-	-
sven1242	-	-	-	Y	-
sven1252	Y	-	-	-	-
sven1262	-	-	-	Y	-
sven1283	-	-	-	Y	-
sven1316	-	-	-	Y	-
sven1372	-	Y	-	-	-
sven1373	-	Y	-	-	-
sven1384	-	-	-	Y	-
sven1396	-	-	-	Y	-
sven1413	-	-	-	Y	-
sven1421	-	-	-	Y	-
sven1432	-	-	-	Y	-
sven1440	-	-	-	Y	-
sven1447	-	Y	-	-	-
sven1448	-	Y	-	-	-
sven1459	-	Y	-	-	-
sven1464	-	-	-	Y	-
sven1472	-	-	-	Y	-
sven1477	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven1478	-	Y	-	-	-
sven1489	-	Y	-	-	-
sven1508	-	-	-	Y	-
sven1517	-	-	-	Y	-
sven1528	-	-	-	Y	-
sven1531	-	-	-	Y	-
sven1536	-	-	-	Y	-
sven1561	-	-	-	Y	-
sven1569	-	-	-	Y	-
sven1593	-	-	-	Y	-
sven1619	-	-	-	Y	-
sven1626	-	-	-	Y	-
sven1638	-	-	-	Y	-
sven1661	-	-	-	Y	-
sven1663	-	-	-	Y	-
sven1669	-	-	-	Y	-
sven1675	-	-	-	Y	-
sven1677	Y	-	-	-	-
sven1690	-	-	-	Y	-
sven1694	Y	-	-	-	-
sven1699	-	-	-	Y	-
sven1735	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven1760	-	-	-	Y	-
sven1774	-	Y	-	-	-
sven1778	-	-	-	Y	-
sven1785	-	-	-	Y	-
sven1792	-	-	-	Y	-
sven1801	-	-	-	Y	-
sven1803	-	-	-	Y	-
sven1805	Y	-	-	-	-
sven1809	Y	-	-	-	-
sven1810	-	-	-	Y	-
sven1812	Y	-	-	-	-
sven1821	Y	-	-	-	-
sven1828	-	-	-	Y	-
sven1829	Y	-	-	-	-
sven1834	-	-	-	Y	-
sven1843	-	-	-	Y	-
sven1857	-	-	-	Y	-
sven1865	-	-	-	Y	-
sven1866	Y	-	-	-	-
sven1868	-	-	-	Y	-
sven1870	-	-	-	Y	-
sven1887	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven1892	Y	-	-	-	-
sven1897	-	-	-	Y	-
sven1933	-	-	-	Y	-
sven1949	-	-	-	Y	-
sven1953	-	-	-	Y	-
sven1961	-	-	-	Y	-
sven1962	-	Y	-	-	-
sven1964	-	-	-	Y	-
sven1981	-	-	-	Y	-
sven1989	-	-	-	Y	-
sven1997	-	-	-	Y	-
sven2033	-	-	-	Y	-
sven2059	-	-	-	Y	-
sven2065	-	-	-	Y	-
sven2067	-	-	-	Y	-
sven2072	-	-	-	Y	-
sven2095	-	-	-	Y	-
sven2151	-	-	-	Y	-
sven2152	-	-	-	Y	-
sven2172	-	-	-	Y	-
sven2175	-	-	-	Y	-
sven2178	Y	-	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven2181	-	-	-	Y	-
sven2190	-	-	-	Y	-
sven2195	-	-	-	Y	-
sven2209	Y	-	-	-	-
sven2210	-	-	-	Y	-
sven2229	-	-	-	Y	-
sven2235	-	-	-	Y	-
sven2262	-	-	-	Y	-
sven2270	-	-	-	Y	-
sven2281	-	-	-	Y	-
sven2284	-	-	-	Y	-
sven2292	-	-	-	Y	-
sven2293	-	-	-	Y	-
sven2297	-	-	-	Y	-
sven2298	-	-	-	Y	-
sven2315	-	-	-	Y	-
sven2325	-	-	-	Y	-
sven2338	-	-	-	Y	-
sven2341	Y	-	-	-	-
sven2365	-	-	-	Y	-
sven2371	Y	-	-	-	-
sven2380	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven2386	Y	-	-	-	-
sven2394	-	-	-	Y	-
sven2405	-	-	-	Y	-
sven2409	-	-	-	Y	-
sven2416	-	-	-	Y	-
sven2419	-	-	-	Y	-
sven2438	-	-	-	Y	-
sven2442	-	-	-	Y	-
sven2450	-	-	-	Y	-
sven2451	-	-	-	Y	-
sven2516	-	-	-	Y	-
sven2518	-	-	-	Y	-
sven2521	-	-	-	Y	-
sven2550	-	-	-	Y	-
sven2552	Y	-	-	-	-
sven2557	-	-	-	Y	-
sven2582	-	-	-	Y	-
sven2586	-	-	-	Y	-
sven2612	-	-	-	Y	-
sven2639	-	Y	-	-	-
sven2642	-	-	-	Y	-
sven2667	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven2675	-	-	-	Y	-
sven2678	Y	Y	-	-	-
sven2683	-	-	-	Y	-
sven2688	-	-	-	Y	-
sven2691	-	-	-	Y	-
sven2696	-	-	-	Y	-
sven2719	-	-	-	Y	-
sven2723	-	-	-	Y	-
sven2750	-	-	-	Y	-
sven2753	-	-	-	Y	-
sven2754	Y	-	-	-	-
sven2760	-	-	-	Y	-
sven2761	-	-	-	Y	-
sven2785	Y	-	-	-	-
sven2793	Y	-	-	-	-
sven2799	-	-	-	Y	-
sven2814	-	-	-	Y	-
sven2829	-	-	-	Y	-
sven2833	Y	-	-	-	-
sven2835	-	-	-	Y	-
sven2837	-	Y	-	-	-
sven2840	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven2850	-	-	-	Y	-
sven2878	-	-	-	Y	-
sven2879	-	-	-	Y	-
sven2886	-	-	-	Y	-
sven2889	-	-	-	Y	-
sven2890	-	-	-	Y	-
sven2895	-	-	-	Y	-
sven2967	-	-	-	Y	-
sven2992	-	-	-	Y	-
sven3000	-	-	-	Y	-
sven3008	-	-	-	Y	-
sven3009	-	-	-	Y	-
sven3029	-	-	-	Y	-
sven3035	-	Y	-	-	-
sven3036	-	Y	-	-	-
sven3037	-	-	-	Y	-
sven3038	-	-	-	Y	-
sven3043	-	-	-	Y	-
sven3044	-	-	-	Y	-
sven3050	-	-	-	Y	-
sven3056	-	-	-	Y	-
sven3064	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven3077	Y	-	-	-	-
sven3080	-	-	-	Y	-
sven3102	-	Y	-	-	-
sven3120	-	-	-	Y	-
sven3126	-	Y	-	-	-
sven3127	-	Y	-	-	-
sven3146	-	-	-	Y	-
sven3148	-	-	-	Y	-
sven3150	-	-	-	Y	-
sven3152	-	-	-	Y	-
sven3160	-	-	-	Y	-
sven3162	-	-	-	Y	-
sven3165	-	-	-	Y	-
sven3175	-	-	-	Y	-
sven3176	-	-	-	Y	-
sven3179	Y	-	-	-	-
sven3180	Y	-	-	Y	-
sven3209	-	-	-	Y	-
sven3224	-	Y	-	-	-
sven3237	-	-	-	Y	-
sven3247	-	-	-	Y	-
sven3254	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven3260	-	-	-	Y	-
sven3273	-	-	-	Y	-
sven3286	-	-	-	Y	-
sven3293	-	-	-	Y	-
sven3300	-	-	-	Y	-
sven3320	-	-	-	Y	-
sven3329	-	-	-	Y	-
sven3338	-	-	-	Y	-
sven3340	-	-	-	Y	-
sven3348	-	Y	-	-	-
sven3349	-	Y	-	-	-
sven3360	-	-	-	Y	-
sven3368	-	-	-	Y	-
sven3370	-	-	-	Y	-
sven3371	Y	-	-	-	-
sven3388	-	-	-	Y	-
sven3423	-	-	-	Y	-
sven3433	Y	-	-	-	-
sven3469	-	-	-	Y	-
sven3473	-	-	-	Y	-
sven3485	Y	-	-	-	-
sven3490	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven3505	-	-	-	Y	-
sven3515	-	-	-	Y	-
sven3525	-	Y	-	-	-
sven3533	-	Y	-	-	-
sven3535	-	-	-	Y	-
sven3551	-	-	-	Y	-
sven3566	Y	-	-	-	-
sven3571	-	-	-	Y	-
sven3577	-	-	-	Y	-
sven3579	-	-	-	Y	-
sven3593	-	-	-	Y	-
sven3606	-	Y	-	-	-
sven3610	-	Y	-	-	-
sven3611	-	-	-	Y	-
sven3613	-	-	-	Y	-
sven3627	Y	-	-	-	-
sven3643	-	Y	-	-	-
sven3644	-	Y	-	-	-
sven3657	-	-	-	Y	-
sven3659	Y	-	-	-	-
sven3696	-	-	-	Y	-
sven3706	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven3711	-	-	-	Y	-
sven3712	-	-	-	Y	-
sven3715	-	-	-	Y	-
sven3717	-	-	-	Y	-
sven3718	-	-	-	Y	-
sven3719	-	-	-	Y	-
sven3721	-	-	-	Y	-
sven3729	-	-	-	Y	-
sven3763	-	-	-	Y	-
sven3768	-	-	-	Y	-
sven3796	-	-	-	Y	-
sven3814	-	-	-	Y	-
sven3837	-	-	-	Y	-
sven3846	-	-	-	Y	-
sven3847	-	-	-	Y	-
sven3854	-	-	-	Y	-
sven3859	-	-	-	Y	-
sven3874	-	-	-	Y	-
sven3891	-	-	-	Y	-
sven3893	-	-	-	Y	-
sven3894	-	-	-	Y	-
sven3895	Y	-	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven3925	Y	-	-	-	-
sven3961	-	-	-	Y	-
sven3963	-	-	-	Y	-
sven3966	-	-	-	Y	-
sven3982	-	-	-	Y	-
sven3983	-	-	-	Y	-
sven4021	-	-	-	Y	-
sven4025	-	-	-	Y	-
sven4033	-	-	-	Y	-
sven4036	-	-	-	Y	-
sven4042	-	-	-	Y	-
sven4043	Y	-	-	-	-
sven4066	-	-	-	Y	-
sven4089	-	-	-	Y	-
sven4090	-	-	-	Y	-
sven4095	-	-	-	Y	-
sven4103	-	-	-	Y	-
sven4130	Y	-	-	-	-
sven4135	-	-	-	Y	-
sven4138	-	-	-	Y	-
sven4149	-	-	-	Y	-
sven4199	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven4204	-	-	-	Y	-
sven4205	-	-	-	Y	-
sven4211	-	-	-	Y	-
sven4227	Y	-	-	-	-
sven4231	-	-	-	Y	-
sven4262	-	-	-	Y	-
sven4267	Y	-	-	-	-
sven4270	Y	-	-	-	-
sven4281	-	-	-	Y	-
sven4283	-	-	-	Y	-
sven4285	-	-	-	Y	-
sven4288	-	-	-	Y	-
sven4296	-	-	-	Y	-
sven4319	Y	-	-	-	-
sven4333	-	-	-	Y	-
sven4337	Y	-	-	-	-
sven4346	Y	-	-	-	-
sven4361	-	-	-	Y	-
sven4364	Y	-	-	-	-
sven4374	-	-	-	Y	-
sven4407	Y	-	-	-	-
sven4426	-	Y	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven4444	-	-	-	Y	-
sven4451	-	-	-	Y	-
sven4456	-	-	-	Y	-
sven4458	Y	-	-	-	-
sven4468	-	Y	-	-	-
sven4469	-	-	-	Y	-
sven4472	-	-	-	Y	-
sven4498	Y	-	-	-	-
sven4518	-	-	-	Y	-
sven4521	Y	-	-	-	-
sven4544	Y	-	-	-	-
sven4590	-	-	-	Y	-
sven4592	-	-	-	Y	-
sven4600	-	-	-	Y	-
sven4609	-	-	-	Y	-
sven4613	-	Y	-	-	-
sven4614	-	Y	-	-	-
sven4621	-	-	-	Y	-
sven4629	-	-	-	Y	-
sven4631	-	-	-	Y	-
sven4671	-	-	-	Y	-
sven4724	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven4747	-	-	-	Y	-
sven4751	-	-	-	Y	-
sven4789	-	Y	-	-	-
sven4792	-	-	-	Y	-
sven4795	-	-	-	Y	-
sven4810	-	-	-	Y	-
sven4813	-	Y	-	-	-
sven4814	-	Y	-	-	-
sven4833	-	-	-	Y	-
sven4844	-	Y	-	-	-
sven4847	Y	-	-	-	-
sven4863	-	-	-	Y	-
sven4869	-	-	-	Y	-
sven4898	-	-	-	Y	-
sven4907	-	-	-	Y	-
sven4943	-	-	-	Y	-
sven4944	-	-	-	Y	-
sven4946	-	Y	-	-	-
sven4954	-	-	-	Y	-
sven4957	-	-	-	Y	-
sven4980	-	-	-	Y	-
sven4983	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven5031	-	-	-	Y	-
sven5041	-	-	-	Y	-
sven5062	-	-	-	Y	-
sven5072	Y	-	-	-	-
sven5093	-	-	-	Y	Butyrolactone
sven5118	-	-	-	Y	-
sven5157	-	-	-	Y	-
sven5160	-	-	-	Y	-
sven5207	-	-	-	Y	-
sven5253	-	-	-	Y	-
sven5256	-	-	-	Y	-
sven5258	-	-	-	Y	-
sven5271	-	-	-	Y	-
sven5283	-	-	-	Y	-
sven5302	-	-	-	Y	-
sven5368	-	-	-	Y	Polyketide
sven5390	-	Y	-	-	-
sven5393	-	-	-	Y	-
sven5395	-	-	-	Y	-
sven5405	-	-	-	Y	-
sven5414	-	-	-	Y	Siderophore
sven5420	-	-	-	Y	Siderophore

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven5426	-	-	-	Y	-
sven5451	-	-	-	Y	-
sven5462	-	-	-	Y	-
sven5480	-	-	-	Y	Siderophore
sven5500	-	-	-	Y	-
sven5522	-	-	-	Y	-
sven5528	-	-	-	Y	-
sven5561	-	-	-	Y	-
sven5571	-	-	-	Y	-
sven5596	-	-	-	Y	-
sven5656	-	-	-	Y	-
sven5682	-	-	-	Y	-
sven5686	-	-	-	Y	-
sven5700	-	-	-	Y	-
sven5709	-	-	-	Y	-
sven5712	-	-	-	Y	-
sven5715	-	-	-	Y	-
sven5721	-	-	-	Y	-
sven5767	-	-	-	Y	-
sven5780	-	-	-	Y	-
sven5798	-	-	-	Y	-
sven5802	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven5809	-	-	-	Y	-
sven5860	-	-	-	Y	-
sven5862	-	-	-	Y	-
sven5871	-	-	-	Y	-
sven5875	-	-	-	Y	-
sven5897	-	-	-	Y	-
sven5904	-	-	-	Y	-
sven5931	-	-	-	Y	-
sven5936	-	-	-	Y	-
sven5955	Y	-	-	-	-
sven5960	-	-	-	Y	-
sven5969	Y	-	-	-	Butyrolactone / polyketide
sven6004	-	-	-	Y	-
sven6027	-	-	-	Y	-
sven6048	-	-	-	Y	-
sven6055	-	-	-	Y	-
sven6104	-	-	-	Y	-
sven6109	-	-	-	Y	-
sven6119	-	-	-	Y	-
sven6142	-	-	-	Y	-
sven6156	-	-	-	Y	-
sven6180	-	-	-	Y	Non-ribosomal

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
					peptide
sven6209	-	-	-	Y	Non-ribosomal peptide
sven6213	-	-	-	Y	Non-ribosomal peptide
sven6264	-	-	-	Y	Non-ribosomal peptide
sven6310	-	-	-	Y	-
sven6361	-	-	-	Y	-
sven6381	-	-	-	Y	-
sven6387	-	-	-	Y	-
sven6396	-	-	-	Y	-
sven6397	-	-	-	Y	-
sven6404	-	-	-	Y	-
sven6430	-	-	-	Y	-
sven6434	-	-	-	Y	-
sven6435	-	Y	-	-	-
sven6437	-	-	-	Y	Terpene
sven6443	-	-	-	Y	Terpene
sven6455	Y	-	-	-	Terpene
sven6456	Y	-	-	-	Terpene
sven6457	-	-	-	Y	Terpene
sven6460	-	-	-	Y	Terpene

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven6464	-	-	-	Y	-
sven6519	-	-	-	Y	-
sven6523	-	-	-	Y	-
sven6526	-	-	-	Y	-
sven6531	-	-	-	Y	Bacteriocin
sven6564	-	-	-	Y	-
sven6580	-	-	-	Y	-
sven6621	-	-	-	Y	-
sven6631	-	-	-	Y	-
sven6657	-	Y	-	-	-
sven6668	-	-	-	Y	-
sven6673	-	-	-	Y	-
sven6751	-	-	-	Y	-
sven6784	-	-	-	Y	-
sven6801	-	-	-	Y	-
sven6803	-	-	-	Y	-
sven6826	Y	-	-	-	-
sven6833	-	-	-	Y	-
sven6840	-	-	-	Y	Melanin
sven6863	-	-	-	Y	-
sven6866	-	-	-	Y	-
sven6872	-	-	-	Y	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven6873	-	-	-	Y	-
sven6885	-	-	-	Y	-
sven6900	-	-	-	Y	-
sven6901	Y	-	-	-	-
sven6935	-	-	-	Y	-
sven6936	-	-	-	Y	-
sven6949	-	-	-	Y	-
sven6957	Y	-	-	-	-
sven7023	-	-	-	Y	-
sven7111	-	-	-	Y	Terpene
sven7226	-	-	-	Y	Polyketide
sven7271	-	-	-	Y	-
sven7272	-	-	-	Y	-
sven7281	-	-	-	Y	-
sven7282	Y	-	-	-	-
sven7289	-	-	-	Y	-
sven7294	-	-	-	Y	-
sven7301	-	-	-	Y	-
sven7316	Y	-	-	-	-
sven7317	-	-	-	Y	-
sven7324	-	-	-	Y	-
sven7335	Y	-	-	-	-

Sense gene[abcd]	cis-antisense RNA	cutoRNA	Divergent transcription	Unclassified*	Biosynthetic clusters[x]
sven7351	-	-	-	Y	-
sven7396	-	-	-	Y	-
sven7418	-	-	-	Y	-
sven7451	-	-	-	Y	Terpene / non-ribosomal peptide
sven7453	-	-	-	Y	-

*Unclassified due to ambiguous transcription start sites and/or uneven coverage

[x] Locations of clusters as per (Nett et al., 2009) for *S. avermitilis* and *S. coelicolor,* and as determined using the program antiSMASH (Medema *et al.*,2011), for *S. venezuelae*.

[a] Previously identified (Vockenhuber et al., 2011)

[b] Previously identified (Swiercz et al., 2008).

[c] Previously identified (D'Alia et al., 2010)

[d] Upregulated (or upregulated cutoRNA partner) in an RNase III mutant (Gatewood et al., 2012)

Table 3.3: Conserved cutoRNA gene pairs.

Conserved cutoRNA gene pairs					Conservation of convergent gene arrangement*				
S. coelicolor	***S. avermitilis***	***S. venezuelae***	**Gene 1 annotation**	**Gene 2 annotation**	***S. griseus***	***S. scabies***	***T. fusca***	***M. tb***	***F. alni***
sco1294 / sco1293	*sav7062 / sav7063*	*sven0868 / sven0867*	putative cystathionine gamma-synthase	hypothetical protein	Y	Y	NH	NH	NH
sco1841 / sco1842	*sav6424 / sav6423*	*sven1488 / sven1489*	hypothetical protein	hypothetical protein	Y	Y	NH	NH	NH
sco2114 / sco2113	*sav6087 / sav6088*	*sven1774 / sven1773*	hypothetical protein	putative bacterioferritin	Y	Y	NH	NH	N
sco2267 / sco2268	*sav5930 / sav5929*	*sven1952 / sven1953*	probable heme oxygenase	conserved hypothetical protein	Y	Y	NH	NH	NH
sco2582 / sco2583	*sav5476 / sav5475*	*sven2364 / sven2365*	conserved hypothetical protein SCC123.20.	putative membrane protein.	Y	Y	NH	NH	NH
sco3134 / sco3133	*sav3572 / sav3571*	*sven2957 / sven2956*	putative two-component system response regulator	putative MarR-family regulator	Y	Y	NH	NH	NH
sco3375 / sco3374	*sav4695 / sav4696*	*sven3225 / sven3224*	putative *lsr2*-like protein	hypothetical proline-rich protein	Y	Y	NH	NH	NH
sco3579 / sco3578	*sav4584 / sav4585*	*sven3349 / sven3348*	*wblA*	putative ion-transporting ATPase	Y	Y	Y	Y	Y
sco3008 / sco3007	*sav5068 / sav5069*	*sven2751 / sven2750*	putative two-component system response regulator	conserved hypothetical protein	Y	Y	NH	NH	NH
sco3190 / sco3191	*sav3681 / sav3682*	*sven3035 / sven3036*	large-conductance mechanosensitive channel	hypothetical protein	N	Y	NH	NH	NH
sco3833 / sco3832	*sav4360 / sav4361*	*sven3607 / sven3606*	transcriptional regulator, TetR family	putative transcriptional regulator	Y	Y	NH	NH	NH
sco4282 / sco4283	*sav3943 / sav3942*	*sven4035 / sven4036*	putative dimeric protein	putative tagatose-6-phosphate kinase	Y	Y	NH	NH	NH
sco4335 / sco4334	*sav3896 / sav3897*	*sven4150 / sven4149*	hypothetical protein	putative hypoxanthine / guanine permease	Y	Y	NH	NH	NH
sco5167 / sco5168	*sav3097 / sav3096*	*sven4813 / sven4814*	hypothetical protein	hypothetical protein	Y	Y	NH	NH	NH
sco5255 / sco5256	*sav2987 / sav2986*	*sven4945 / sven4946*	signal peptidase protein	hypothetical protein	Y	Y	NH	NH	NH
sco5390 / sco5389	*sav2865 / sav2866*	*sven5041 / sven5040*	putative luciferase-family protein	hypothetical protein	Y	Y	NH	NH	N
sco6750 / sco6749	*sav1663 / sav1664*	*sven6435 / sven6434*	isopentenyl-diphosphate delta-isomerase	hypothetical protein	Y	Y	NH	NH	NH
sco6771 / sco6772	*sav1643 / sav1642*	*sven6456 / sven6457*	putative small hydrophobic secreted protein	putative protein tyrosine phosphatase	Y	Y	NH	NH	NH
sco7252 / sco7253	*sav1236 / sav1235*	*sven0266 / sven0265*	putative regulatory protein	hypothetical protein	Y	Y	NH	NH	NH

***** Yes (Y); No (N); At least one of the two genes do not have homologues in the given species (NH)

Table 3.4: Intergenic sRNAs in *S. coelicolor*, *S. avermitilis*, and *S. venezuelae*.

Species	sRNA name	Left Gene	Right Gene	Strand	Start	End	Size	Located within biosynthetic cluster[y]	Previously identified or predicted
S. coelicolor									
	scr0636	*sco0635*	*sco0636*	+	677033	677230	198		
	scr0792	*sco0791*	*sco0792*	+	838750	838863	114		
	scr0959	*sco0958*	*sco0959*	+	1008731	1008802	72		
	scr0999	*sco0998*	*sco0999*	-	1054736	1054647	90		
	scr1104	*sco1103*	*sco1104*	-	1161700	1161580	121		Vockenhuber et al (2011)
	scr1363	*sco1362*	*sco1363*	-	1441015	1440924	92		
	scr1390*	*sco1389*	*sco1390*	-	1474104	1473910	195		
	scr1424	*sco1423*	*sco1424*	-	1519780	1519565	216		
	scr1434	*sco1433*	*sco1434*	+	1529418	1529480	63		
	scr1594	*sco1593*	*sco1594*	+	1703984	1704169	186		
	scr1792*	*sco1791*	*sco1792*	-	1922085	1921800	286		Panek et al (2008)
	scr1805	*sco1804*	*sco1805*	-	1933162	1932972	191		
	scr1961	*sco1960*	*sco1961*	+	2099076	2099145	70		
	scr2101	*sco2100*	*sco2101*	-	2258212	2257982	231		Swiercz et al (2008)
	scr2107	*sco2106*	*sco2107*	+	2264717	2264823	107		
	scr2117	*sco2116*	*sco2117*	+	2276540	2276611	72		
	scr2136	*sco2135*	*sco2136*	-	2297658	2297564	95		
	scr2147	*sco2146*	*sco2147*	-	2308772	2308678	95		
	scr2257	*sco2256*	*sco2257*	+	2424206	2424263	58		
	scr2288	*sco2287*	*sco2288*	+	2458925	2458996	72		
	scr2446	*sco2445*	*sco2446*	+	2625429	2625650	222		
	scr2589	*sco2588*	*sco2589*	+	2725406	2725465	60		
	scr2528	*sco2527*	*sco2528*	+	2725412	2725465	54		
	scr2634	*sco2633*	*sco2634*	+	2860742	2860810	69		
	scr2736	*sco2735*	*sco2736*	+	2982223	2982670	448		Swiercz et al (2008); Vockenhuber et al (2011)

	scr2750	*sco2749*	*sco2750*	-	2996439	2996338	102		
	scr2822	*sco2821*	*sco2822*	+	3082275	3082345	71		Panek et al (2008)
	scr2922	*sco2921*	*sco2922*	-	3174924	3174429	496		
	scr2943	*sco2942*	*sco2943*	+	3198596	3198780	185		
	scr2952	*sco2951*	*sco2952*	+	3208719	3208802	84		Vockenhuber et al (2011)
	scr3020*	*sco3019*	*sco3020*	-	3302281	3302024	258		
	scr3077	*sco3076*	*sco3077*	-	3370220	3369991	230		
	scr3097	*sco3096*	*sco3097*	+	3391994	3392073	80		
	scr3202	*sco3201*	*sco3202*	-	3510186	3510094	93		Vockenhuber et al (2011)
	scr3253	*sco3252*	*sco3253*	+	3605514	3605596	83		
	scr3334*	*sco3333*	*sco3334*	-	3691139	3690886	254		Panek et al (2008)
	scr3410	*sco3409*	*sco3410*	-	3775694	3775649	46		
	scr3437	*sco3436*	*sco3437*	-	3799361	3799139	223		
	scr3558	*sco3557*	*sco3558*	-	3933668	3933527	142		Panek et al (2008); Swiercz et al (2008); Vockenhuber et al (2011)
	scr3559	*sco3558*	*sco3559*	+	3934693	3934919	227		Panek et al (2008); Vockenhuber et al (2011)
	scr3580	*sco3579*	*sco3580*	-	3958172	3957843	330		Vockenhuber et al (2011)
	scr3603	*sco3602*	*sco3603*	-	3979771	3979385	387		
	scr3679	*sco3678*	*sco3679*	-	4061931	4061681	251		
	scr3716	*sco3715*	*sco3716*	-	4091239	4091112	128		
	scr3901	*sco3900*	*sco3901*	+	4295743	4296186	444		
	scr3920	*sco3919*	*sco3920*	+	4315358	4315489	132		Vockenhuber et al (2011)
	scr3929	*sco3928*	*sco3929*	+	4323850	4324377	528		
	scr3931	*sco3930*	*sco3931*	-	4326774	4326705	70		
	scr4076*	*sco4075*	*sco4076*	-	4470108	4470036	73		Panek et al (2008)
	scr4108	*sco4107*	*sco4108*	+	4507134	4507300	167		
	scr4123*	*sco4122*	*sco4123*	+	4530078	4530310	233		

	scr4135	*sco4134*	*sco4135*	-	4550061	4549825	237		
	scr4139	*sco4138*	*sco4139*	-	4553616	4553538	79		
	scr4220	*sco4219*	*sco4220*	-	4627309	4627161	149		
	scr4233	*sco4232*	*sco4233*	-	4637485	4637326	160		
	scr4339	*sco4338*	*sco4339*	+	4753312	4753401	90		
	scr4389	*sco4388*	*sco4389*	-	4805789	4805580	210		Swiercz et al (2008); Vockenhuber et al (2011)
	scr4395	*sco4394*	*sco4395*	-	4811827	4811713	115		
	scr4412	*sco4411*	*sco4412*	-	4829635	4829349	287		
	scr4524	*sco4523*	*sco4524*	-	4944655	4944415	241		
	scr4619	*sco4618*	*sco4619*	+	5043143	5043220	78		
	scr4628	*sco4627*	*sco4628*	+	5049482	5049538	57		
	scr4632	*sco4631*	*sco4632*	+	5054815	5055060	246		Vockenhuber et al (2011)
	scr4677	*sco4676*	*sco4677*	+	5108129	5108199	71		Swiercz et al (2008)
	scr4684	*sco4683*	*sco4684*	+	5113596	5113760	165		
	scr4689	*sco4688*	*sco4689*	-	5117055	5116751	305		
	scr4909	*sco4908*	*sco4909*	+	5341263	5341370	108		
	scr5028	*sco5027*	*sco5028*	+	5462580	5462628	49		Swiercz et al (2008)
	scr5112	*sco5111*	*sco5112*	-	5556709	5556524	186		
	scr5239	*sco5238*	*sco5239*	+	5699820	5700035	216		
	scr5346	*sco5345*	*sco5346*	-	5814859	5814784	76		
	scr5439	*sco5438*	*sco5439*	-	5912258	5912163	96		Panek et al (2008)
	scr5583	*sco5582*	*sco5583*	+	6085733	6085815	83		
	scr5633	*sco5632*	*sco5633*	+	6133070	6133458	389		
	scr5676	*sco5675*	*sco5676*	-	6176413	6176160	254		Swiercz et al (2008); Vockenhuber et al (2011)
	scr5710	*sco5709*	*sco5710*	-	6221571	6221499	73		
	scr5839	*sco5838*	*sco5839*	+	6392117	6392286	170		
	scr5917	*sco5916*	*sco5917*	+	6484345	6484908	564		

	scr5922	*sco5921*	*sco5922*	-	6489972	6489791	182		
	scr5999	*sco5998*	*sco5999*	-	6575212	6574790	423		
	scr6106	*sco6105*	*sco6106*	+	6706591	6706667	77		Vockenhuber et al (2011)
	scr6154	*sco6153*	*sco6154*	+	6755084	6755550	467		
	scr6204	*sco6203*	*sco6204*	-	6818847	6818680	168		
	scr6214	*sco6213*	*sco6214*	+	6828082	6828120	39		
	scr6287	*sco6286*	*sco6287*	-	6946592	6946547	46	Yellow cryptic polyketide	
	scr6925	*sco6924*	*sco6925*	+	7688336	7688467	132		Vockenhuber et al (2011)
	scr7098	*sco7097*	*sco7098*	+	7888419	7888550	132		
	scr7204	*sco7203*	*sco7204*	+	8006662	8007032	371		
	scr7213	*sco7212*	*sco7213*	+	8016644	8016820	177		
	scr7214	*sco7213*	*sco7214*	+	8017756	8017850	95		
	scr7434	*sco7433*	*sco7434*	-	8247159	8247075	85		
	scr7776	*sco7775*	*sco7776*	+	8604074	8604158	85		
S. avermitilis									
	sar0048	*sav47*	*sav48*	+	57774	57862	89		
	sar0059	*sav58*	*sav59*	+	67588	67733	146		
	sar0088	*sav87*	*sav88*	+	98161	98540	380		
	sar0170	*sav169*	*sav170*	+	189927	189977	51		
	sar0212	*sav211*	*sav212*	+	247123	247236	114		
	sar0245	*sav244*	*sav245*	-	282713	282594	120		
	sar0250	*sav249*	*sav250*	-	288680	288607	74		
	sar0273	*sav272*	*sav273*	-	320477	320276	202		
	sar0287	*sav286*	*sav287*	-	336631	336553	79		
	sar7594[x]	*sav312*	*sav7594*	+	370213	370258	46		
	sar0324_1	*sav323*	*sav324*	-	382268	382160	109		
	sar0324_2	*sav323*	*sav324*	-	382428	382371	58		
	sar0324_3	*sav323*	*sav324*	-	382612	382447	166		
	sar0351	*sav350*	*sav351*	-	412386	412056	331		
	sar0353	*sav352*	*sav353*	-	414792	414750	43		
	sar0444	*sav443*	*sav444*	-	598309	598225	85		
	sar0466	*sav465*	*sav466*	+	616811	616926	116		
	sar7577[x]	*sav504*	*sav7577*	-	648062	647767	296		

	sar0625	*sav624*	*sav625*	-	777160	777061	100		
	sar0634	*sav633*	*sav634*	-	788238	788113	126		
	sar0709	*sav708*	*sav709*	-	864406	864174	233		
	sar0742	*sav741*	*sav742*	-	890982	890920	63		
	sar0778	*sav777*	*sav778*	+	927129	927220	92		
	sar0805	*sav804*	*sav805*	-	958110	958026	85		
	sar0823	*sav822*	*sav823*	+	973852	973976	125		
	sar0829	*sav828*	*sav829*	+	982914	982983	70		
	sar0959	*sav958*	*sav959*	-	1219321	1219274	48		
	sar0964	*sav963*	*sav964*	-	1223354	1223170	185		
	sar0967	*sav966*	*sav967*	-	1225858	1225751	108		
	sar0994	*sav993*	*sav994*	-	1259585	1259487	99		
	sar1055	*sav1054*	*sav1055*	-	1332741	1332671	71		
	sar1077	*sav1076*	*sav1077*	-	1357266	1357224	43		
	sar1106	*sav1105*	*sav1106*	-	1391110	1391029	82		
	sar1129	*sav1128*	*sav1129*	-	1417372	1417296	77		
	sar1190	*sav1189*	*sav1190*	+	1480461	1480530	70		
	sar1249	*sav1248*	*sav1249*	+	1548899	1548993	95	Polyketide / non-ribosomal peptide hybrid	
	sar1289	*sav1288*	*sav1289*	-	1597349	1597206	144		
	sar1474	*sav1473*	*sav1474*	-	1817130	1817100	31		
	sar1503	*sav1502*	*sav1503*	-	1843553	1843465	89		
	sar1553	*sav1552*	*sav1553*	-	1912470	1912379	92	Polyketide	
	sar1571	*sav1570*	*sav1571*	+	1935799	1935844	46		
	sar1655	*sav1654*	*sav1655*	+	2027150	2027205	56	Hopene	
	sar1689	*sav1688*	*sav1699*	+	2066680	2066772	93		
	sar1711	*sav1710*	*sav1711*	-	2091025	2090892	134		
	sar1814	*sav1813*	*sav1814*	+	2216915	2217244	330		
	sar2017	*sav2016*	*sav2017*	-	2458516	2458411	106		
	sar2028	*sav2027*	*sav2028*	+	2468943	2469157	215		
	sar2087	*sav2086*	*sav2087*	-	2541468	2541337	132		
	sar2124	*sav2123*	*sav2124*	-	2590455	2590363	93		
	sar2127	*sav2126*	*sav2127*	+	2594347	2594434	88		
	sar2199	*sav2198*	*sav2199*	+	2673499	2673564	66		
	sar2203	*sav2202*	*sav2203*	+	2682862	2683133	272		

	sar2245	*sav2244*	*sav2245*	+	2732474	2732605	132		
	sar2254	*sav2253*	*sav2254*	+	2748973	2749062	90		
	sar2261	*sav2260*	*sav2261*	+	2758394	2758465	72		
	sar2301	*sav2300*	*sav2301*	+	2803270	2803624	355		
	sar2386	*sav2385*	*sav2386*	-	2911641	2911540	102	Aromatic polyketide	
	sar2406	*sav2405*	*sav2406*	+	2932336	2932772	437		
	sar2421	*sav2420*	*sav2421*	-	2949746	2949603	144		
	sar2515*	*sav2514*	SAV_*2515*	-	3079265	3079081	185		
	sar2517	*sav2516*	*sav2517*	-	3082189	3082129	61		
	sar2548	*sav2547*	*sav2548*	+	3126939	3127014	76		
	sar2551	*sav2550*	*sav2551*	+	3128837	3128898	62		
	sar2567	*sav2566*	*sav2567*	-	3148409	3148337	73		
	sar2627	*sav2626*	*sav2627*	+	3226335	3226392	58		
	sar2652	*sav2651*	*sav2652*	-	3250602	3250521	82		
	sar2765	*sav2764*	*sav2765*	-	3392358	3392270	89		
	sar2770	*sav2769*	*sav2770*	-	3399050	3398950	101		
	sar2808	*sav2807*	*sav2808*	+	3451152	3451231	80		
	sar2924	*sav2923*	*sav2924*	+	3659378	3659414	37		
	sar2966	*sav2965*	*sav2966*	+	3716515	3716601	87		
	sar3006	*sav3005*	*sav3006*	-	3760718	3760638	81		
	sar3017	*sav3016*	*sav3017*	+	3771883	3771800	84		
	sar3018	*sav3017*	*sav3018*	-	3773826	3773483	344		
	sar3107	*sav3106*	*sav3107*	+	3879684	3880075	392		
	sar3140	*sav3139*	*sav3140*	+	3912959	3913036	78		
	sar3145	*sav3144*	*sav3145*	-	3917556	3917496	61		
	sar3164	*sav3163*	*sav3164*	+	3941458	3941806	349	Non-ribosomal peptide	
	sar3238	*sav3237*	*sav3238*	-	4036188	4036140	49		
	sar3338	*sav3337*	*sav3338*	-	4152217	4152150	68		
	sar3386	*sav3385*	*sav3386*	+	4207298	4207426	129		
	sar3482	*sav3481*	*sav3482*	+	4314599	4314690	92		
	sar3510	*sav3509*	*sav3510*	-	4344147	4344001	147		
	sar3573	*sav3572*	*sav3573*	-	4423000	4422844	157		
	sar3677	*sav3676*	*sav3677*	+	4553774	4553929	156		
	sar3679	*sav3678*	*sav3679*	+	4555649	4555702	54		
	sar3694_1	*sav3693*	*sav3694*	+	4571212	4571486	275		

	sar3694_2	*sav3693*	*sav3694*	-	4571405	4571266	140		
	sar3698	*sav3697*	*sav3698*	-	4575349	4575256	94		
	sar3708_1*	*sav3707*	*sav3708*	+	4590173	4590242	70		
	sar3708_2*	*sav3707*	*sav3708*	-	4590259	4590179	81		
	sar3713	*sav3712*	*sav3713*	+	4594744	4594818	75		
	sar3721	*sav3720*	*sav3721*	+	4602366	4602436	71		
	sar3730	*sav3729*	*sav3730*	+	4617509	4617600	92		
	sar3855	*sav3854*	*sav3855*	+	4762016	4762130	115		
	sar3858	*sav3857*	*sav3858*	+	4765693	4765915	223		
	sar3861	*sav3860*	*sav3861*	-	4769708	4769638	71		
	sar3892	*sav3891*	*sav3892*	-	4808188	4808105	84		
	sar3948	*sav3947*	*sav3948*	+	4872344	4872418	75		
	sar3970	*sav3969*	*sav3970*	+	4895793	4896030	238		
	sar3975	*sav3975*	*sav3976*	-	4901101	4901011	91		
	sar3980	*sav3979*	*sav3980*	+	4906190	4906277	88		
	sar3982	*sav3982*	*sav3983*	+	4909537	4909620	84		
	sar4024	*sav4023*	*sav4024*	+	4945054	4945399	346		
	sar4076	*sav4075*	*sav4076*	+	4996048	4996124	77		
	sar4105_1*	*sav4104*	*sav4105*	-	5024050	5024095	46		
	sar4105_2*	*sav4104*	*sav4105*	-	5029867	5029660	208		
	sar4113	SAV_4112	*sav4113*	-	5043566	5043450	117		
	sar4119*	*sav4118*	*sav4119*	-	5051673	5051507	167		
	sar4142*	*sav4141*	*sav4142*	+	5081598	5081712	115		
	sar4154	*sav4153*	*sav4154*	+	5102609	5102792	184		
	sar4183	*sav4182*	*sav4183*	-	5132493	5132428	66		
	sar4221	*sav4220*	*sav4221*	-	5176975	5176902	74		
	sar4227	*sav4226*	*sav4227*	+	5182448	5182494	47		
	sar4246	*sav4245*	*sav4246*	-	5207192	5207050	143		
	sar4274	*sav4273*	*sav4274*	-	5240107	5240034	74		
	sar4280	*sav4279*	*sav4280*	+	5246032	5246079	48		
	sar4323	*sav4322*	*sav4323*	-	5298534	5298439	96		
	sar4325	*sav4324*	*sav4325*	+	5300359	5300585	227		
	sar4449	*sav4448*	*sav4449*	-	5435097	5434767	331		
	sar4464	*sav4463*	*sav4464*	+	5452987	5453201	215		
	sar4466	*sav4465*	*sav4466*	-	5454893	5454787	107		
	sar4518	*sav4517*	*sav4518*	+	5519730	5520278	549		

	sar4599_1	*sav4598*	*sav4599*	+	5613649	5613695	47		
	sar4599_2	*sav4598*	*sav4599*	-	5613740	5613701	40		
	sar4604	*sav4603*	*sav4604*	-	5622151	5621921	231		
	sar4605	*sav4604*	*sav4605*	+	5623152	5623293	142		
	sar4634	*sav4633*	*sav4634*	+	5657853	5657894	42		
	sar4635	*sav4634*	*sav4635*	-	5659414	5659293	122		
	sar4691	*sav4690*	*sav4691*	+	5717781	5717832	52		
	sar4726*	*sav4725*	*sav4726*	+	5758724	5758952	229		
	sar4755_1	*sav4754*	*sav4755*	+	5798158	5798235	78		
	sar4755_2	*sav4754*	*sav4755*	-	5798244	5798178	67		
	sar4784	*sav4783*	*sav4784*	+	5827430	5827521	92		
	sar5002_1	*sav5001*	*sav5002*	+	6051621	6051861	241		
	sar5002_2	*sav5001*	*sav5002*	-	6052035	6051886	150		
	sar5018	*sav5017*	*sav5018*	-	6075746	6075660	87		
	sar5019	*sav5018*	*sav5019*	-	6076816	6076706	111		
	sar5020	*sav5019*	*sav5020*	-	6077813	6077707	107		
	sar5033	*sav5032*	*sav5033*	-	6092394	6092349	46		
	sar5057*	*sav5056*	*sav5057*	+	6123837	6124050	214		
	sar5126	*sav5125*	*sav5126*	-	6225412	6225322	91		
	sar5162	*sav5161*	*sav5162*	-	6270115	6270026	90		
	sar5217	*sav5216*	*sav5217*	-	6322900	6322737	164		
	sar5282	*sav5281*	*sav5282*	-	6399268	6399222	47		
	sar5316	*sav5315*	*sav5316*	+	6443023	6443119	97		
	sar5331_1	*sav5330*	*sav5331*	-	6460234	6460136	99		
	sar5331_2	*sav5330*	*sav5331*	-	6460569	6460260	310		
	sar5366	*sav5365*	*sav5366*	-	6506449	6506370	80		
	sar5413	*sav5412*	*sav5413*	-	6563407	6563315	93		
	sar5523	*sav5522*	*sav5523*	+	6694234	6694613	380		
	sar5549	*sav5548*	*sav5549*	-	6712176	6712127	50		
	sar5562	*sav5561*	*sav5562*	-	6850201	6850110	92		
	sar5751	*sav5750*	*sav5751*	+	6953178	6953422	245		
	sar5782	*sav5781*	*sav5782*	+	6987407	6987469	63		
	sar5828	*sav5827*	*sav5828*	+	7034388	7034568	181		
	sar5855	*sav5854*	*sav5855*	-	7063523	7063435	89		
	sar5912	*sav5911*	*sav5912*	+	7121977	7122048	72		
	sar5914	*sav5913*	*sav5914*	-	7125394	7125232	163		

	sar6002	*sav6001*	*sav6002*	+	7229111	7229269	159		
	sar6057	*sav6056*	*sav6057*	+	7291229	7291400	172		
	sar6063	*sav6062*	*sav6063*	+	7297145	7297192	48		
	sar6085	*sav6084*	*sav6085*	-	7322414	7322306	109		
	sar6087	*sav6086*	*sav6087*	-	7325372	7325302	71		
	sar6096	*sav6095*	*sav6096*	-	7332956	7332846	111		
	sar6104	*sav6103*	*sav6104*	+	7340361	7340426	66		
	sar6191	*sav6190*	*sav6191*	+	7441504	7441585	82		
	sar6224	*sav6223*	*sav6224*	+	7476616	7476662	47		
	sar6284	*sav6283*	*sav6284*	-	7542085	7542006	80		
	sar6287	*sav6286*	*sav6287*	+	7546861	7546922	62		
	sar6350_1	*sav6349*	*sav6350*	+	7617994	7618066	73		
	sar6350_2	*sav6349*	*sav6350*	+	7618072	7618377	306		
	sar6383	*sav6382*	*sav6383*	-	7656436	7656340	97		
	sar6459	*sav6458*	*sav6459*	-	7736351	7736291	61		
	sar6487*	*sav6486*	*sav6487*	+	7764906	7765176	271		
	sar6642	*sav6641*	*sav6642*	-	7946216	7946167	50		
	sar6650_1	*sav6649*	*sav6650*	+	7954775	7954869	95		
	sar6650_2	*sav6649*	*sav6650*	+	7954913	7954999	87		
	sar6673	*sav6672*	*sav6673*	+	7984373	7984448	76		
	sar6697	*sav6696*	*sav6697*	-	8009565	8009483	83		
	sar6711	*sav6710*	*sav6711*	-	8026341	8026244	98		
	sar6745	*sav6744*	*sav6745*	-	8063459	8063309	151		
	sar6875_1	*sav6874*	*sav6875*	+	8204040	8204175	136		
	sar6875_2	*sav6874*	*sav6875*	+	8204215	8204302	88		
	sar6912	*sav6911*	*sav6912*	-	8246202	8246140	63		
	sar6923	*sav6922*	*sav6923*	+	8259274	8259573	300		
	sar6976*	*sav6975*	*sav6976*	+	8328102	8328240	139		
	sar6989	*sav6988*	*sav6989*	-	8348863	8348818	46		
	sar6990	*sav6989*	*sav6990*	-	8349781	8349748	34		
	sar6991	*sav6990*	*sav6991*	+	8351549	8351628	80		
	sar6998	*sav6997*	*sav6998*	+	8358793	8358966	174		
	sar7100	*sav7099*	*sav7100*	-	8461862	8461743	120		
	sar7183	*sav7182*	*sav7183*	+	8552575	8552613	39		
	sar7207	*sav7206*	*sav7207*	-	8583516	8583427	90		
	sar7256	*sav7255*	*sav7256*	-	8652122	8652078	45		

	sar7310	*sav7209*	*sav7310*	+	8718717	8718789	73		
	sar7393	*sav7392*	*sav7393*	+	8820487	8820529	43		
	sar7513	*sav7512*	*sav7513*	+	8956302	8956421	120		
	sar7536	*sav7535*	*sav7536*	+	8977882	8977950	69		
S. venezuelae									
	svr0090	*sven0089*	*sven0090*	+	86665	86870	206		
	svr0155	*sven0154*	*sven0155*	+	161764	161816	53		
	svr0164	*sven0163*	*sven0164*	+	171837	172019	183		
	svr0205	*sven0204*	*sven0205*	-	219518	219415	104		
	svr0308	*sven0307*	*sven0308*	-	327545	327508	38		
	svr0357	*sven0356*	*sven0357*	+	383067	383154	88		
	svr0359	*sven0358*	*sven0359*	+	386316	386401	86		
	svr0413	*sven0412*	*sven0413*	+	439342	439378	37		
	svr0441	*sven0440*	*sven0441*	+	473451	473516	66		
	svr0461	*sven0460*	*sven0461*	+	501812	502021	210		
	svr0475	*sven0474*	*sven0475*	+	514567	514606	40	Polyketide / non-ribosomal peptide hybrid	
	svr0552	*sven0551*	*sven0552*	-	633652	633532	121	Terpene	
	svr0651	*sven0650*	*sven0651*	-	763534	763245	290		
	svr0658	*sven0657*	*sven0658*	-	774825	774738	88		
	svr0735_1	*sven0734*	*sven0735*	+	846062	846129	68		
	svr0735_2	*sven0734*	*sven0735*	-	846191	846065	127		
	svr0785	*sven0784*	*sven0785*	-	905788	905581	208		
	svr0823	*sven0822*	*sven0823*	+	950241	950395	155		
	svr0825	*sven0824*	*sven0825*	+	952484	952535	52		
	svr0974_1*	*sven0973*	*sven0974*	-	1104401	1104137	265		
	svr0974_2	*sven0973*	*sven0974*	-	1104518	1104425	94		
	svr1004	*sven1003*	*sven1004*	-	1146315	1146248	68		
	svr1013_1	*sven1012*	*sven1013*	+	1155885	1155923	39		
	svr1013_2	*sven1012*	*sven1013*	-	1155919	1155881	39		
	svr1013_3	*sven1012*	*sven1013*	-	1155980	1155941	40		
	svr1015	*sven1014*	*sven1015*	-	1157731	1157534	198		
	svr1031	*sven1030*	*sven1031*	+	1171874	1171946	73		
	svr1032	*sven1031*	*sven1032*	+	1174057	1174096	40		

	svr1076	*sven1075*	*sven1076*	-	1219934	1219898	37		
	svr1078	*sven1077*	*sven1078*	-	1222806	1222766	41		
	svr1134	*sven1133*	*sven1134*	+	1281666	1281744	79		
	svr1153	*sven1152*	*sven1153*	+	1335028	1335071	44		
	svr1186	*sven1185*	*sven1186*	+	1335028	1335071	44		
	svr1188	*sven1187*	*sven1188*	+	1337216	1337329	114		
	svr1240	*sven1239*	*sven1240*	+	1393042	1393093	52		
	svr1249	*sven1248*	*sven1249*	-	1403844	1403774	71		
	svr1270	*sven1269*	*sven1270*	+	1431298	1431342	45		
	svr1310	*sven1309*	*sven1310*	-	1469417	1469377	41		
	svr1341	*sven1340*	*sven1341*	+	1503166	1503321	156		
	svr1427	*sven1426*	*sven1427*	-	1598148	1597742	407		
	svr1428*	*sven1427*	*sven1428*	-	1600880	1600603	278		
	svr1577	*sven1576*	*sven1577*	-	1758249	1758215	35		
	svr1663	*sven1662*	*sven1663*	-	1860448	1860336	113		
	svr1676	*sven1675*	*sven1676*	-	1871497	1871416	82		
	svr1699	*sven1698*	*sven1699*	-	1896981	1896944	38		
	svr1760	*sven1759*	*sven1760*	+	1965201	1965241	41		
	svr1766	*sven1765*	*sven1766*	+	1969717	1969824	108		
	svr1769	*sven1768*	*sven1769*	+	1972038	1972076	39		
	svr1778	*sven1777*	*sven1778*	+	1982548	1982633	86		
	svr1804	*sven1803*	*sven1804*	-	2009995	2009848	148		
	svr1820	*sven1819*	*sven1820*	-	2026024	2025909	116		
	svr1875	*sven1874*	*sven1875*	+	2086548	2086588	41		
	svr1914	*sven1913*	*sven1914*	-	2133063	2133003	61		
	svr1920	*sven1919*	*sven1920*	-	2135820	2135747	74		
	svr1942	*sven1941*	*sven1942*	+	2158399	2158459	61		
	svr2034	*sven2033*	*sven2034*	+	2246355	2246432	78		
	svr2100	*sven2099*	*sven2100*	-	2291833	2291786	48		
	svr2113	*sven2112*	*sven2113*	-	2299104	2299047	58		
	svr2128	*sven2127*	*sven2128*	+	2307576	2307608	33		
	svr2153	*sven2152*	*sven2153*	-	2323486	2323419	68		
	svr2173_1*	*sven2172*	*sven2173*	+	2338858	2338912	55		
	svr2173_2*	*sven2172*	*sven2173*	+	2339004	2339094	91		
	svr2184	*sven2183*	*sven2184*	-	2350942	2350908	35		
	svr2186	*sven2185*	*sven2186*	+	2354566	2354657	92		

	svr2213	*sven2212*	*sven2213*	-	2382190	2382148	43		
	svr2222	*sven2221*	*sven2222*	-	2391316	2391229	88		
	svr2286	*sven2285*	*sven2286*	+	2466170	2466321	152		
	svr2309	*sven2308*	*sven2309*	+	2492736	2492818	83		
	svr2325	*sven2324*	*sven2325*	-	2508047	2507989	59		
	svr2374	*sven2373*	*sven2374*	+	2564947	2564990	44		
	svr2380	*sven2379*	*sven2380*	+	2573116	2573306	191		
	svr2404	*sven2403*	*sven2404*	+	2605276	2605326	51		
	svr2416	*sven2415*	*sven2416*	+	2617740	2617829	90		
	svr2475	*sven2474*	*sven2475*	+	2692459	2692540	82		
	svr2477	*sven2476*	*sven2477*	+	2694888	2694962	75		
	svr2486	*sven2485*	*sven2486*	+	2701949	2702025	77		
	svr2501	*sven2500*	*sven2501*	-	2712672	2712634	39		
	svr2503	*sven2502*	*sven2503*	+	2713691	2713763	73		
	svr2535	*sven2534*	*sven2535*	+	2753750	2753848	99		
	svr2569	*sven2568*	*sven2569*	-	2795716	2795658	59	Siderophore	
	svr2605	*sven2604*	*sven2605*	+	2840678	2840755	78		
	svr2658	*sven2657*	*sven2658*	-	2893438	2893355	84		
	svr2670	*sven2669*	*sven2670*	+	2904517	2904558	42		
	svr2676	*sven2675*	*sven2676*	-	2909883	2909847	37		
	svr2691	*sven2690*	*sven2691*	+	2925203	2925300	98		
	svr2719	*sven2718*	*sven2719*	-	2960718	2960571	148		
	svr2763*	*sven2762*	*sven2763*	-	3026714	3026501	214		
	svr2831	*sven2830*	*sven2831*	-	3105245	3105096	150		
	svr3031	*sven3030*	*sven3031*	+	3316921	3317100	180		
	svr3058	*sven3057*	*sven3058*	-	3349419	3349351	69		
	svr3068	*sven3067*	*sven3068*	+	3362151	3362221	71		
	svr3195_1*	*sven3194*	*sven3195*	-	3501490	3501255	236		
	svr3195_2*	*sven3194*	*sven3195*	+	3501555	3501600	46		
	svr3209	*sven3208*	*sven3209*	-	3515843	3515777	67		
	svr3316	*sven3315*	*sven3316*	+	3631774	3631883	110		
	svr3328	*sven3327*	*sven3328*	-	3644356	3644215	142		
	svr3329	*sven3328*	*sven3329*	+	3645319	3645545	227		
	svr3373	*sven3372*	*sven3373*	-	3692244	3692163	82		
	svr3390	*sven3389*	*sven3390*	-	3703742	3703697	46		
	svr3412	*sven3411*	*sven3412*	+	3729165	3729223	59		

	svr3468	*sven3467*	*sven3468*	+	3780761	3781260	500		
	svr3504	*sven3503*	*sven3504*	+	3813879	3814075	197		
	svr3556	*sven3555*	*sven3556*	-	3862804	3862721	84		
	svr3584	*sven3583*	*sven3584*	-	3895923	3895890	34		
	svr3641	*sven3640*	*sven3641*	-	3954043	3953989	55		
	svr3699	*sven3698*	*sven3699*	+	4016611	4016686	76		
	svr3707	*sven3706*	*sven3707*	+	4022640	4022733	94		
	svr3746	*sven3745*	*sven3746*	-	4070199	4070157	43		
	svr3778	*sven3777*	*sven3778*	-	4105389	4105357	33		
	svr3789*	*sven3788*	*sven3789*	+	4113664	4113824	161		
	svr3832*	*sven3831*	*sven3832*	+	4157390	4157458	69		
	svr3869	*sven3868*	*sven3869*	+	4202938	4203060	123		
	svr3877*	*sven3876*	*sven3877*	+	4217760	4217964	205		
	svr3967	*sven3966*	*sven3967*	-	4308327	4307973	355		
	svr3995	*sven3994*	*sven3995*	+	4331751	4331835	85		
	svr3999	*sven3998*	*sven3999*	+	4336011	4336100	90		
	svr4012	*sven4011*	*sven4012*	+	4345906	4346105	200		
	svr4031	*sven4030*	*sven4031*	+	4365408	4365444	37		
	svr4046	*sven4045*	*sven4046*	+	4382673	4382911	239		
	svr4090	*sven4089*	*sven4090*	+	4433744	4433822	79		
	svr4155	*sven4154*	*sven4155*	+	4496853	4496938	86		
	svr4204	*sven4203*	*sven4204*	+	4548689	4548728	40		
	svr4210	*sven4209*	*sven4210*	-	4555638	4555522	117		
	svr4236	*sven4235*	*sven4236*	+	4580872	4580998	127		
	svr4279	*sven4278*	*sven4279*	+	4632238	4632272	35		
	svr4346	*sven4345*	*sven4346*	-	4703066	4702976	91		
	svr4464	*sven4463*	*sven4464*	-	4811777	4811740	38		
	svr4465	*sven4464*	*sven4465*	+	4812758	4812839	82		
	svr4477	*sven4476*	*sven4477*	-	4828765	4828614	152		
	svr4511	*sven4510*	*sven4511*	+	4865853	4865892	40		
	svr4633	*sven4632*	*sven4633*	+	4987729	4987852	124		
	svr4647	*sven4646*	*sven4647*	-	5004978	5004937	42	Melanin	
	svr4693	*sven4692*	*sven4693*	+	5048583	5048630	48		
	svr4737	*sven4736*	*sven4737*	+	5093394	5093590	197		
	svr4824	*sven4823*	*sven4824*	-	5188990	5188941	50		
	svr4904	*sven4903*	*sven4904*	-	5282887	5282654	234		

	svr5026	*sven5025*	*sven5026*	-	5410945	5410906	40		
	svr5058	*sven5057*	*sven5058*	+	5443073	5443117	45		
	svr5092	*sven5091*	*sven5092*	-	5480122	5480023	100	Butyrolactone	
	svr5117	*sven5116*	*sven5117*	-	5524420	5524299	122	Thiopeptide	
	svr5272	*sven5271*	*sven5272*	-	5695125	5694800	326		
	svr5279	*sven5278*	*sven5279*	+	5706907	5706987	81		
	svr5283	*sven5282*	*sven5283*	-	5713220	5713165	56		
	svr5316	*sven5315*	*sven5316*	+	5744324	5744401	78		
	svr5393	*sven5392*	*sven5393*	+	5837098	5837146	49		
	svr5428	*sven5427*	*sven5428*	-	5884275	5884213	63		
	svr5510	*sven5509*	*sven5510*	+	5985187	5985241	55		
	svr5514	*sven5513*	*sven5514*	+	5992941	5993104	164		
	svr5535	*sven5534*	*sven5535*	+	6017922	6017963	42		
	svr5558	*sven5557*	*sven5558*	+	6039597	6039637	41		
	svr5635	*sven5634*	*sven5635*	-	6121371	6121303	69		
	svr5651	*sven5650*	*sven5651*	-	6133699	6133654	46		
	svr5664	*sven5663*	*sven5664*	+	6147492	6147624	133		
	svr5679	*sven5678*	*sven5679*	+	6160952	6160989	38		
	svr5681	*sven5680*	*sven5681*	+	6161818	6161861	44		
	svr5700	*sven5699*	*sven5700*	+	6184424	6184564	141		
	svr5729	*sven5728*	*sven5729*	+	6213966	6214077	112		
	svr5830	*sven5829*	*sven5830*	-	6349417	6349380	38		
	svr5922	*sven5921*	*sven5922*	+	6449278	6449447	170		
	svr5940	*sven5939*	*sven5940*	-	6464189	6464112	78		
	svr5941	*sven5940*	*sven5941*	+	6468076	6468164	89		
	svr6086	*sven6085*	*sven6086*	-	6629464	6629262	203		
	svr6110	*sven6109*	*sven6110*	-	6654343	6654245	99		
	svr6112	*sven6111*	*sven6112*	+	6656954	6657029	76		
	svr6179	*sven6178*	*sven6179*	+	6730240	6730368	129	Non-ribosomal peptide	
	svr6310	*sven6309*	*sven6310*	-	6886369	6886288	82		
	svr6402	*sven6401*	*sven6402*	-	6987310	6987166	145		
	svr6475	*sven6474*	*sven6475*	-	7064808	7064763	46		
	svr6552	*sven6551*	*sven6552*	+	7160290	7160453	164		
	svr6559	*sven6558*	*sven6559*	+	7167725	7167830	106		
	svr6576	*sven6575*	*sven6576*	+	7185047	7185087	41		

	svr7049	*sven7048*	*sven7049*	-	7715443	7715260	184	Non-ribosomal peptide	
	svr7129	*sven7128*	*sven7129*	+	7825187	7825254	68		
	svr7278	*sven7277*	*sven7278*	+	8008644	8008754	111		
	svr7297	*sven7296*	*sven7297*	-	8027121	8026909	213		
	svr7351	*sven7350*	*sven7351*	+	8083855	8083893	39		

*: sRNA found adjacent to a tRNA or rRNA gene

[x]: unusual gene order in *S. avermitilis*

[y]: locations of clusters as per Nett *et al.* (2009) for *S. avermitilis* and *S. coelicolor,* and using the program antiSMASH (Medema *et al.,* 2011) for *S. venezuelae.*

Table 3.5: Conserved intergenic sRNAs in *Streptomyces*

	S. coelicolor	***S. avermitilis***	***S. venezuelae***	**Potential for encoding a short conserved ORF**
Conserved in three species				
	scr1104	sar1503	svr0735_2	-
	scr1434	sar6912	svr1031	-
	scr1594	sar6745	svr1188	-
	scr1792	sar6487	svr1428	-
	scr2107	sar6096	svr1766	-
	scr2117	sar6085	svr1778	Y (18 amino acids)
	scr2147	sar6057	svr1804	-
	scr2634	sar5413	svr2416	-
	scr2736	sar5331_1	svr2535	-
	scr2952	sar5126	svr2719	-
	scr3077	sar3510	svr2831	-
	scr3334	sar4726	svr3195_1	-
	scr3558	sar4605	svr3328	-
	scr3559	sar4604	svr3329	-
	scr3920	sar4274	svr3699	-
	scr4076	sar4142	svr3832	-
	scr4123	sar4105_1	svr3328	-
	scr4339	sar3892	svr4155	-
	scr4395	sar3855	svr4210	-
	scr4619	sar4221	svr1924	-
	scr4677	sar3140	svr3556	-
	scr5239	sar3018	svr4904	-
	scr5583	sar2652	svr5279	-
	scr5839	sar2421	svr5514	-
	scr5917	sar0324_1	svr3468	-
	scr6214	sar2017	svr6086	-
Conserved in *S. avermitilis* and *S. venezuelae*				
		sar4154	svr0461	-
		sar6191	svr1676	
		sar5855	svr2034	-
		sar5162	svr2670	-
		sar3677	svr3031	-
		sar4755_1	svr3556	-
		sar4227	svr3746	-
		sar3861	svr4204	-
		sar5020	svr4465	-
		sar2517	svr5393	-
		sar1689	svr6402	-
Conserved in *S. coelicolor* and *S. avermitilis*				

	scr0959	sar7256		-
	scr1363	sar6998		Y (12-13 amino acids)
	scr1390	sar6976		-
	scr1424	sar6923		-
	scr1961	sar6284		-
	scr2101	sar6103		-
	scr2136	sar6063		-
	scr2822	sar0959		-
	scr3020	sar5057		Y (54-55 amino acids)
	scr3202	sar3694_2		-
	scr3679	sar4464		-
	scr4108	sar4119		Y (17 amino acids)
	scr4139	sar4076		-
	scr4233	sar3970		-
	scr4389	sar3858		-
	scr4689	sar0088		-
	scr5439	sar2808		-
	scr5710	sar2548		-
	scr6106	sar2124		-
Conserved in *S. coelicolor* and *S. venezuelae*				
	scr4628		svr1914	-
	scr4684		svr3504	-

Table 5.1: The most widely distributed hairpins using *E. coli* as the query species

Focus species/gene	gene length	position	ΔG	sequence	# homologous genes	# conserved hairpins	%	Others species/gene
>NC_000913.3_dnaE	3888	1261	-14.3	TACCGGTAGGGCCAGGCCGTGGCTCCGGTG	11	8	0.73	>NC_010572.1_SGR_RS27130,>NC_013929.1_SCAB_RS32605,>NC_003155.5_SAVERM_RS31635,>NC_000964.3_dnaE,>NC_002516.2_dnaE,>NC_000907.1_dnaE,>NC_003143.1_dnaE,>NC_003198.1_STY0254,
>NC_000913.3_ybiT	1995	153	-12	CTCGGCGGCGACCTTGAGCCGACGCTGGGT	10	7	0.7	>NC_010572.1_SGR_RS28190,>NC_003155.5_SAVERM_RS33040,>NC_013929.1_SCAB_RS33955,>NC_003143.1_ybiT,>NC_003198.1_ybiT,>NC_000964.3_ykpA,>NC_002737.2_SPy_2210,
>NC_000913.3_rep	2697	274	-15.8	GCCCCTGCCTGGTGCTGGCGGGCGCGGGTT	11	7	0.64	>NC_000915.1_HP1478,>NC_002516.2_rep,>NC_000907.1_rep,>NC_003143.1_rep,>NC_003198.1_STY3642,>NC_000964.3_pcrA,>NC_002737.2_pcrA,
>NC_000913.3_trxB	1989	1022	-13.8	AGGCCCGGCGGGATACACCGCTGCTGTCTA	12	7	0.58	>NC_002737.2_nox1,>NC_000915.1_HP0825,>NC_000964.3_trxB,>NC_003155.5_SAVERM_RS22255,>NC_003888.3_SCO3890,>NC_000907.1_trxB,>NC_003198.1_STY0956,
>NC_000913.3_sdhA	2280	58	-17.6	GTGCCGGTGGCGCAGGTATGCGCGCGGCGC	9	6	0.67	>NC_000964.3_sdhA,>NC_000915.1_HP0192,>NC_000907.1_frdA,>NC_010572.1_SGR_RS13265,>NC_003143.1_sdhA,>NC_003198.1_STY0777,
>NC_000913.3_ftsZ	1452	435	-14.3	GGTGGTGGTACCGGTACAGGTGCAGCACCA	10	6	0.6	>NC_000915.1_HP0979,>NC_002737.2_ftsZ,>NC_000964.3_ftsZ,>NC_000907.1_ftsZ,>NC_002516.2_ftsZ,>NC_003143.1_ftsZ,
>NC_000913.3_proV	1350	595	-15.8	AACGTGTGGGATTAGCCCGCGCGTTAGCGA	10	6	0.6	>NC_013929.1_SCAB_RS03260,>NC_000964.3_opuAA,>NC_002737.2_opuAA,>NC_003888.3_SCO1621,>NC_002516.2_PA5094,>NC_003143.1_proV,
>NC_000913.3_clpP	693	275	-14.1	CCCAGGCGGGGTGATCACTGCCGGGATGTC	11	6	0.55	>NC_003888.3_SCO7281,>NC_002737.2_clpP,>NC_000964.3_clpP,>NC_000907.1_clpP,>NC_003143.1_clpP,>NC_003198.1_STY0490,
>NC_000913.3_nrdE	3243	1408	-11.4	ACCAACTCGGCGCTCGTCAGGGGGCTGGTG	11	6	0.55	>NC_002737.2_nrdE.1,>NC_003143.1_nrdE,>NC_003198.1_STY2932,>NC_003888.3_SCO5226,>NC_003155.5_SAVERM_RS15715,>NC_013929.1_SCAB_RS14360,
>NC_000913.3_pgk	1263	1119	-11.6	ATCGCTGGCGGCGGCGACACTCTGGCAGCA	11	6	0.55	>NC_000915.1_pgk,>NC_000907.1_pgk,>NC_003155.5_SAVERM_RS32395,>NC_003888.3_pgk,>NC_010572.1_SGR_RS27800,>NC_013929.1_pgk,
>NC_000913.3_prfA	1131	382	-12.9	GAACCGGCGGCGACGAAGCGGCGCTGTTCG	11	6	0.55	>NC_010572.1_SGR_RS10695,>NC_003888.3_prfA,>NC_000964.3_prfA,>NC_002737.2_prfA,>NC_000915.1_prfA,>NC_003143.1_prfA,

>NC_000913.3_tufA	2415	512	-11.4	CCCGGGCGACGACACTCCGATCGTTCGTGG	13	6	0.46	>NC_003143.1_tuf,>NC_002737.2_tufA,>NC_000907.1_tuf, >NC_000907.1_tuf,>NC_000915.1_tuf,>NC_003198.1_tufA,
>NC_000913.3_tufA	2415	1181	-16	CCGTGAAGGCGGCCGTACCGTTGGCGCGGG	13	6	0.46	>NC_002737.2_tufA,>NC_000907.1_tuf,>NC_000907.1_tuf, >NC_003198.1_tufA,>NC_003155.5_SAVERM_RS25400,>N C_013929.1_tuf,
>NC_000913.3_tufB	2415	512	-11.4	CCCGGGCGACGACACTCCGATCGTTCGTGG	13	6	0.46	>NC_003143.1_tuf,>NC_002737.2_tufA,>NC_000907.1_tuf, >NC_000907.1_tuf,>NC_000915.1_tuf,>NC_003198.1_tufB,
>NC_000913.3_tufB	2415	1181	-16	CCGTGAAGGCGGCCGTACCGTTGGCGCGGG	13	6	0.46	>NC_002737.2_tufA,>NC_000907.1_tuf,>NC_000907.1_tuf, >NC_003198.1_tufB,>NC_003155.5_SAVERM_RS25400,>N C_013929.1_tuf,
>NC_000913.3_dcm	1878	311	-11	CGGAATTGGCGGCATCCGTCGCGGTTTTGA	6	5	0.83	>NC_010572.1_SGR_RS36285,>NC_000915.1_HP0051,>NC _000907.1_HI1041,>NC_000964.3_mtbP,>NC_003198.1_ST Y2200,
>NC_000913.3_mnmG	1908	71	-16.8	AGGCACCGAGGCCGCGATGGCCGCGGCGCG	7	5	0.71	>NC_000915.1_gidA,>NC_003143.1_gidA,>NC_003198.1_S TY3904,>NC_000964.3_gidA,>NC_002737.2_gidA,
>NC_000913.3_artQ	1710	895	-15.7	CAAGCGCCGCCGGGATGACCGTCGGCCTTG	8	5	0.62	>NC_000907.1_artQ,>NC_003143.1_artQ,>NC_003198.1_S TY0922,>NC_000915.1_HP0939,>NC_000964.3_tcyB,
>NC_000913.3_cpdB	4647	1502	-9.9	TGCCGCACCGTTTAAAGTCGGTGGTCGCAA	9	5	0.56	>NC_002737.2_SPy_0872,>NC_000964.3_yfkN,>NC_00090 7.1_cpdB,>NC_003143.1_cpdB,>NC_003198.1_cpdB,
>NC_000913.3_edd	2037	1781	-14.3	ACCAGAAGCCTACGATGGCGGGCTGCTGGC	9	5	0.56	>NC_000964.3_ilvD,>NC_000907.1_ilvD,>NC_000915.1_HP 1100,>NC_003143.1_edd,>NC_003198.1_STY2093,
>NC_000913.3_moeB	1200	152	-17.5	GGGCCTGGGCGGCCTCGGCTGTGCAGCCTC	9	5	0.56	>NC_000964.3_moeB,>NC_010572.1_SGR_RS11560,>NC_0 03155.5_SAVERM_RS16000,>NC_013929.1_SCAB_RS14650 ,>NC_003198.1_STY0884,
>NC_000913.3_mppA	1845	87	-9.6	GTCAGCAGCATTTCTCTTTCGTATGCTGCA	9	5	0.56	>NC_010572.1_SGR_RS29185,>NC_000915.1_HP0298,>NC _003888.3_SCO6451,>NC_000964.3_oppA,>NC_003143.1_ mppA,
>NC_000913.3_proA	1290	541	-13.8	CCTGCGGCTTACCGGCGGGTGCCGTGCAGG	9	5	0.56	>NC_002516.2_proA,>NC_010572.1_SGR_RS24720,>NC_00 3155.5_SAVERM_RS28255,>NC_013929.1_SCAB_RS28765, >NC_000907.1_proA,
>NC_000913.3_dnaA	2220	1212	-15.4	CTGGCGCGCGCGGCGGCTCGCCAGGTGGCG	10	5	0.5	>NC_002737.2_dnaA,>NC_002516.2_dnaA,>NC_003143.1_ dnaA,>NC_003198.1_STY3940,>NC_000964.3_dnaA,
>NC_000913.3_gpsA	1053	44	-17.4	GATCGGTGCCGGCTCGTACGGCACCGCTCT	10	5	0.5	>NC_000964.3_gpsA,>NC_002737.2_gpsA,>NC_000907.1_g psA,>NC_003143.1_gpsA,>NC_003198.1_gpsA,
>NC_000913.3_mpl	1497	365	-13.6	CTGGGTGCTGGCCGTTGCCGGTACACACGG	10	5	0.5	>NC_000915.1_murC,>NC_000964.3_murC,>NC_002516.2_ mpl,>NC_003143.1_mpl,>NC_003198.1_mpl,
>NC_000913.3_pheS	1209	247	-13.7	AGCTGCCGCCAGAAGAGCGTCCGGCAGCTG	10	5	0.5	>NC_000915.1_pheS,>NC_002516.2_pheS,>NC_000907.1_ pheS,>NC_003143.1_pheS,>NC_003198.1_STY1773,
>NC_000913.3_secY	1422	1115	-11.9	GAAGTCCGGTGCATTTGTACCAGGAATTCG	10	5	0.5	>NC_000915.1_secY,>NC_000964.3_secY,>NC_013929.1_S CAB_RS17415,>NC_003143.1_secY,>NC_003198.1_prlA,
>NC_000913.3_ybiT	1995	97	-14.9	ACCGTTACGGCCTGATTGGCGCGAACGGTA	10	5	0.5	>NC_000907.1_HI0658,>NC_000915.1_HP0853,>NC_00314 3.1_ybiT,>NC_003198.1_ybiT,>NC_002737.2_SPy_2210,

>NC_000913.3_dnaE	3888	1757	-14.1	CGGTAAGCACGCCGGTGGGGTGGTTATCGC	11	5	0.45	>NC_002737.2_dnaE,>NC_000915.1_dnaE,>NC_002516.2_dnaE,>NC_000907.1_dnaE,>NC_003198.1_STY0254,
>NC_000913.3_nrdA	3213	1398	-13.3	GGCGGTGTGCGCGGCGGTGCGGCAACGCTG	11	5	0.45	>NC_002737.2_nrdE.1,>NC_003198.1_STY2506,>NC_003888.3_SCO5226,>NC_003155.5_SAVERM_RS15715,>NC_013929.1_SCAB_RS14360,
>NC_000913.3_recA	1191	396	-14.5	CTGTGCTCCCAGCCGGACACCGGCGAGCAG	11	5	0.45	>NC_000964.3_recA,>NC_002737.2_recA,>NC_000907.1_recA,>NC_003143.1_recA,>NC_003198.1_STY2950,
>NC_000913.3_rpsR	501	420	-10.3	CTGGCTCGCGCTATCAAACGCGCTCGCTAC	11	5	0.45	>NC_000915.1_rpsR,>NC_002516.2_rpsR,>NC_003198.1_rpsR,>NC_000907.1_rpsR,>NC_000964.3_rpsR,
>NC_000913.3_thrS	3999	3674	-11.2	CTTCCCGACCTGGCTTGCGCCGGTTCAGGT	12	5	0.42	>NC_003888.3_thrS,>NC_003888.3_thrS,>NC_003198.1_STY1778,>NC_000964.3_thrZ,>NC_002737.2_thrS,
>NC_000913.3_gapA	2103	1667	-14	CTGGCGCGGCGGCCGCGGCGCTTCCCAGAA	13	5	0.38	>NC_000915.1_HP0921,>NC_002737.2_plr,>NC_000907.1_gapdH,>NC_003143.1_gapA,>NC_003198.1_STY1825,
>NC_000913.3_tufA	2415	51	-10.9	GTTAACGTTGGTACTATCGGCCACGTTGAC	13	5	0.38	>NC_003143.1_tuf,>NC_002737.2_tufA,>NC_000907.1_tuf,>NC_000907.1_tuf,>NC_003198.1_tufA,
>NC_000913.3_tufB	2415	51	-10.9	GTTAACGTCGGTACTATCGGCCACGTTGAC	13	5	0.38	>NC_003143.1_tuf,>NC_002737.2_tufA,>NC_000907.1_tuf,>NC_000907.1_tuf,>NC_003198.1_tufB,
>NC_000913.3_adiA	2367	529	-16.1	CAGCGCCAGGCCACCAGGGCGGCGTTGGTT	5	4	0.8	>NC_000964.3_speA,>NC_000907.1_speF,>NC_003143.1_YPO1201,>NC_003198.1_adi,
>NC_000913.3_bamA	2865	255	-11.7	GGCCTTCAGCGTGTCGCCGTTGGTGCGGCC	5	4	0.8	>NC_000915.1_HP0655,>NC_000907.1_HI0917,>NC_003143.1_YPO1052,>NC_003198.1_yaeT,
>NC_000913.3_lplA	1380	514	-14.5	GCCTGGCGCGGCGCAGTAGCGGTGGCGGCG	5	4	0.8	>NC_003143.1_lplA,>NC_003198.1_lplA,>NC_000964.3_yhfJ,>NC_002737.2_lplA,
>NC_000913.3_lpxB	1230	81	-19.3	GTCGCCGGAGAAACCTCCGGCGATATCCTG	5	4	0.8	>NC_000915.1_lpxB,>NC_000907.1_lpxB,>NC_003143.1_lpxB,>NC_003198.1_STY0252,
>NC_000913.3_nrdD	2352	742	-14.8	AGGTTGCCAGCCATATTTATGGCGGCACCA	5	4	0.8	>NC_002737.2_nrdD,>NC_000907.1_nrdD,>NC_003143.1_nrdD,>NC_003198.1_nrdD,
>NC_000913.3_rpoN	1536	68	-12.7	GCTCCAACAGGCAATTCGTCTGTTGCAGTT	5	4	0.8	>NC_000915.1_HP0714,>NC_000964.3_sigL,>NC_003143.1_rpoN,>NC_003198.1_STY3499,
>NC_000913.3_thiI	1488	894	-12.7	CTGGTCACCGGCGAAGCGCTCGGCCAGGTG	5	4	0.8	>NC_000964.3_ytbJ,>NC_002737.2_thiI,>NC_003143.1_thiI,>NC_003198.1_STY0464,
>NC_000913.3_yrdD	2325	1449	-15.7	GTCCGTCCTCTGAAATCTTCAGCGGATGGA	6	4	0.67	>NC_003143.1_YPO0245,>NC_003198.1_yrdD,>NC_000915.1_HP0116,>NC_000964.3_topA,
>NC_000913.3_glnD	2877	341	-12.8	GGCATTGGTCGCCGTCGGTGGCTACGGTCG	7	4	0.57	>NC_002516.2_glnD,>NC_000907.1_glnD,>NC_003143.1_glnD,>NC_003198.1_STY0237,
>NC_000913.3_hofB	1794	1016	-13.5	GGTGCTGGTAACTGGCCCTACCGGCAGCGG	7	4	0.57	>NC_002737.2_comYA,>NC_000964.3_comGA,>NC_003143.1_hofB,>NC_003198.1_STY0165,
>NC_000913.3_lacA	846	635	-7	CGTCGTGGCGGCTGGCGTTCCTTGTCGGGT	7	4	0.57	>NC_002737.2_SPy_1065,>NC_003888.3_SCO0261,>NC_000964.3_maa,>NC_003198.1_maA,
>NC_000913.3_lldP	1821	542	-13.9	CACCGCGCCAGTGGCATTTGGTGCGATGGG	7	4	0.57	>NC_000907.1_lctP,>NC_000964.3_yvfH,>NC_000915.1_HP0140,>NC_003198.1_STY4104,
>NC_000913.3_narY	3186	679	-12.6	CCAGCTGCGTGGCGACCTGCCCAAGCGGCG	7	4	0.57	>NC_000907.1_nrfC,>NC_003143.1_dmsB,>NC_003143.1_dmsB,>NC_003198.1_STY1487,

>NC_000913.3_nrdF	2100	301	-9.7	TTATCGGCGCGCCTTCTCTGATGCCCGATG	7	4	0.57	>NC_000964.3_nrdF,>NC_000964.3_nrdF,>NC_003143.1_nrdF,>NC_003198.1_STY2933,
>NC_000913.3_acs	2076	545	-15.9	GCTGGCCTGCGCCCGCATTGGCGCGGTGCA	8	4	0.5	>NC_000964.3_acsA,>NC_000915.1_HP1045,>NC_003143.1_acs,>NC_003198.1_acs,
>NC_000913.3_ascB	1653	159	-15.1	TTATGGGGCGGCGCGCTTGCCGCCAACCAG	8	4	0.5	>NC_003143.1_bglA,>NC_003198.1_STY3207,>NC_002737.2_bglA,>NC_000964.3_bglH,
>NC_000913.3_bglA	1644	762	-11.1	CAGTTTGTCGCCAGCGCCCTGGCGGTGAAA	8	4	0.5	>NC_002737.2_bglA,>NC_000964.3_bglA,>NC_003143.1_bglA,>NC_003198.1_STY3207,
>NC_000913.3_maa	840	605	-11.1	AGATGTCCCGGACAACGTTGTCGTGGGCGG	8	4	0.5	>NC_003155.5_SAVERM_RS36165,>NC_003888.3_SCO0261,>NC_000964.3_maa,>NC_003198.1_maA,
>NC_000913.3_mdtC	3423	365	-14.2	CGGCGCAGCGCGTGATGTGCAGGCGGCGAT	8	4	0.5	>NC_000907.1_acrB,>NC_002516.2_PA2526,>NC_003143.1_yegO,>NC_003198.1_STY2341,
>NC_000913.3_napA	4908	63	-21.5	GCCGTTGCGGCCGCTGCGGCGGCTGCCGGT	8	4	0.5	>NC_000907.1_dmsA,>NC_002516.2_napA,>NC_003143.1_napA,>NC_003198.1_STY2485,
>NC_000913.3_ygbM	1689	1090	-12.7	CGGGGGAATGGGGATTATCCGCCCTTCCCG	8	4	0.5	>NC_003143.1_YPO1238,>NC_003198.1_hyi,>NC_002516.2_PA0550,>NC_000907.1_HI1013,
>NC_000913.3_yhdZ	816	519	-14.6	CAGCAACGCGTTGCCATTGCGCGTTCGCTG	8	4	0.5	>NC_000964.3_glnQ,>NC_002516.2_PA1339,>NC_003143.1_gltL,>NC_003198.1_STY0707,
>NC_000913.3_acnA	2826	1172	-12.6	CCTGGCAGGGCCTAAACGCCCACAGGATCG	9	4	0.44	>NC_000907.1_leuC,>NC_000964.3_citB,>NC_003143.1_acnA,>NC_003198.1_STY1339,
>NC_000913.3_carA	1233	869	-13.8	TCAGCTGCTGGCGCTGGCGAGCGGTGCGAA	9	4	0.44	>NC_003143.1_carA,>NC_003198.1_STY0076,>NC_000964.3_pyrAA,>NC_002737.2_carA,
>NC_000913.3_glpF	984	325	-13.8	AAGTTGCCGGCGCTTTCTGTGCTGCGGCTT	9	4	0.44	>NC_000907.1_glpF,>NC_003143.1_glpF,>NC_003198.1_STY3783,>NC_000964.3_glpF,
>NC_000913.3_gltD	3069	1986	-14.5	GGCGCAGGCCCGGCAGGTCTGGCGTGTGCG	9	4	0.44	>NC_002516.2_gltD,>NC_003143.1_gltD,>NC_000964.3_gltD,>NC_003198.1_STY3511,
>NC_000913.3_ilvD	1992	135	-14.2	GGTGCTCGTGCGCTGTGGCGCGCCACCGGA	9	4	0.44	>NC_010572.1_SGR_RS20530,>NC_000907.1_ilvD,>NC_003143.1_ilvD,>NC_003198.1_STY3653,
>NC_000913.3_rffG	2073	42	-17.3	GGTGCCGGGTTTATTGGCTCGGCGCTGGTG	9	4	0.44	>NC_002737.2_cpsFQ,>NC_000964.3_spsJ,>NC_000907.1_rffG,>NC_003143.1_rffG,
>NC_000913.3_rlmC	1518	1291	-13	AGCTGGTGCTGGTTAACCCGCCGCGCCGCG	9	4	0.44	>NC_000907.1_rumB,>NC_003143.1_rumB,>NC_000964.3_yefA,>NC_002737.2_SPy_1606,
>NC_000913.3_sstT	2976	1156	-15.7	CCTGTGGCGCATCCGGCGTGGCGGGGGGGT	9	4	0.44	>NC_002737.2_SPy_0324,>NC_000907.1_HI1545,>NC_002516.2_PA2042,>NC_000964.3_gltT,
>NC_000913.3_sthA	1872	427	-16.7	GCCCCGGCGGCGAAGGCGCTGCAATGGGCC	9	4	0.44	>NC_003198.1_STY3748,>NC_003143.1_sthA,>NC_002737.2_acoL,>NC_000964.3_pdhD,
>NC_000913.3_tdh	1080	79	-12.9	CCGACGTTCCTGTACCGGAACTCGGGCATA	9	4	0.44	>NC_002737.2_adhA,>NC_003143.1_tdh,>NC_003198.1_tdh,>NC_003155.5_SAVERM_RS08630,
>NC_000913.3_trpB	1293	400	-15.8	GCGTGGCGTCGGCCCTTGCCAGCGCCCTGC	9	4	0.44	>NC_000915.1_HP1278,>NC_000907.1_trpB,>NC_003143.1_trpB,>NC_000964.3_trpB,
>NC_000913.3_trpB	1293	727	-12.9	ATGCCGTTATCGCCTGTGTTGGCGGCGGTT	9	4	0.44	>NC_000915.1_HP1278,>NC_000907.1_trpB,>NC_003198.1_STY1325,>NC_003155.5_SAVERM_RS31800,

>NC_000913.3_yadG	1095	174	-15.1	GCGCTTCTCGGGCCGAACGGGGCCGGGAAA	9	4	0.44	>NC_002516.2_PA2812,>NC_003143.1_yadG,>NC_000964.3_yfiL,>NC_002737.2_SPy_0744,
>NC_000913.3_yecC	816	507	-14.1	CAGCGTGTTGCGATTGCGCGTGCGCTGGCA	9	4	0.44	>NC_000907.1_HI1078,>NC_003143.1_yecC,>NC_003198.1_STY2159,>NC_000964.3_artR,
>NC_000913.3_znuC	1230	160	-14.6	CTTTACTTGGGCCAAATGGCGCAGGTAAGT	9	4	0.44	>NC_002737.2_potA,>NC_000907.1_yebM,>NC_003143.1_znuC,>NC_003198.1_STY2100,
>NC_000913.3_accA	1719	1229	-12.4	CACCCCGGGGGCTTATCCTGGCGTGGGCGC	10	4	0.4	>NC_000915.1_HP0557,>NC_002737.2_accA,>NC_000907.1_accA,>NC_003143.1_accA,
>NC_000913.3_aroK	1686	74	-12	GGGTGCCGGAAAAAGCACTATTGGGCGCCA	10	4	0.4	>NC_000915.1_aroK,>NC_002737.2_aroK,>NC_000964.3_aroK,>NC_003198.1_aroK,
>NC_000913.3_atpA	1629	323	-12.7	TGGCCGTGGCCTGCTGGGCCGTGTGGTTAA	10	4	0.4	>NC_003198.1_STY3911,>NC_003155.5_SAVERM_RS14980,>NC_013929.1_SCAB_RS13650,>NC_000915.1_HP1134,
>NC_000913.3_clpX	1425	810	-12.2	CCACCGCAAGGTGGGCGTAAACATCCGCAG	10	4	0.4	>NC_000915.1_clpX,>NC_003143.1_clpX,>NC_002737.2_clpX,>NC_000964.3_clpX,
>NC_000913.3_dapE	1629	1409	-12.7	CGGAACGTCCGACGGGCGCTTTATTGCCCG	10	4	0.4	>NC_000964.3_yodQ,>NC_000915.1_HP0212,>NC_000907.1_dapE,>NC_003198.1_STY2721,
>NC_000913.3_dmsA	4941	3158	-11.8	AGATGCCGCACTGGTTAACGGTCTGGCGTA	10	4	0.4	>NC_003198.1_STY0962,>NC_000964.3_yoaE,>NC_003143.1_dmsA,>NC_000907.1_dmsA,
>NC_000913.3_eno	1341	979	-11.9	TGGTTGGTGACGACCTGTTCGTAACCAACA	10	4	0.4	>NC_000907.1_eno,>NC_003198.1_STY3081,>NC_000964.3_eno,>NC_002737.2_eno,
>NC_000913.3_ettA	2022	108	-12.6	CCTGGGGCAAAAATTGGTGTCCTGGGTCTG	10	4	0.4	>NC_000915.1_HP0853,>NC_000964.3_ydiF,>NC_003198.1_STY4928,>NC_000907.1_HI1252,
>NC_000913.3_ftsW	1410	918	-13	CAATCGCTGATGGCGTTTGGTCGCGGCGAA	10	4	0.4	>NC_000915.1_HP0743,>NC_002737.2_ftsW,>NC_000907.1_ftsW,>NC_003143.1_ftsW,
>NC_000913.3_lptB	1026	145	-14	GTCTGCTGGGGCCAAACGGTGCCGGTAAGA	10	4	0.4	>NC_000964.3_yfiL,>NC_002737.2_SPy_0744,>NC_000907.1_HI1148,>NC_003143.1_yhbG,
>NC_000913.3_metN	1137	129	-12.4	GGCGTTATCGGTGCCTCAGGCGCGGGTAAG	10	4	0.4	>NC_002737.2_SPy_0320,>NC_000915.1_metN,>NC_003143.1_metN,>NC_003198.1_STY0274,
>NC_000913.3_metQ	930	838	-11.9	AATTCGTCCAGGCTTATCAGTCTGACGAAG	10	4	0.4	>NC_002737.2_SPy_0319,>NC_003143.1_metQ,>NC_003198.1_STY0272,>NC_013929.1_SCAB_RS35625,
>NC_000913.3_mraY	1185	1101	-15.4	CCGGAACCGCGCGTCATTGTGCGTTTCTGG	10	4	0.4	>NC_002516.2_mraY,>NC_000907.1_mraY,>NC_003143.1_mraY,>NC_003198.1_STY0145,
>NC_000913.3_murC	1557	95	-14.4	TGGTGGTGCCGGTATGGGCGGTATTGCCGA	10	4	0.4	>NC_002737.2_murC,>NC_000907.1_murC,>NC_003143.1_murC,>NC_003198.1_STY0149,
>NC_000913.3_rplV	399	323	-12.8	GCGCACCAGCCACATCACTGTGGTTGTGTC	10	4	0.4	>NC_003198.1_rplV,>NC_003143.1_rplV,>NC_000907.1_rplV,>NC_002737.2_rplV,
>NC_000913.3_rpoD	2175	575	-11.2	CCCGGTACGCATGTACATGCGTGAAATGGG	10	4	0.4	>NC_000915.1_HP0088,>NC_000907.1_rpoD,>NC_003143.1_rpoD,>NC_003198.1_STY3390,
>NC_000913.3_rpsQ	294	215	-10.1	CGGTGACGTGGTTGAAATCCGCGAATGCCG	10	4	0.4	>NC_002516.2_rpsQ,>NC_000907.1_rpsQ,>NC_003143.1_rpsQ,>NC_003198.1_rpsQ,
>NC_000913.3_sapB	1050	743	-13	AGCGGCGGCGACCCGCGGTTTGTCGCGCTT	10	4	0.4	>NC_002737.2_dppB,>NC_003888.3_SCO6113,>NC_003198.1_STY1368,>NC_000964.3_oppB,

>NC_000913.3_sdaA	1506	1306	-13.8	TTAACGCCGCGCGGATGGCTCTGCGCCGCA	10	4	0.4	>NC_002737.2_sdhA,>NC_000907.1_sdaA,>NC_003143.1_sdaA,>NC_010572.1_SGR_RS10070,
>NC_000913.3_sdaB	1482	1289	-16.1	CGCACGTATGGCGCTGCGCCGTACCAGCGA	10	4	0.4	>NC_002737.2_sdhA,>NC_000907.1_sdaA,>NC_003143.1_sdaA,>NC_010572.1_SGR_RS10070,
>NC_000913.3_tdcG	1494	1111	-14.2	TTGGCGTGGCCTGTTCAATGGCGGCGGCAG	10	4	0.4	>NC_002737.2_sdhA,>NC_010572.1_SGR_RS10070,>NC_000907.1_sdaA,>NC_003143.1_sdaA,
>NC_000913.3_tyrS	1350	240	-20	CCGGTTGCGCTGGTAGGCGGCGCGACGGGT	10	4	0.4	>NC_003143.1_tyrS,>NC_003198.1_STY1673,>NC_002737.2_tyrS,>NC_000964.3_tyrS,
>NC_000913.3_xseA	1440	351	-16.4	ATGCAGCCGGCCGGTGAAGGGCTGCTGCAA	10	4	0.4	>NC_000915.1_xseA,>NC_003143.1_xseA,>NC_002737.2_xseA,>NC_000964.3_xseA,
>NC_000913.3_ychF	1128	937	-11.2	CGGTTGGAGCAACCGCGCCGCAGGCAGCGG	10	4	0.4	>NC_000915.1_ychF,>NC_003143.1_ychF,>NC_003198.1_STY1910,>NC_000964.3_yyaF,
>NC_000913.3_yihW	1827	173	-16.3	GATTACGCGCCATCACGGTGGCGCGGGTCG	10	4	0.4	>NC_003143.1_glpR,>NC_000907.1_glpR,>NC_000907.1_glpR,>NC_002516.2_glpR,
>NC_000913.3_ynfE	5082	3179	-13.4	CACCGATGCCGCGCTGGTTGCGGGTATTGC	10	4	0.4	>NC_000964.3_yoaE,>NC_003143.1_dmsA,>NC_000907.1_dmsA,>NC_002516.2_PA2714,
>NC_000913.3_hisS	1392	67	-14.1	ACCTGCCTGGCGAAACGGCCATCTGGCAGC	11	4	0.36	>NC_003198.1_STY2767,>NC_003143.1_hisS,>NC_000964.3_hisS,>NC_002737.2_hisS,
>NC_000913.3_nrdA	3213	1287	-13.4	GCCGGGCGTATTCGTGCGCTGGGTAGCCCG	11	4	0.36	>NC_000964.3_nrdE,>NC_000907.1_nrdA,>NC_003143.1_nrdA,>NC_002516.2_nrdA,
>NC_000913.3_nrdE	3243	1262	-13.2	ACTGCAGCTGTCGAAACGCGGCGGCGGCGT	11	4	0.36	>NC_000964.3_nrdE,>NC_003143.1_nrdE,>NC_003198.1_STY2932,>NC_000915.1_HP0680,
>NC_000913.3_proS	1965	1170	-15.6	GCCGGTCCGGGTTCACTGGGTCCGGTAAAC	11	4	0.36	>NC_000915.1_HP0238,>NC_002516.2_proS,>NC_000907.1_proS,>NC_000964.3_proS,
>NC_000913.3_rpoC	9294	5342	-14.2	CGGTCGTCGCGGTCGTGCGATCACCGGTTC	11	4	0.36	>NC_003143.1_rpoC,>NC_003198.1_STY3731,>NC_003888.3_SCO4655,>NC_002737.2_rpoC,
>NC_000913.3_secA	2973	1076	-15.1	GCAGGGCCGTCGCTGGTCCGATGGTCTGCA	11	4	0.36	>NC_000915.1_secA,>NC_000907.1_secA,>NC_003143.1_secA,>NC_000964.3_secA,
>NC_000913.3_fusA	4218	3032	-13.3	GGCTGAACGTCACGCAAGTGATGACGAGCC	12	4	0.33	>NC_002737.2_fus,>NC_000907.1_fusA,>NC_003198.1_fusA,>NC_003143.1_fusA,

Appendix 2: Perl script for Chapter 5

```
#!/usr/bin/perl

open (ANALYSIS,">analysis.out");

#this subroutine accepts 4 variables, 1) fasta_name, which is the fasta_header, 2)
#nucleotide sequence to analyze, 3) ΔG value cutoff, 4) window_size
#the program divides the sequence into sliding widows based on the window size and
#a slide size of 1 base, and calls RNAFold, reporting all ΔG values below the cutoff

sub hairpinfinder {
                    #four variables are passed into hairpinfinder; the first is
                    #the fasta name, the second is the sequence, and the third is the ΔG
                    #cutoff, below which a hairpin is determined found, the fourth
                    #is the window size with which to divide the mRNA

                    @return=();
                    $fasta_name = shift;
                    $cur_fasta_sequence = shift;
                    $cutoff = shift;
                    $window_size = shift;

                    #hard code in the slide_size to be 1; this means that each window will
                    #be shifted one nucleotide at each iteration

                    $slide_size = 1;

                    #generate sub-sequences based on window and slide sizes
                    #create variables for left and right ends of the window
                    $cur_left=0;
                    $cur_right=$cur_left+$window_size;
                    $count = 1;
                    while ($cur_right <= length($cur_fasta_sequence)) {

                                        $center_position = $cur_left + ($window_size/2);
                                        $cur_window_name= $fasta_name;

                                        #define the current window sequence
                                        $cur_window_seq = substr($cur_fasta_sequence, $cur_left, $window_size);


                                        #call ViennaRNA RNAfold program and report if current
                                        #sequence folds more stabily than ΔG cutoff
                                        $RNAFoldOutput = `/usr/local/ViennaRNA/Progs/RNAfold --noPS <<< "$cur_window_seq"`;

                                        #report ΔG, sequence and position if ΔG less than
                                        #cutoff
                                        $RNAFoldOutput =~ m/([-.]{1}[\d]{1,2}[\.]{1}[\d]{2})([\)]{1}$)/;
                                        $cur_G = $1;


                                        #sliding around a hairpin by 1 nucelotide will result
                                        #in multiple hairpin hits in one region; this next section of code
                                        #reports only the most stable ΔG region in a local area

                                        #for the first hairpin, simply remmember the variables
                                        #needed; store these in variables with the prefix
                                        #'pos', for 'possible'

                                        if ($cur_G < $cutoff) {
                                                            if ($pos_window_name eq "") {
                                                                                $pos_window_name = $cur_window_name;
                                                                                $pos_location = $center_position;
                                                                                $pos_G = $cur_G;
                                                                                $pos_sequence=$cur_window_seq;
                                                            }
                                                            else {
                                                                                if (($center_position - $pos_location <= $window_size) && ($cur_G < $pos_G)) {
                                                                                    $pos_window_name=$cur_window_name;
                                                                                    $pos_location = $center_position;
                                                                                    $pos_G = $cur_G;
                                                                                    $pos_sequence=$cur_window_seq;
                                                                                }
                                                            }
                                                            $count++;
                                        }
                                        #if we've made it far enough away from last hairpin to
                                        #report and there is a hairpin in que to print
                                        #else {
                                                            if  ((($center_position - $pos_location) > $window_size)&&($pos_window_name ne "")) {
                                                                                push @return, ($pos_window_name,$pos_location,length($cur_fasta_sequence),$pos_G,$pos_sequence);

                                                                                $pos_window_name="";
                                                                                $pos_location="";
                                                                                $pos_G="";
                                                                                $pos_sequence="";
                                                                                $count++;
                                                            }
                                        #}

                                        #increment left and right cutoff variables
                                        $cur_left++;
                                        $cur_right++;
                    }
                    return @return;
```

```
foreach $i (@referenceoutput) {
                        if ($mycounter == 1) {
                                        push @reference_name, $i;
                        }
                        elsif ($mycounter == 2) {
                                        push @reference_hairpin_location, $i;
                        }
                        elsif ($mycounter == 3) {
                                        push @reference_gene_length, $i;
                        }
                        elsif ($mycounter ==4) {
                                        push @reference_G,$i;
                        }
                        elsif ($mycounter == 5) {
                                        push @reference_sequence, $i;
                                        $mycounter=0;
                        }
                        else {
                                        print "Error";
                        }
                        $mycounter++;
}

$mycounter=0;

@homolog_name=();
@homolog_hairpin_location=();
@homolog_gene_length=();
@homolog_G=();
@homolog_sequence=();

foreach $j (@homologoutput) {
                        $mycounter++;
                        if ($mycounter == 1) {
                                        push @homolog_name, $j;
                        }
                        elsif ($mycounter ==2){
                                        push @homolog_hairpin_location, $j;
                        }
                        elsif ($mycounter ==3) {
                                        push @homolog_gene_length, $j;
                        }
                        elsif ($mycounter ==4) {
                                        push @homolog_G, $j;
                        }
                        elsif ($mycounter ==5) {
                                        push @homolog_sequence, $j;
                                        $mycounter=0;
                        }
}
$u=0;
$v=0;
$ref_hairpin_count=@reference_name;
$homolog_hairpin_count=@homolog_name;
#loop through all hairpins in reference and check if homolog hairpin
#is at same location
while ($u<=$ref_hairpin_count-1) {
                        #$v is a counter for the hairpins in each of the homologs
                        $v=0;
                        print ANALYSIS "$reference_name[$u]\t";
                        print ANALYSIS "$reference_gene_length[$u]\t";
                        print ANALYSIS "$reference_hairpin_location[$u]\t";
                        print ANALYSIS "$reference_G[$u]\t";
                        print ANALYSIS "$reference_sequence[$u]\t";
                        print ANALYSIS "$number_of_homologs\t";

                        $is_conserved=0;
                        $homolog_list="";
                        $no_hairpin_list="";
                                        while ($v < $homolog_hairpin_count) {
                                                        $distance=abs($reference_hairpin_location[$u]-$homolog_hairpin_location[$v]);
                                                        if ($distance <= 10 && $reference_hairpin_location[$u]!= "" && $homolog_hairpin_location[$v] != "" ) {
                                                                        if ($homolog_list !~ /$homolog_name[$v]/) {
                                                                                        $is_conserved++;
                                                                                        $homolog_list=$homolog_list . $homolog_name[$v] . ",";
                                                                        }
                                                        }
                                                        $v++;
                                        }


                                        print ANALYSIS "$is_conserved\t";
                                        if ($number_of_homologs > 0) {
                                                        $percentage=$is_conserved/$number_of_homologs;
                                        }
                                        else {
                                            $percentage=0;
                                        }
                                        printf ANALYSIS ("%.2f","$percentage");
                                        print ANALYSIS "\t";
                                        print ANALYSIS "$homolog_list\t";
                                        print ANALYSIS "$no_hairpin_list";
                                        $u++;
                                        print ANALYSIS "\n";
```

```
                                                $shifted_nucleotide=$shifted_nucleotide.substr($unshifted_nucleotide,$current_loc,3);
                                                $current_loc=$current_loc+3;
                                        }
                                }
                                $fasta[$fasta_counter]=$shifted_nucleotide;
                                $fasta_counter++;
                        }
}
$b=0;
$unshifted_nucleotide="";
$shifted_nucleotide="";


$b=0;
$unshifted_nucleotide="";
$shifted_nucleotide="";

while ($b<$other_headers_count) {
                        foreach $other (@others_array) {
                                $other =~m/(.+)\t(.+)\t(.+)\t(.+)\t(.+)\t(.+)$/;

                                if ($otherheaders[$b] eq ">$1\_$2") {
                                                $unshifted_nucleotide=$5;

                                                $current_loc=0;
                                                $shifted_nucleotide="";
                                                foreach $char (split //, $othersequence[$b]) {
                                                                if ($char eq "-") {
                                                                        $shifted_nucleotide=$shifted_nucleotide."---";
                                                                }
                                                                else {
                                                                   $shifted_nucleotide=$shifted_nucleotide.substr($unshifted_nucleotide,$current_loc,3);
                                                                   $current_loc=$current_loc+3;
                                                                }

                                                }


                                                $fasta[$fasta_counter]=$otherheaders[$b];
                                                $fasta_counter++;
                                                $fasta[$fasta_counter]=$shifted_nucleotide;
                                                $fasta_counter++;
                                }
                        }

    $b++;
}

@referenceoutput=();
@homologoutput=();
$x=1;
foreach $fasta_line (@fasta) {
                        #first two lines in array contain the referecne fasta name and
                        #sequence, #number of homologs can be calculated here based on length of
                        #fasta file

                        $number_of_homologs = (@fasta -2)/2;
                        if ($x<=2) {

                                        if ($x==1) {
                                                        $reference_fasta_name=$fasta_line;
                                        }
                                        elsif ($x==2) {
                                                        $reference_fasta_sequence=$fasta_line;
                                                        $cutoff=GCutoffCalculator($reference_fasta_sequence,30);
                                                        push @referenceoutput, hairpinfinder($reference_fasta_name,$reference_fasta_sequence,$cutoff,30);
                                        }
                        }
                        elsif ($x>2) {
                                        if ($x % 2 == 1) {
                                                        $cur_compare_fasta_name=$fasta_line;
                                        }
                                        else {
                                                        $cur_compare_fasta_sequence=$fasta_line;
                                                        $cutoff=GCutoffCalculator($cur_compare_fasta_sequence,30);
                                                        push @homologoutput, hairpinfinder($cur_compare_fasta_name,$cur_compare_fasta_sequence,$cutoff,30);
                                        }
                        }
                        $x++;
}


#both the references hairpin array and homolog hairpin array arranged
#in repeating pattern as follows - the first item is the name, second
#is location, third is length of gene, fourth is ΔG of hairpin, fifth
#is the hairpin sequence

#first, create separate arrays for each of these
$mycounter=1;

@reference_name=();
@reference_hairpin_location=();
@reference_gene_length=();
@reference_G=();
@reference_sequence=();
```

```
                                }
                        }
                        if ($need_to_remove eq "yes") {
                                pop @next_print_seq;
                                pop @next_print_other_headers;
                                $need_to_remove = "no";
                        }
                        push @next_print_other_headers, ">$next_hom_spec\_$next_hom_gene";
                        foreach $other (@others_array) {
                                $other =~m/^(.+)\t(.+)\t(.+)\t(.+)\t(.+)\t(.+)$/;
                                if ($1 eq $next_hom_spec && $2 eq $next_hom_gene) {
                                        push @next_print_seq, $6;
                                }
                        }
                }
    $i++;
}
$foc_spec = $next_foc_spec;


@alignment_file_list=();
print "@homolog_file_list\n";
print "\n\n\n";
$z=0;
#open all homologs and align using MUSCLE
foreach $homolog_file (@homolog_file_list) {
                        `muscle -quiet -fasta -in $homolog_file -out alignment$z`;
                        $alignment_file_name=$homolog_file."aln";
                        push @alignment_file_list, "alignment$z";
                        $z++;
                        print "$homolog_file\talignment$z\n";
}

#for every alignment find hairpins in query and other species, and then
#compare locations
foreach $alignment_file (@alignment_file_list) {
                        $other_headers_count=0;
                        $other_sequence_count=0;
                        $focus_species_header="";
                        my @otherheaders=();
                        my @othersequence=();
                        open(my $tempalignaa, '<', $alignment_file) or die "Could not open file '$alignment_file' $!";
                        while (my $line = <$tempalignaa>) {
                                        chomp $line;

                                        #if it's the focus species header,
                                        #print in the first position;
                                        if ($line =~ />$foc_spec/ && $foc_spec ne "") {
                                                        $line=~s/\n//g;
                                                        $focus_species_header=$line;
                                                        $at_focus="yes";
                                                        $focus_sequence="";
                                        }
                                        #if its not at the header for focus species but does have a ">"
                                        #then must be at a header for other
                                        elsif ($line =~ />/) {
                                                        $line=~s/\n//g;
                                                        push @otherheaders,$line;
                                                        $at_focus = "no";
                                                        $other_headers_count++;
                                        }
                                        #if we just were at the focus species header, then the first sequecne we encounter
                                        #must be the focus species sequence
                                        elsif ($at_focus eq "yes") {
                                                        $line =~s/\n//g;
                                                        $focus_sequence=$focus_sequence.$line;
                                        }
                                        else {
                                            $line =~s/\n//g;
                                            $othersequence[$other_headers_count-1]=$othersequence[$other_headers_count-1].$line;

                                        }
                        }


                        #this is the part of the program that takes the above
                        #amino acid sequence alignments and uses it to re-print a nuceleotide
                        #sequence
                        $fasta_counter=0;
                        @fasta=();
                        $fasta[$fasta_counter]="$focus_species_header";
                        $fasta_counter++;

                        foreach $reference (@reference_array) {
                        $reference =~m/(.+)\t(.+)\t(.+)\t(.+)\t(.+)\t(.+)$/;
                                if ($focus_species_header eq ">$1\_$2")                 {
                                                        $unshifted_nucleotide=$5;
                                        $shifted_nucleotide="";
                                                        $current_loc=0;
                                                        foreach $char (split //, $focus_sequence) {
                                                                        if ($char eq "-") {
                                                                                        $shifted_nucleotide=$shifted_nucleotide."---";
                                                                        }
                                                                        elsif ($char eq "A" || "G" || "C" || "T" || "N" || "U" || "a" || "g" || "c" || "u"
                                                                                        || "t" || "n") {
```

```
#occurs below
open (TEXTOUT,">text.out");
foreach $reference (@reference_array) {
                        open(TEMPFOCUSFASTA,">tempfocus.fa");
                        $reference =~m/(.+)\t(.+)\t(.+)\t(.+)\t(.+)\t(.+)$/;
                        $fasta_header=">lcl|$1_$2|";
                        $fasta_sequence=$6;
                        print TEMPFOCUSFASTA "$fasta_header\n$fasta_sequence\n";
                        close TEMPFOCUSFASTA;
                        $other_species_limit=$species_count*2;
                        `blastp -evalue 0.000000001 -max_hsps 1 -max_target_seqs $other_species_limit -query tempfocus.fa -db otherdb2.fa -outfmt "6 qseqid sseqid pident qlen length mismatch gapope evalue bitscore" -task blastp >>text.out`;
}
close TEXTOUT;

#sort the above text.out file based on columns 1 (species) and 2 (gene)
`sort -k 1,1 -k 2,2 ./text.out > text_sorted.out`;

open(my $test, '<', "./text_sorted.out") or die "Could not open file '$filename' $!";
$line_count=0;
$gene_count=0;

#further filtering of potential homologs occurs below
open (HOMOLOGS,">homologs.out");
while (my $row = <$test>) {
                        chomp $row;
                        $row=~m/^lcl\|(.+\..)_(.+)\|\t(.+\..)_(.+)\t(.+)\t(.+)\t(.+)\t(.+)\t(.+)\t(.+)$/;
                        $percent_coverage=$7/$6;
                        $query_species=$1;
                        $query_gene_name=$2;
                        $other_species=$3;
                        $other_gene_name=$4;
                        $percent_identity=$5;
                        $percent_coverage=$7/$6;

                        print HOMOLOGS "$query_species\t$query_gene_name\t$other_species\t$other_gene_name\t$percent_identity\t$percent_coverage\n";
}
close HOMOLOGS;

open(my $homologs, '<', "./homologs.out") or die "Could not open file '$filename' $!";
$i=0;
$homolog_file_list_counter=0;
@homolog_file_list=();
while (my $row = <$homologs>) {
                        chomp $row;
                        $row=~m/^(.+)\t(.+)\t(.+)\t(.+)\t(.+)\t(.+)$/;
                        $old_hom_spec=$next_hom_spec;
                        $old_foc_gene=$next_foc_gene;
                        $old_hom_gene=$next_hom_gene;
                        $old_foc_spec=$next_foc_spec;
                        $next_foc_spec=$1;
                        $next_foc_gene=$2;
                        $next_hom_spec=$3;
                        $next_hom_gene=$4;
                        $next_id=$5;
                        $next_cov=$6;


                        if ( $old_foc_gene ne $next_foc_gene && $i != 0) {
                                        #at a new gene so print the old homologs
                                        $n=0;
                                        while ($n<=@next_print_other_headers) {
                                            print TEMP_FASTA_AA_HOMO "$next_print_other_headers[$n]\n$next_print_seq[$n]\n";
                                            $n++;
                                        }
                                        @next_print_other_headers=();
                                        @next_print_seq=();

                                        #freshly opened fasta_aa_homo file
                                        close TEMP_FASTA_AA_HOMO;
                                        open (TEMP_FASTA_AA_HOMO,">temp_fasta_aa_homo$i.out");
                                        push @homolog_file_list, "temp_fasta_aa_homo$i.out";
                                        #$homolog_file_list[$homolog_file_list_counter]="temp_fasta_aa_homo$i.out";
                                        $homolog_file_list_counter++;
                                        print TEMP_FASTA_AA_HOMO ">$next_foc_spec\_$next_foc_gene\n";
                                        print ">$next_foc_spec\_$next_foc_gene\n";
                                        $ready_for_analysis="yes";
                                        $last_header_printed = ">$next_foc_spec\_$next_foc_gene\n";
                                        #print amino acid sequence for focus species
                                        foreach $reference (@reference_array) {
                                                        $reference =~ m/^(.+)\t(.+)\t(.+)\t(.+)\t(.+)\t(.+)$/;
                                                        if ($2 eq $next_foc_gene) {
                                                                        $ref_amino=$6;
                                                                        print TEMP_FASTA_AA_HOMO "$ref_amino\n";
                                                        }
                                        }
                        }
                        if ($next_hom_spec ne $old_hom_spec || $old_foc_gene ne $next_foc_gene) {
                            $max_score=0;
                        }
                        $cur_score = $next_id*$next_cov;
                        if ($cur_score>$max_score) {
                            $max_score=$cur_score;
                            foreach $known_headers (@next_print_other_headers) {
                                        if ($known_headers =~ /$next_hom_spec/) {
                                                        $need_to_remove="yes";
```

```
                        $cur_species=$1;
                        $cur_gene_name=$3;
                        $cur_protein_name=$4;
                        $cur_protein_id=$5;
                    }
                    else {
                                        $row=~s/\n//;
                                        $nucleotide_sequence=$nucleotide_sequence.$row;
                    }
                    $line_count++;
  }


#read in all files in 'others' directory and print an equivalent amino acid sequence fasta file,
#very similar to what was done for focus species above

opendir(DIR,"./others");
my @others= readdir(DIR);
closedir DIR;
$gene_count=0;
foreach $other (@others) {
                    open(my $othersfasta, '<', "./others/$other") or die "Could not open file '$other' $!";
                    $line_count=0;

                    while (my $row = <$othersfasta>) {
                                        chomp $row;
                                        if ($row =~ />/) {
                                                            $cur_amino_acid_sequence=translate($nucleotide_sequence);
                                                            $others_array[$gene_count]="$cur_species\t$cur_gene_name\t$cur_protein_name\t$cur_protein_id\t$nucleotide_sequence\t$cur_amino_acid_sequence\n";
                                                            $nucleotide_sequence="";
                                                            $cur_amino_acid_sequence="";
                                                            $gene_count++;
                                                    $cur_header=$row;
                                                    $cur_header=~m/^>lcl\|(.+)_cds_(.+)\[gene=(.+)\] \[protein=(.+)\] \[protein_id=(.+)\](.+)$/;
                                                    $cur_species=$1;
                                                    $cur_gene_name=$3;
                                                    $cur_protein_name=$4;
                                                    $cur_protein_id=$5;
                                        }
                                        else {
                                                    $row=~s/\n//;
                                                    $nucleotide_sequence=$nucleotide_sequence.$row;
                                        }
                                        $line_count++;
                    }
}


#make a file that is a BLASTable database
open(OTHERDBFASTA,">otherdb.fsa");

#record all unique species names in a array called unique species
@unique_species=();
$species_count=0;

foreach $other (@others_array) {
    $other =~ m/(.+)\t(.+)\t(.+)\t(.+)\t(.+)\t(.+)$/;
    $fasta_header=">lcl|$1_$2";

    if ($species_count==0) {
                    $unique_species[$species_count]=$1;
                    $species_count++;

    }
    else {
                    $already_known="no";
                    foreach $known (@unique_species) {
                                        if ($known eq $1)  {
                                                            $already_known="yes";
                                        }
                    }
                    if ($already_known eq "no") {
                                        $unique_species[$species_count]=$1;
                                        $species_count++;
                    }
    }

    $fasta_sequence=$6;
    print OTHERDBFASTA "$fasta_header\n$fasta_sequence\n";
}
close OTHERDBFASTA;

#this removes any duplicated fasta headers, which can cause errors during
#database creation (note: if errors still persist manually remove any duplicates
#which arrise due to upper/lower case non-uniquness

`../fasta_split_window/NEW/uniquify-fasta.pl otherdb.fsa >otherdb2.fsa`;

#this makes a blast db based on amino acid sequences
`makeblastdb -in otherdb2.fsa -hash_index -parse_seqids -dbtype prot -max_file_sz '2GB'`;

#go through the reference (focus) file and create temp fasta files
#that are used to blast the database created with the other sequence

#the below section roughly finds homologs based on blast searches. Since output format 6
#is not compatible with the one hit per species option, further filtering of these hits
```

```
	$amino_acid_sequence="";
	while ($a <= ($b)) {
		$codon = substr($nucleotide_sequence,$a-1,3);
		$a=$a+3;
		if ($codon eq "ATT" || $codon eq "ATC" || $codon eq "ATA") {
		    $cur_aa="I";
		}
		elsif ($codon eq "CTT" || $codon eq "CTC" || $codon eq "CTA" || $codon eq "CTG" || $codon eq "TTA" || $codon eq "TTG"){
		    $cur_aa="L";
		}
		elsif ($codon eq "GTT" || $codon eq "GTC" || $codon eq "GTA" || $codon eq "GTG") {
		    $cur_aa="V";
		}
		elsif ($codon eq "TTT" || $codon eq "TTC") {
		    $cur_aa = "F";
		}
		elsif ($codon eq "ATG") {
		    $cur_aa="M";
		}
		elsif ($codon eq "TGT" || $codon eq "TGC") {
		    $cur_aa="C";
		}
		elsif ($codon eq "GCT" || $codon eq "GCC" || $codon eq "GCA" || $codon eq "GCG") {
		    $cur_aa = "A";
		}
		elsif ($codon eq "GGT" || $codon eq "GGC" || $codon eq "GGA" || $codon eq "GGG") {
		    $cur_aa = "G";
		}
		elsif ($codon eq "CCT" || $codon eq "CCC" || $codon eq "CCA" || $codon eq "CCG") {
		    $cur_aa = "P";
		}
		elsif ($codon eq "ACT" || $codon eq "ACC" || $codon eq "ACA" || $codon eq "ACG") {
		    $cur_aa = "T";
		}
		elsif ($codon eq "TCT" || $codon eq "TCC" || $codon eq "TCA" || $codon eq "TCG" || $codon eq "AGT" || $codon eq "AGC") {
		    $cur_aa="S";
		}
		elsif ($codon eq "TAT" || $codon eq "TAC") {
		    $cur_aa = "Y";
		}
		elsif ($codon eq "TGG") {
		    $cur_aa = "W";
		}
		elsif ($codon eq "CAA" || $codon eq "CAG") {
		    $cur_aa = "Q";
		}
		elsif ($codon eq "AAT" || $codon eq "AAC") {
		    $cur_aa="N";
		}
		elsif ($codon eq "CAT" || $codon eq "CAC") {
		    $cur_aa="H";
		}
		elsif ($codon eq "GAA" || $codon eq "GAG") {
		    $cur_aa = "E";
		}
		elsif ($codon eq "GAT" || $codon eq "GAC") {
		    $cur_aa="D";
		}
		elsif ($codon eq "AAA" || $codon eq "AAG") {
		    $cur_aa="K";
		}
		elsif ($codon eq "CGT" || $codon eq "CGC" || $codon eq "CGA" || $codon eq "CGG" || $codon eq "AGA" || $codon eq "AGG") {
		    $cur_aa="R";
		}
		elsif ($codon eq "TAA" || $codon eq "TAG" || $codon eq "TGA") {
		    $cur_aa="X";
		}
		$amino_acid_sequence=$amino_acid_sequence.$cur_aa;
	}
	return $amino_acid_sequence;
}

#read in the focus file, currently hardcoded as S. coelicolor, but this can be changed
#it then parses all of the information in the fasta sequence into multiple variables such as
#the gene name, protein name, and translates the nucleotide sequence into a amino acid sequence
#using the translate subroutine described above

my $focusfasta = './focus/Streptomyces_coelicolor_nucleotide.fa';
open(my $focusfasta, '<', $focusfasta) or die "Could not open file '$filename' $!";
  $line_count=0;
  $gene_count=0;
  while (my $row = <$focusfasta>) {
        chomp $row;
	print "$row\n";
	if ($row =~ />/) {
		#first, since it's a new line need to print the last
		#entire sequence
		if ($line_count>0) {
		    $cur_amino_acid_sequence=translate($nucleotide_sequence);
		    $reference_array[$gene_count]="$cur_species\t$cur_gene_name\t$cur_protein_name\t$cur_protein_id\t$nucleotide_sequence\t$cur_amino_acid_sequence\n";
		    $nucleotide_sequence="";
		    $gene_count++;
		}
	    $cur_header=$row;
	    $cur_header=~m/^>lcl\|(.+)_cds_(.+)\[gene=(.+)\] \[protein=(.+)\] \[protein_id=(.+)\](.+)$/;
```

```
}

#this subroutine calculates a ΔG cutoff based on factor of the average
#ΔG across an mRNA
sub GCutoffCalculator {
                        #pass in an mRNA sequence and a window size
                        $mRNA_sequence = shift;
                        $window_size = shift;
                        $Gtotal=0;
                        $Gaverage=0;
                        $mRNA_sequence=~s/-//g;

                        #hard code in the slide_size to be 1; this means that each window will
                        #be shifted one nucleotide at each iteration
                        $slide_size = 1;

                        #generate sub-sequences based on window and slide sizes
                        #create variables for left and right ends of the window
                        $cur_left=0;
                        $cur_right=$cur_left+$window_size;
                        $Gcount = 0;

                        while ($cur_right <= length($mRNA_sequence)) {
                                        $center_position = $cur_left + ($window_size/2);

                                        #define the current window sequence
                                        $cur_window_seq = substr($mRNA_sequence, $cur_left, $window_size);
                                        #call ViennaRNA RNAfold program and report if current
                                        #sequence folds more stabily than ΔG cutoff
                                        $RNAFoldOutput = `/usr/local/ViennaRNA/Progs/RNAfold --noPS <<< "$cur_window_seq"`;

                                        #add ΔG to ΔG total for this position
                                        $RNAFoldOutput =~ m/([- ]{1}[\d]{1,2}[\.]{1}[\d]{2})([\)]{1}$)/;
                                        $cur_G = $1;
                                        $Gtotal=$Gtotal+$cur_G;
                                        $Gcount++;
                                        $cur_left++;
                                        $cur_right++;
                        }
                        if ($Gcount != 0) {
                                        $Gaverage=$Gtotal/$Gcount;
                                        $cutoff=$Gaverage*2.5;
                        }
                        return $cutoff;
}

#both the references hairpin array and homolog hairpin array arranged
#in repeating pattern as follows - the first item is the name, second
#is location, third is length of gene, fourth is ΔG of hairpin, fifth
#is the hairpin sequence

#first, create separate arrays for each of these
$mycounter=1;


$x=0;
$y=0;
$ref_hairpin_count=@reference_name;
$homolog_hairpin_count=@homolog_name;

#loop through all hairpins in reference and check if homolog hairpin
#is at same location
while ($x<$ref_hairpin_count) {
                        #$y is a counter for the hairpins in each of the homologs
                        $y=0;
                        print "$reference_name[$x]\t";
                        print "$reference_gene_length[$x]\t";
                        print "$reference_hairpin_location[$x]\t";
                        print "$reference_G[$x]\t";
                        print "$reference_sequence[$x]\t";
                        print "$number_of_homologs\t";
                        $is_conserved=0;
                        $homolog_list="";
                        while ($y < $homolog_hairpin_count) {
                                        $distance=abs($reference_hairpin_location[$x]-$homolog_hairpin_location[$y]);
                                        if ($distance <= 10) {
                                                        $is_conserved++;
                                                        $homolog_list=$homolog_list . $homolog_name[$y] . ",";
                                        }
                                        $y++;
                        }
                        print "$is_conserved\t";
                        $percentage=$is_conserved/$number_of_homologs;
                        printf ("%.2f","$percentage");
                        print "\t";
                        print "$homolog_list";
                        $x++;
                        print "\n";
}
# this subroutine translates a nucleotide fasta file into an
#amino acid fasta file
sub translate {
                        $nucleotide_sequence = shift;
                        $a = 1;
                        $b = length($nucleotide_sequence);
```