CHARACTERIZING G-QUADRUPLEXES, A NOVEL REGULATORY ELEMENT, IN *STREPTOMYCES* BACTERIA

TITLE: CHARACTERIZING G-QUADRUPLEXES, A NOVEL REGULATORY ELEMENT, IN *STREPTOMYCES* BACTERIA

By: SAVANNAH COLAMECO, B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science

McMaster University MASTER OF SCIENCE (2018) Hamilton, Ontario (Biology)

TITLE: Characterizing G-quadruplexes, a novel regulatory element, in *Streptomyces* bacteria

AUTHOR: Savannah Colameco, B.Sc. (McMaster University)

SUPERVISOR: Dr. Marie A. Elliot

NUMBER OF PAGES: xi,90

Abstract

Less is known about the mechanisms that govern gene regulation in GC-rich bacteria than in the more AT-rich model organisms like *Escherichia coli* and *Bacillus subtilis*. G-quadruplexes (GQs) are stable structures that form in G-rich nucleic acid sequences, and have the potential to be important regulators of gene expression – particularly in GC-rich organisms. *Streptomyces* are extremely GC-rich bacteria with the capacity to produce a vast range of antibiotic compounds. There are still many gaps in our understanding of gene regulation in these bacteria, and yet GQ sequences have never been investigated for their regulatory potential in *Streptomyces*, even though they are known to play important roles in various eukaryotic systems. Here, we performed an in-depth *in silico* analysis of the *S. venezuelae* genome and found an abundance of GQ sequences in these genomes. We discovered that these sequences were enriched in putative regulatory regions and in antibiotic biosynthetic clusters. We followed up this *in silico* analysis with reporter assays that demonstrated that GQ sequences affected gene expression in *Streptomyces*. We also took steps towards elucidating the mechanism of action for an observed increase in reporter activity in the presence of the GQ sequence. Finally, we discovered two proteins, SVEN_2656 and SVEN_3866, with the potential to interact with GQ sequences and may function in preventing adverse effects of GQ structures. These results indicate that GQs have the potential to act as important regulators of gene expression in *Streptomyces* bacteria, and future work on these systems could lead to a broader understanding of gene regulation in all bacteria, and in *Streptomyces* specifically, could be employed to stimulate antibiotic production.

Acknowledgements

There are so many people to thank who all played a part in making this experience so memorable for me. First and foremost, thank you to my supervisor, Dr. Marie Elliot your continuous support, guidance, and seemingly endless patience. I feel incredibly lucky to have had the chance to work in your lab. Thank you also to my committee members, Dr. Brian Golding and Dr. Yingfu Li, for your support throughout this process. Your insightful comments and suggestions were extremely valuable.

Thanks to Albert for being 'there' for me even though you weren't *here* most of the time. It's been a long journey, and I can't believe we've finally made it. I am so excited for us to start our life together in Ottawa.

To my family, you know what they say: you can't choose your family. While that may be true, I don't think it would have been possible to choose a group of people to make me feel more loved and supported than you have. To my parents, you've given me so much to be grateful for that if I were to list everything, it would be longer than this thesis. I guess I can look past the fact that you never took me to Disneyland. To my brother, Lucas Colameco, thanks for being my best friend. Every time I go home, it's like we're kids again because of the ridiculous shenanigans we get ourselves into.

To all past and present members of the Elliot lab, I was fortunate enough to have spent this time with an incredible group of people, and I'm sure I'm leaving the lab with a number of lifelong friends. A few special shoutouts: To my twin, Xiafei (Fei) Zhang, I grew up thinking I only had a brother, but I always wished I had a twin sister. Now I know I had one, you were just on a different continent. To my cousin, labmate, roommate, and *friend*, David Crisante, we won't always be labmates or roommates, and while we will always be cousins, I hope we will remain friends too. To Dr. Stephanie Jones, I'm sorry if I never lived up to Emma as a benchmate, but I want you to know that you are my Emma. To Danielle Sexton, I am never taking movie recommendations from you again, but luckily having the same taste in movies is not a requirement for friendship. To Dr. Matthew Moody, I don't think anyone has ever made me laugh as much as you have. I will be telling the story about the time you stuck your toe into a stranger's coffee for a long time. To Rachel Young, thank you for teaching me the ropes when I first joined the lab. You were an incredible mentor, and I'm lucky to still count you as one of my friends. Finally, to my mentees, I still can't believe Marie left me in charge of anyone, but thanks for teaching me more than I probably taught you (and sorry?)! To everyone else, I am left with extremely fond memories of all the times we spent having lunch together, going to the Phoenix, going to Pancake House, and of course, in the lab (some of which was spent doing work).

Table of Contents

List of Figures

List of Tables

List of Abbreviations

$\%_{v/v}$	number of mL per 100 mL
$\%_{w/v}$	number of grams per 100 mL
°	degree
µCi	microcurie
µg	microgram
µL	microlitre
µM	micromolar
^{32}P	phosphorus-32 isotope
A	adenine
ANOVA	analysis of variance
ATP	adenosine triphosphate
Bio	biotin
bp	base pair
C	cytosine; Celsius
CD	circular dichroism
cm	centimeter
CTP	cytidine triphosphate
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
dNTP	deoxynucleoside triphosphates
EMSA	electrophoretic mobility shift assay
G	guanine
g	gravity
GC	guanine-cytosine
GFP	green fluorescent protein
GQ	guanine-quadruplex
GTP	guanosine triphosphate
Gus	β-glucuronidase
h	hour
HQ	hybrid-quadruplex
HSD	honest significant difference
IPTG	isopropyl β-D-1-thiogalactopyranoside
IVT	*in vitro* transcription
kb	kilobase
kV	kilovolt
LB	lysogeny broth
LC	liquid chromatography
M	molar
Mb	megabase
mCi	millicurie
mg	milligram
min	minute
mL	millilitre
mM	millimolar

mRNA	messenger ribonucleic acid
MS	mannitol-soy flour; mass spectrometry
MYM	maltose-yeast extract-malt extract
N	any nucleotide
ng	nanogram
nM	nanomolar
nm	nanometer
nt	nucleotide
NTP	nucleoside triphosphates
NTS	non-template strand
OD	optical density
P	promoter
PCR	polymerase chain reaction
pM	picomolar
qPCR	quantitative polymerase chain reaction
RPKM	reads per kilobase of transcript per million mapped reads
RNA	ribonucleic acid
RNase	ribonuclease
rpm	revolutions per minute
rRNA	ribosomal ribonucleic acid
RT	reverse-transcription
s	second
SEM	standard error of the mean
T	thymine
Taq	*Thermus aquaticus*
tRNA	transfer ribonucleic acid
TS	template strand
TSB	tryptone soya broth
TSS	transcription start site
U	uracil; units
UTP	uridine triphosphate
UTR	untranslated region
UV	ultra violet
W	watt
WT	wild type
YEME	yeast extract-malt extract
YT	yeast extract-tryptone
α	alpha
β	beta
γ	gamma
Δ	deletion mutant
ΔG	Gibbs free energy

1. Introduction

1.1 Bacterial genomic GC content

Among bacteria, there is an incredible diversity in genomic GC content, ranging from 13% up to 75%, but the cause of this diversity remains unknown (1). While things like mutational biases and environmental factors (temperature, UV exposure, oxygen concentration) have been proposed, these factors fail to account for the observed variability (1). Nevertheless, variation in GC content has serious implications for the field of microbial genetics. Despite this huge diversity, most bacterial genetic studies have focussed on a few bacteria that have little variation in their genomic GC content. Consequently, there remains tremendous scope for uncovering additional regulatory mechanisms associated with more diverse and complex bacterial genomes.

Notably, some of the best-studied model species, including *Escherichia coli* and *Bacillus subtilis*, represent a relatively narrow range in GC content (50% and 43% respectively), with both being near the middle of the GC spectrum. Consequently, the genetic implications of extreme GC content have been seriously understudied to date, even though high GC bacteria are known to have major differences in many of the genetic signals regulating gene expression. By way of example, intrinsic transcription terminator sequences in *E. coli* usually consist of a GC-rich hairpin structure followed by a U-rich tail sequence. However, in GC-rich bacteria, these canonical intrinsic termination signals are rare, with other non-canonical structures serving the same function (2). Similarly, as genomic GC content rises, promoter sequences tend to diverge from the well-described *E. coli* -10 and -35 sites. In *Mycobacteria* for example, the -10 site is similar to that found in *E. coli*, but the -35 site is highly variable, making it difficult to predict promoter sequences in this organism based on what we know from classical bacterial promoters (3). Therefore, studies of more diverse bacteria have the potential to reveal alternative mechanisms of chromosome organization and genetic control.

Many medically, agriculturally, and industrially relevant organisms are bacteria with high GC content. These include, *Pseudomonas aeruginosa* (66% GC content), which is an opportunistic respiratory pathogen that affects many cystic fibrosis patients; *Mycobacterium tuberculosis* (65% GC content), which is the leading cause of death worldwide from bacterial infections, and whose extensive drug resistance is a global health concern; *Sinorhizobium* species (62-63% GC content), which are used in agriculture for their ability to fix nitrogen for plants; and members of the *Streptomyces* genus (>70% GC content), which are prolific antibiotic producers.

1.2 G-quadruplex structures

One important feature of G-rich nucleic acid sequences is their ability to form stable non-canonical structures called guanine-quadruplexes (GQs). These structures form in sequences that contain four tracts of at least three consecutive guanines, each separated by short sequences of any nucleotide (Figure 1.1a). These structures form due to the ability of guanine bases to hydrogen bond with two other guanines, interacting at 90-degree angles through Hoogsteen base pairing. When four guanines interact with each other in this way, they form a square planar structure called a guanine-quartet (Figure 1.1b). When multiple guanine-quartets stack directly on top of each other, they form a stable GQ structure (Figure 1.1c). As the carbonyl groups from all four guanines are facing inwards in this conformation, a monovalent metal cation helps to stabilize the structure by coordinating the oxygen atoms (Figure 1.1d). Due to their size, potassium ions have the highest affinity for GQ structures, while sodium and ammonium ions also have stabilizing effects (4). In contrast, lithium ions have a negligible effect on GQ stability (4). Most studies on GQ function to date focus on intramolecular GQs (GQs that form within a single DNA or RNA molecule), but intermolecular GQs have also been observed (5–8). These can be composed of up to four molecules of DNA:DNA, RNA:RNA, or DNA:RNA hybrid-quadruplexes (HQs).

1.3 Biological importance of GQs

The extremely stable nature of GQs raises many questions about the effects they may have on cellular processes, as their formation has the potential to impact DNA replication, transcription, and translation. To date, most studies into the biological functions of GQs have been carried out in eukaryotic systems (yeast and human cell lines) because of their potential roles in human diseases.

1.3.1 GQs in telomeres and cancer

Some the most well-known and best-studied GQ sequences are those found in the telomeres of eukaryotic chromosomes. Telomeres are repetitive G-rich sequences at the ends of eukaryotic chromosomes that protect the chromosomes from damage. The G-rich strand of the telomere forms a single-stranded 3’ overhang in which GQ structures can form (9). Some proposed roles for GQ structures in telomere function include enhancing the binding of telomere-binding proteins, and preventing nuclease-mediated degradation of the ends of the chromosome (10, 11). It is, however, also possible for telomere-associated GQs to adversely affect telomere function by inhibiting telomere synthesis by the enzyme telomerase (12). Notably, telomerase is not active in most cell types; however, many cancer cells activate telomerase as a way of escaping cell death. It is therefore possible for GQ-stabilizing

ligands to have cancer therapeutic potential, by stabilizing telomeric GQ structures, inhibiting telomerase activity, and ultimately promoting cancer cell death (13).

Another proposed application for GQ-stabilizing ligands in cancer therapies is in the control of cancer-promoting gene (oncogene) expression. Genome mining of sequences with the potential to form GQs has revealed that these sequences are enriched near promoter regions in eukaryotes, including the promoters of many oncogenes (14). One of the best-studied oncogenes having a promoter GQ is the *c-myc* gene. This gene is upregulated in many cancers, but when cells are cultured with GQ-stabilizing ligands, *c-myc* expression decreases and cell survival is reduced (15). Thus, GQ-stabilizing ligands have the potential to be used as cancer therapeutics by reducing oncogene expression.

1.3.2 GQs in DNA replication

Beyond replication of telomere sequences, GQs can broadly influence DNA replication, and it appears that there are helicases encoded by many organisms that function to resolve GQ sequences. Recent work in the yeast *Saccharomyces cerevisiae* revealed that loss of the Pif1 helicase interfered with the fidelity of DNA replication as shown by the rapid expansion of GQ-forming tandem repeats (16). In humans, disrupting the function of GQ-resolving helicases has been associated with a range of diseases including Fanconi anaemia (FANCJ helicase) (17), Werner syndrome (WRN helicase) (18, 19), and Bloom syndrome (BLM helicase) (19). In addition to impeding DNA replication, there is also mounting evidence that GQ sequences can promote replication initiation, as many eukaryotic origins of replication coincide with G-rich sequences that have the potential to form GQs (20).

1.3.3 GQs in gene expression

GQ sequences can have a variety of effects on gene expression, and these have been attributed to both DNA and RNA GQs. At the transcriptional level, DNA GQs can affect gene expression either positively or negatively, possibly by assisting with opening the DNA duplex, or causing RNA polymerase stalling along the transcript (21). At the translational level, RNA GQs can inhibit both ribosome loading (22–24) and ribosome progression along the transcript (25). These effects have been extensively studied in human and yeast cells, with far fewer studies in bacteria. Bacterial studies to date have largely focused on *E. coli* (7, 26, 27), with a few exceptions. The only work done in GC-rich organisms, including *Mycobacterium* and *Deinococcus*, focused on the possibility of GQ sequences affecting promoter function (28, 29). Therefore, it is possible that other mechanisms of GQ-mediated gene expression could be uncovered by conducting a broader analysis of GQs in GC-rich bacteria whose genomes contain many GQ sequences.

1.4 Detection of GQs

Given the broad range of biological functions that are being ascribed to GQs, there is considerable interest in developing experimental approaches to detect these nucleic acid structures. These methods range from *in silico* approaches for identifying potential GQ-forming sequences in genomes, to biophysical experiments aimed at characterizing their structure, and techniques that detect their formation both *in vitro* and *in vivo*.

1.4.1 *in silico* prediction of GQ sequences

Several algorithms have been developed to find predicted GQ sequences based on the appearance of sequence motifs such as G_3N_{1-7} G_3N_{1-7} G_3N_{1-7} G_3. These programs include G4Hunter (30), QGRS Mapper (31), and QuadBase (32), and they can be used to search the genomes of a wide variety of species for putative GQ sequences. In this way, over 300,000 putative GQ sequences were found in the human genome (33). Upon analyzing the locations of these GQ sequences, it was found that they were enriched in regulatory regions, with GQ sequences being found within 1 kb upstream of 50% of genes in the human genome (34). A similar study that evaluated the number of GQ sequences in 18 bacterial species revealed that GQ sequences were also prevalent in the putative regulatory regions, defined as within 200 bp upstream of coding sequences, of bacteria (35).

Experimentally, GQ sequences have considerably more flexibility than ascribed by the (G_3N_{1-7}) repeating motif. For example, GQs can form with larger loop regions (up to 30 nt in some cases), with bulges in the structure such that the G-tracts can be interrupted by one or two nucleotides, and with as few as two G-quartets (G_2N_{1-7}…) (36). Challenges remain in predicting which GQ sequences will actually form GQ structures. A more recent GQ-finding tool uses a scoring system to predict whether a particular GQ sequence will form a GQ based on experimental data sets and allows for imperfections in the structure (bulges and longer loop sequence) (37). However, the algorithm only found approximately 60% of experimentally validated GQ structures, suggesting there is still much that we do not understand about the flexibility of GQ structures.

1.4.2 Biophysical methods of GQ characterization

To determine whether specific sequences can form a GQ structure, several biophysical techniques can be used, including circular dichroism (CD), UV melting, and the binding of fluorescent probes (38). All of these methods involve using short oligonucleotides, and rely on the unique properties of GQ structures compared with other DNA conformations. CD is one of the most commonly used techniques, as it

provides information both about the conformation of a GQ structure – for example, GQs with parallel and anti-parallel conformations have different characteristic spectra (39) – and for determining the conditions required for GQ folding – for example, comparing the effects of different ions at varying concentrations on GQ folding (40). The limitation of CD, as well as UV melting and fluorescent probe binding, however, is that these methods can only be used on short oligonucleotides They are therefore useful for providing information on which sequences can form GQ structures, but they do not provide insight into which ones will form in particular genomic contexts or in complex RNA molecules.

One point of note is that GQ structures are slow to form, and therefore the extent of their biological relevance has been questioned (11). It is currently not known if there would be time for them to form during the opening of the DNA duplex during replication and transcription. However, recent evidence has suggested that the presence of nascent mRNA, as well as confined space (such as being in the exit channel of a polymerase) can both facilitate GQ folding (41, 42). In an attempt to address this question of biological relevance, techniques have been developed for studying GQ sequences in more physiologically relevant environments – such as in their native genomic context – alongside other *in vivo* methods.

1.4.3 The use of GQ-specific antibodies

In an effort to detect GQs in more biologically meaningful contexts (*e.g.* in genomic DNA), several structure-specific antibodies have been developed that preferentially bind GQ DNA over duplex or single-stranded DNA (43–45). These antibodies have been used to immunoprecipitate GQ-containing genomic DNA, followed by deep sequencing to map the locations of GQ structures throughout the genome (46). This work helped to confirm that GQ structures can form in genomic DNA, and in doing so, provided much-needed experimental support for the proposal that GQs can indeed form *in vivo*. In addition to pull-down experiments, GQ-specific antibodies have also been used in immunofluorescence experiments to detect GQ structures within cells; these experiments have ultimately confirmed that GQ structures form in telomeric DNA (43).

To date, the antibodies employed in these immunoprecipitation/immunofluorescence studies have been specific for select GQ conformations. Consequently, not all GQ conformations are captured. Thus far, studies using GQ-specific antibodies have reported far fewer GQ structures than what would be expected based on *in silico* predictions, although whether this is because of antibody specificity, or because these structures are simply not as common in the DNA as would be expected based on simple sequence predictions, is not yet known.

1.4.4 Polymerase stop assays

Recent years have led to the development of more global methods of assessing GQ formation, with the polymerase stop assay having been particularly powerful (47). This assay can be used to determine whether a GQ structure is able to form at any particular genetic locus, and in theory, could provide information on all possible GQ structures within a given genome. This assay is based on the idea that GQ structures can impede polymerase progression, and therefore the existence of a GQ structure can lead to polymerase pausing or stopping. When combined with next generation sequencing, this method can be used as a high-throughput technique to identify GQ structures in genomic DNA (36) and in cellular RNA (48). When this method was applied to human genomic DNA, 73% of predicted GQ sequences corresponded to an observed GQ, but surprisingly, more than half of the polymerase stops were not associated with predicted, conventional GQs, but instead with GQ sequences that had long loop regions and bulges in their structure (36). Using this technique to evaluate the human transcriptome, there was far less overlap between the observed and predicted GQs than there was for DNA GQs (48). This indicates that the conditions required for RNA GQ formation are still poorly understood. While this method provides useful insights into the accuracy of *in silico* predictions and the diverse sequences that can form GQ structures, they are still inherently *in vitro* techniques (the DNA/RNA being analyzed had been folded in vitro), and thus the question remains as to whether such structures form *in vivo*.

An interesting modification to the polymerase stop assay has recently been developed. DMS-seq is used for probing *in vivo* RNA structure (49), and can be coupled with the polymerase stop assay to identify RNA-associated GQ structures. This DMS-seq technique was used on mouse and yeast cells, and in these organisms, no RNAs were found to meet the cut-off needed to be considered official GQs. This suggests that RNA GQs are unfolded in eukaryotic cells (50). When the same experiment was performed on *E. coli*, RNA GQs were tolerated, but it was determined that these sequences were generally depleted in the genomes of *E. coli* as well as the two other species examined, *Pseudomonas putida* and *Synechococcus* sp. WH8102. This paper is somewhat controversial, in that the results are inconsistent with other studies that have found specific functions for RNA GQs in regulating gene expression *in vivo* (see section 1.3.3 and chapter 4 introduction). There is still a considerable body of evidence supporting an *in vivo* role for RNA GQs, but given these conflicting results, there remain many unknowns regarding their formation and regulatory roles *in vivo*.

1.5 *Streptomyces* as model organisms

To date, GQ research has focused heavily on eukaryotic systems, with some work being done in *E. coli*. Shifting this focus to more GC-rich bacteria may help to uncover novel mechanisms of GQ-mediated gene regulation, due to the large number of predicted GQ sequences in high GC genomes.

Streptomyces is a genus of Gram-positive soil-dwelling bacteria that are studied for their complex life cycle and their biosynthetic capabilities. When *Streptomyces* were first isolated from the soil, they were thought to be fungi because of their filamentous, multicellular life cycle (Figure 1.2). This unusual and complex life cycle starts with the germination of dormant spores, with the subsequent germ tubes gorwing to form a network of branching, vegetative mycelia. In response to some – as yet unknown – environmental cue, the colony transitions to its reproductive phase of growth, which begins with the raising of aerial hyphae, which go on to differentiate into chains of dormant spores (51).

Industrial interest in this genus began in the 1940's with the discovery of the antibiotics streptothricin and streptomycin (52). Since then, countless other antibiotic, anti-cancer, anti-fungal, and anti-parasitic agents have been discovered in these bacteria. Currently, over two-thirds of clinically used antibiotics are produced by *Streptomyces* species, earning them the title of biosynthetic powerhouses. Despite this incredible biosynthetic diversity, genome mining has revealed that they have far greater biosynthetic potential than anticipated. While most *Streptomyces* species produce a few known compounds, most genomes contain 20-50 biosynthetic clusters with potentially novel products; the majority of these clusters are transcriptionally silent in laboratory environments. These observations suggest that this genus is still far from being 'tapped out' with respect to its metabolic potential, despite the declining number of new antibiotic compounds being discovered in *Streptomyces* (52, 53). Part of the problem is that there is still much that we do not understand about gene regulation in these complex bacteria. For example, heterologous expression or modifying promoter sequences within 'cryptic' (not-normally expressed) clusters has only been successful in stimulating the expression of a handful of clusters (54), indicating that the regulation of these clusters is more complex than we realize. Without a better understanding of gene regulation in *Streptomyces* it will be impossible to fully exploit their biosynthetic potential.

1.6 Aims of this study

In an attempt to fill some of these knowledge gaps in *Streptomyces* gene regulation, this study aims to investigate the prevalence and function of GQ sequences in the genomes of *Streptomyces* species, and

to elucidate any roles these may play in gene regulation. Given the high GC content of *Streptomyces* genomes (>70%), GQ sequences are expected to be abundant and have the potential to play important regulatory roles in the cell. GQ sequences have never been studied in *Streptomyces,* and this study may provide new insights into gene regulation in these bacteria that could be applied to understanding antibiotic biosynthesis. Our understanding of GQ function in gene regulation in the streptomycetes could also be applied to other bacterial species since so little is known about the functions of GQ structures in bacteria. The main goals of this study are to:

1) Identify GQ sequences to determine their abundance and their relative position and orientation in the chromosome
2) Explore/probe regulatory functions of GQ sequences in *Streptomyces*, with a focus on their transcriptional effects, and their interaction with GQ-binding proteins

1.7 Figures

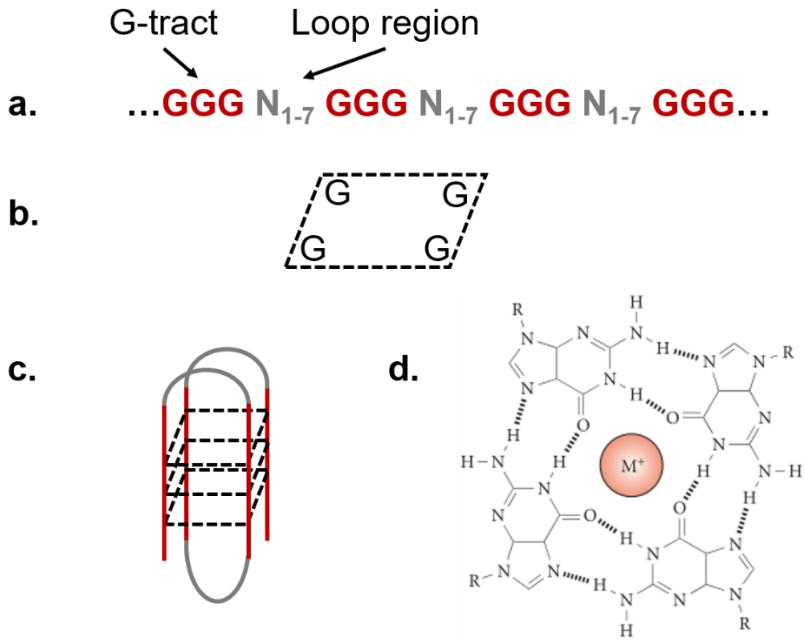

Figure 1.1: G-quadruplex structure. a. GQ-forming sequence with four G-tracts and three short loop regions of any nucleotide (N). **b.** A G-quartet in which each guanine hydrogen bonds with two other guanines in a square planar conformation. **c.** A GQ structure composed of four G-quartets with the G-tracts highlighted in red and the loop regions highlighted in grey. **d.** The structure of a G-quartet in which the guanines interact through Hoogsteen base pairing and a metal cation in the center stabilizes the structure. Modified from Bochman et al. 2012.

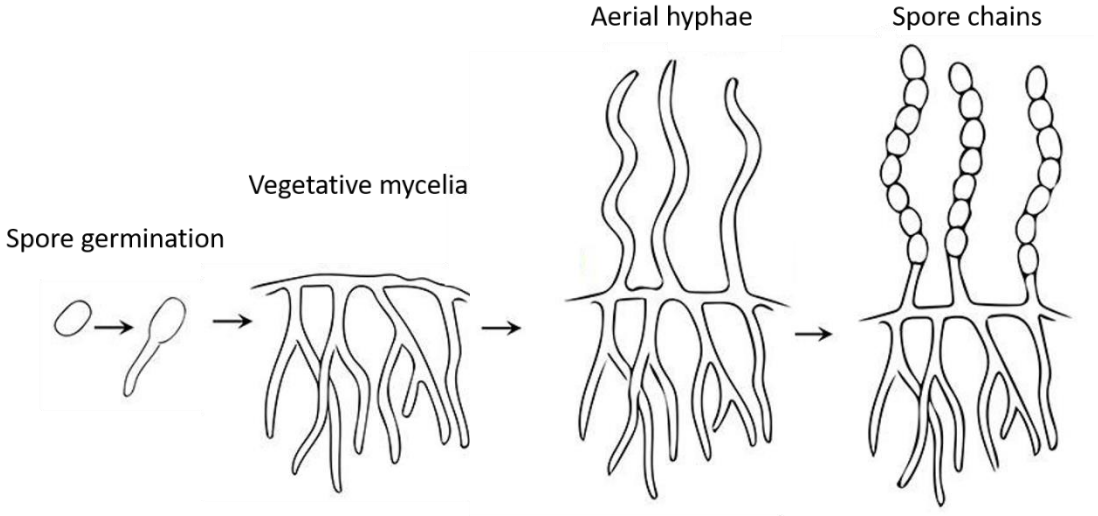

Figure 1.2: Classical *Streptomyces* life cycle. Life cycle begins with spore germination, followed by the growth of a network of branching vegetative mycelia into the growth medium. Growth transitions to aerial development with the formation of aerial hyphae, followed by the formation of spore chains. Modified from Jones et. al. 2017.

2. Materials and Methods

2.1 *In silico* analysis of sequence data

2.1.1 Searching for GQ sequences

GQ search data was generated using either a custom Python script (Appendix A and B) or a PowerShell script (developed by S. Miller, unpublished) depending on the type of information that was required. For simply identifying the number of GQ sequences in a given sequence file, the script in Appendix A was used. If more information on the locations of GQ sequences was required, either the script in Appendix B or the script from S. Miller was used.

2.1.2 Searching for GQ sequences in shuffled genomic DNA sequences

To iteratively shuffle and search a genome for GQ sequences, the script in Appendix C was used. This script made use of the program uShuffle (55), which was specifically designed for shuffling biological sequences, and provided the option of maintaining the frequency of duplet, triplet, or, more broadly, *k*-let nucleotide sequences. For our purposes, we used *k*-let size = 3 and shuffled the sequence 1,000 times (n = 1,000).

2.1.3 Identifying untranslated region (UTR) GQ sequences

Untranslated region (UTR) GQ sequences (those within transcribed but untranslated regions, often found at the 5' and 3' ends of coding sequences) were identified using a custom Python script (Appendix D). The script made use of the output data from the Rockhopper RNA-seq analysis tool, to generate a list of the genomic locations of all UTRs defined as: predicted transcription start to translation start (5' UTR) and translation stop to predicted transcription stop (3' UTR). Once all UTR positions were defined, the script was used to look for overlap between the genomic locations of GQ sequences and UTRs, and then generated a list of all UTR GQ sequences. The data were then curated manually.

2.1.4 Identifying GQ sequences in proximity to transcription start sites

Locations of transcription start sites (TSSs) were identified from the data files obtained from M. Bush and M. Buttner (unpublished) using a custom Python script that went through the file line by line to determine the genomic locations of regions where the number of reads went from less than or equal to 2 to greater than or equal to 20 in the next line, indicating the location of a TSS (Appendix E). Eight data files in total were analyzed in this way, one for each strand, for each of four time points. These files were then run through another custom Python script that determined those sequences that were within 100 nt of a predicted GQ start, where both input data files were sorted by genomic location (smallest to

largest) before being run through the script (Appendix F). The output files were then compiled and duplicates were removed.

2.2 Bacterial strains, plasmids, and culturing conditions

A list of all bacterial strains used in this study can be found in Table 2.1, while the details of all plasmids can be found in Table 2.2. Unless otherwise specified, all liquid cultures were grown shaking at 200 rpm. *Streptomyces venezuelae* was grown in maltose-yeast extract-malt extract (MYM) medium at 30°C, while *S. coelicolor* was grown in a 1:1 mixture of yeast extract-malt extract/tryptone soya broth (YEME/TSB) medium at 30°C with a metal spring. *E. coli* cultures were grown in Lysogeny Broth (LB) medium at 37°C.

When required, media were supplemented with the following antibiotics at the indicated final concentrations (µg/mL): ampicillin (100), apramycin (50), chloramphenicol (25), hygromycin B (50), kanamycin (50 for *E. coli*, 25 for *Streptomyces*), nalidixic acid (20), spectinomycin (100), streptomycin (5).

Spore stocks of *Streptomyces* strains were made by streaking out a lawn of the strain onto MYM agar (*S. venezuelae*) or MS agar overlaid with a cellophane disc (*S. coelicolor*) without selection. After 2-3 days (*S. venezuelae*) or 4-5 days (*S. coelicolor*) of incubation at 30°C, all biomass was scraped into 10 mL of sterile water, then sonicated to dislodge spores from the mycelium. The entire suspension was then passed through a 10 mL syringe with a cotton filter. The resulting solution was centrifuged for 5 min at 2,200 x*g* at 4°C. The spores were then resuspended in an equal volume of 40%$_{v/v}$ glycerol and stored at -20°C. *E. coli* stocks were made by mixing equal parts of an overnight culture and 40%$_{v/v}$ glycerol, before being stored at -80°C.

2.3 Oligonucleotides

All oligonucleotides used in this work are listed in Table 2.3.

2.4 Molecular cloning

2.4.1 Polymerase Chain Reaction (PCR)

All diagnostic PCRs were done in using 20 µL reactions volumes with either NEB Quick-Load 2× Taq Master Mix or Taq DNA polymerase in the reaction, using the cycling conditions outlined in Table 2.4. For high fidelity PCRs, Phusion DNA polymerase was used as outlined in Table 2.4. As required, PCR products were purified using PureLink™ PCR Purification Kit (Invitrogen) or visualized on an agarose gel, excised from the gel and then purified using a gel extraction kit (New England Biolabs Monarch® DNA Gel Extraction Kit or Omega Bio-tek EZNA® Gel Extraction Kit).

2.4.2 Digestion and dephosphorylation/phosphorylation of DNA

DNA digestions were performed in 50 µL reactions with 1-5 U of *Bam*HI, *Bgl*II, *Eco*RI, *Eco*RV, *Kpn*I, *Spe*I, *Xba*I (New England Biolabs enzymes) in the appropriate manufacturer-specified buffer. Reactions were incubated at 37°C for 1-2 h. When required, DNA was dephosphorylated after being digested by adding 20 U Calf Intestinal Phosphatase (New England Biolabs) directly to digestion reactions and incubating at 37°C for 1 h. When necessary, DNA was phosphorylated in 20 µL reaction volumes using 10 U T4 DNA Polynucleotide Kinase (New England Biolabs), T4 DNA Polynucleotide Kinase reaction buffer, and 1 mM ATP. Reactions were incubated for 30 min at 37°C. Following digestion, dephosphorylation, or phosphorylation, DNA was either purified directly using PureLink™ PCR Purification Kit (Invitrogen), or visualized on an agarose gel, excised and then purified using a gel extraction kit (New England Biolabs Monarch® DNA Gel Extraction Kit or Omega Bio-tek EZNA® Gel Extraction Kit).

2.4.3 DNA Ligations

Prior to DNA ligation, DNA was visualized on an agarose gel to determine the relative concentrations of insert and vector DNA, such that a 3:1 (insert:vector) molar ratio could be added to the reaction. DNA ligations were performed using the Rapid DNA Ligation Kit (Sigma Aldrich) as per the manufacturer's instructions. Briefly, Buffer 1, Buffer 2, and DNA ligase were added to the appropriate amounts of insert and vector DNA. Reactions were incubated at room temperature for 5 min, before being transformed by heat shocking into Subcloning Efficiency™ DH5α™ *E. coli* cells (Invitrogen).

2.4.4 Transfer of DNA into *E. coli*

For heat shocking, DNA was added to a 50 µL aliquot of Subcloning Efficiency™ DH5α™ *E. coli* cells (Invitrogen), and the mixture was incubated on ice for 30 min. The cell-DNA mixture was then transferred to a 37°C water bath for 30 s, then returned to ice for 2 min. One milliliter of LB medium was added to the tube, and cells were shaken at 37°C at 200 rpm for 1 h, after which they were plated on selective media and incubated at 37°C overnight.

For electroporation, DNA was added to a 50 µL aliquot of electrocompetent *E. coli* cells. The mixture was then transferred to a cold 0.2 cm electroporation cuvette, which was then placed in the BioRad MicroPulser™ Electroporation Apparatus and pulsed at on the EC2 setting (2.5 kV). One milliliter of cold LB was added to the cuvette and the transformed cells were then transferred into a clean 1.5 mL Eppendorf tube. Cells were shaken at 37°C at 200 rpm for 1 h, before being plated on selective media and incubated at 37°C overnight.

2.4.5 Isolating plasmid DNA from *E. coli*

Plasmid DNA was extracted from overnight *E. coli* cultures using PureLink™ Quick Plasmid Mini Prep Kit (Invitrogen) as per the manufacturer's instructions.

2.4.6 Transferring DNA into *Streptomyces*

DNA was introduced into *Streptomyces* by conjugation using the *E. coli* ET 12567/pUZ8002-containing conjugative strain. *E. coli* cells were grown to an OD_{600} of approximately 0.4, then washed three times in fresh LB liquid medium. The resulting cell suspension was then collected and resuspended in 500 µL fresh LB liquid. Separately, approximately 1×10^8 *Streptomyces* spores were added to 500 µL yeast extract-tryptone (YT) broth, then either heat shocked at 55°C for 5 min, then incubated on ice (*S. coelicolor*) or placed directly on ice (*S. venezuelae*). The washed *E. coli* cell suspension was then mixed with the *Streptomyces* spore suspension and plated on MS agar. After 7-13 h (*S. venezuelae*) or 16-20 h (*S. coelicolor*), plates were overlaid with nalidixic acid and the plasmid-specific selective antibiotic.

2.4.7 Isolating genomic DNA from *Streptomyces*

Genomic DNA was isolated from overnight cultures of *Streptomyces* using the Bacterial Genomic DNA Isolation Kit (Norgen) as per the manufacturer's instructions for Gram-positive bacteria.

2.4.8 Deleting the gene encoding Rho

The gene encoding Rho, *sven_5009*, was deleted from *S. venezuelae* using the previously described PCR-targeted gene replacement protocol for *Streptomyces* gene deletions (56). The deletion cosmid, in which *sven_5009* was replaced with an apramycin-resistance cassette, was obtained from R.J. St-Onge, then introduced into *E. coli* ET 12567/pUZ8002 for conjugative transfer into *S. venezuelae*. Once the cosmid was in *S. venezuelae*, ex-conjugants were screened for apramycin-resistance and kanamycin-sensitivity, which would indicate that the chromosomal *sven_5009* had been successfully replaced by the apramycin-resistance cassette and that the rest of the cosmid had been lost. Colonies with the desired resistance profile were then checked by PCR looking for a lack of product with the SV5009F/SV5009intR primer combination, as well as appearance of a product of the appropriate size with the SV5009F/SV5009KOR primer combination to ensure that the gene was deleted and replaced with the apramycin resistance cassette.

2.5 Reporter assays

2.5.1 β-glucuronidase (Gus) reporter assays

β-glucuronidase (Gus) reporter assays were performed in *Streptomyces* as described by Myronovskyi et al (57). Briefly, liquid cultures of the strains to be assayed were grown for 16 h (*S. venezuelae*) or 24 h (*S. coelicolor*), after which 1 mL of culture was spun down and used for the assay. The cells were

resuspended in lysis buffer (50 mM phosphate buffer [pH 7.0], 0.27%$_{v/v}$ β-mercaptoethanol, 0.1%$_{v/v}$ Triton X-100, 1 mg/ml lysozyme) and the cell mixture was incubated at 37°C for 30 min. After the incubation, the cell lysate was then centrifuged to remove debris, and the supernatant was used in the assay. Thirty to fifty microlitres of lysate were used in 200 µL reactions in a 96 well plate. The Gus substrate, 4-nitrophenyl-β-D-glucuronide, was added to the reactions to a final concentration of 600 µg/mL. Gus activity was determined by measuring the absorbance at 420 nm of the reactions and normalized to OD_{600} (*S. venezuelae*) or dry weight (*S. coelicolor* and *S. venezuelae* Δ*sven_5009* strains).

2.5.2 GFP assays

The GFP construct pUC19 + P*em7*-120 nt-GQ-*gfp* was synthesized by GenScript. All other iterations of the construct were generated by digesting and religating the vector with appropriate restriction enzymes (Figure 2.1). Assays were conducted in *E. coli* DH5α cells which were grown overnight in LB liquid supplemented with ampicillin. The cells were then washed with M9 medium and subcultured into fresh M9 medium with ampicillin. After 5 h of incubation, cultures were transferred to a 96 well plate and fluorescence intensity was measured with excitation/emission wavelengths of 485 nm/514 nm. Fluorescence was normalized to the OD_{600} of the cultures.

2.6 Circular dichroism (CD)

Samples were prepared for CD by heating a 1 µM oligonucleotide solution (100 mM KCl, 50 mM Tris-HCl, pH 7.4) to 95°C for 5 min, followed by slow cooling to room temperature. Spectra were measured at 1 nm wavelength intervals between 220 nm and 320 nm in a 1 cm quartz cuvette. A DNA-free negative control was treated in the same way as the samples, and the associated values were subtracted from all DNA-containing samples.

2.7 RNA techniques

2.7.1 Extraction of total RNA

Total RNA was extracted from liquid-grown *S. venezuelae* cultures. 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%$_{v/v}$ β-mercaptoethanol), after which they were mixed with an equal volume of phenol-chloroform-isoamyl alcohol solution (50:50:1). Aqueous and organic phases were separated by centrifugation, and the aqueous phase extracted and treated twice more with phenol-chloroform-isoamyl alcohol solution. RNA was precipitated at -20°C in a 10:1 isopropanol-sodium acetate solution (3 M sodium acetate, pH 6). The RNA was then pelleted and washed with 70%$_{v/v}$ ethanol before being resuspended in nuclease-free water. Contaminating DNA was

removed from the RNA preparation using 20 U TURBO™ DNase (Invitrogen), and incubating at 37°C for 1 h. The DNase was then removed using a phenol-chloroform-isoamyl alcohol extraction, followed by RNA precipitation as described above. RNA purity and concentration were assessed by NanoDrop™ and RNA quality was assessed by agarose gel electrophoresis. RNA was confirmed to be DNA-free by PCR check, where DNA-free was defined as lack of a band following electrophoresis of reactions after a 35 cycle PCR.

2.7.2 RT-qPCR

RNA was reverse-transcribed either with gene-specific reverse primers (SuperScript RT III; Invitrogen) or with random oligonucleotides (LunaScript™ RT; New England Biolabs) according to the manufacturer's instructions (see Table 2.3 for oligonucleotide sequence information). qPCR was performed using either PerfeCTa™ SYBR Green Super Mix (Quantabio) or Luna® Universal qPCR Master Mix (New England Biolabs), with the BioRad CFX96™ Real-Time PCR machine. Data were normalized to either 5S rRNA (RNA stability) or *rpoB* (qPCR).

2.7.3 *In vitro* transcription

In vitro transcription was performed either with *E. coli* RNA polymerase holoenzyme (New England Biolabs; *in vitro* transcription experiment) or with T7 RNA polymerase (MEGAshortscript™ T7 Transcription Kit; Ambion; synthesizing RNA for *in vitro* RNA stability assays). For *E. coli* RNA polymerase, template was generated by PCR using a primer that contained the *lysC* promoter sequence at its 5' end, as per St-Onge et al (58). Approximately 20 ng template DNA was added to each 10 µL reaction along with 3 µM each ATP, GTP, and UTP, and 1 µL CTP-[α-^{32}P] (Perkin Elmer, 10 mCi/mL), *E. coli* RNA polymerase reaction buffer, and 1 U *E. coli* RNA polymerase. Reactions were incubated at 37°C for 30 min, before being separated in the BioRad SequiGen™ GT Nucleic Acid Electrophoresis Cell with a SequaGel – UreaGel 6 (National Diagnostics) urea-polyacrylamide gel. The gel was run for 90 min at 60 W, and the gel was visualized by using a phosphorimager.

For T7 RNA polymerase, template DNA was amplified by PCR using a forward primer with a T7 promoter sequence at its 5' end, after which it was purified by gel extraction. The quantity and purity of template DNA was assessed using a NanoDrop™ spectrophotometer. In a 40 µL reaction, 50 nM template was added together with 7.5 nM each NTP, T7 reaction buffer, and 4 U of T7 enzyme. Reactions were incubated at 37°C for 4 h, at which point 8 U TURBO™ DNase was added to the reaction, which was then incubated at 37°C for another 1 h. RNA was then purified by gel extraction. Briefly, products were separated using the BioRad SequiGen™ GT Nucleic Acid Electrophoresis Cell with a SequaGel – UreaGel

6 (National Diagnostics) urea-polyacrylamide gel. The gel was run for 90 min at 60 W, at which point the gel fragment containing the product of interest was excised by UV shadowing with a Fluor-coated TLC plate (Invitrogen). The gel fragment was transferred to a clean 1.5 mL tube, and was then soaked three times for 30 min in 500 µL crush-soak buffer (200 mM NaCl, 10 mM Tris, 1 mM EDTA, pH 8). RNA was recovered by precipitation in isopropanol-sodium acetate solution as described above for RNA extractions. Once resuspended in nuclease-free water, RNA quantity and purity were verified using NanoDrop™ spectrophotometry.

2.7.4 RNA stability assays

RNA stability assays were performed both *in vitro* and *in vivo*. For *in vitro* RNA stability assays, RNA was *in vitro* transcribed and gel purified. The RNA was radiolabeled at its 3' end using T4 RNA ligase (New England Biolabs) in a 30 µL reaction containing T4 RNA ligase buffer, 1 mM ATP, 10% DMSO, ~1 pmol RNA, 10 U T4 RNA ligase, and 2 µL 5'[^{32}P]pCp (10 mCi/mL). Labeling reactions were incubated overnight at 16°C. Radiolabeled RNA was purified using NucAway™ spin column (Ambion). Lysates from wild type and Δ*rnj S. venezuelae* 50 mL liquid-grown cultures after 16 h of growth were obtained by sonication. Cells were collected by centrifugation at 2,200 ×*g* for 10 min, then resuspended in 10 mL lysis buffer (10 mM Tris, 0.1 M NaCl, 5%$_{w/v}$ glycerol, pH 7). Cell suspension was then lysed by sonication (Branson Sonifier Cell Disruptor350, 10× 30 s pulses at 50% duty cycle, output control = 4). A Bradford assay (59) was performed on the lysates to ensure equal amounts of protein were being added to the reactions. Fifteen microliters of end-labeled, purified RNA and lysate (6 µg protein) were mixed, and reactions were flash frozen in a liquid nitrogen bath at time-points ranging from 0-90 min. Reactions were then separated on a 9 cm SeqaGel – UreaGel 6 urea-polyacrylamide gel and visualized using a phosphorimager.

In vivo RNA stability assays were performed on 100 mL liquid-grown *S. venezuelae* cultures after 16 h of growth exposed to the RNA polymerase-targeting antibiotic rifampicin (500 µg/mL), which was expected to stop the initiation of RNA synthesis. Samples were taken at 0 min (before rifampicin addition), and every 2 min, up to 10 min after the addition of rifampicin. At each time-point, a 13.5 mL sample of the culture was removed from the total culture volume and was mixed with 1.5 mL of a cold 9:1 ethanol-phenol solution before being incubated on ice for the remainder of the time course. Cells were collected by centrifugation at 2,200 ×*g* for 5 min, and were then processed using the RNA extraction protocol described above. Relative quantities of different RNAs of interest were determined by RT-qPCR.

2.8 Biotin pulldowns

Biotin pulldowns were performed using 5' biotinylated oligonucleotides obtained from IDT® Integrated DNA Technologies. These biotinylated probes were folded in a 500 µL, 0.5 µM solution by heating to 95°C for 5 min, followed by slow cooling to room temperature in GQ folding buffer (100 mM KCl, 50 mM Tris-HCl pH 7.4). One milligram of streptavidin magnetic beads (Sigma Aldrich) were washed three times in 500 µL of GQ folding buffer. The washed beads were then mixed with 300 µL of folded DNA probes to allow the DNA to anneal to the beads by slow shaking at room temperature for 30 min. The DNA-bead mixture was washed three times with 500 µL GQ folding buffer before being mixed with 5 mL cell lysate and incubated shaking at 200 rpm at 30°C for 1 h. The cell lysate was obtained through sonication (Branson Sonifier Cell Disruptor350, 12x 40 s pulses at 50% duty cycle, output control = 4) of 1 L of 18 h liquid-grown *S. venezuelae* cells resuspended in 20 mL lysis buffer (10 mM Tris, 0.1 M NaCl, 5%$_{v/v}$ glycerol, pH 7). Protein-bound beads were recovered using a magnet, and were then washed with 200-400 µL elution solution (20 mM Tris, 1 mM EDTA, 10% glycerol, 75-500 mM NaCl) containing increasing concentrations of salt to elute proteins from the DNA. Eluted proteins were separated on a 10% polyacrylamide gel, and were visualized by silver staining.

Silver staining was performed by first soaking gels in MilliQ water for 30 min shaking at room temperature. Water was replaced with 300 mL fixing solution (50%$_{v/v}$ ethanol, 10%$_{v/v}$ glacial acetic acid) and incubated shaking for 10 min, followed by application of a 300 mL rinse solution (50%$_{v/v}$ ethanol) for 5 min. In turn, the rinse solution was replaced with 300 mL sensitizing solution (0.02%$_{w/v}$ sodium thiosulphate), which was applied for 2 min. To wash the gel, it was incubated in MilliQ water for 2 min, followed by 300 mL cold staining solution (0.1%$_{w/v}$ $AgNO_3$). The gel was then incubated in this cold staining solution for at least 20 min while shaking at 4°C. The gel was then quickly rinsed twice with MilliQ water and developed using 500 mL developer solution (2%$_{w/v}$ sodium carbonate, 0.015%$_{w/v}$ formaldehyde, and 4%$_{v/v}$ sensitizer solution). The developing reaction was allowed to continue until the desired resolution was achieved (no more than 10 min), after which the gel was placed in 300 mL stop solution (1%$_{v/v}$ acetic acid).

2.9 Mass spectrometry

Mass spectrometry (MS) analysis was performed on bands excised from silver-stained gels that were more abundant in the GQ sample than in the control sample for the biotin pulldown experiments. Excised bands were placed in 1%$_{v/v}$ acetic acid solution and sent to the SPARC BioCentre Mass Spectrometry Facility at SickKids Hospital, Toronto, Ontario, Canada for analysis. Samples were digested

with trypsin, and were analyzed by LC-MS on the Orbitrap Elite™ hybrid ion trap-Orbitrap mass spectrometer to identify proteins by *de novo* peptide sequencing.

2.10 Protein overexpression and purification

Overexpression constructs for SVEN_2656 and SVEN_3866 were made by cloning the entire coding region of each gene into pET15b. The coding sequence was PCR amplified from genomic DNA using primers SVEN_2656F/R and SVEN_3866F/R. Both forward primers contained *Nde*I restriction sites at the and the reverse primer for SVEN_2656 contained a *Bam*HI restriction site while the SVEN_3866 reverse primer contained a *Bgl*II restriction site which were all used for cloning into pET15b digested with *Bam*HI and *Nde*I such that they were inserted into the same reading frame as the 6× His-tag. The overexpression constructs were then transformed by electroporation into *E. coli* Rosetta2 cells, and selected on LB plates supplemented with ampicillin and chloramphenicol. Ten millilitre overnight cultures from single colonies were used to start 500 mL volume subcultures in LB medium supplemented with ampicillin and chloramphenicol. Once the OD_{600} of the subcultures reached 0.4, overexpression was induced with 1 mM IPTG. After 3 h of induction at 30°C, cells were collected by centrifugation and stored at -80°C. One millilitre pre- and post-induction samples were separated on a 12% polyacrylamide gel and stained with Coomassie Brilliant Blue to test for overexpression.

For His-tag purification, cell pellets were thawed on ice, and were then resuspended in 5 mL lysis buffer (300 mM NaCl, 50 mM NaH_2PO_4, 10 mM imidazole, pH 8) containing one cOmplete™ Mini Protease Inhibitor Cocktail Tablet (Sigma Aldrich) for every 10 mL lysis buffer. Cells were lysed by sonication (Branson Sonifier Cell Disruptor350, 20× 30 s pulses at 50% duty cycle, output control = 4), after which the lysates were centrifuged for up to 1 h at 9,600× *g*. The resulting cell-free lysates were then mixed with 1 mL Ni resin (Ni-NTA agarose resin) and incubated slowly shaking at 4°C for 1 h. Proteins were purified by loading samples on chromatography columns, followed by washes with lysis buffer containing increasing concentrations of imidazole (10 mM – 2 M) to first wash and then elute proteins. The washes were done in 5 mL volumes, while the elution steps were done in 0.5 – 1 mL volumes. All washes and elutions, as well as crude soluble and crude insoluble fractions, were separated on a 12% polyacrylamide gel and stained with Coomassie Brilliant Blue to check the solubility and purity of the protein. Proteins were transferred into protein-specific storage buffers detailed below by dialysis using Slide-A-Lyzer® MINI Dialysis Devices with 3.5 kDa molecular weight cut-off (Thermo Scientific) overnight at 4°C. SVEN_2656 was stored in 25 mM Tris-HCl, 0.5 mM EDTA, 2 mM β-mercaptoethanol, 50%$_{v/v}$ glycerol, pH 7.6 at -20°C, as described for *E. coli* ribonuclease (RNase) PH (60). SVEN_3866 was stored in

5 mM Tris-HCl, 50 mM NaCl, pH 8 at -80°C, as described for *E. coli* TrmB (61). Protein concentrations was determined using a Bradford assay (59).

2.11 EMSAs

DNA oligonucleotide probes were 5' end-labeled in a 20 µL reaction with T4 polynucleotide kinase buffer, 10 U T4 polynucleotide kinase (New England Biolabs), 0.1 µM oligonucleotide, and 5 µL [γ-^{32}P]ATP (Perkin Elmer, 0.4 mCi/mL). Reactions were incubated at 37°C for 30 min, after which the labeled DNA was purified using NucAway spin columns. The labeled probes were diluted to 10 nM in GQ folding buffer, and heated to 95°C for 5 min followed by slow cooling to room temperature to allow for proper folding. One nanomolar probe DNA was added to each reaction, together with protein ranging in concentration from 2-150 µM in a 20 µL reaction volume. Probe and protein were incubated together in GQ folding buffer at room temperature for 30 min. They were then loaded onto a 10% non-denaturing acrylamide gel that was run for 30 min at 150 V. Gels were visualized using a phosphorimager.

2.12 Figures and Tables

Table 2.1: Bacterial strains

Species	Strain	Description/use	Reference
E. coli	DH5α	Cloning strain	Invitrogen
	ET12567/pUZ8002	Methylation deficient strain used for conjugative transfer of DNA into *Streptomyces*	(62)
	Rosetta 2	Overexpression strain that contains pRARE2	Novagen
S. venezuelae	ATCC 1072	Wild type	(63)
	Δ*rnj*	RNase J deletion mutant	(61)
	Δ*rho*	Rho deletion mutant	This work
S. coelicolor	M145	Wild type	(64)

Table 2.2 Plasmids

Plasmid	Description/use	Reference
pGUS	Integrative *Streptomyces*-specific reporter vector for transcriptional fusions with the *gusA* gene	(57)
pGUS + P*ermE**	Strong *Streptomyces* promoter, P*ermE**, upstream of *gusA*	R.J. St-Onge (unpublished)
pGUS + P*ermE**-30 bp-4G NTS	P*ermE**, 30 bp spacer sequence, and 4G GQ sequence on the NTS upstream of *gusA*	This work
pGUS + P*ermE**-30 bp-4G TS	P*ermE**, 30 bp spacer sequence, and 4G GQ sequence on the TS upstream of *gusA*	This work
pGUS + P*ermE**-30 bp-4G HQ	P*ermE**, 30 bp spacer sequence, and 4G HQ seqeunce upstream of *gusA*	This work
pGUS + P*ermE**-30 bp-5G NTS	P*ermE**, 30 bp spacer sequence, and 5G GQ sequence on the NTS upstream of *gusA*	This work
pGUS + P*ermE**-30 bp-5G TS	P*ermE**, 30 bp spacer sequence, and 5G GQ sequence on the TS upstream of *gusA*	This work
pGUS + P*ermE**-30 bp-5G HQ	P*ermE**, 30 bp spacer sequence, and 5G HQ sequence upstream of *gusA*	This work
pGUS + P*ermE**-30 bp-6G NTS	P*ermE**, 30 bp spacer sequence, and 6G GQ sequence on the NTS upstream of *gusA*	This work
pGUS + P*ermE**-30 bp-6G TS	P*ermE**, 30 bp spacer sequence, and 6G GQ sequence on the TS upstream of *gusA*	This work
pGUS + P*ermE**-30 bp-6G HQ	P*ermE**, 30 bp spacer sequence, and 6G HQ sequence upstream of *gusA*	This work
pGUS + P*ermE**-5G NTS	P*ermE** and 5G GQ sequence on the NTS upstream of *gusA*	This work
pGUS + P*ermE**-5G TS	P*ermE** and 5G GQ sequence on the TS upstream of *gusA*	This work
pGUS + P*ermE**-5G HQ	P*ermE** and 5G HQ sequence upstream of *gusA*	This work
pGUS + P*ermE**-120 bp	P*ermE** and 120 bp spacer sequence upstream of *gusA*	This work
pGUS + P*ermE**-120 bp-5G NTS	P*ermE**, 120 bp spacer, and 5G GQ sequence on NTS upstream of *gusA*	This work

pGUS + P*ermE**-120 bp-5G TS	P*ermE**, 120 bp spacer, and 5G GQ sequence on TS upstream of *gusA*	This work
pGUS + P*ermE**-120 bp-5G HQ	P*ermE**, 120 bp spacer, and 5G HQ sequence upstream of *gusA*	This work
pGUS + P*ermE**-115 bp	P*ermE** and 115 bp spacer sequence upstream of *gusA*	This work
pGUS + P*ermE**-115 bp-5G NTS	P*ermE**, 115 bp spacer, and 5G GQ sequence on NTS upstream of *gusA*	This work
pGUS + P*ermE**-115 bp-5G TS	P*ermE**, 115 bp spacer, and 5G GQ sequence on TS upstream of *gusA*	This work
pGUS + P*ermE**-115 bp-5G HQ	P*ermE**, 115 bp spacer, and 5G HQ sequence upstream of *gusA*	This work
Pl1_H4 SVEN_5009::aac(3)IV	Rho deletion construct	R.J. St-Onge (unpublished)
pUC19	*E. coli* cloning vector	(61)
pUC19 + *gfp*	pUC19 with *gfp*	This work
pUC19 + P*em7-gfp*	pUC19 with P*em7* upstream of *gfp*	This work
pUC19 + P*em7*-5G GQ-*gfp*	pUC19 with P*em7* and 5G GQ sequence on NTS upstream of *gfp*	This work
pUC19 + P*em7*-120 bp-*gfp*	pUC19 with P*em7* and 120 bp spacer sequence upstream of *gfp*	This work
pUC19 + P*em7*-120 bp-5G GQ-*gfp*	pUC19 with P*em7*, 120 bp spacer, and 5G GQ sequence on NTS upstream of *gfp*	GenScript
pRARE2	Contains tRNAs for 7 rare codons (AGA, AGG, AUA, CUA, GGA, CCC, and CGG) in *E. coli*.	Novagen
pET15b	Overexpression vector	Novagen
pET15b + *sven_2656*	Overexpression of SVEN_2656	This work
pET15b + *sven_3866*	Overexpression of SVEN_3866	This work

Table 2.3 Oligonucleotides

Name	Sequence (5’ - 3’)	Description
ermEF	GCA CTT CTA GAA GCC CGA CCC GAG CAC GCG C	Sequencing and PCR checks
ermER	GCA CTG GTA CCG ATC CTA CCA ACC GGC ACG A	Sequencing and PCR checks
gusAR2	TCG ATA CCG CAG TTC TCC	Sequencing, PCR checks, generating template for IVT
GfpR	CAA GTG TTG GCC ATG GAA CAG G	Sequencing and PCR checks
em7F	TCG AAC GTG TTG ACA ATT AAT CAT CGG CAT AGT ATA TCG GCA TAG TAT AAT ACG	Sequencing and PCR checks
SV5009KOF	CAG ATG GCC GAC GTC CGC TCC AGG GAA GGA CCC TTC GTG ATT CCG GGG ATC CGT CGA CC	PCR checking knockout strain
SV5009KOR	GGG CCC GGC TCT CGT CGT GTG ACG CGG CTC TCA CCG TCA TGT AGG CTG GAG CTG CTT C	PCR checking knockout strain

SV5009intR	GAG CTT GAG CTC CAT GTT G	PCR checking knockout strain
SV5009F	GCA CTC ATA TGA GCG ACA CCA CCG ATC T	PCR checking knockout strain
SV5009R	GCA CTG GAT CCT CAG TCG TTG TTG GCG GCA C	PCR checking knockout strain
T7 fwd	TAA TAC GAC TCA CTA TAG GG	Sequencing and PCR checks
T7 term	GCT AGT TAT TGC TCA GCG G	Sequencing and PCR checks
PlysC-120 nt F	**TAC GAC AAA TTG CAA AAA TAA TGT TGT CCT TTT AAA TAA GAT CTG ATA AAA TGT GAA CT**G GTA CCG CGC TAT CCG GTG	*In vitro* transcription (*E. coli* RNAP)
T7 prom-120 nt F	**TAA TAC GAC TCA CTA TAG** GGT ACC GCG CTA TCC GGT G	*In vitro* transcription (T7 RNAP)
gusA F	GGC AGC TTC AAC GAC CAG TT	qPCR
gusA R	CTG ATA CCA GAC GTT CCC CG	qPCR
rpoB F	TCA AGG AGT TCT TCG GCA CC	qPCR
rpoB R	ACC GAT CAG ACC GAT GTT CG	qPCR
5S rRNA F	CGG TGG TCA TAG CGT TAG GG	qPCR
5S rRNA R	GAA AGG CTT AGC TCC CGG GT	qPCR
30 nt spacer F	GGG AGA GGG AGG AAG GAG GGA GGG AAG GAC GTA C	Cloning reporter constructs
30 nt spacer R	GTC CTT CCC TCC CTC CTT CCT CCC TCT CCC GTA C	Cloning reporter constructs
120 nt spacer	CAT CAT *GGT ACC* GCG CTA TCC GGT GAT CTC CAA ATT AGA ACA TAC CGC CCC ACG AGG GCT AGA ATT ACC TAC CGG CCT CCA CCA TGC CTG CGC TAT ACG CGC CCA CTC TCC CGT T*GG TAC C*CA TCA T	Cloning reporter constructs
115 nt spacer	GGC ATC AT*G GTA CC*C CCT AAT ATG ACA TCA TTA GTG GCC AAA TGC CAC TCC CAA AAT TCT GCC CAG AAG CGT TTA GGT CCG CCC CAC TGA AGC TGC CTA AAA CGA CCA CCA A*GG TAC C*CA TCA TGG	Cloning reporter constructs
4G GQ F	CTA GTG GGG TGG GGT GGG GTG GGG	Cloning reporter constructs
4G GQ R	CTA GCC CCA CCC CAC CCC ACC CCA	Cloning reporter constructs
4G HQ F	CTA GTG GGG TGG GG	Cloning reporter constructs
4G HQ R	CTA GCC CCA CCC CA	Cloning reporter constructs
5G GQ F	CTA GTG GGG GTG GGG GTG GGG GTG GGG G	Cloning reporter constructs
5G GQ R	CTA GCC CCC ACC CCC ACC CCC ACC CCC A	Cloning reporter constructs
5G HQ F	CTA GTG GGG GTG GGG G	Cloning reporter constructs
5G HQ R	CTA GCC CCC ACC CCC A	Cloning reporter constructs

6G GQ F	CTA GTG GGG GGT GGG GGG TGG GGG GTG GGG GG	Cloning reporter constructs
6G GQ R	CTA GCC CCC CAC CCC CCA CCC CCC ACC CCC CA	Cloning reporter constructs
6G HQ F	CTA GTG GGG GGT GGG GGG	Cloning reporter constructs
6G HQ R	CTA GCC CCC CAC CCC CCA	Cloning reporter constructs
GQ probe	Bio-TAT ATA GGG AGG GCG GGA GGG	Biotin pulldown
Mut probe	Bio-TAT ATA GGG AGA GCG AGA GGG	Biotin pulldown control probe
SVEN_2656 F	CAT*C ATA TG*A TGT CTC GTA TCG ACG GCC GT	Cloning overexpression constructs
SVEN_2656 R	CAT CAT *GGA TCC* TCA GAG GGT TCC TTC GAG GGC	Cloning overexpression constructs
SVEN_3866 F	CAT CAT *CAT ATG* ATG AAC AGC TCG CAG TCC GT	Cloning overexpression constructs
SVEN_3866 R	CAT CAT *AGA TCT* TCA CTT CCT CCG GAA GAG GA	Cloning overexpression constructs

_Spe_I- or _Kpn_I-compatible overhangs, **promoter sequence (T7 or P*lysC*), _restriction enzyme sites_ (*Kpn*I, *Bam*HI, *Bgl*II, or *Nde*I)

Table 2.4: PCR reaction and cycling conditions.

Polymerase	**Reaction conditions**	**Cycling conditions**
Taq (GeneDirex)	1× Taq Buffer 5% DMSO 0.2 mM dNTPs 0.5 µM fwd and rev primers 0.5 U Taq DNA polymerase DNA from colony or 1-5 ng template DNA	1) 95°C 3 min 2) 95°C 30 s 3) 45-65°C 30 s 4) 72°C 1 min per kb of DNA 5) Repeat steps 2-4 29× 6) 72°C 5 min
Quick-load 2× Taq Master Mix (New England Biolabs)	1× Taq Master Mix 5% DMSO 0.5 µM fwd and rev primers DNA from colony or 1-5 ng template DNA	Same as above
Phusion (New England Biolabs)	1× Phusion Buffer 5% DMSO 0.2 mM dNTPs 0.5 µM fwd and rev primers 0.5 U Phusion DNA polymerase 1-5 ng template DNA	1) 98°C 3 min 2) 98°C 10 s 3) 45-65°C 30 s 4) 72°C 30 sec per kb of DNA 5) Repeat steps 2-4 29× 6) 72°C 5 min

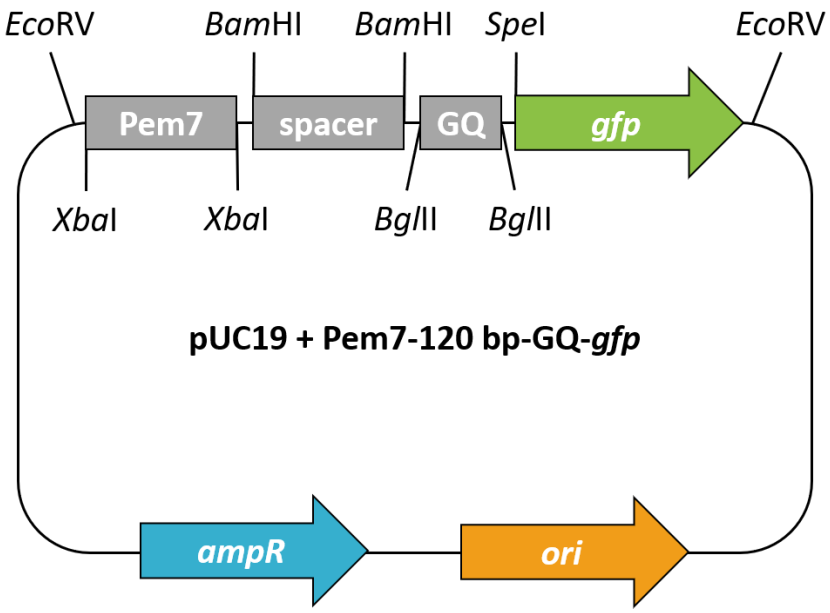

Figure 2.1: Schematic representation of pUC19-*gfp* vector synthesized by GenScript. All restriction enzyme sites that were used to make different iterations of the vector are listed. The vector also contains an ampicillin resistance gene (*ampR*) as well as a ColE1 origin of replication (*ori*). The Pem7 promoter, which is a synthetic constitutive promoter was used.

3. *In silico* characterization of putative GQ forming sequences in *Streptomyces*

3.1 Introduction

As discussed briefly in Chapter 1, several algorithms have been developed for searching primary sequences for motifs with the potential to form DNA and RNA GQ structures. These algorithms commonly use the sequence motif: $G_3N_{1-7}G_3N_{1-7}G_3N_{1-7}G_3$, or variations on this motif, to identify potential GQs. One of the first investigations into the genome-wide prevalence of GQ sequences focused on the human genome (33). This study identified over 300,000 GQ sequences in human genomic DNA, many of which were located in putative regulatory regions (33). It was later discovered that there was at least one GQ sequence within 1 kb upstream of almost 50% of human genes (65), suggesting that GQ structures may have major roles as global regulators of gene expression. Analogous bioinformatic analyses have since been carried out to identify putative GQ sequences in the genomes of plants (66, 67), yeast (68), bacteria (24, 26, 35), and viruses (69).

One of the largest studies of bacterial GQ-forming sequences searched the genomes of 18 bacterial species for sequences with the potential to form GQs (35). This work, which included *Streptomyces coelicolor* in the analysis, also found that GQ motifs were enriched in putative regulatory regions – defined as regions up to 200 nt upstream of the start of an open reading frame. *S. coelicolor*, which had the highest genomic GC content of all the organisms investigated, also had the highest frequency of genomic GQ sequences. The authors also proposed a role for GQ structures in global gene regulation in *E. coli,* given that GQ sequences coincided with the binding sites of major regulators of gene expression, including the housekeeping sigma factor (σ^{70}) (35). However, experimental evidence to support this proposal is still needed.

Recent work has suggested that many *M. tuberculosis* promoter regions contained GQ sequences (29). Given this, and the mounting evidence for the potential roles of GQs in promoter function, GQ sequences are now being used to predict the locations of promoters in GC-rich organisms (70). However, we still know very little about the prevalence and function(s) of GQs in bacteria. *Streptomyces* are extremely high-GC organisms that are predicted to have an abundance of GQ sequences, making them an excellent system for answering these questions. Here, we aim to analyze the genomes of three model *Streptomyces* species to determine the prevalence and relative positions of GQ sequences in the genomes of these GC-rich bacteria, then make use of these results to deduce potential functions based on their locations. We propose to follow up on these findings experimentally and use these data to

guide current and future experimental investigations into the regulatory roles of GQ structures in bacteria.

3.2 Results

3.2.1 Abundance of GQ sequences in *Streptomyces* genomes

To begin characterizing the number of GQ sequences in *Streptomyces*, we focused our attention on three well-studied *Streptomyces* species: *S. venezuelae, S. coelicolor,* and *Streptomyces avermitilis*. We searched the genomes of these species for the sequence motif: $G_3N_{1-7}G_3N_{1-7}G_3N_{1-7}G_3$, which is generally accepted as the consensus sequence for GQs. We found 2,986 GQ sequences in *S. venezuelae*, 2,657 in *S. coelicolor*, and 2,430 in *S. avermitilis*. These sequences were then classified as being either intragenic (within coding regions) or intergenic (between coding regions). In all three species, the intergenic GQ sequences represented between 21% and 26% of the total number of GQ sequences (Table 3.1). Given the 88.9% coding density of the *S. coelicolor* chromosome (71), this was approximately double what we would expect to find in intergenic regions if GQ sequences were randomly distributed throughout the genome. Based on these results, we concluded that GQ sequences were highly abundant in *Streptomyces* genomes. However, due to their extremely high GC content, many of these sequences could have arisen by chance.

To determine whether the number of GQ sequences in *S. venezuelae* was higher than the number we would expect to find by chance alone, we used a custom Python script to randomly shuffle the entire *S. venezuelae* genome while maintaining triplet frequencies using the program uShuffle (55). The number of GQ sequences was then counted in the shuffled genome, before the shuffling and searching process was repeated for a designated number (*n*) of times. After 1,000 iterations of this process, we found an average expected number of 1,309 GQ sequences, which was far lower than the observed 2,986 GQ sequences ($p<0.001$) (Figure 3.1). Similar trends were observed when the same script was run on *S. coelicolor* and *S. avermitilis* genomic DNA. This indicated that there is likely selective pressure to maintain these sequences, given that the number of GQ sequences in diverse *Streptomyces* genomes was higher than what would have been expected based on chance alone, even when considering their extremely high GC content.

3.2.2 Distribution of GQ sequences in *Streptomyces* genomes

To better understand the function of GQ sequences in the streptomycetes, we focused our attention on *S. venezuelae.* In examining the distribution of GQ sequences within the *S. venezuelae* chromosome,

there was no apparent pattern: they appeared to be uniformly distributed throughout the chromosome (Figure 3.2a). However, when the sequences were separated based on strand position (positive [top] strand or negative [bottom] strand) on the chromosome, we observed an interesting GQ sequence enrichment on the negative strand on the left arm of the chromosome, and on the positive strand on the right arm of the chromosome (Figure 3.2b). *Streptomyces* chromosomes are linear, with the origin of replication in the center, and thus this configuration represented an enrichment on the lagging strand of DNA synthesis (*i.e.* the template for leading strand synthesis) (Figure 3.2c). Plotting the *S. coelicolor* and *S. avermitilis* GQ sequences in the same way, we found the same trend for *S. coelicolor* but not *S. avermitilis* (Figure 3.2d and 3.2e). Expanding our search to include other *Streptomyces* species, alongside other GC-rich bacteria, we found that approximately half of all species we analyzed followed the same pattern as *S. venezuelae* and *S. coelicolor* while the others showed no obvious enrichment. Among the species that displayed this pattern were the only two non-*Streptomyces* species we investigated: *M. tuberculosis* and *P. aeruginosa* (Table 3.2). Given that the pattern was conserved between multiple species, some of which were distantly related to our model species, this indicated that there may be some functional role associated with this stranded enrichment of GQ sequences.

3.2.3 Analysis of GQ sequences in putative regulatory regions

Knowing that over 20% of the identified GQ sequences were located in intergenic regions, we decided to further determine how many of these were found in possible regulatory regions, by looking specifically for GQ sequences in untranslated regions (UTRs), near transcription start sites, and between convergently and divergently oriented genes. GQ sequences found in UTRs could have roles in transcriptional, post-transcriptional, or translational regulation, while those near transcription start sites could have either positive or negative effects on transcription initiation. GQ sequences between convergently oriented genes could prevent transcriptional readthrough into the next gene, while those found between divergently oriented genes could facilitate transcription: if one gene is highly expressed, the GQ could help keep the DNA open to facilitate transcription of the other gene.

To assess the number of GQ sequences in UTRs, we generated a list of all UTRs of transcribed genes using a custom Python script that used RNA-seq data (unpublished data from E. Sherwood) to determine UTR boundaries. We then compared this to our list of GQ sequences to identify those that overlapped with the UTRs using another custom Python script. We ultimately identified 97 GQ sequences that overlapped with the UTRs of transcribed genes. Of these, the vast majority (72 of 97) were in 3′ UTRs, while only ~26% (25 of 97) were in 5′ UTRs (Appendix G).

Despite the low number of GQ sequences within 5' UTRs, when we expanded our search to include regions outside of UTRs by looking for GQ sequences in proximity to transcription start sites (TSSs), we found many more. We used a custom Python script to identify GQ sequences near TSSs of transcribed genes using 5'-tag RNA-seq data (unpublished data from M. Bush and M. Buttner) to identify TSSs. We analyzed the data for RNA isolated at four time points spanning the complete developmental cycle of *S. venezuelae* (10 h, 14 h, 18 h, and 24 h). We found a total of unique 146 TSSs that were within 100 bp, either upstream or downstream, of a GQ sequence. Of these 146 GQ sequences, almost half (66 of them) were on the same strand as their associated gene (Appendix H). When we looked at the expression level of the associated gene (unpublished RNA-seq data obtained from E. Sherwood), we found the average expression level in RPKM for genes on the same strand as their associated GQ sequence was 322.9, compared with 126.6 for genes that were on the strand opposite their associated GQ sequence (Figure 3.3). While this difference was not statistically significant (p=0.17), it suggested that GQ sequences on the same strand as their associated gene may exert a more positive effect on gene expression than those on the opposite strand.

To look for GQ sequences between convergently and divergently oriented genes, we used a custom Perl script (developed by S. Jones, unpublished) to extract all pairs of divergently oriented and convergently oriented genes from the *S. venezuelae* genome. We then found all GQ sequences that overlapped with these intergenic regions using the BEDTools intersect utility. Of the 646 intergenic GQ sequences, 139 were located between convergently oriented genes (Appendix I), while 175 were located between divergently oriented genes (Appendix J). We then plotted the expression level of gene 1 (leftmost gene) as a function of the expression level of gene 2 (rightmost gene) based on the RPKM values obtained from RNA-seq data (unpublished data obtained from E. Sherwood) to determine whether there was any correlation between the expression of the two genes for either data set (Figure 3.4). We found that there was no correlation between the expression of gene 1 and the expression of gene 2 for either divergently or convergently oriented genes (Figure 3.4). However, these data may help guide future studies into specific regulatory functions of GQ sequences by identifying the GQ sequences between convergently or divergently oriented genes.

3.2.4 Analysis of intragenic GQ sequences

Since the majority of identified GQ sequences were intragenic in our three model streptomycetes, we investigated whether these sequences were enriched in particular regions within genes. Looking at their relative positions within genes, we found that they were highly enriched in the first 5% of the gene, and

slightly enriched in the last 15% of the gene (Figure 3.5a). When we separated them into coding and non-coding strand GQ sequences, we found that for both strands, the enrichment at the beginning of the gene was conserved, but the enrichment at the end of the gene was much more pronounced for coding strand GQ sequences (Figure 3.5b,c). Coding strand GQ sequences seemed to be somewhat depleted compared with their non-coding strand counterparts, with only 647 of the intragenic GQ sequences found on the coding strand, while 1,693 were found on the non-coding strand.

3.2.5 GQ sequences in secondary metabolic clusters

While mapping the genomic positions of GQ sequences, we noted that there were localized regions with very high GQ sequence density in some organisms, including *S. avermitilis* and *Streptomyces pristinaespiralis* (Figure 3.6). In these two organisms, we found that the regions with the highest densities of GQ sequences corresponded to the biosynthetic gene clusters for the antibiotics avermectin in *S. avermitilis* and pristinamycin in *S. pristinaespiralis*. Strikingly, of the 102 GQ sequences within the avermectin cluster, 60 were located in only two genes, both of which encode type 1 polyketide synthases. However, the GQ sequences in the pristinamycin cluster were more evenly spaced out throughout the cluster. We then searched for GQ sequences in the 33 antiSMASH-predicted secondary metabolic gene clusters in *S. venezuelae* (72). We found a total of 378 GQ sequences in these clusters, with at least one GQ sequence in every cluster. The density of GQ sequences per kb in secondary metabolic clusters was 0.38 compared to 0.34 in the entire genome.

3.3 Conclusions

Based on our global analysis of GQ sequences in *Streptomyces* genomes, we concluded that GQ sequences were highly abundant in these organisms. Even though *Streptomyces* have extremely high GC content, there were still many more GQ sequences than would have been expected by chance alone. These GQ sequences were not randomly distributed, as they were enriched in intergenic regions, with many areas where they could exert potential regulatory function (in UTRs and near TSSs). Based on these analyses, we hypothesize that coding strand-associated GQ sequences may promote or facilitate transcription, given that genes associated with a GQ sequence on the same strand tended to be more highly expressed than genes associated with GQ sequences on the opposite strand. Intragenic GQ sequences were enriched at the beginnings and ends of coding sequences, with more being found on the non-coding strand than on the coding strand, indicating that they may be depleted in mRNAs. However, even GQ sequences on the non-coding strand have the potential to act as regulators at the transcriptional level, for example by causing polymerase stalling to reduce gene expression. We also

found that there were many GQ sequences in secondary metabolic clusters, particularly in those encoding avermectin and pristinamycin. If GQ sequences are involved in the regulation of these clusters, GQs could potentially be used to manipulate the production of antibiotics.

These analyses have provided new insights into possible roles for a novel regulatory element in *Streptomyces* bacteria, although it is important to remember that these are only sequences that are predicted to form GQs, and experimental validation is required for more definitive conclusions to be drawn.

Table 3.1: Number of predicted GQ sequences in three model streptomycetes. Total numbers of GQ motifs were determined, after which sequences were classified as either intragenic (defined as within protein coding regions) or intergenic.

	Number of GQ sequences		
Species	**Intergenic**	**Intragenic**	**Total**
S. venezuelae	646 (21.6%)	2,340 (78.4%)	2,986
S. coelicolor	577 (21.7%)	2,080 (78.3%)	2,657
S. avermitilis	635 (26.1%)	1,795 (73.9%)	2,430

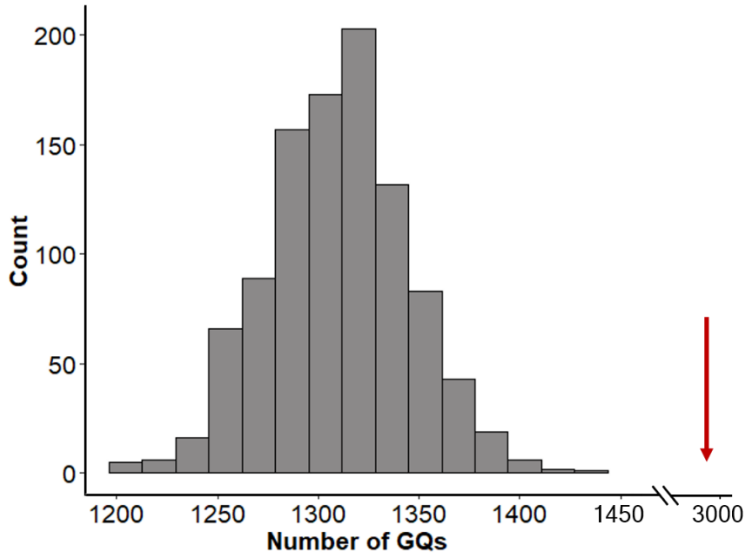

Figure 3.1: Number of expected vs. observed GQ sequences in *S. venezuelae*. The *S. venezuelae* genome was shuffled 1,000 times using uShuffle (*k*-let size = 3) and searched for GQ sequences after each re-shuffling. The histogram depicts the distribution of the number of GQ sequences found in the shuffled genomes, which represents the expected number of GQ sequences (mean=1,309; *n*=1,000). The red arrow demonstrates the number of observed GQ sequences in the *S. venezuelae* genome, 2,986, which is significantly higher than the expected number ($p<0.001$).

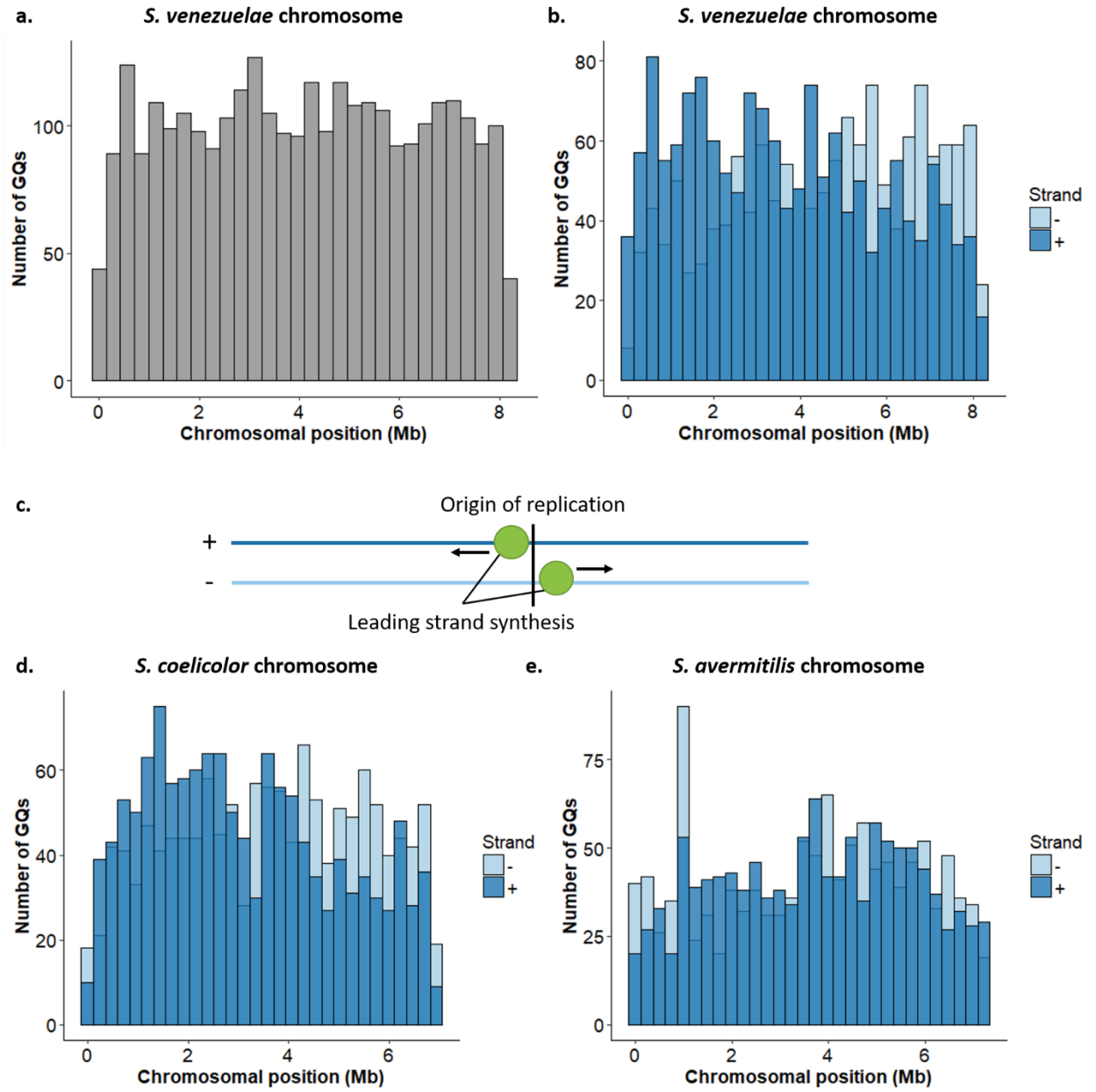

Figure 3.2: Distribution of GQ sequences in the genomes of model *Streptomyces* species. a. Distribution of all GQ sequences along the *S. venezuelae* chromosome. **b.** Distribution of GQ sequences along the *S. venezuelae* chromosome separated by positive strand (dark blue) and negative strand (light blue). **c.** Schematic of the linear *Streptomyces* chromosome with the origin of replication in the centre and DNA polymerase initiating leading strand synthesis bidirectionally from the origin. **d.** Distribution of GQ sequences in *S. coelicolor* separated by strand. **e.** Distribution of GQ sequences in *S. avermitilis* separated by strand.

Table 3.2: GQ sequences in other *Streptomyces* species and other GC-rich bacteria.

Species	Pattern?[1] (Y/N)	Number of GQ sequences	Genome size (Mb)	Number of GQ per Mb	GC content
Mycobacterium tuberculosis	Y	426	4.4	97.2	65.6
Pseudomonas aeruginosa	Y	328	6.6	49.7	66.2
Streptomyces albus	N	3776	6.8	553.7	71.0
Streptomyces avermitilis	N	2430	9.1	266.4	70.7
Streptomyces bingchenggensis	N	3812	11.9	319.2	70.7
Streptomyces cattleya	Y	2514	8.1	310.7	73.0
Streptomyces clavuligerus	Y	3832	8.6	447.7	72.3
Streptomyces coelicolor	Y	2657	9.0	293.6	72.0
Streptomyces collinus	Y	2824	8.3	341.5	72.6
Streptomyces davawensis	N	2512	9.5	265.3	70.6
Streptomyces fulvissimus	Y	2966	7.9	375.0	71.5
Streptomyces griseoflavus	Y	2174	8.0	270.1	65.6
Streptomyces griseus	Y	3449	8.7	395.1	72.1
Streptomyces himastatinicus	N	2887	11.0	261.7	67.6
Streptomyces hygroscopicus	Y	3146	11.0	285.2	67.6
Streptomyces laurentii	Y	3429	8.0	427.0	72.3
Streptomyces lividans	Y	2669	8.5	314.4	71.4
Streptomyces pristinaespiralis	N	2198	8.1	270.4	66.8
Streptomyces roseosporus	Y	2728	7.8	348.8	69.0
Streptomyces scabiei	N	3703	10.1	364.8	71.4
Streptomyces sp. PAMC26508	Y	1728	7.5	229.8	71.1
Streptomyces sviceus	N	2474	9.3	265.7	68.5
Streptomyces violaceusniger	N	3405	10.8	315.3	70.9
Streptomyces viridochromogenes	Y	2333	8.6	269.7	70.3

[1]Refers to enrichment on the lagging strand that was observed in *S. venezuelae* and *S. coelicolor*

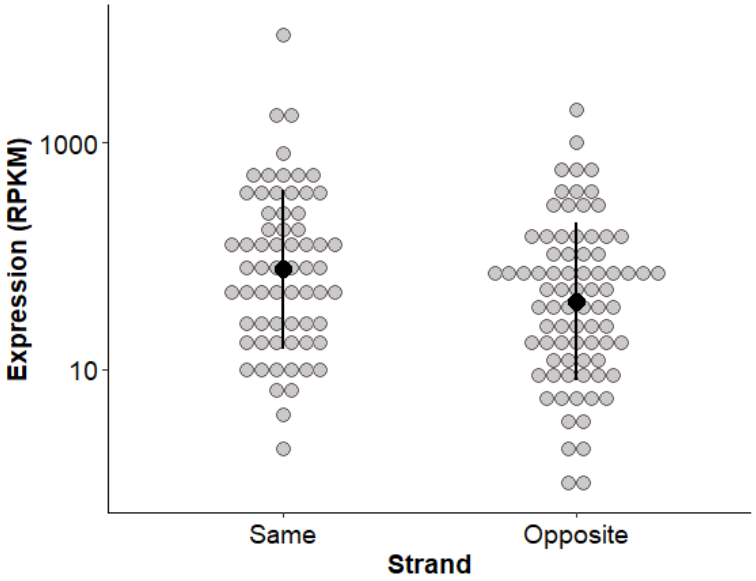

Figure 3.3: Expression of genes with GQ sequences in proximity to TSSs. Expression level (RPKM) of genes with a GQ sequence within 100 nt of their putative TSS. The data were separated based on whether the GQ sequence was on the same strand or opposite strand of the associated gene. Black dots indicate mean of each group: Same = 322.9 (*n* = 66), Opposite = 126.6 (*n* = 80), black lines indicate standard deviation. The difference between the means is not statistically significant (*p* = 0.13).

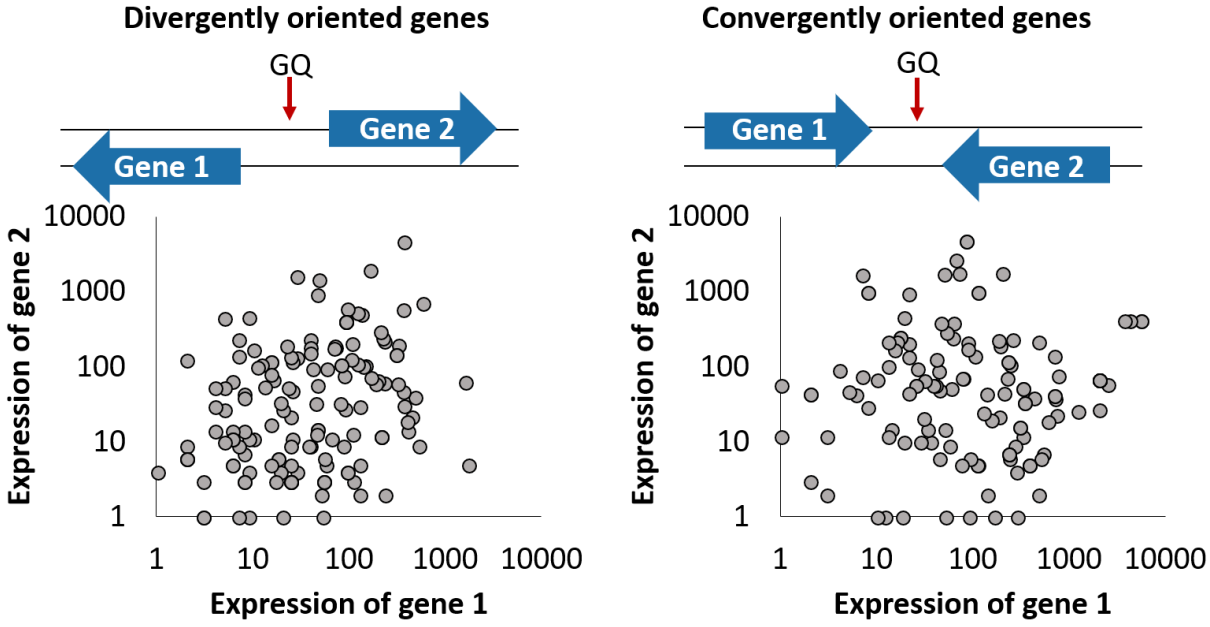

Figure 3.4: GQ sequences between convergently and divergently oriented genes. Correlation between the expression level of gene 1 and gene 2 in the 175 divergently oriented gene pairs and the 139 convergently oriented gene pairs with GQ sequences in between them.

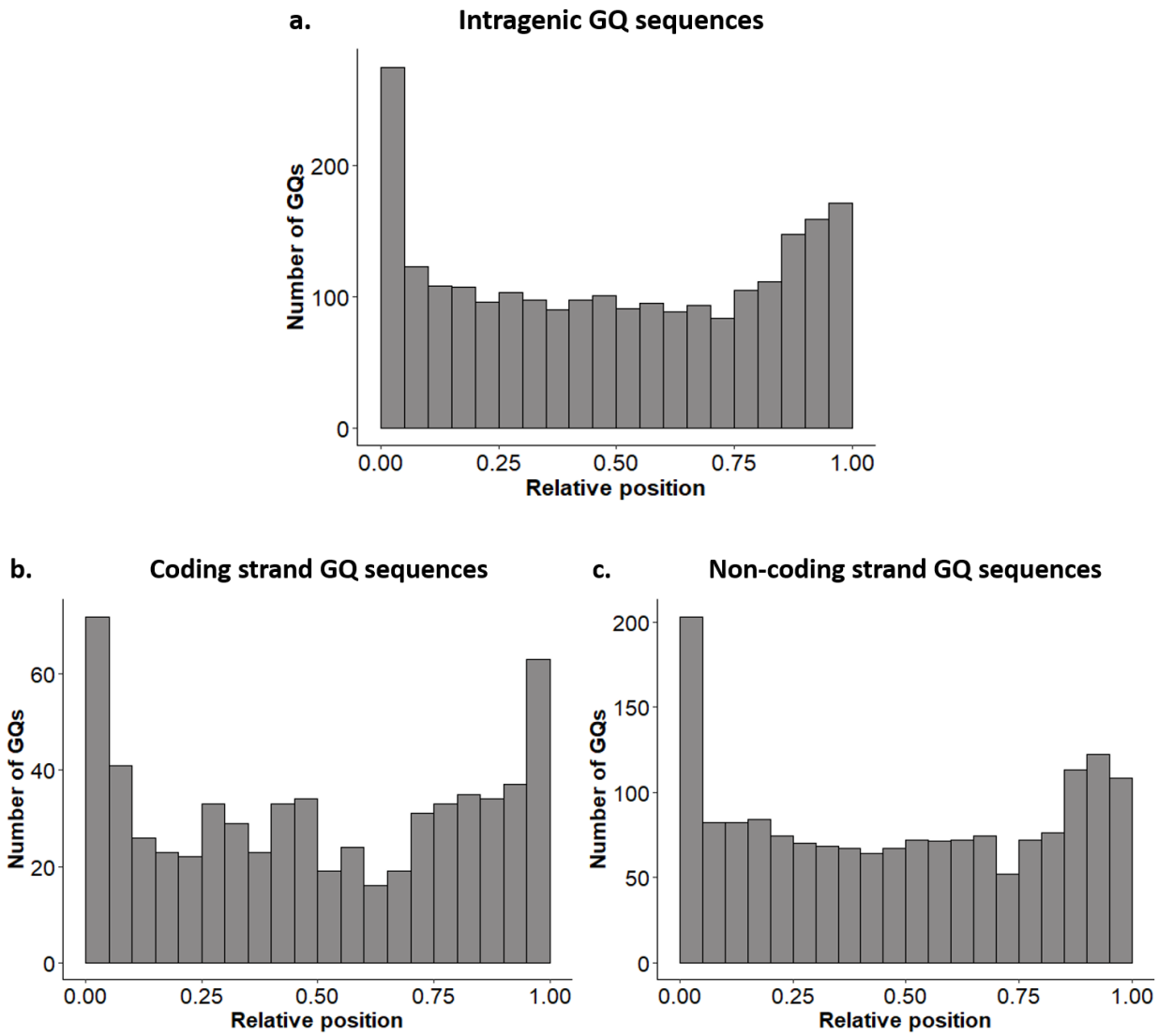

Figure 3.5: Relative positions of intragenic GQ sequences. Relative positions of all **a.** intragenic, **b.** coding strand, and **c.** non-coding strand GQ sequences in *S. venezuelae*.

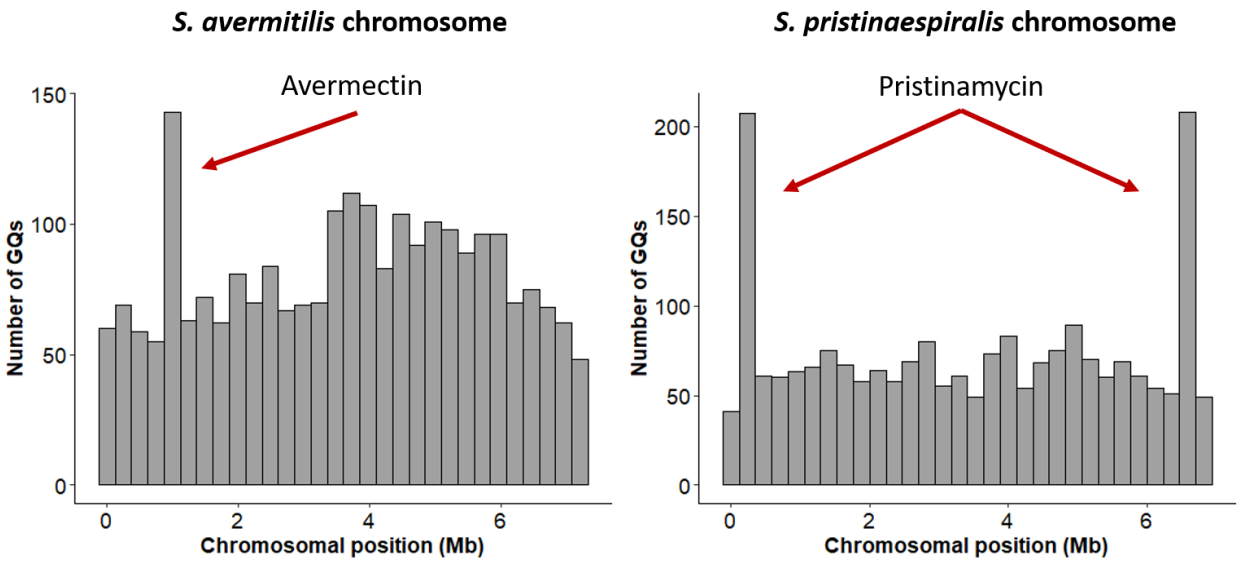

Figure 3.6: Enrichment of GQ sequences in secondary metabolic clusters. Distribution of GQ sequences in the genomes of *S. avermitilis* and *S. pristinasepiralis*. Red arrows indicate the avermectin and pristinamycin clusters that are enriched for GQ sequences.

4. Investigating mechanisms of transcriptional regulation by GQs

4.1 Introduction

There is substantial evidence that GQ sequences affect gene expression at the transcriptional level; however, these studies have not converged on a specific function for GQ sequences, with negative and positive effects having been observed for GQ sequences on both strands. It has generally been observed that GQ sequences on the template strand (TS) cause decreased gene expression – possibly due to polymerase stalling (27, 73). Further evidence for this comes from the observation that small molecules that stabilize GQ structures have been found to cause decreased gene expression (73), most notably in the case of oncogenes with promoter GQ sequences (14, 15, 74). In general, it has also been observed that GQ sequences on the non-template strand (NTS) have less of an effect on gene expression (27, 73). However, exceptions to these trends have also been observed and the underlying mechanisms of action remain unknown. For example, it was shown that when the GQ sequence was on the NTS of DNA, the formation of DNA:RNA HQs resulted in transcription termination in bacteria (7, 8), while the insertion of GQ sequences on the TS at specific locations in the promoter resulted in increased gene expression (27).

Since NTS GQ sequences also appear in the mRNA, there is the potential for these sequences to influence translation. Indeed, in bacteria (75) and in eukaryotes (23), GQ sequences near the ribosome binding site have been shown to negatively impact gene expression by interfering with translation initiation. This *in vivo* evidence for specific roles of RNA GQs in influencing gene expression goes against the finding that RNA GQ structures are globally unfolded *in vivo* (50). These conflicting results emphasize our lack of understanding of the complex regulatory roles that may be mediated by these structures.

While there have been some *in silico* analyses of GQ sequences in GC-rich bacteria (29, 35), the only studies into the regulatory functions of these structures in bacteria were conducted in *E. coli* (7, 24, 27). Here, we have demonstrated that *Streptomyces* species have an abundance of sequences with the potential to form GQs and that these sequences could play important roles as global regulators of gene expression (see Chapter 3). Given the sheer number of these sequences in *Streptomyces* genomes compared to *E. coli*, studying GQs in the streptomycetes has the potential to uncover novel roles for these structures, and provide insight into how organisms deal with the potential negative effects on cellular processes. Here, we systematically assessed the effects of GQ sequences on gene expression in *Streptomyces* bacteria. We observed a variety of effects on gene expression, some of which were consistent with what had previously been observed in *E. coli*. We also found a highly positive effect on

gene expression under certain circumstances and have worked to elucidate the mechanism of action underlying this.

4.2 Results

4.2.1 Monitoring the effects of GQ sequences on gene expression using transcriptional reporters

To determine whether GQ sequences could influence gene expression in *Streptomyces*, we constructed a series of transcriptional reporters using the Gus reporter system. Our reporter constructs all contained the same strong promoter, P*ermE**, driving the expression of *gusA*, the gene encoding the β-glucuronidase, 'Gus', enzyme. We introduced three types of GQ sequences upstream of *gusA,* together with a GC-rich 30 bp spacer sequence between the promoter and the GQ sequence. Specifically, we examined the effects of intramolecular GQ sequences on the NTS, intramolecular GQ sequences on the TS, and intermolecular – HQ – sequences (Figure 4.1) on downstream *gusA* transcription. The intramolecular GQ sequences all contained four G-tracts. In contrast, the HQ sequences only had two G-tracts such that they could only form DNA:RNA HQ structures between the NTS of DNA and the nascent RNA, where such structures have been shown previously to promote transcription termination in *E. coli* (7). We also varied the number of G's in each G-tract to include either four, five, or six G's, to assess the effects of different GQ stabilities on gene expression (Figure 4.1), as longer G-tracts are generally associated with more stable structures (11).

When we tested these transcriptional reporters in *S. venezuelae* and compared the relative Gus activity to the promoter alone control, we found that the effects varied substantially depending on both the length of the GQ sequence and the type of GQ sequence being tested (Figure 4.2). Most GQs led to significantly reduced Gus activity (suggesting reduced transcription or translation); however, there were exceptions observed for each type of GQ. Notably, only the 5G G-tract length had a consistently (and significantly) repressive effect on Gus activity in all three types of GQ sequence. We therefore decided to move forward using 5G-containing constructs for further testing.

We next wanted to probe the effect of GQ position relative to the promoter, and thus varied the length of the random spacer sequence separating the GQ from the promoter. We had initially selected a 30 bp spacer because it was long enough to allow RNA polymerase to enter elongation phase before encountering the GQ sequence. We reasoned that any effect of the GQ sequence on gene expression would therefore not be due to interfering with transcription initiation. However, given the large number of GQ sequences in proximity to promoter sequences (Chapter 3) and mounting evidence for the effects

of GQ sequences on promoter function, we decided to test the effects of GQs on Gus activity without a spacer sequence separating the promoter and the GQ sequence. The effects of longer spacers on Gus activity were also tested. Collectively, we sought to analyze the effects of GQ sequences on transcription initiation (no spacer), early transcription elongation (30 bp spacer), and late elongation (100+ bp spacer).

When we tested our no spacer constructs in *S. venezuelae*, we found that the presence of all three types of GQ sequences led to decreased Gus activity compared to the promoter alone construct, similar to what was observed with the 30 nt spacer (Figure 4.3). Interestingly, the strain with the GQ on the NTS had higher Gus activity than the strain with the GQ on the TS. While this difference was not statistically significant in *S. venezuelae*, the trend was reproducible, and when tested in *S. coelicolor,* was statistically significant (Figure 4.4). This trend was also consistent with our observation in Chapter 3 that genes with GQ sequences in proximity to their promoters tended to be more highly expressed if the GQ sequence was on the same strand as the gene (*i.e.* the NTS) than if the GQ sequence was on the opposite strand (Figure 3.3).

When we increased the spacer length to 120 bp and tested the effects of the three types of GQ sequences, we found that the spacer sequence alone led to a significant decrease in reporter activity. Interestingly, when the GQ sequence on the NTS was added downstream of the spacer, Gus activity was restored to promoter-alone levels (Figure 4.3). To rule out the possibility that this was a spacer sequence-specific effect, we repeated the assays with an alternative spacer sequence of approximately the same length, and we saw the same results (Figure 4.3). When we tested the same spacer sequence in *S. coelicolor*, again we observed the same trend (Figure 4.4). To determine whether this response was specific to organisms with a high GC DNA content, we tested whether the same effect was observed in *E. coli,* using GFP as our reporter construct output. Again, we observed a similar trend, where having the GQ sequence immediately after the promoter on the NTS led to reduced GFP levels, while addition of the 120 nt spacer sequence caused a large reduction in GFP activity (Figure 4.5). We frequently observed increased GFP activity with the GQ on the NTS compared to the spacer alone (in two of three replicates), however, this effect was not statistically significant.

The spacer sequences had 60% GC content, and were designed using a random sequence generator. Candidate sequences were then screened for any significant secondary structures using a program that determined the folding free energy, ΔG, for a sliding window of 30 nt that moved through the spacer sequence 1 nt at a time (program developed by M.J. Moody, unpublished). We found no significant

secondary structures in our selected sequence (Figure 4.6). The most stable structure within the 120 nt spacer sequence had a ΔG value of -7, with the longest stem in this structure being only 4 bp long (Figure 4.6). We also verified that the inserted GQ sequence was capable of forming a GQ structure using CD, and found that the CD spectrum for the 5G GQ sequence was consistent with what has been seen for parallel GQ structures previously (59), with a minimum at 240 nm and a peak at 265 nm (Figure 4.7).

Since increased Gus activity in the presence of a GQ on the NTS with longer spacer sequences was conserved between both streptomycetes analyzed, with multiple spacer sequences, and was somewhat recapitulated in *E. coli*, we were interested in determining the molecular mechanism underlying this observation. We therefore focused our attention on the increased Gus activity observed with the GQ on the NTS. Since the GQ on the TS did not affect Gus activity with the longer spacer sequences, we used this construct as a control. The HQ reporter did not cause reduced Gus activity with the longer spacer sequences, suggesting that, contrary to previous reports (6, 7), it did not cause transcription termination in our system.

4.2.2 Analysis of GQ effects at the transcriptional level

To probe the mechanism behind the decrease in Gus activity when the longer spacer sequences were added between the promoter and *gusA,* and the restoration to promoter-alone levels when the GQ sequence was added to the NTS of DNA, we started by looking at RNA levels within the cell to determine whether increased Gus activity corresponded to increased transcript abundance. We analyzed transcript abundance using reverse-transcription quantitative PCR (RT-qPCR), focusing on both the spacer alone and the spacer with GQ transcripts. We found that there was an approximately 2-fold increase in transcript abundance for the spacer with GQ transcript compared with the spacer alone transcript (Figure 4.8a), suggesting that the presence of the GQ enhanced transcript levels.

To test whether this was the result of increased transcription, we conducted *in vitro* transcription using both spacer alone, and spacer with GQ as template, in association with an *E. coli*-specific promoter. *In vitro* transcription was carried out using *E. coli* RNA polymerase, and the resulting products were separated on a sequencing gel, to determine whether the presence of the GQ sequence alone was sufficient to promote increased transcription *in vitro*. Unexpectedly, we found that the GQ sequence adversely affected transcription *in vitro,* as there was no full-length transcript detected in the spacer with GQ sample, but full-length transcript was observed in the spacer alone sample (Figure 4.8b). Therefore, it seemed that either there was some factor missing from our *in vitro* transcription assay that

promotes increased transcript abundance of GQ-containing transcript *in vivo*, or that increased transcription was not what was responsible for the increased transcript levels associated with GQ sequences.

4.2.3 Testing the stability of GQ mRNAs

Given our *in vitro* transcription results, we hypothesized that the GQ may affect mRNA stability by preventing RNase-mediated degradation, rather than exerting its effects at the transcriptional level. In *Streptomyces*, one of the major RNases is RNase J which can act as a 5'-3' exoribonuclease. Since the GQ sequence was located in the 5' UTR of our reporter genes, we proposed that the GQ structure could be preventing degradation from the 5' end of the transcript by RNase J, resulting in higher Gus activity in the presence of the GQ sequence. To test this hypothesis, we conducted our original reporter assays in an *S. venezuelae* RNase J mutant background (Δ*rnj*; obtained from S.E. Jones). In the Δ*rnj* background, there was no longer any difference between the spacer alone and the GQ on the NTS samples (Figure 4.9a), which supported our hypothesis that GQs may function to protect RNAs from degradation by RNase J.

To test this hypothesis more directly, we performed an RNA stability assay with 3'-end labeled *in vitro* transcribed spacer alone, and spacer with GQ transcripts, incubated with wild type and Δ*rnj* cell-free lysates. We found that there was no difference in the degradation rates for the two RNAs exposed to either lysate over the 90 min time course of the assay (Figure 4.9b)

We also examined the relative stabilities of our spacer and spacer with GQ transcripts using an *in vivo* RNA stability assay, to ensure that the results we obtained *in vitro* (with the cell lysates) were representative of what was happening *in vivo*. To do this, we grew liquid cultures of *S. venezuelae* with both the spacer alone and the spacer with GQ reporter constructs, and then exposed the cultures to the RNA polymerase-targeting antibiotic rifampicin, over a 10 min time course. We extracted RNA from each sample and performed RT-qPCR to measure transcript abundance of the spacer alone and the spacer with GQ transcripts at time zero (before rifampicin addition), and to monitor their degradation over time. We observed no difference in the rate of decay between the two transcripts, suggesting that the GQ sequence did not influence mRNA stability under these assay conditions (Figure 4.9c).

4.2.4 Testing the anti-termination capabilities of GQs

Given that our RNA stability hypothesis was not strongly supported by our data, we next wanted to determine whether GQs could cause increased transcript abundance by preventing premature

transcription termination. Given that there were no stable secondary structures in the spacer sequence (Figure 4.4), intrinsic termination was not expected to be a factor, so we instead focused our attention on the activity of the transcription termination factor Rho. We hypothesized that Rho might be targeting the spacer sequence for premature transcription termination and that the GQ sequence might block this activity. In this way, the GQ sequence may function as an anti-terminator, ultimately resulting in increased gene expression. We deleted *sven_5009*, the gene encoding Rho in *S. venezuelae,* and examined our reporter constructs with the 120 bp spacer sequence in this deletion background. In the Δ*sven_5009* strains, we found no difference in reporter activity between the spacer alone and the promoter alone, indicating that Rho may indeed target the spacer transcript as we predicted. However, the reporter with the GQ on the NTS still had higher Gus activity than the spacer alone, indicating that there were additional factors leading to increased reporter activity (Figure 4.10).

4.3 Conclusions

Here, we demonstrated that GQ sequences could profoundly affect gene expression in *Streptomyces*. We tested a variety of GQ sequences positioned at different distances relative to the promoter, and found that both position and type of GQ sequence impacted gene expression. One of our most interesting and unexpected findings was that GQ sequence on the NTS positioned at sites upwards of 100 nt from the promoter, led to large increases in expression of the associated gene. While we have not yet definitively elucidated the mechanism of action underlying this increased expression, the fact that the same trend was observed in *S. coelicolor* suggests that the mechanism of action may be conserved within *Streptomyces*. Similar trends were also observed in *E. coli*, suggesting that the GQ-mediated effect that we observed on gene expression may be broadly conserved across bacteria.

4.4 Figures

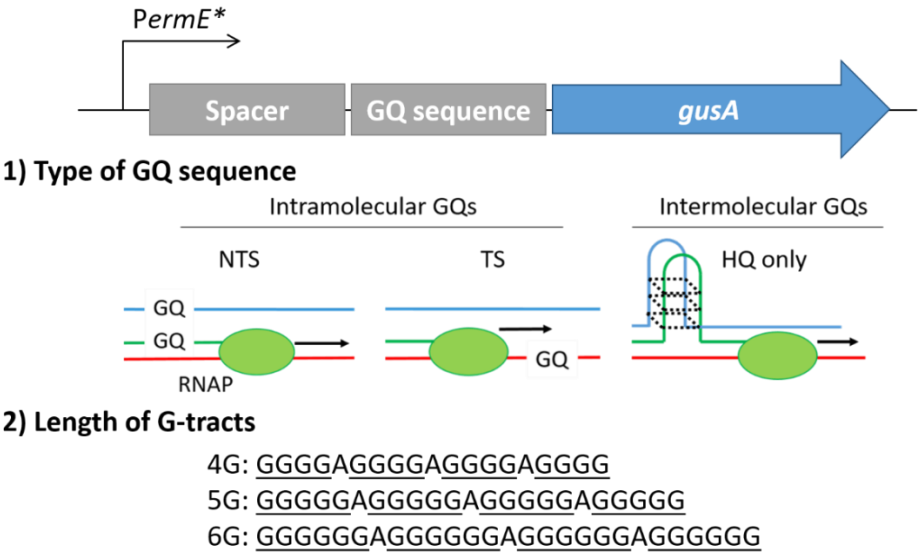

Figure 4.1: Construction of transcriptional reporters. All transcriptional reporters involved P*ermE** driving the expression of the reporter gene *gusA*. A variety of GQ sequence types and lengths were placed upstream of *gusA*, along with a 30 bp random spacer sequence inserted between the promoter and the GQ sequence. In the schematics representing the different types of GQ sequences, the blue strand represents the NTS of DNA, the red strand represents the TS of DNA, and the green strand represents the nascent RNA.

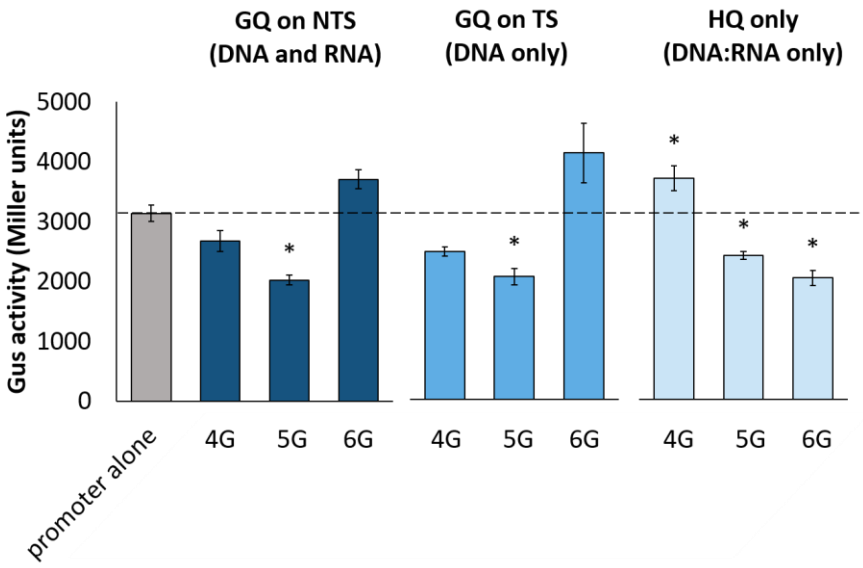

Figure 4.2: Transcriptional reporters to detect the effects of various GQ sequences on gene expression. Gus activity for reporter strains grown for 16 h in MYM liquid culture. All constructs contained a 30 bp spacer sequence between the promoter and the GQ sequence. Dark blue bars (left) represent GQ sequences on the NTS of DNA, medium blue bars represent GQs on the TS of DNA (centre), and light blue bars show sequences that can only form DNA:RNA HQs (right). The x-axis indicates the length of the G-tract in each GQ sequences. Statistical significance was determined using a two-way analysis of variance (ANOVA) Tukey honest significant difference (HSD) test for calculating all pairwise comparisons and is denoted for all relevant comparisons by asterisks which indicate $p<0.05$ compared to the promoter alone control. Error bars represent standard error of the mean (SEM) between three biological replicates.

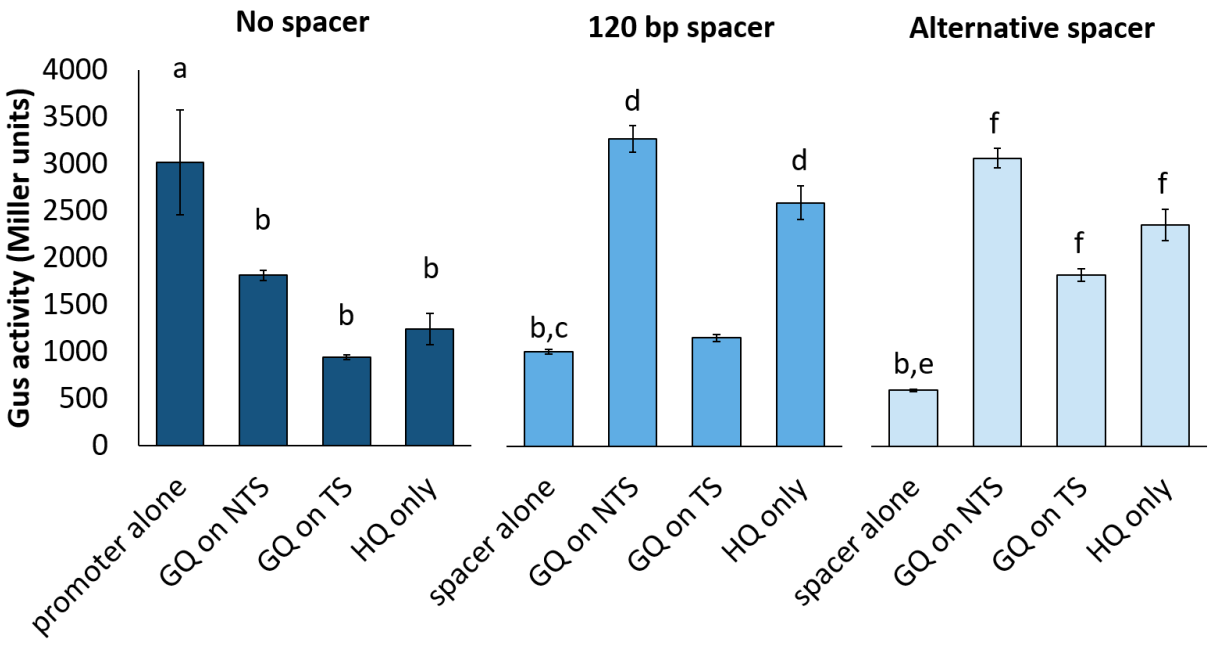

Figure 4.3: Effect of varying spacer length on Gus reporter activity. Gus activity for reporter strains grown in MYM liquid medium for 16 h. The 5G GQ constructs from Figure 4.2 were tested with no spacer, a 120 bp spacer, and an alternative spacer of approximately equal length. Dark blue bars (left) show the no spacer constructs, medium blue bars (middle) depict the 120 bp spacer constructs, and light blue bars (right) show the alternative (also 120 bp) spacer constructs. Letters denote statistical significance as follows: b is significantly different from a, d is significantly different from c, and f is significantly different from e. Statistical significance (p<0.05) was determined using a two-way ANOVA with Tukey HSD test for calculating all pairwise comparisons; however, only the biologically relevant comparisons are shown. Error bars represent SEM for three biological replicates.

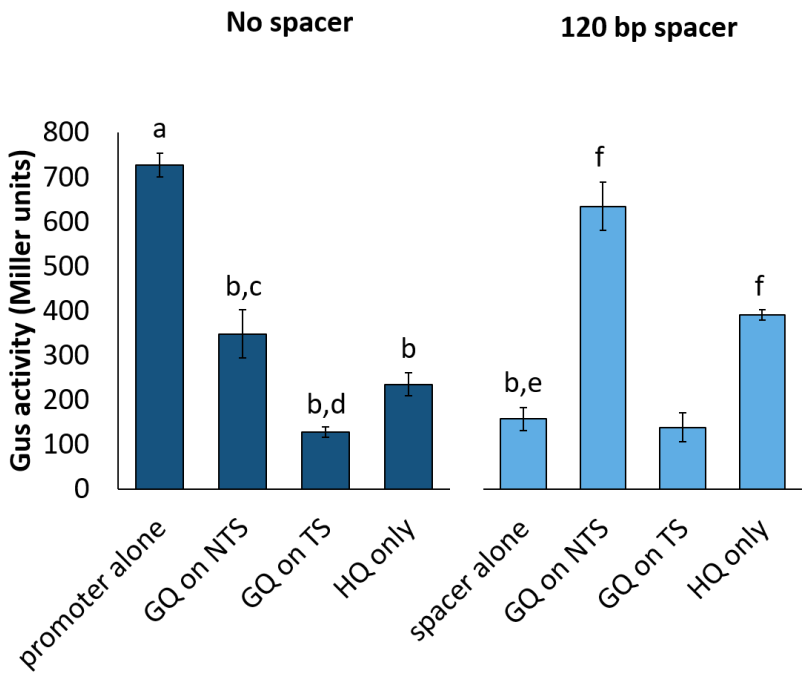

Figure 4.4: Effects of GQ sequences on Gus activity in *S. coelicolor*. The same reporter constructs as in the first two panels of Figure 4.3 were transferred into *S. coelicolor*. Assays were performed on cultures grown for 24 h in liquid YEME/TSB medium. Dark blue bars (left) represent the no spacer constructs, medium blue bars represent 120 bp spacer constructs (right). Letters denote statistical significance as follows: b is significantly different from a, d is significantly different from c, and f is significantly different from e. Statistical significance (p<0.05) was determined using a two-way ANOVA with Tukey HSD test for calculating all pairwise comparisons; only the biologically relevant comparisons are shown. Error bars represent SEM for three biological replicates.

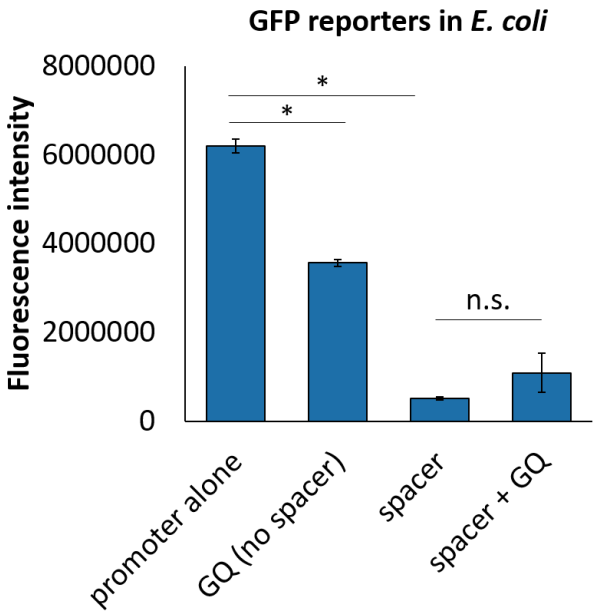

Figure 4.5: *E. coli* GFP reporter assays. *Streptomyces* reporter constructs were recapitulated in an *E. coli* plasmid vector, replacing *gusA* with *gfp* as the reporter gene. Gene expression was tested in *E. coli* using GFP fluorescence intensity normalized to cell density (OD_{600}) as a readout. These constructs included the promoter alone, GQ on the with no spacer, 120 bp spacer alone, and 120 bp spacer with GQ on the NTS. Statistical significance was determined using an ANOVA with Tukey's HSD for all pairwise comparisons. Asterisks indicate $p<0.05$; error bars represent SEM for three biological replicates.

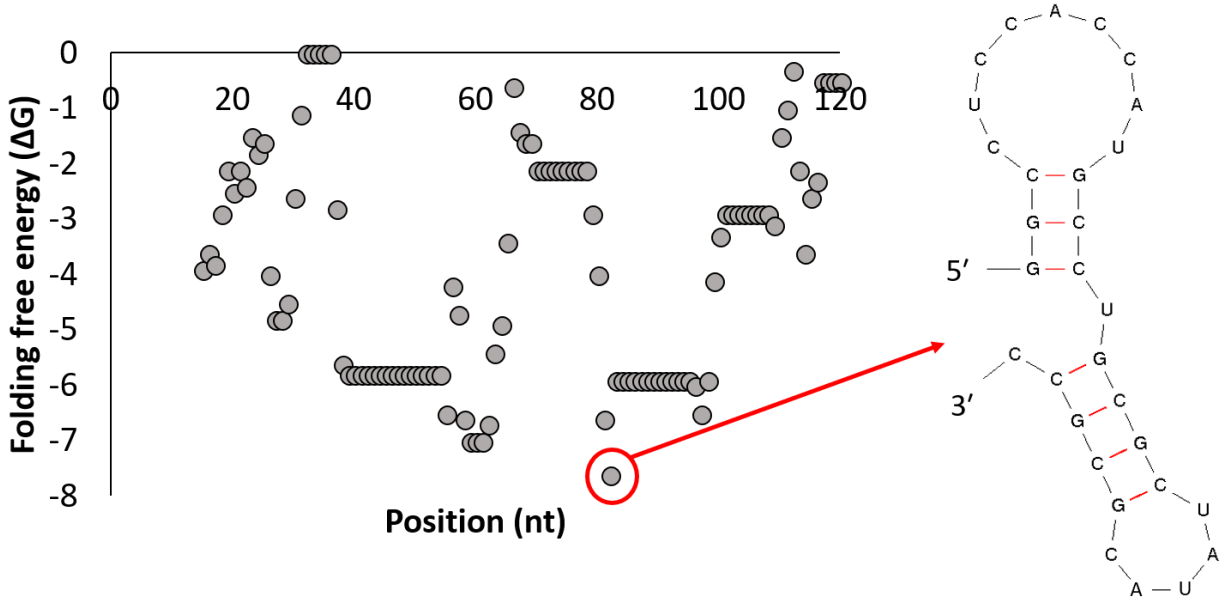

Figure 4.6: Determination of secondary structures in the 120 nt spacer sequence. Folding free energy, ΔG, was calculated for every 30 nt window in the 120 nt spacer sequence (left). The most stable structure (red circle) was visualized with the use of mFold web server (right).

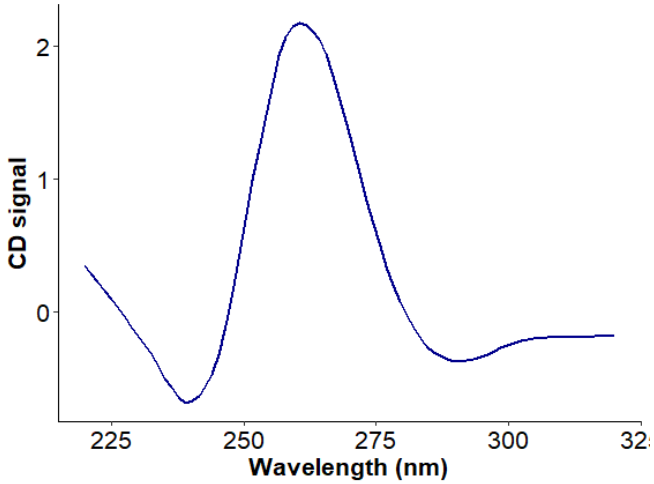

Figure 4.7: CD of 5G GQ sequence. CD was performed on 5G GQ sequence in 100 mM KCl, 50 mM Tris buffer. CD was performed at 25°C from 220 to 320 nm in 1 nm increments.

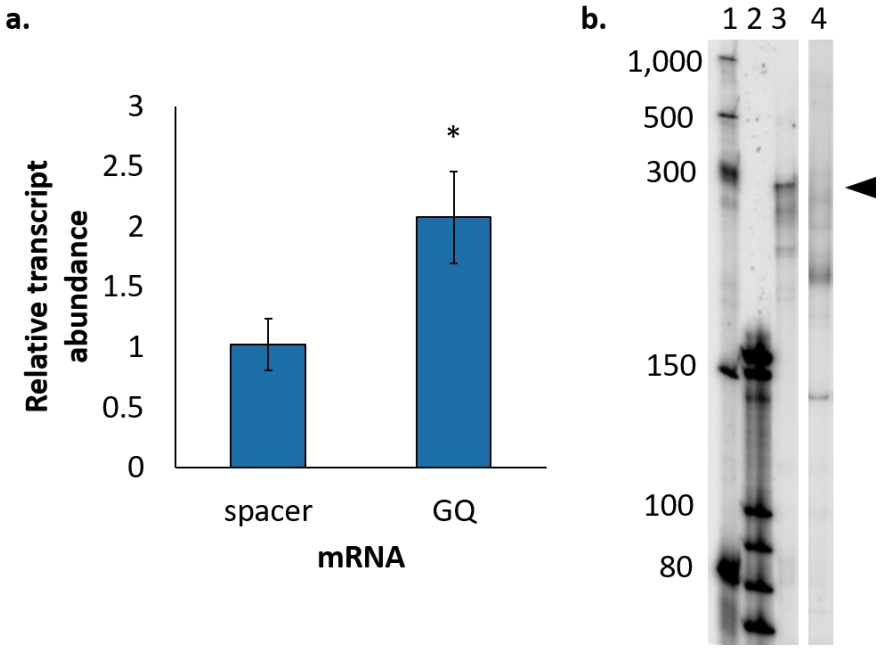

Figure 4.8: Transcriptional analysis of GQ-mediated effects on gene expression. a. RT-qPCR analysis of transcript abundance for the spacer alone and spacer with GQ mRNAs. Results were from three biological replicates and values were normalized to the *rpoB* transcript. Statistical significance was determined using t-test (p=0.027) and is represented by an asterisk. Error bars represent SEM. **b.** *In vitro* transcription of the spacer alone and spacer with GQ templates using *E. coli* RNA polymerase. Lanes 1 and 2: RNA ladders (nt length is shown to the left); lane 3: spacer alone reaction; lane 4: spacer + GQ reaction. The black arrow indicates the location of the expected full-length transcript.

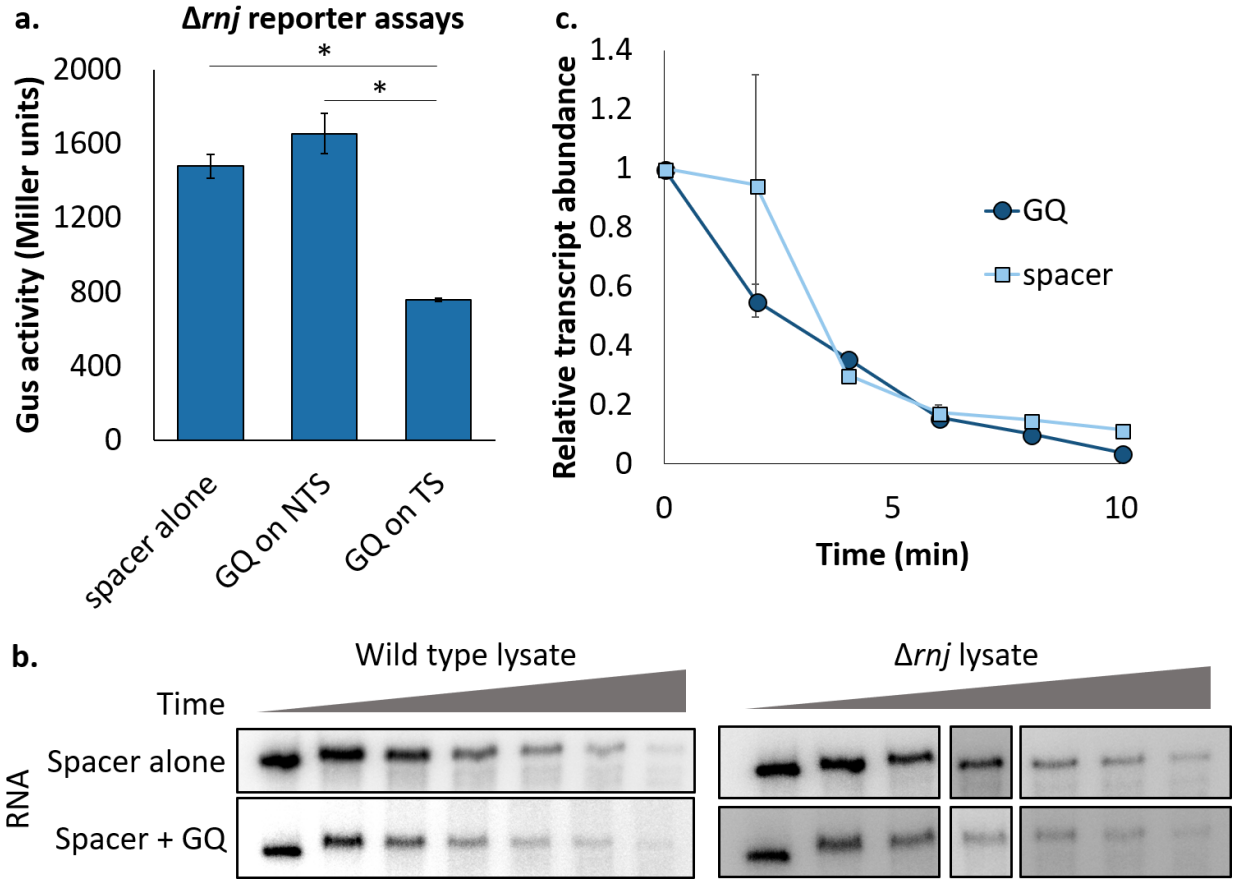

Figure 4.9: Effects of GQ sequences on mRNA stability. a. Gus reporter assays in *S. venezuelae* Δ*rnj* mutant background with the 120 nt spacer constructs. Statistical significance was determined using an ANOVA. Error bars represent SEM between three biological replicates. **b.** *In vitro* RNA stability assays from 0 to 90 min on *in vitro* synthesized RNAs with 3'-end labeling. **c.** *In vivo* RNA stability assays on wild type *S. venezuelae* strains containing the 120 nt spacer alone (spacer) and spacer with GQ (GQ) constructs. Shown is the average of two biological replicates with values normalized to the 5S rRNA transcript through RT-qPCR. Error bars show standard deviation.

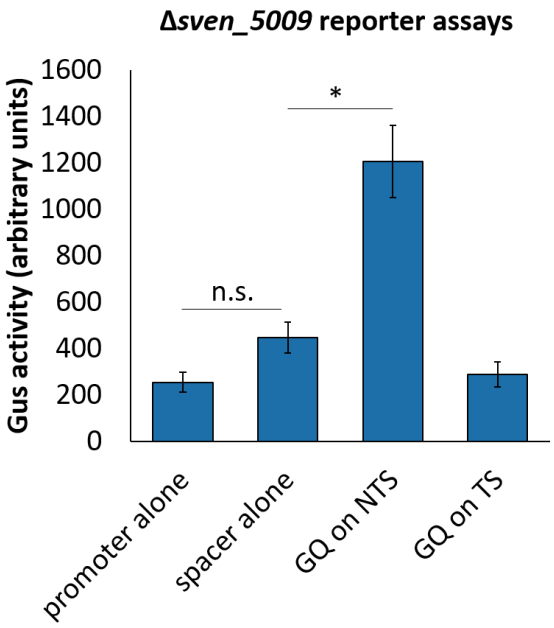

Figure 4.10: Gus reporter assay in Δ*sven_5009* background. The reporter constructs with the 120 bp spacer sequence were tested in the Rho deletion (Δ*sven_5009*) background. The averages of three biological replicates are shown, with error bars representing SEM. Statistical significance was determined using an ANOVA. Asterisk indicates $p<0.05$ for relevant comparisons and n.s. indicates that the difference is not statistically significant. Note that these values cannot be compared to the values of previous Gus assays because they were normalized to dry weight instead of cell density (OD_{600}).

5. Identifying novel GQ-binding proteins

5.1 Introduction

Several proteins have been identified that can either stabilize GQ structures or disrupt them. These proteins can be conformation-specific (recognizing only parallel or anti-parallel GQs) or display structure specificity (binding GQ structures generally over duplex DNA) (76). Most of the work on GQ-interacting proteins has been done in eukaryotes, with the only characterized GQ-protein interactions in bacteria being helicases required for their unwinding (77, 78). In eukaryotes, the known functions of GQ-interacting proteins are much broader with important roles in telomere function and gene expression. A comprehensive database, G4IPDB, has been compiled containing all known GQ-interacting proteins (79).

In human cells, the protein POT1 specifically resolves telomeric GQ structures, allowing for telomere elongation by the telomerase enzyme (80). Other human telomeric proteins bind GQ sequences as part of the shelterin complex that protects the ends of telomeres from DNA damage (76). DNA GQ-stabilizing proteins, such as PARP1 and nucleolin, can bind to the promoter-associated GQ sequences of oncogenes, preventing their expression (81). GQ-specific helicases and nucleases have also been identified, several of which have been associated with human diseases, including the FANCJ, WRN, and BLM helicases (as mentioned in Chapter 1). In addition, RNA-specific GQ-interacting proteins have been discovered in eukaryotes. These include RNA GQ helicases (78), ribosomal proteins (82), and RNA GQ-stabilizing proteins (78). Some RNA GQ-stabilizing proteins, such as the protein FMRP2 which is involved in Fragile X syndrome, inhibit translation when they bind 5’ UTR GQs (83).

Given the number of GQ sequences in *Streptomyces*, there may be many proteins with the ability to bind GQs. Here we used a biotin-based pulldown strategy involving a GQ DNA probe to specifically pull out GQ-binding proteins, followed by MS analysis to identify any binding proteins. Since relatively little is known about GQ-interacting proteins in bacteria, these findings have the potential to provide unique insights into the factors influencing GQs in bacteria.

5.2 Results

5.2.1 Identification of GQ binding proteins

To identify *Streptomyces* proteins that specifically bound GQ structures, we used a biotin pulldown approach with wild type *S. venezuelae* protein lysate and a 5’-biotinylated probe with the ability to form GQ sequences. As a control, the same probe was used with two point mutations that prevented GQ formation. To confirm that the GQ probe was indeed capable of forming a GQ structure, and to ensure that the negative control probe did not, we performed CD analysis on both probes (Figure 5.1). After

ensuring that they both behaved as expected, *S. venezuelae* lysate was incubated with one or the other of the two biotinylated probes, after which streptavidin magnetic beads were used to specifically pull out any proteins interacting with either probe. The protein was then eluted from the DNA with a salt concentration gradient, and the elutions were separated on a polyacrylamide gel and visualized using silver staining. We identified a protein of approximately 30 kDa that was more intense in the samples associated with the GQ probe than in the ones associated with the control probe (Figure 5.2). We excised this protein from the gel and performed MS to identify it. Upon repeating this experiment using independent protein lysates, we identified several proteins that were common to both pull-down experiments (Table 5.1). Of these proteins, SVEN_2656 and SVEN_3866 were two of the proteins with the highest coverage in both experiments and both had predicted RNA binding domains. It is therefore conceivable that they could also bind DNA, given that some GQ-interacting proteins bind both DNA and RNA GQs (76). In prioritizing proteins for follow-up investigation, we focused on these two proteins. SVEN_2656, also known as RNase PH, has been implicated in the degradation of structured transcripts in *E. coli* (84), while SVEN_3866, or TrmB, functions to modify the guanine-N7 position of tRNA (85). Importantly, this N7 position of guanine is critical for GQ formation.

5.2.2 Validating MS hits

To begin validating the *in vitro* interaction of SVEN_2656 and SVEN_3866 with GQ sequences, we overexpressed and purified both proteins from *E. coli* (Figure 5.3). We then performed EMSAs using GQ and mutant (negative control) sequences as probes. Preliminary EMSAs revealed a potential shift of the GQ probe when incubated with increasing concentrations of both proteins (Figure 5.4). However, we were unable to detect the control probe on a gel, and were therefore unable to draw any conclusions about the specificity of the protein-GQ DNA interaction at this point.

5.3 Conclusions

We have identified several proteins with the potential to interact with GQ structures specifically *in vitro*. The two prioritized proteins have not previously been identified as GQ-binding proteins, and are of particular interest given their broad conservation and their demonstrated ability to bind nucleic acids. We speculate that SVEN_2656 may have a role in degrading transcripts with RNA GQs, while TrmB could methylate DNA or RNA GQ structures, preventing their interference with cellular processes.

5.4 Figures

GQ probe: Bio–TATATAGGGAGGGCGGGAGGG

Control probe: Bio–TATATAGGGAGAGCGAGAGGG

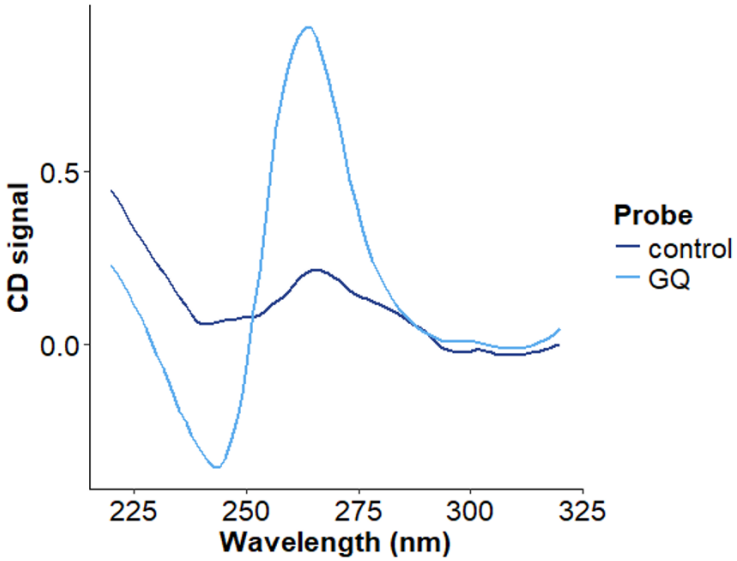

Figure 5.1: CD on control and GQ probes for biotin pulldown. CD was performed on GQ and control (mutated GQ) oligonucleotides in 100 mM KCl, 50 mM Tris buffer. CD was performed at 25°C from 220 to 320 nm in 1 nm increments. Probe sequences are indicated above the chart with G-tracts participating in GQ underlined and asterisks indicating the point mutations that were introduced into the second probe.

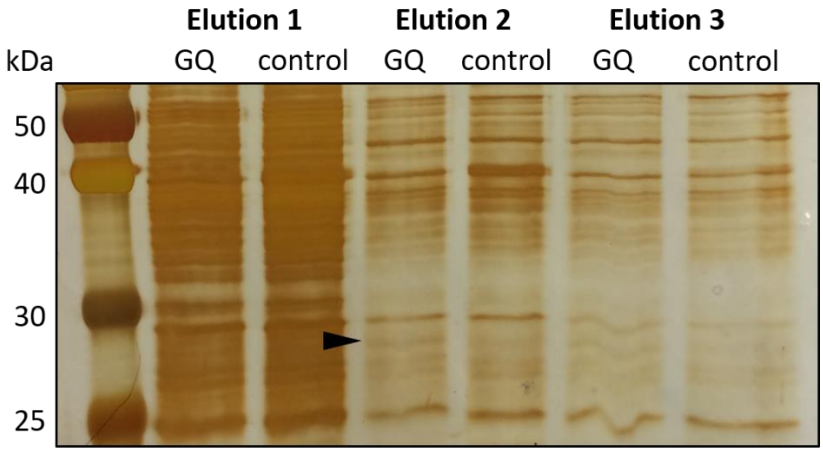

Figure 5.2: Elutions from biotin pulldowns. Biotin pulldowns on *S. venezuelae* lysate were performed using GQ and control 5'-biotinylated probes. Proteins were eluted from DNA with washes of increasing salt concentrations. After silver staining, the band indicated by the black arrowhead above was consistently observed to be more intense in the GQ lane than in the control lane, and was therefore excised and sent for MS analysis.

Table 5.1: Proteins identified through MS analysis of biotin pulldown samples. Proteins that were identified to be more abundant in the GQ samples than in the control sample in two independent rounds of biotin pulldown/MS analysis. Average % coverage refers to the average protein sequence coverage of each protein between the two experiments.

SVEN number	Annotation	Average % coverage
SVEN_2656	**Ribonuclease PH**	**43.5**
SVEN_1574	Triosephosphate isomerase	37.5
SVEN_3866	**tRNA (guanine-N(7)-)-methyltransferase**	**24.0**
SVEN_2300	Isoprenyl transferase	23.8
SVEN_4449	Putative dehydrogenase	19.6
SVEN_5303	30S ribosomal protein S2	14.4
SVEN_1843	Dihydrolipoamide acetyltransferase component of pyruvate dehydrogenase complex	8.9
SVEN_4524	Transcriptional regulator, TetR family	7.7
SVEN_5360	Translation initiation factor IF-2	2.1

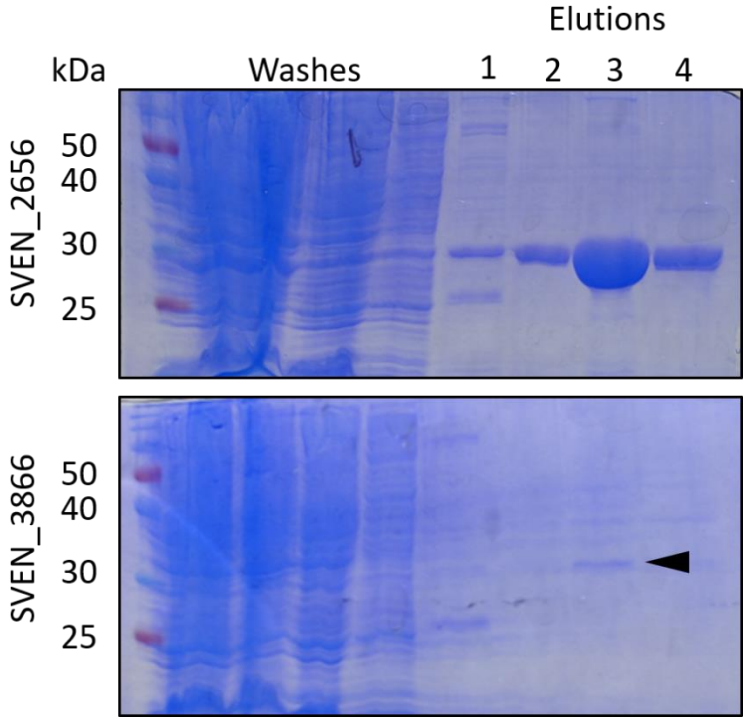

Figure 5.3: Purification of SVEN_2656 and SVEN_3866 from *E. coli*. SVEN_2656 and SVEN_3866 were overexpressed in *E. coli* Rosetta2 cells and purified by His-tag purification. Bound proteins were washed, and then eluted from the column with increasing concentrations of imidazole-containing buffer (ranging from 250 mM to 2 M).

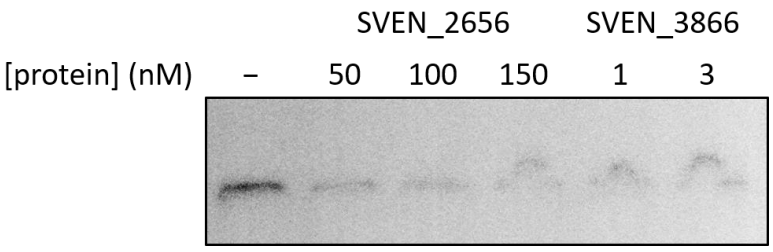

Figure 5.4: EMSA testing the binding of SVEN_2656 and SVEN_3866 to GQ DNA probe. Two of the proteins identified through MS analysis, SVEN_2656 and SVEN_3866, were overexpressed and purified from *E. coli*, then used in an EMSA to confirm the GQ DNA-protein interaction. The DNA probe was the GQ probe used for the biotin pulldowns.

6. Discussion, conclusions, and future directions

In this study, we conducted an extensive *in silico* analysis of the potential GQ sequences in the *S. venezuelae* genome. We found that GQ sequences were highly abundant in *Streptomyces* species and that many were located in putative regulatory regions. We then followed this up with an experimental analysis of the potential regulatory functions of GQ structures in *Streptomyces* and focused on elucidating the mechanism that resulted in a large increase in gene expression in the presence of a GQ sequence. Finally, we searched for potential GQ-interacting proteins, and identified two candidates that we prioritized for further study.

6.1 *in silico* analysis of GQ sequences

Our *in silico* analysis of GQ sequences in *Streptomyces* genomes made use of the consensus sequence: $G_3N_{1-7}G_3N_{1-7}G_3N_{1-7}G_3$ to identify putative GQ sequences. While it is known that this motif is not representative of the full range of GQ-forming sequences, it was a necessary starting point that allowed us to lay the foundations for this research. The only other analysis of GQ sequences in *Streptomyces* was done in *S. coelicolor* alongside 17 other species, and did not take an in-depth look at *Streptomyces* genomes specifically (35).

We found that there was an abundance (almost 3,000 in *S. venezuelae*) of GQ sequences in these organisms and that represented more than would be expected by chance alone based on our genome shuffling experiment (Figure 3.1). This finding was important because with increasing GC content, the likelihood that these sequences would arise by chance would increase. Our data revealed that these sequences were enriched in *Streptomyces* genomes, even when taking their extremely high GC content into consideration. This indicated that there was selective pressure to maintain them, presumably due to the biological functions they serve. While most GQ sequences were located within genes, we found that more than 20% were located in intergenic regions (Table 3.1) – double what we would have expected. Since most regulatory elements are found in between genes, this meant that GQ sequences were probably enriched in regulatory regions. This was consistent with previous findings of GQ sequence enrichment in regulatory regions of other organisms (35, 65, 68).

When we looked at the genome-wide distribution of these sequences, we found an unexpected enrichment on the lagging strand of DNA synthesis for approximately half of the species we tested (Figure 3.2, Table 3.2). There are several ways of interpreting these results, although experimental data is ultimately needed to determine which if any of these interpretations appears to be the case. It is possible that GQ sequences could affect DNA replication either positively or negatively, given the strand-

based asymmetry that was centered around the origin of replication. Since it is unclear whether our results indicated a depletion of GQ sequences on one strand or an enrichment on the other strand, there are two ways of thinking about these results: (1) GQ sequences were depleted on the leading strand of DNA (template for lagging strand synthesis) because they negatively affect this strand during DNA replication; (2) GQ sequences were enriched on the lagging strand (template for leading strand synthesis) because they positively affect this strand during DNA replication. In terms of negative effects, GQ sequences can inhibit polymerase progression, which could explain why they were depleted on one strand (16). However, it has recently been shown that disruptions in DNA replication can also be beneficial since there is evidence that the lagging strand develops fewer errors due to the discontinuous nature of its replication (86). Given these observations, it is possible that the presence of GQ sequence on the template for leading strand synthesis results in more discontinuous, but more accurate, leading strand synthesis. There is also evidence that the reciprocal C-rich structures, i-motifs, are more of an impediment to DNA polymerase than GQ sequences (87). It is possible that GQ sequences are enriched on the template for leading strand synthesis because the formation of i-motifs would be more deleterious on this strand. The leading strand is also more C-rich (71), so another possibility is that GQ sequences simply arise more frequently on the opposite strand as a result of it being more G-rich.

This pattern is unlike anything that has been described in the literature, and so it is difficult to speculate on the purpose of it or even if it is biologically relevant in the absence of experimental data. The fact that the same pattern was observed for approximately half of the *Streptomyces* species assayed as well as some distantly related species – *M. tuberculosis* and *P. aeruginosa* – provided some support for this being a biologically meaningful observation.

Our analysis of GQ sequences in biosynthetic clusters revealed that they were slightly more abundant in biosynthetic clusters than in the rest of the *S. venezuelae* genome, with some species (*e.g. S. avermitilis* and *S. pristinaespiralis*) displaying drastic increases in the density of GQ sequences in specific antibiotic clusters (Figure 3.6). While we are unable to deduce any functional role for the enrichment of these sequences in biosynthetic clusters at this time, the fact that there were so many within them is interesting and warrants further investigation given the current antibiotic crisis. Researchers have long struggled with stimulating antibiotic production from cryptic clusters in *Streptomyces*, and further investigation into the potential regulatory roles of these structures in the context of antibiotic production may be part of the solution. The same way GQ-interacting ligands are being investigated as

cancer therapeutics, they could potentially be used to stimulate silent clusters in *Streptomyces* if we were to uncover a role for GQ sequences in regulating antibiotic production.

Given that no in-depth *in silico* analyses of GQ sequences in such GC-rich organisms has been conducted previously, the findings here are laying important groundwork for future investigations. There were similarities between our findings and what is known based on work in other organisms, but there were also important differences. We found an abundance of GQ sequences in putative regulatory regions, which had been previously observed in eukaryotes and in bacteria. However, the strand-based enrichment that we uncovered had never been reported, highlighting the importance of studying GC-rich organisms.

6.2 Regulatory roles of GQs in *Streptomyces*

One of the main goals of this study was to provide further insights into the mechanisms controlling gene expression in *Streptomyces* by investigating a novel regulatory element. While we have not yet established a definitive mechanism of action for GQs in *Streptomyces*, we showed that there were an abundance of sequences with the potential to form GQs and that many of them were enriched in putative regulatory regions. When we looked for GQ sequences within UTRs, we found many located in 3'UTRs, where they could be involved in transcription termination or post-transcriptional regulation. We also found several in 5' UTRs, in the beginning coding regions, as well as in close proximity to transcription start sites. This indicated that GQ sequences in the 5' regions of genes may have important regulatory functions, and our subsequent experiments have supported this proposal.

Possibly the most important takeaway from our transcriptional analyses was that GQ sequences could profoundly influence gene expression in *Streptomyces*. Our data showed varying effects on gene expression based on the type of GQ sequence and the position of the GQ sequence relative to the transcription start site/promoter. Some of these effects were consistent with what had been observed for other species. For example, we found that GQ sequences adjacent to the promoter had a negative effect on gene expression, similar to that observed previously in human cells (73) and in *E. coli* (27). We also determined that GQ sequences near the promoter had a less negative effect on gene expression when they were on the NTS than on the TS in *S. venezuelae* and *S. coelicolor* (Figure 4.3 and Figure 4.4). These results were consistent with our *in silico* analysis of GQ sequences in proximity to transcription start sites, which revealed that genes with GQs on the NTS tended to be more highly expressed than genes with GQs on the TS (Figure 3.3). While we only looked at GQ sequences downstream of the promoter in our reporter assays, it is conceivable that GQ sequences upstream of promoter regions

could have different, possibly positive, effects on gene expression. To investigate this further, we could leverage our *in silico* data to determine whether the trend is stronger for GQ sequences located upstream of the TSS compared to those that are downstream of it. Since our *in silico* results did not show a statistically significant difference based on the strand of the GQ sequence, it could be that narrowing down our analysis to specific types of GQ sequences or certain regions (upstream vs. downstream of the promoter) would give us more clear-cut results. There are, however, many factors influencing gene expression and GQ sequences, and if they play any role at all, they are only part of the puzzle.

Given the highly negative effect we observed in our reporter assays for GQ sequences directly adjacent to the promoter – especially for GQ sequences on the TS – and the large number of GQ sequences located in antibiotic clusters, it would be interesting to determine whether this is one of the mechanisms that functions to silence antibiotic clusters. Any genes in transcriptionally silent clusters would not have been included in our analysis of GQ sequences in proximity to TSSs since the gene needed to be transcribed in the experimental conditions in order for a TSS to have been observed. Further investigation of the GQ sequences in immediate upstream region of genes would be needed to determine whether there are many GQ sequences near the promoters of genes in these clusters. If GQ sequences do play a role in downregulating the expression of antibiotic clusters, mutating these sequences could lead to the upregulation of several clusters, particularly those with no known transcriptional repressors or those in which previous attempts to upregulate the cluster have failed.

We also discovered that the presence of a GQ sequence on the NTS of DNA could yield increased reporter activity (Figure 4.3). This finding was contrary to other findings in the literature that indicated that GQ sequences mostly have a negative effect on gene expression, especially mRNA GQs which have been shown to inhibit translation (23, 25, 27, 73). We showed that this increase in reporter activity was associated with a corresponding increase in mRNA levels (Figure 4.8), but it was not caused by differential mRNA stability (Figure 4.9) or inhibition of Rho-dependent termination activity (Figure 4.10). The mechanism through which the GQ sequence enhanced reporter activity remains unknown. The fact that it was partially restored in *E. coli* suggests that whatever factor(s) contributed to this increase in gene expression could be broadly conserved.

The variability of the effects of GQ sequences on gene expression emphasizes the versatility of these structures in regulating gene expression, with the capacity to act as both positive and negative regulators. However, few distinct mechanisms of action have been elucidated for these structures, and

when mechanisms are successfully established, they seem to be highly context-dependent (*e.g* the same GQ sequence had highly variable effects depending on its orientation and distance from the promoter in the current study, with similar variability being observed in other studies as well (27, 73)). The function of these structures remains enigmatic, but the current work highlights the extent of their regulatory potential.

6.3 Potential GQ-protein interactions

Given the importance and variety of interactions between GQ sequences and proteins in eukaryotes, and the fact that essentially nothing was known about these interactions in bacteria, we wanted to take an unbiased experimental approach to screening for potential GQ-interacting proteins. Through our analysis, we identified two candidates for further study: SVEN_2656, annotated as RNase PH, and SVEN_3866, annotated as TrmB. RNase PH is a 3'-5' exoribonuclease that is involved in the 3'-end processing of transcripts (60) and has also been implicated in the degradation of structured RNAs in *E. coli* (84). Given this role in the degradation of structured RNA, we propose that it may function to degrade RNA GQs in *Streptomyces*.

TrmB is a tRNA methyltransferase that modifies tRNAs at the N7 position of guanine (85). The N7 position is required to form G-quartets but is not involved in G-C base-pairing. If SVEN_3866 was able to bind and modify GQ DNA or RNA, its activity would prevent GQ formation by these sequences.

There is evidence that GQ structures can interact with methyltransferases *in vitro* (88), and multiple studies have shown that methylation in close proximity to or within GQ sequences can cause changes to the stability of the GQ structure (89–93). It is well-established that N7-methylguanine cannot participate in GQ formation, but that this modified base can participate in normal G-C base-pairing (94, 95). However, this has never been investigated as a way of preventing GQ-mediated effects on cellular processes. In bacteria, little is known about the way methylation marks affect gene expression, with the best-studied methylation marks being 6-methyladenine, 4-methylcytosine, and 5-methylcytosine (96). In humans, two methyltransferases have been described with the ability to bind both DNA and RNA, both of which have tRNAs as their RNA substrates (97). It is tantalizing to speculate that SVEN_3866 could bind DNA as well as RNA. It is also possible that its associated gene/protein product has been misannotated as a tRNA methyltransferase, when in fact its preferred substrate is DNA. The first step to investigating either of these roles for these two proteins will be to confirm their interaction with GQ sequences through EMSAs since our preliminary results were inconclusive due to the absence of successfully labelled control probes.

6.4 Conclusions

This work represents a comprehensive investigation into the regulatory functions of GQ sequences in a GC-rich bacterium. Our results have laid the foundation for better understanding these sequences in GC-rich bacteria, where these GQ sequences are highly abundant. Our experimental and *in silico* results indicate that GQs may have a wide range of regulatory impacts - some of which are consistent with observations in other organisms, while others have not previously been reported. The fact that we observed completely new effects for GQ sequences, underscores the diverse regulatory potential of these structures. Our GQ-protein interaction experiments led to the identification of several proteins that may function to alleviate potential negative consequences associated with GQ sequences. While we have not yet defined a specific mechanism of action underlying our observations, this work has provided a number of interesting directions for future investigation.

6.5 Future directions

One of the most exciting areas for further work is to follow up investigations into the potential GQ-interacting proteins that we identified. An important first step will be to validate these interactions with EMSAs – and both RNA and DNA targets – to ensure that they do indeed bind specifically to GQ structures. Once validated, these studies open the door to exploring the vastly understudied area of bacterial epigenetics.

In addition to better understanding the proteins that can associate with GQ sequences, it will also be important to determine which GQ sequences can form GQ structures *in vivo*. This question could be addressed using polymerase stop assays, either in a high-throughput manner to discover all possible GQ structures, or focusing on a subset of prioritized GQ sequences. A more sophisticated *in silico* analysis would also be a useful complement to this work, in assessing imperfections in the consensus motif, and accounting for the surrounding base composition.

Our reporter assays have provided us with strong evidence that GQ sequences affect gene expression in *Streptomyces*. We have taken steps toward establishing the mechanism of action for a specific example (GQs associated with the NTS, differentially positioned downstream of an active promoter), and to this point have eliminated several mechanistic possibilities. Future work could investigate the possibility that the increased expression in the presence of the GQ sequence was due to increased transcription, even though our *in vitro* transcription data contradict this hypothesis or the possibility that the increased reporter activity resulted from increased translation of the GQ mRNA.

Overall, this study has provided a foundational understanding of GQ-mediated gene regulation in *Streptomyces*. This is a completely unexplored area that has the potential to influence our understanding of gene regulation in these bacteria, and in other GC-rich microbes.

References

1. Agashe,D. and Shankar,N. (2014) The evolution of bacterial DNA base composition. *J. Exp. Zool. Part B Mol. Dev. Evol.*, **322**, 517–528.

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

3. Newton-Foot,M. and Gey Van Pittius,N.C. (2013) The complex architecture of mycobacterial promoters. *Tuberculosis*, **93**, 60–74.

4. You,J., Li,H., Lu,X.-M., Li,W., Wang,P.-Y., Dou,S.-X. and Xi,X.-G. (2017) Effects of monovalent cations on folding kinetics of G-quadruplexes. *Biosci. Rep.*, **37**, BSR20170771.

5. Zheng,K., Xiao,S., Liu,J., Zhang,J., Hao,Y. and Tan,Z. (2013) Co-transcriptional formation of DNA:RNA hybrid G-quadruplex and potential function as constitutional cis element for transcription control. *Nucleic Acids Res.*, **41**, 5533–5541.

6. Zheng,K., Wu,R., He,Y., Xiao,S., Zhang,J., Liu,J., Hao,Y. and Tan,Z. (2014) A competitive formation of DNA:RNA hybrid G-quadruplex is responsible to the mitochondrial transcription termination at the DNA replication priming site. *Nucleic Acids Res.*, **42**, 10832–10844.

7. Wu,R.Y., Zheng,K.W., Zhang,J.Y., Hao,Y.H. and Tan,Z. (2015) Formation of DNA:RNA hybrid G-quadruplex in bacterial cells and its dominance over the intramolecular DNA G-quadruplex in mediating transcription termination. *Angew. Chemie - Int. Ed.*, **54**, 2447–2451.

8. Xiao,S., Zhang,J.Y., Wu,J., Wu,R.Y., Xia,Y., Zheng,K.W., Hao,Y.H., Zhou,X. and Tan,Z. (2014) Formation of DNA:RNA hybrid G-quadruplexes of two G-quartet layers in transcription: Expansion of the prevalence and diversity of G-quadruplexes in genomes. *Angew. Chemie - Int. Ed.*, **53**, 13110–13114.

9. Ambrus,A., Chen,D., Dai,J., Bialis,T., Jones,R.A. and Yang,D. (2006) Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution. *Nucleic Acids Res.*, **34**, 2723–2735.

10. Rhodes,D. and Lipps,H.J. (2015) G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.*, **43**, 8627–8637.

11. Bochman,M.L., Paeschke,K. and Zakian,V.A. (2012) DNA secondary structures: Stability and function of G-quadruplex structures. *Nat. Rev. Genet.*, **13**, 770–780.

12. Zahler,A.M., Williamson,J.R., Cech,T.R. and Prescott,D.M. (1991) Inhibition of telomerase by G-quartet DNA structures. *Nature*, **350**, 718–720.

13. Tauchi,T., Shin-Ya,K., Sashida,G., Sumi,M., Okabe,S., Ohyashiki,J.H. and Ohyashiki,K. (2006) Telomerase inhibition with a novel G-quadruplex-interactive agent, telomestatin: *in vitro* and *in vivo* studies in acute leukemia. *Oncogene*, **25**, 5719–5725.

14. Brooks,T.A., Kendrick,S. and Hurley,L. (2010) Making sense of G-quadruplex and i-motif functions in oncogene promoters. *FEBS J.*, **277**, 3459–3469.

15. Chen,B., Wu,Y., Tanaka,Y. and Zhang,W. (2014) Small molecules targeting c-Myc oncogene: Promising anti-cancer therapeutics. *Int. J. Biol. Sci.*, **10**, 1084–1096.

16. Lopes,J., Piazza,A., Bermejo,R., Kriegsman,B., Colosio,A., Teulade-Fichou,M.-P., Foiani,M. and Nicolas,A. (2011) G-quadruplex-induced instability during leading-strand replication. *EMBO J.*, **30**, 4033–4046.

17. London,T.B.C., Barber,L.J., Mosedale,G., Kelly,G.P., Balasubramanian,S., Hickson,I.D., Boulton,S.J. and Hiom,K. (2008) FANCJ is a structure-specific DNA helicase associated with the maintenance of genomic G/C tracts. *J. Biol. Chem.*, **283**, 36132–36139.

18. Tang,W., Robles,A.I., Beyer,R.P., Gray,L.T., Nguyen,G.H., Oshima,J., Maizels,N., Harris,C.C. and Monnat,R.J. (2016) The Werner syndrome RECQ helicase targets G4 DNA in human cells to modulate transcription. *Hum. Mol. Genet.*, **25**, 2060–2069.

19. Li,J.L., Harrison,R.J., Reszka,A.P., Brosh,R.M., Bohr,V.A., Neidle,S. and Hickson,I.D. (2001) Inhibition of the Bloom's and Werner's syndrome helicases by G-quadruplex interacting ligands. *Biochemistry*, **40**, 15194–15202.

20. Valton,A.-L. and Prioleau,M.-N. (2016) G-Quadruplexes in DNA Replication: A Problem or a Necessity? *Trends Genet.*, **32**, 697–706.

21. Armas,P., David,A. and Calcaterra,N. (2016) Transcriptional control by G-quadruplexes: *in vivo* roles and perspectives for specific intervention. *Transcription*, **8**, 21–25.

22. Agarwala,P., Pandey,S. and Maiti,S. (2015) The tale of RNA G-quadruplex. *Org. Biomol. Chem.*, **13**, 5570–5585.

23. Bugaut,A. and Balasubramanian,S. (2012) 5'-UTR RNA G-quadruplexes: Translation regulation and targeting. *Nucleic Acids Res.*, **40**, 4727–4741.

24. Wieland,M. and Hartig,J.S. (2009) Investigation of mRNA quadruplex formation in *Escherichia coli*. *Nat. Protoc.*, **4**, 1632–1640.

25. Endoh,T. and Sugimoto,N. (2016) Mechanical insights into ribosomal progression overcoming RNA G-quadruplex from periodical translation suppression in cells. *Sci. Rep.*, **6**, 22719.

26. Du,X., Wojtowicz,D., Bowers,A.A., Levens,D., Benham,C.J. and Przytycka,T.M. (2013) The genome-wide distribution of non-B DNA motifs is shaped by operon structure and suggests the transcriptional importance of non-B DNA structures in *Escherichia coli*. *Nucleic Acids Res.*, **41**, 5965–5977.

27. Holder,I.T. and Hartig,J.S. (2014) A matter of location: Influence of G-quadruplexes on *Escherichia coli* gene expression. *Chem. Biol.*, **21**, 1511–1521.

28. Kota,S., Dhamodharan,V., Pradeepkumar,P.I. and Misra,H.S. (2015) G-quadruplex forming structural motifs in the genome of *Deinococcus radiodurans* and their regulatory roles in promoter functions. *Appl. Microbiol. Biotechnol.*, **99**, 9761–9769.

29. Perrone,R., Lavezzo,E., Riello,E., Manganelli,R., Palù,G., Toppo,S., Provvedi,R. and Richter,S.N. (2017) Mapping and characterization of G-quadruplexes in *Mycobacterium tuberculosis* gene promoter regions. *Sci. Rep.*, **7**, 5743.

30. Bedrat,A., Lacroix,L. and Mergny,J.L. (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.*, **44**, 1746–1759.

31. Kikin,O., Antonio,L.D. and Bagga,P.S. (2018) QGRS Mapper: A web-based server for predicting G-

quadruplexes in nucleotide sequences. *Nucleic Acids Res.*, **34**, 676–682.

32. Yadav,V. kumar, Abraham,J.K., Mani,P., Kulshrestha,R. and Chowdhury,S. (2008) QuadBase: Genome-wide database of G4 DNA - Occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res.*, **36**, 381–385.

33. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.

34. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.

35. Rawal,P., Kummarasetti,V.B.R., Ravindran,J., Kumar,N., Halder,K., Sharma,R., Mukerji,M., Das,S.K. and Chowdhury,S. (2006) Genome-wide prediction of G4 DNA as regulatory motifs: Role in *Escherichia coli* global regulation. *Genome Res.*, **16**, 644–655.

36. Chambers,V.S., Marsico,G., Boutell,J.M., Di Antonio,M., Smith,G.P. and Balasubramanian,S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.*, **33**, 1–7.

37. Zendulka,J. and Lexa,M. (2018) Sequence analysis pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics*, **33**, 3373–3379.

38. Kwok,C.K. and Merrick,C.J. (2017) G-Quadruplexes: Prediction, characterization, and biological application. *Trends Biotechnol.*, **35**, 997–1013.

39. Paramasivan,S., Rujan,I. and Bolton,P.H. (2007) Circular dichroism of quadruplex DNAs: Applications to structure, cation effects and ligand binding. *Methods*, **43**, 324–331.

40. Satpathi,S., Kulkarni,M., Mukherjee,A. and Hazra,P. (2016) Ionic liquid induced G-quadruplex formation and stabilization: spectroscopic and simulation studies. *Phys. Chem. Chem. Phys.*, **18**, 29740–29746.

41. Shrestha,P., Jonchhe,S., Emura,T., Hidaka,K., Endo,M., Sugiyama,H. and Mao,H. (2017) Confined space facilitates G-quadruplex formation. *Nat. Nanotechnol.*, **12**, 582–588.

42. Shrestha,P., Xiao,S., Dhakal,S., Tan,Z. and Mao,H. (2014) Nascent RNA transcripts facilitate the formation of G-quadruplexes. *Nucleic Acids Res.*, **42**, 7236–46.

43. Liu,H.-Y., Zhao,Q., Zhang,T.-P., Wu,Y., Xiong,Y.-X., Wang,S.-K., Ge,Y.-L., He,J.-H., Lv,P., Ou,T.-M., *et al.* (2016) Conformation selective antibody enables genome profiling and leads to discovery of parallel G-quadruplex in human telomeres. *Cell Chem. Biol.*, **23**, 1261–1270.

44. Fernando,H., Rodriguez,R. and Balasubramanian,S. (2008) Selective recognition of a DNA G-quadruplex by an engineered antibody. *Biochemistry*, **47**, 9365–9371.

45. Biffi,G., Tannahill,D., Mccafferty,J. and Balasubramanian,S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182–186.

46. Lam,E.Y.N., Beraldi,D., Tannahill,D. and Balasubramanian,S. (2013) G-quadruplex structures are stable and detectable in human genomic DNA. *Nat. Commun.*, **4**, 1796.

47. Sun,D. and Hurley,L.H. (2010) Biochemical techniques for the characterization of G-quadruplex

structures: EMSA, DMS footprinting, and DNA polymerase stop assay. *Methods Mol. Biol.*, **608**, 65–79.

48. Kwok,C.K., Marsico,G., Sahakyan,A.B., Chambers,V.S. and Balasubramanian,S. (2016) rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat. Methods*, **13**, 841–844.

49. Kwok,C.K., Tang,Y., Assmann,S.M. and Bevilacqua,P.C. (2015) The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends Biochem. Sci.*, **40**, 221–232.

50. Guo,J.U. and Bartel,D.P. (2016) RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science (80-.).*, **353**, 5371–5371.

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

52. de Lima Procópio,R.E., da Silva,I.R., Martins,M.K., de Azevedo,J.L. and de Araújo,J.M. (2012) Antibiotics produced by *Streptomyces*. *Brazilian J. Infect. Dis.*, **16**, 466–471.

53. Chaudhary,A.K., Dhakal,D. and Sohng,J.K. (2013) An insight into the "-omics" based engineering of streptomycetes for secondary metabolite overproduction. *Biomed Res. Int.*, **2013**, 1–15.

54. Rutledge,P.J. and Challis,G.L. (2015) Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nat. Rev. Microbiol.*, **13**, 509–523.

55. Jiang,M., Anderson,J., Gillespie,J. and Mayne,M. (2008) uShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, **9**, 192–202.

56. 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 fo the sesquiterpene soil odor geosmin. *PNAS*, **100**, 1541–1546.

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

58. St-Onge,R.J. and Elliot,M.A. (2017) Regulation of a muralytic enzyme-encoding gene by two non-coding RNAs. *RNA Biol.*, **14**, 1592–1605.

59. Bradford,M.M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.*, **72**, 248–254.

60. Kelly,K.O., Reuven,N.B., Li,Z.W. and Deutscher,M.P. (1992) RNase PH is essential for transfer RNA processing and viability in RNase-deficient *Escherichia coli* cells. *J Biol Chem*, **267**, 16015–16018.

61. Liu,Q., Gao,Y., Yang,W., Zhou,H., Gao,Y., Zhang,X., Teng,M. and Niu,L. (2008) Crystallization and preliminary crystallographic analysis of tRNA (m7G46) methyltransferase from *Escherichia coli*. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, **64**, 743–745.

62. MacNeil,D.J., Gewain,K.M., Ruby,C.L., Dezeny,G., Gibbons,P.H. and MacNeil,T. (1992) Analysis of *Streptomyces avermitilis* genes required for avermectin biosynthesis utilizing a novel integration vector. *Gene*, **111**, 61–68.

63. Bibb,M.J., Domonkos,Á., Chandra,G. and Buttner,M.J. (2012) Expression of the chaplin and rodlin hydrophobic sheath proteins in *Streptomyces venezuelae* is controlled by σ BldN and a cognate

anti-sigma factor, RsbN. *Mol. Microbiol.*, **84**, 1033–1049.

64. Kieser,T., Bibb,M.J., Buttner,M.J., Chater,K.F. and Hopwood,D.A. (2000) Practical *Streptomyces* Genetics The John Innes Foundation, Norwich.

65. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.

66. Mullen,M.A., Olson,K.J., Dallaire,P., Assmann,S.M. and Bevilacqua,P.C. (2018) RNA G-Quadruplexes in the model plant species *Arabidopsis thaliana*: prevalence and possible functional roles. *Nucleic Acids Res.*, **38**, 8149–8163.

67. Garg,R., Aggarwal,J. and Thakkar,B. (2016) Genome-wide discovery of G-quadruplex forming sequences and their functional relevance in plants. *Sci. Rep.*, **6**, 31–35.

68. Hershman,S.G., Chen,Q., Lee,J.Y., Kozak,M.L., Yue,P., Wang,L. and Johnson,F.B. (2018) Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **36**, 144–156.

69. Biswas,B., Kandpal,M., Jauhari,U.K. and Vivekanandan,P. (2016) Genome-wide analysis of G-quadruplexes in herpesvirus genomes. *BMC Genomics*, **17**, 1–16.

70. Salvo,M. Di, Pinatel,E., Talà,A., Fondi,M., Peano,C. and Alifano,P. (2018) G4PromFinder: An algorithm for predicting transcription promoters in GC-rich bacterial genomes based on AT-rich elements and G-quadruplex motifs. *BMC Bioinformatics*, **19**, 1–11.

71. Bentley,S., Chater,K., 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.

72. Blin,K., Medema,M.H., Kazempour,D., Fischbach,M.A., Breitling,R., Takano,E. and Weber,T. (2013) antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.*, **41**, 204–212.

73. Agarwal,T., Roy,S., Kumar,S., Chakraborty,T.K. and Maiti,S. (2014) In the sense of transcription regulation by G-quadruplexes: Asymmetric effects in sense and antisense strands. *Biochemistry*, **53**, 3711–3718.

74. Luedtke,N.W. (2009) Targeting G-quadruplex DNA with small molecules. *Chim. Int. J. Chem.*, **63**, 134–139.

75. Wieland,M. and Hartig,J.S. (2007) RNA quadruplex-based modulation of gene expression. *Chem. Biol.*, **14**, 757–763.

76. Kaushik,M., Kaushik,S., Bansal,A., Saxena,S. and Kukreti,S. (2011) Structural diversity and specific recognition of four stranded G-quadruplex DNA. *Curr Mol Med*, **11**, 744–769.

77. Thakur,R.S., Desingu,A., Basavaraju,S., Subramanya,S., Rao,D.N. and Nagaraju,G. (2014) *Mycobacterium tuberculosis* DinG is a structure-specific helicase that unwinds G4 DNA: Implications for targeting G4 DNA as a novel therapeutic approach. *J. Biol. Chem.*, **289**, 25112–25136.

78. Wu,X. and Maizels,N. (2001) Substrate-specific inhibition of RecQ helicase. *Nucleic Acids Res.*, **29**, 1765–1771.

79. Mishra,S.K., Tawani,A., Mishra,A. and Kumar,A. (2016) G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci. Rep.*, **6**, 38144.

80. Zaug,A.J., Podell,E.R. and Cech,T.R. (2005) Human POT1 disrupts telomeric G-quadruplexes allowing telomerase extension *in vitro*. *Proc. Natl. Acad. Sci.*, **102**, 10864–10869.

81. Brázda,V., Hároníková,L., Liao,J. and Fojta,M. (2014) DNA and RNA quadruplex-binding proteins. *Int. J. Mol. Sci.*, **15**, 17493–17517.

82. Von Hacht,A., Seifert,O., Menger,M., Schütze,T., Arora,A., Konthur,Z., Neubauer,P., Wagner,A., Weise,C. and Kurreck,J. (2014) Identification and characterization of RNA guanine-quadruplex binding proteins. *Nucleic Acids Res.*, **42**, 6630–6644.

83. Bensaid,M., Melko,M., Bechara,E.G., Davidovic,L., Berretta,A., Catania,M.V., Gecz,J., Lalli,E. and Bardoni,B. (2009) FRAXE-associated mental retardation protein (FMR2) is an RNA-binding protein with high affinity for G-quartet RNA forming structure. *Nucleic Acids Res.*, **37**, 1269–1279.

84. Jain,C. (2012) Novel role for RNase PH in the degradation of structured RNA. *J. Bacteriol.*, **194**, 3883–3890.

85. Purta,E., Van Vliet,F., Tricot,C., De Bie,L.G., Feder,M., Skowronek,K., Droogmans,L. and Bujnicki,J.M. (2005) Sequence-structure-function relationships of a tRNA (m7G46) methyltransferase studied by homology modeling and site-directed mutagenesis. *Proteins Struct. Funct. Genet.*, **59**, 482–488.

86. Maslowska,K.H., Makiela-Dzbenska,K., Mo,J.-Y., Fijalkowska,I.J. and Schaaper,R.M. (2018) High-accuracy lagging-strand DNA replication mediated by DNA polymerase dissociation. *Proc. Natl. Acad. Sci.*, **115**, 4212–4217.

87. Takahashi,S., Brazier,J.A. and Sugimoto,N. (2017) Topological impact of noncanonical DNA structures on Klenow fragment of DNA polymerase. *Proc. Natl. Acad. Sci.*, **114**, 9605–9610.

88. Cree,S.L., Fredericks,R., Miller,A., Pearce,F.G., Filichev,V., Fee,C. and Kennedy,M.A. (2016) DNA G-quadruplexes show strong interaction with DNA methyltransferases *in vitro*. *FEBS Lett.*, **590**, 2870–2883.

89. Lin,J., Hou,J., Xiang,H., Yan,Y., Gu,Y., Tan,J., Li,D., Gu,L., Ou,T. and Huang,Z. (2013) Stabilization of G-quadruplex DNA by C-5-methyl-cytosine in *bcl-2* promoter: Implications for epigenetic regulation. *Biochem. Biophys. Res. Commun.*, **433**, 368–373.

90. François,M., Leifert,W., Tellam,R. and Fenech,M. (2015) G-quadruplexes: A possible epigenetic target for nutrition. *Mutat. Res. - Rev. Mutat. Res.*, **764**, 101–107.

91. Yoshida,W., Yoshioka,H., Bay,D.H., Iida,K., Ikebukuro,K., Nagasawa,K. and Karube,I. (2016) Detection of DNA methylation of G-quadruplex and i-motif-forming sequences by measuring the initial elongation efficiency of polymerase chain reaction. *Anal. Chem.*, **88**, 7101–7107.

92. Halder,R., Halder,K., Sharma,P., Garg,G., Sengupta,S. and Chowdhury,S. (2010) Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide. *Mol. Biosyst.*, **6**, 2439.

93. Zamiri,B., Mirceta,M., Bomsztyk,K., Macgregor,R.B. and Pearson,C.E. (2015) Quadruplex formation by both G-rich and C-rich DNA strands of the C9orf72 (GGGGCC)8•(GGCCCC)8 repeat: Effect of CpG methylation. *Nucleic Acids Res.*, **43**, 10055–10064.

94. Kou,Y., Koag,M.C. and Lee,S. (2015) N7 methylation alters hydrogen-bonding patterns of guanine in

duplex DNA. *J. Am. Chem. Soc.*, **137**, 14067–14070.

95. Ezaz-nikpay,K. and Verdine,G.L. (1994) The effects of N7-methylguanine on duplex DNA structure. *Chem. Biol.*, **1**, 235–240.

96. Sánchez-Romero,M.A., Cota,I. and Casadesús,J. (2015) DNA methylation in bacteria: From the methyl group to the methylome. *Curr. Opin. Microbiol.*, **25**, 9–16.

97. Rana,A.K. and Ankri,S. (2016) Reviving the RNA world: An insight into the appearance of RNA methyltransferases. *Front. Genet.*, **7**, 1–9.

Appendices

Appendix A: Python script for determining the number of GQ sequences in a given sequence file

```
#!/usr/bin/python
import sys, fileinput, re, csv, ushuffle
sequence = ""
file = fileinput.input()

for line in file:
    if line[0] == ">":
        title = line[1:]
    else:
        sequence = sequence + line
sequence = sequence.upper().replace("\n", "")

GQ_for = re.findall("GGG[ATGCN]{1,7}CCC[ATGCN]{1,7}GGG[ATGCN]{1,7}CCC",
sequence)
GQ_rev = re.findall("CCC[ATGCN]{1,7}GGG[ATGCN]{1,7}CCC[ATGCN]{1,7}GGG",
sequence)
GQ = GQ_for + GQ_rev
print(len(GQ))
```

Appendix B: Python script for determining the locations of GQ sequences in a given sequence file

```
#!/usr/bin/python
import sys, fileinput, re, csv, ushuffle
sequence = ""
file = fileinput.input()

for line in file:
    if line[0] == ">":
        title = line[1:]
    else:
        sequence = sequence + line
sequence = sequence.upper().replace("\n", "")

p1 = re.compile("GGG[ATGCN]{1,7}GGG[ATGCN]{1,7}GGG[ATGCN]{1,7}GGG")
p2 = re.compile("CCC[ATGCN]{1,7}CCC[ATGCN]{1,7}CCC[ATGCN]{1,7}CCC")
shuff = ushuffle.shuffle(sequence, len(sequence), 6)

with open (<output file name>, 'wb') as file:
    writer = csv.writer(file)
    writer.writerow(['Chromosome', 'Start', 'End', 'Strand'])

for m in p1.finditer(sequence):
    with open(<output file name>, 'a') as file:
        writer = csv.writer(file)
        writer.writerow(['chr1', m.start(), m.end(), '+'])
for m in p2.finditer(sequence):
    with open(<output file name>, 'a') as file:
        writer = csv.writer(file)
        writer.writerow(['chr1', m.start(), m.end(), '-'])
```

Appendix C: Python script for determining the number of GQ sequences after *n* re-shufflings of a genomic sequence

```
#!/usr/bin/python
import sys, fileinput, ushuffle, re, csv
sequence = ""
num_GQ= []

for line in fileinput.input():
    if line[0] == ">":
        title = line[1:]
    else:
        sequence = sequence + line

sequence = sequence.upper().replace("\n", "")

shuff = ushuffle.shuffle(sequence, len(sequence), 3)

count = 0
while count <= n:
    shuff = ushuffle.shuffle(shuff, len(shuff), 3)
    GQ_for = re.findall("GGG[ATCGN]{1,7}GGG[ATGCN]{1,7}GGG[ATCGN]{1,7}GGG",
shuff)
    GQ_rev = re.findall("CCC[ATCGN]{1,7}CCC[ATGCN]{1,7}CCC[ATGCN]{1,7}CCC",
shuff)
    GQ = GQ_for + GQ_rev
    num_GQ.append(len(GQ))
    count = count + 1

writes data to csv file
with open(<output file name>, 'w') as output:
    writer = csv.writer(output, lineterminator = '\n')
    for val in num_GQ:
        writer.writerow([val])
```

Appendix D: Python script for identifying all UTR GQs.

```
#!/usr/local/bin/python3
import sys, fileinput, re
import numpy as np
import pandas as pd

## Importing data from csv files as Pandas dataframes
RNA_seq = pd.DataFrame.from_csv(<input RNA-seq data file>, header = 0, sep =
",", index_col=0)
RNA_seq = pd.DataFrame.dropna(RNA_seq)
cols = ['Transcription Start', 'Translation Start', 'Translation Stop',
'Transcription Stop']
RNA_seq[cols] = RNA_seq[cols].applymap(np.int64)
GQ_seq = pd.DataFrame.from_csv(<input GQ locations data file>, header = 0,
sep = ",", index_col=0)

## Defining all UTRs in RNA seq data as ranges from genomic start to stop
positions
UTRs = []
for i, row in RNA_seq.iterrows():
```

```
    if RNA_seq.loc[i, "Translation Start"] > RNA_seq.loc[i, "Transcription
Start"]:
        UTR = range(RNA_seq.loc[i, "Transcription Start"], RNA_seq.loc[i,
"Translation Start"])
        UTRs.append(UTR)
    elif RNA_seq.loc[i, "Transcription Start"] > RNA_seq.loc[i, "Translation
Start"]:
        UTR = range(RNA_seq.loc[i, "Translation Start"], RNA_seq.loc[i,
"Transcription Start"])
        UTRs.append(UTR)
    elif RNA_seq.loc[i, "Transcription Stop"] > RNA_seq.loc[i, "Translation
Stop"]:
        UTR = range(RNA_seq.loc[i, "Translation Stop"], RNA_seq.loc[i,
"Transcription Stop"])
        UTRs.append(UTR)
    elif RNA_seq.loc[i, "Translation Stop"] > RNA_seq.loc[i, "Transcription
Stop"]:
        UTR = range(RNA_seq.loc[i, "Transcription Stop"], RNA_seq.loc[i,
"Translation Stop"])
        UTRs.append(UTR)

## Creates a list of all GQ start positions form the Pandas dataframe
GQ_starts = []
for j, row in GQ_seq.iterrows():
    GQ_starts.append(GQ_seq.loc[j, 'Start'])

## Defines intersect as a function that takes two lists and returns values
that are found in both lists
def intersect(a, b):
    return list(set(a) & set(b))

## Finds all GQs that start in UTRs
for x in UTRs:
    UTR_GQs.append(intersect(x, GQ_starts))

## Removes empty values
while [] in UTR_GQs:
    UTR_GQs.remove([])

## Flattens UTR_GQs and saves it as new_UTR_GQs so that it is now a list of
integers instead of a list of lists of integers
new_UTR_GQs = []
def my_fun(temp_list):
    for ele in temp_list:
        if type(ele) == list:
            my_fun(ele)
        else:
            new_UTR_GQs.append(ele)
my_fun(UTR_GQs)

## Takes values from the list of UTR GQs and finds them in the Pandas
dataframe, then takes that row from the dataframe and saves it to a new file
final = GQ_seq[GQ_seq['Start'].isin(new_UTR_GQs)]
final.to_csv(<output file name>)
```

Appendix E: Python script that identifies all TSSs from a .wig file.

```
#!/usr/bin/python
import sys, fileinput, csv

n = 1
TSS_start = []
data = []

for line in fileinput.input():
     data.append(line.rstrip())

for x in data:
     if float(x) > 20 and float(data[int(n-2)]) <= 2:
          TSS_start.append(n)
          n = n+1
     else:
          n = n+1

with open(<output file name>, 'w') as output:
     writer = csv.writer(output, lineterminator = '\n')
     for val in TSS_start:
          writer.writerow([val])
```

Appendix F: Python script that uses list of TSSs generated in appendix 2.5 and determines which ones have predicted GQ starts within 100 nt of them.

```
#!/usr/bin/python
import sys, fileinput, csv
import numpy as np

with open(<GQ sequence data file>, 'rb') as file1:
     reader = csv.reader(file1)
     data1 = list(reader)

with open(<TSS positions data file>, 'rb') as file2:
     reader = csv.reader(file2)
     data2 = list(reader)

GQ_starts = []
TSSs = []

def my_fun1(temp_list):
    for ele in temp_list:
        if type(ele) == list:
            my_fun1(ele)
        else:
            GQ_starts.append(int(ele))
my_fun1(data1)

def my_fun2(temp_list):
    for ele in temp_list:
        if type(ele) == list:
            my_fun2(ele)
        else:
            TSSs.append(int(ele))
```

```
my_fun2(data2)

def find_closest(alist, target):
    return min(alist, key=lambda x:abs(x-target))


def list_matching(list1, list2):
    list1_copy = list1[:]
    pairs = []
    for i, e in enumerate(list2):
        elem = find_closest(list1_copy, e)
        pairs.append(list1.index(elem))
        list1_copy.remove(elem)
    with open(<output file name>, 'w') as output:
        writer = csv.writer(output, lineterminator = '\n')
        for val in pairs:
            writer.writerow([val])

list_matching(GQ_starts, TSSs)

with open(<output file name>, 'rb') as file3:
     reader = csv.reader(file3)
     data3 = list(reader)

TSS_GQ_index = []

def my_fun3(temp_list):
    for ele in temp_list:
        if type(ele) == list:
            my_fun3(ele)
        else:
            TSS_GQ_index.append(int(ele))
my_fun3(data3)

TSS_GQ = []

for x in TSS_GQ_index:
    TSS_GQ.append(GQ_starts[x])

with open(<output file name>, 'w') as output:
    writer = csv.writer(output, lineterminator = '\n')
    for val in TSS_GQ:
        writer.writerow([val])
```

Appendix G: UTR GQs

Gene	Gene Strand	GQ start (genomic locus)	GQ Strand	5′ or 3′ UTR?
sven_0139	+	141001	-	3′
sven_0268	+	282378	-	5′
sven_0301	+	317788	+	3′
sven_0648	-	762239	+	5′
sven_0667	+	782865	-	3′
sven_0725	-	837525	-	5′
sven_0766	+	885491	+	3′
sven_1015	-	1157521	+	3′
sven_1069	-	1210499	-	3′
sven_1069	-	1210528	-	3′
sven_1106	-	1251431	+	3′
sven_1444	-	1616529	-	5′
sven_1505	-	1682479	+	3′
sven_1698	+	1896922	-	3′
sven_1764	-	1967892	+	3′
sven_1779	-	1984589	-	3′
sven_1854	+	2066994	-	3′
sven_2062	+	2274234	+	3′
sven_2177	-	2345921	+	5′
sven_2304	+	2487090	-	3′
sven_2305	+	2489005	+	3′
sven_2324	-	2507899	+	5′
sven_2342	+	2530468	-	3′
sven_2342	+	2530543	-	3′
sven_2439	-	2647068	-	3′
sven_2455	+	2668765	+	3′
sven_2455	+	2668567	-	3′
sven_2517	-	2730171	+	3′
sven_2578	-	2807400	+	5′
sven_2684	-	2919298	-	3′
sven_2719	-	2960782	-	3′
sven_2752	-	3007918	+	3′
sven_2770	-	3036228	+	3′
sven_2834	+	3110345	-	3′
sven_2834	+	3110402	-	3′
sven_2838	-	3112331	-	3′
sven_2926	-	3193430	-	3′
sven_2992	+	3269605	-	5′
sven_2999	+	3276339	-	3′
sven_3002	+	3282853	-	3′
sven_3003	-	3282754	+	3′

sven_3036	-	3321053	+	3′
sven_3043	-	3327447	+	3′
sven_3057	-	3348088	+	3′
sven_3147	+	3446745	+	5′
sven_3166	-	3462460	+	5′
sven_3211	+	3519257	+	5′
sven_3212	-	3520887	-	3′
sven_3337	+	3657197	+	3′
sven_3340	-	3659288	+	3′
sven_3355	+	3674507	+	3′
sven_3549	-	3853686	+	3′
sven_3791	-	4115497	+	3′
sven_3794	-	4117283	+	3′
sven_4014	+	4347354	+	5′
sven_4128	-	4468026	-	3′
sven_4152	+	4495525	+	3′
sven_4204	+	4548824	-	5′
sven_4225	-	4569767	+	3′
sven_4362	-	4720465	+	3′
sven_4364	-	4721237	+	3′
sven_4421	+	4764646	+	5′
sven_4454	+	4793154	+	5′
sven_4498	+	4853739	+	5′
sven_4608	+	4964250	+	3′
sven_4683	+	5041752	-	5′
sven_4771	-	5137023	-	3′
sven_5008	+	5392079	-	5′
sven_5049	-	5436732	-	3′
sven_5078	+	5467202	-	3′
sven_5247	+	5675372	-	3′
sven_5265	+	5690820	-	3′
sven_5272	+	5695094	-	5′
sven_5301	+	5729578	-	3′
sven_5425	+	5882790	+	3′
sven_5425	+	5882844	+	3′
sven_5476	+	5942764	-	5′
sven_5739	-	6231746	+	3′
sven_6014	-	6544750	-	3′
sven_6014	-	6544812	-	3′
sven_6038	-	6570297	+	3′
sven_6042	+	6575673	+	3′
sven_6043	+	6576067	-	3′
sven_6047	+	6578862	-	5′
sven_6145	+	6702037	-	3′

sven_6288	+	6861515	-	3′
sven_6337	+	6919946	-	3′
sven_6343	+	6924691	+	5′
sven_6463	+	7049879	+	3′
sven_6518	+	7121596	+	3′
sven_6518	+	7121627	+	3′
sven_6525	+	7127504	+	3′
sven_6616	+	7232031	+	3′
sven_6803	+	7449939	-	5′
sven_6891	+	7537361	+	3′
sven_7203	+	7919172	-	5′
sven_7236	-	7962149	-	5′

Appendix H: GQs in proximity to TSSs

Gene	TSS (genomic locus)	Gene Strand	GQ position (genomic locus)	GQ Strand	Distance to TSS[1]	Expression (RPKM)
sven_0110	105424	-	105379	+	45	324
sven_0139	133237	+	133155	-	82	39
sven_0184	186098	-	186091	-	7	16
sven_0215	219884	+	219799	+	85	321
sven_0268	273983	+	273990	-	87	75
sven_0341	357694	+	357608	+	86	23
sven_0354	374087	+	373992	-	95	21
sven_0412	428768	+	428798	+	30	461
sven_0502	545282	+	545355	+	73	53
sven_0525	586137	-	586156	+	19	8
sven_0565	644082	-	644154	-	55	19
sven_0616	705093	-	705038	-	55	11
sven_0616	704976	-	704984	+	8	11
sven_0707	813175	+	813206	-	63	9
sven_0735	837302	+	837249	+	52	52
sven_0764	875197	-	875217	-	81	28
sven_0864	979875	+	979932	-	57	282
sven_0878	1001244	+	1001187	+	57	799
sven_0891	1013271	-	1013180	+	91	3
sven_0964	1084826	-	1084920	-	94	23
sven_1021	1160615	-	1160531	-	84	51
sven_1140	1283315	+	1283279	+	36	8862
sven_1199	1346805	-	1346705	+	100	8
sven_1222	1372629	-	1372645	-	16	11
sven_1368	1531786	-	1531821	-	35	7
sven_1396	1557636	+	1557549	-	82	31
sven_1487	1660641	+	1660684	+	43	129
sven_1505	1682267	-	1682346	+	79	119
sven_1631	1824889	-	1824972	-	83	134
sven_1652	1848312	-	1848402	+	90	19
sven_1678	1884446	-	1884466	-	77	190
sven_1722	1918256	+	1918160	+	96	219
sven_1778	1978204	+	1978169	+	35	10
sven_1835	2039572	+	2039526	-	22	11
sven_1859	2068115	+	2068049	-	66	553
sven_1881	2088580	+	2088502	+	78	53
sven_1929	2141841	+	2141795	-	46	67
sven_1967	2181687	-	2181748	-	61	126
sven_2016	2226878	-	2226812	+	66	108
sven_2228	2392216	+	2392143	+	73	232
sven_2232	2401531	-	2401567	+	36	6

sven_2291	2470809	+	2470716	-	93	23
sven_2296	2474899	+	2474818	+	81	48
sven_2372	2552816	+	2552726	-	74	154
sven_2500	2706845	+	2706877	+	32	16
sven_2559	2779696	+	2779699	+	3	11
sven_2583	2806153	+	2806057	-	96	9
sven_2624	2859386	-	2859461	+	77	68
sven_2638	2873247	-	2873246	-	9	10
sven_2639	2872621	+	2872534	-	87	173
sven_2641	2874895	+	2874917	-	22	58
sven_2694	2924590	-	2924615	-	25	52
sven_2698	2928492	+	2928447	-	45	11
sven_2716	2945758	+	2945677	-	81	164
sven_2855	3116561	+	3116494	-	67	45
sven_2916	3170948	-	3170939	+	9	26
sven_2935	3201822	-	3201829	+	7	4
sven_2969	3231404	-	3231311	+	93	138
sven_2969	3244661	-	3244674	+	53	138
sven_3002	3274498	+	3274424	-	100	77
sven_3045	3329424	-	3329497	+	52	299
sven_3211	3514935	+	3514975	+	40	254
sven_3290	3596110	+	3596016	-	94	81
sven_3350	3661752	+	3661813	-	79	387
sven_3356	3669086	+	3669121	+	35	338
sven_3363	3677310	-	3677269	-	41	24
sven_3529	3831693	+	3831644	+	49	6
sven_3612	3920938	-	3920958	-	20	28
sven_3792	4104696	+	4104613	+	83	322
sven_3813	4127859	-	4127940	+	32	10
sven_3823	4136080	+	4136033	+	47	108
sven_3833	4149981	+	4149904	+	77	582
sven_3891	4229416	+	4229321	+	95	579
sven_3959	4293082	+	4292982	+	100	19
sven_3965	4298592	-	4298682	-	90	364
sven_4005	4327815	+	4327824	+	9	47
sven_4014	4339182	+	4339157	+	25	152
sven_4200	4537426	+	4537512	-	21	17
sven_4221	4560978	-	4560972	+	6	57
sven_4223	4557787	+	4557738	+	49	69
sven_4250	4591897	-	4591818	-	79	2
sven_4273	4616544	+	4616502	-	1	244
sven_4416	4750063	-	4750137	+	74	89
sven_4452	4783682	-	4783752	+	45	1946
sven_4498	4844289	+	4844245	+	44	1927

sven_4510	4857460	-	4857540	-	80	397
sven_4535	4879466	-	4879500	+	67	39
sven_4543	4885741	+	4885826	-	85	44
sven_4574	4920761	-	4920796	+	35	98
sven_4609	4956503	-	4956434	+	96	6
sven_4611	4957893	-	4957882	+	46	394
sven_4672	5021116	-	5021033	+	14	15
sven_4785	5140737	+	5140806	+	69	131
sven_4792	5148409	-	5148342	-	67	111
sven_4795	5150336	+	5150356	+	20	340
sven_4806	5161116	+	5161216	-	100	599
sven_4837	5201271	+	5201184	-	7	81
sven_4838	5202467	+	5202388	-	41	51
sven_4839	5203991	-	5204068	-	93	1568
sven_4845	5209417	-	5209340	-	7	190
sven_4907	5277539	+	5277447	-	92	61
sven_4977	5337495	+	5337493	+	2	144
sven_5062	5438465	+	5438450	-	15	60
sven_5080	5460230	+	5460153	+	77	4
sven_5104	5490562	+	5490592	-	42	62
sven_5130	5529907	+	5529830	-	77	79
sven_5202	5613964	-	5613906	+	58	2
sven_5218	5631702	-	5631647	+	55	31
sven_5374	5808503	-	5808591	-	88	138
sven_5439	5888567	+	5888654	-	85	6
sven_5472	5931587	-	5931659	+	98	18
sven_5479	5943042	-	5943056	+	16	1001
sven_5485	5948841	+	5948930	+	89	44
sven_5635	6105020	+	6105111	+	91	0
sven_5664	6141100	+	6141131	+	31	73
sven_5783	6272989	+	6272904	+	85	74
sven_5908	6430235	+	6430216	-	94	21
sven_5909	6430895	+	6430911	-	15	6
sven_5946	6468211	-	6468198	+	13	358
sven_6039	6562810	-	6562845	+	35	20
sven_6043	6567351	+	6567283	+	67	453
sven_6067	6588571	+	6588601	+	30	19
sven_6120	6656327	-	6656250	-	45	92
sven_6146	6693405	+	6693340	-	65	128
sven_6335	6908875	-	6908867	+	8	37
sven_6343	6915673	+	6915696	+	23	71
sven_6353	6927101	+	6927019	-	95	5
sven_6415	6990462	+	6990546	+	84	27
sven_6453	7029660	+	7029573	+	76	530

sven_6504	7098943	+	7098944	-	1	151
sven_6515	7110702	-	7110630	+	72	17
sven_6520	7113992	-	7113901	+	91	9
sven_6539	7142213	-	7142189	+	24	74
sven_6594	7197833	+	7197753	+	80	9
sven_6628	7236929	-	7236897	+	32	2
sven_6684	7309637	-	7309652	-	58	18
sven_6709	7331034	+	7331086	-	19	1
sven_6714	7332442	+	7332468	-	30	17
sven_6871	7507793	+	7507737	-	79	40
sven_6992	7638087	+	7638108	-	61	75
sven_7070	7743418	-	7743495	+	77	566
sven_7280	7999685	+	7999661	+	24	91
sven_7299	8018977	+	8018924	-	53	28
sven_7378	8119936	-	8119873	+	63	1
sven_7379	8120671	+	8120573	-	98	0
sven_7433	8196431	+	8196348	-	92	13

[1]Defined as the absolute value of the difference between the TSS and the GQ position.

Appendix I: GQs between convergently oriented genes

Gene 1	Gene 2	GQ position (genomic location)	Gene 1 expression (RPKM)	Gene 2 expression (RPKM)
sven_0215	*sven_0216*	229798	321	12
sven_0425	*sven_0426*	455385	41	128
sven_0542	*sven_0543*	618543	0	3
sven_0549	*sven_0550*	631867	232	7
sven_0549	*sven_0550*	632211	232	7
sven_0549	*sven_0550*	632243	232	7
sven_0692	*sven_0693*	807843	472	2
sven_0901	*sven_0902*	1030804	31	66
sven_0939	*sven_0940*	1068748	12	1
sven_0963	*sven_0964*	1092211	728	23
sven_0987	*sven_0988*	1122097	7	76
sven_1012	*sven_1013*	1155969	5418	416
sven_1012	*sven_1013*	1155932	5418	416
sven_1056	*sven_1057*	1200016	19	461
sven_1105	*sven_1106*	1251430	70	1789
sven_1138	*sven_t3*	1287857	5	47
sven_1246	*sven_1247*	1400104	153	20
sven_1337	*sven_1338*	1500131	8	1014
sven_1337	*sven_1338*	1500296	8	1014
sven_1358	*sven_1359*	1522645	61	249
sven_1483	*sven_1484*	1658033	83	4859
sven_1483	*sven_1484*	1658208	83	4859
sven_1500	*sven_1501*	1677411	65	2681
sven_1504	*sven_1505*	1682368	228	119
sven_1504	*sven_1505*	1682478	228	119
sven_1594	*sven_1595*	1780206	58	52
sven_1607	*sven_1608*	1796368	50	15
sven_1619	*sven_1620*	1808029	188	193
sven_1698	*sven_1699*	1896921	532	7
sven_1731	*sven_1732*	1933221	197	1796
sven_1778	*sven_1779*	1984588	10	68
sven_1848	*sven_1849*	2062114	588	19
sven_1968	*sven_1969*	2186114	30	21
sven_1973	*sven_1974*	2189888	0	21
sven_2062	*sven_2063*	2274233	56	9
sven_2215	*sven_2216*	2386128	145	0
sven_2217	*sven_2218*	2389518	18	1
sven_2231	*sven_2232*	2401877	91	6
sven_2293	*sven_2294*	2477655	87	174
sven_2342	*sven_2343*	2530610	1999	68
sven_2342	*sven_2343*	2531231	1999	68

sven_2342	*sven_2343*	2530467	1999	68
sven_2342	*sven_2343*	2530542	1999	68
sven_2342	*sven_2343*	2531303	1999	68
sven_2342	*sven_2343*	2531411	1999	68
sven_2438	*sven_2439*	2647067	49	1750
sven_2573	*sven_2574*	2801837	270	9
sven_2588	*sven_2589*	2822194	286	1
sven_2683	*sven_2684*	2919297	52	293
sven_2834	*sven_2835*	3110344	682	42
sven_2834	*sven_2835*	3110401	682	42
sven_2903	*sven_2904*	3163238	13	219
sven_2925	*sven_2926*	3193429	43	89
sven_2933	*sven_2934*	3205318	110	5
sven_2974	*sven_2975*	3256370	126	25
sven_2999	*sven_3000*	3276338	697	38
sven_3002	*sven_3003*	3282753	77	72
sven_3056	*sven_3057*	3348087	7	1707
sven_3060	*sven_3061*	3352058	26	96
sven_3096	*sven_3097*	3396939	165	1
sven_3147	*sven_3148*	3447351	298	16
sven_3147	*sven_3148*	3447381	298	16
sven_3170	*sven_3171*	3464560	21	206
sven_3197	*sven_3198*	3504381	223	72
sven_3211	*sven_3212*	3520886	254	235
sven_3287	*sven_3288*	3599483	25	57
sven_3287	*sven_3288*	3599515	25	57
sven_3337	*sven_3338*	3657196	239	107
sven_3468	*sven_3469*	3781630	321	52
sven_3468	*sven_3469*	3781674	321	52
sven_3522	*sven_3523*	3833856	76	72
sven_3560	*sven_3561*	3870089	87	211
sven_3581	*sven_3582*	3893284	179	231
sven_3623	*sven_3624*	3937603	21	45
sven_3961	*sven_3962*	4303107	21	966
sven_3987	*sven_3988*	4328653	1	0
sven_4015	*sven_4016*	4351179	44	6
sven_4129	*sven_4130*	4470383	760	77
sven_4152	*sven_4153*	4495524	207	45
sven_4188	*sven_4189*	4532220	1	12
sven_4224	*sven_4225*	4569766	46	386
sven_4238	*sven_4239*	4584064	6	43
sven_4278	*sven_4279*	4632251	138	2
sven_4283	*sven_4284*	4637464	13	104
sven_4287	*sven_4288*	4643690	1212	26

sven_4361	*sven_4362*	4720464	62	391
sven_4438	*sven_4439*	4777903	234	6
sven_4463	*sven_4464*	4812279	3	12
sven_4486	*sven_4487*	4840604	3687	421
sven_4511	*sven_4512*	4867258	14	15
sven_4523	*sven_4524*	4879495	103	141
sven_4541	*sven_4542*	4895346	13	12
sven_4545	*sven_4546*	4899610	21	138
sven_4575	*sven_4576*	4929540	423	39
sven_4608	*sven_4609*	4964249	495	6
sven_4723	*sven_4724*	5083362	476	219
sven_4741	*sven_4742*	5101508	2	3
sven_4801	*sven_4802*	5166455	16	217
sven_5086	*sven_5087*	5476562	33	15
sven_5183	*sven_5184*	5599991	15	171
sven_5213	*sven_5214*	5635387	36	10
sven_5247	*sven_5248*	5675371	2467	59
sven_5341	*sven_5342*	5774056	19	10
sven_5425	*sven_5426*	5882789	373	5
sven_5425	*sven_5426*	5882843	373	5
sven_5425	*sven_5426*	5882883	373	5
sven_5478	*sven_5479*	5947449	111	1001
sven_5478	*sven_5479*	5947434	111	1001
sven_5482	*sven_5483*	5952386	686	141
sven_5779	*sven_5780*	6276049	51	1
sven_5859	*sven_5860*	6388947	44	50
sven_6004	*sven_6005*	6533866	17	252
sven_6004	*sven_6005*	6533999	17	252
sven_6013	*sven_6014*	6544749	2	44
sven_6013	*sven_6014*	6544811	2	44
sven_6037	*sven_6038*	6570296	1	57
sven_6055	*sven_6056*	6596990	90	1
sven_6275	*sven_6276*	6848673	0	0
sven_6275	*sven_6276*	6848703	0	0
sven_6275	*sven_6276*	6848748	0	0
sven_6275	*sven_6276*	6848778	0	0
sven_6292	*sven_6293*	6866445	109	5
sven_6292	*sven_6293*	6866211	109	5
sven_6463	*sven_6464*	7049878	279	4
sven_6506	*sven_6507*	7111015	10	1
sven_6518	*sven_6519*	7121595	337	34
sven_6518	*sven_6519*	7121626	337	34
sven_6525	*sven_6526*	7127503	184	22
sven_6564	*sven_6565*	7174169	40	57

sven_6747	*sven_6748*	7388822	3	2
sven_6826	*sven_6827*	7474935	4159	423
sven_6862	*sven_6863*	7509063	1996	27
sven_6928	*sven_6929*	7581125	4	91
sven_7081	*sven_7082*	7762661	75	5
sven_7084	*sven_7085*	7767590	135	44
sven_7117	*sven_7118*	7806999	8	29
sven_7369	*sven_7370*	8112016	21	0
sven_7438	*sven_7439*	8210462	28	10
sven_t23	*sven_2615*	2854408	38	57

Appendix J: GQs between divergently oriented genes

Gene 1	Gene 2	GQ position (genomic locus)	Gene 1 expression (RPKM)	Gene 2 expression (RPKM)
sven_0038	*sven_0039*	36256	50	2
sven_0046	*sven_0047*	43900	0	7
sven_0085	*sven_0086*	82546	8	39
sven_0107	*sven_0108*	109635	41	95
sven_0133	*sven_0134*	135538	4	30
sven_0145	*sven_0146*	148771	5	455
sven_0149	*sven_0150*	153534	9	1
sven_0162	*sven_0163*	169920	56	5
sven_0190	*sven_0191*	203036	0	3
sven_0190	*sven_0191*	203339	0	3
sven_0190	*sven_0191*	203471	0	3
sven_0243	*sven_0244*	258218	5	53
sven_0252	*sven_0253*	267941	6	14
sven_0292	*sven_0293*	306986	23	53
sven_0318	*sven_0319*	340144	12089	1
sven_0411	*sven_0412*	437740	9	461
sven_0477	*sven_0478*	517071	443	22
sven_0523	*sven_0524*	592109	20	1
sven_0535	*sven_0536*	610043	0	29
sven_0548	*sven_0549*	630872	39	232
sven_0616	*sven_0617*	713538	11	101
sven_0690	*sven_0691*	806676	131	514
sven_0706	*sven_0707*	821965	7	9
sven_0725	*sven_0726*	837524	144	105
sven_0741	*sven_0742*	852869	45	13
sven_0753	*sven_0754*	865513	110	3
sven_0778	*sven_0779*	899611	76	0
sven_0912	*sven_0913*	1039639	12	108
sven_0956	*sven_0957*	1083270	5	27
sven_1019	*sven_1020*	1163560	0	20
sven_1019	*sven_1020*	1163624	0	20
sven_1061	*sven_1062*	1204467	127	2
sven_1149	*sven_1150*	1296561	15	118
sven_1149	*sven_1150*	1296688	15	118
sven_1187	*sven_1188*	1337442	25	118
sven_1271	*sven_1272*	1433105	524	9
sven_1368	*sven_1369*	1535416	7	1
sven_1441	*sven_1442*	1614313	71	184
sven_1486	*sven_1487*	1661472	104	129
sven_1505	*sven_1506*	1683622	119	112
sven_1533	*sven_1534*	1714584	8	14

sven_1539	*sven_1540*	1721319	68	181
sven_1584	*sven_1585*	1768629	0	2
sven_1584	*sven_1585*	1768774	0	2
sven_1652	*sven_1653*	1850405	19	34
sven_1667	*sven_1668*	1864717	0	1
sven_1730	*sven_1731*	1929530	318	197
sven_1754	*sven_1755*	1959168	24	5
sven_1814	*sven_1815*	2020819	0	23
sven_1814	*sven_1815*	2020869	0	23
sven_1814	*sven_1815*	2020907	0	23
sven_1894	*sven_1895*	2111964	395	19
sven_1928	*sven_1929*	2144994	194	67
sven_1940	*sven_1941*	2157421	228	223
sven_1967	*sven_1968*	2185425	126	30
sven_2026	*sven_2027*	2238664	46	15
sven_2026	*sven_2027*	2239195	46	15
sven_2177	*sven_2178*	2345920	28	133
sven_2227	*sven_2228*	2396237	7	232
sven_2295	*sven_2296*	2479451	357	48
sven_2324	*sven_2325*	2507898	54	6
sven_2341	*sven_2342*	2530042	162	1999
sven_2412	*sven_2413*	2613740	108	13
sven_2491	*sven_2492*	2706726	211	12
sven_2491	*sven_2492*	2706757	211	12
sven_2491	*sven_2492*	2706927	211	12
sven_2502	*sven_2503*	2713747	17	3
sven_2558	*sven_2559*	2784180	25	11
sven_2578	*sven_2579*	2807399	93	4
sven_2578	*sven_2579*	2807881	93	4
sven_2638	*sven_2639*	2877401	10	173
sven_2640	*sven_2641*	2879784	46	58
sven_2694	*sven_t25*	2929484	52	1
sven_2697	*sven_2698*	2933316	64	11
sven_2746	*sven_2747*	3001592	53	3
sven_2746	*sven_2747*	3001683	53	3
sven_2746	*sven_2747*	3001721	53	3
sven_2763	*sven_2764*	3027854	6	65
sven_2772	*sven_2773*	3038483	138	103
sven_2966	*sven_2967*	3246752	4	14
sven_3001	*sven_3002*	3279360	87	77
sven_3045	*sven_3046*	3334295	299	149
sven_3045	*sven_3046*	3334433	299	149
sven_3057	*sven_3058*	3349525	1707	5
sven_3146	*sven_3147*	3446744	207	298

sven_3209	*sven_3210*	3517742	165	73
sven_3283	*sven_3284*	3594775	85	9
sven_3289	*sven_3290*	3601208	15	81
sven_3363	*sven_3364*	3682654	24	9
sven_3482	*sven_3483*	3793788	8	3
sven_3482	*sven_3483*	3793822	8	3
sven_3528	*sven_3529*	3837624	18	6
sven_3528	*sven_3529*	3837673	18	6
sven_3612	*sven_3613*	3927922	28	1626
sven_3763	*sven_t38*	4090738	13	55
sven_3832	*sven_3833*	4158019	357	582
sven_3945	*sven_3946*	4287499	46	944
sven_3965	*sven_3966*	4306887	364	4771
sven_4010	*sven_t51*	4345160	127	5
sven_4052	*sven_4053*	4388671	19	4
sven_4178	*sven_4179*	4520469	24	3
sven_4178	*sven_4179*	4520495	24	3
sven_4178	*sven_4179*	4521294	24	3
sven_4178	*sven_4179*	4521353	24	3
sven_4203	*sven_4204*	4548823	2	125
sven_4222	*sven_4223*	4566532	16	69
sven_4247	*sven_4248*	4597744	3	1
sven_4247	*sven_4248*	4597792	3	1
sven_4360	*sven_4361*	4719421	229	62
sven_4365	*sven_4366*	4722675	232	2
sven_4369	*sven_4370*	4724797	70	192
sven_4375	*sven_4376*	4732307	25	49
sven_4416	*sven_4417*	4759520	89	28
sven_4436	*sven_4437*	4776657	24	139
sven_4510	*sven_4511*	4865977	397	14
sven_4526	*sven_4527*	4881381	8	7
sven_4535	*sven_4536*	4889020	39	186
sven_4576	*sven_4577*	4930274	39	156
sven_4672	*sven_4673*	5030638	15	17
sven_4724	*sven_4725*	5084182	219	247
sven_4805	*sven_4806*	5169605	93	599
sven_4839	*sven_4840*	5212742	1568	64
sven_4893	*sven_4894*	5269997	0	5
sven_4893	*sven_4894*	5270349	0	5
sven_4906	*sven_4907*	5286176	186	61
sven_4944	*sven_4945*	5318648	312	60
sven_4958	*sven_4959*	5330831	363	31
sven_5039	*sven_5040*	5424788	48	1489
sven_5065	*sven_5066*	5451388	16	0

sven_5065	*sven_5066*	5451511	16	0
sven_5112	*sven_5113*	5520090	24	22
sven_5128	*sven_5129*	5536637	0	1
sven_5140	*sven_5141*	5551995	3	3
sven_5146	*sven_5147*	5558684	79	33
sven_5207	*sven_5208*	5627837	2	9
sven_5342	*sven_5343*	5774493	10	11
sven_5417	*sven_5418*	5874528	15	5
sven_5438	*sven_5439*	5897274	2	6
sven_5535	*sven_5536*	6018427	575	723
sven_5571	*sven_5572*	6058956	39	9
sven_5634	*sven_5635*	6121190	0	0
sven_5883	*sven_5884*	6409838	44	33
sven_5939	*sven_5940*	6464221	28	4
sven_5961	*sven_5962*	6493146	4	54
sven_6006	*sven_6007*	6534936	9	11
sven_6035	*sven_6036*	6568243	2	6
sven_6035	*sven_6036*	6568333	2	6
sven_6174	*sven_6175*	6728640	0	5
sven_6352	*sven_6353*	6936013	6	5
sven_6414	*sven_6415*	6999541	20	27
sven_6469	*sven_6470*	7060047	58	97
sven_6498	*sven_6499*	7097696	14	0
sven_6505	*sven_6506*	7109804	5	10
sven_6507	*sven_6508*	7112607	1	4
sven_6526	*sven_6527*	7128852	22	195
sven_6560	*sven_6561*	7170526	6	11
sven_6560	*sven_6561*	7170615	6	11
sven_6593	*sven_6594*	7206735	37	9
sven_6716	*sven_6717*	7347847	0	0
sven_6828	*sven_6829*	7476819	90	411
sven_6828	*sven_6829*	7476921	90	411
sven_6828	*sven_6829*	7476972	90	411
sven_6828	*sven_6829*	7477041	90	411
sven_6870	*sven_6871*	7518026	472	40
sven_7202	*sven_7203*	7919171	8	44
sven_7258	*sven_7259*	7986924	9	4
sven_7378	*sven_7379*	8130891	1	0
sven_7397	*sven_7398*	8159838	0	0
sven_7397	*sven_7398*	8159895	0	0
sven_7405	*sven_7406*	8166096	6	5
sven_t11	*sven_2173*	2339403	7	141
sven_t33	*sven_3627*	3940584	120	529
sven_t40	*sven_3823*	4144105	81	108

sven_t40	*sven_3823*	4144145	81	108
sven_t46	*sven_3852*	4179112	105	207