Annotation and evolution of bacterial ncRNA genes

Bethany R. Jose

a thesis submitted for the degree of

Doctor of Philosophy

at the University of Otago, Dunedin,

New Zealand.

13th of December, 2019

Abstract

Successful pathogenic bacteria must alter gene expression in response to changing and hostile environments. Non-coding RNAs (ncRNAs) contribute to adaptability and pathogenicity by forming complex regulatory networks, and include riboswitches, *cis*-regulatory elements and sRNAs. Despite their important biological function, the annotation and discovery of ncRNAs is hindered by a lack of sequence conservation or other distinguishing sequence features.

Studies of the evolutionary dynamics and origins of sRNA genes have been hindered by poor sequence conservation, which makes annotation via sequence homology challenging. The short length and relative simplicity of sRNA genes also make them interesting candidates for observing *de novo* gene formation from transcriptional noise, or exaptation from existing elements.

We have used a pipeline based on profile hidden Markov models to study the conservation patterns of sRNA genes from *Salmonella* Typhimurium. Our results show that sRNAs are both rapidly acquired and exhibit rapid sequence turnover. We found that horizontal gene transfer is the main driver of sRNA acquisition in *Salmonella*, and identified *Salmonella*-specific sRNAs that appear to be derived from phage control systems, and other mobile genetic elements, as well as Type I toxin-antitoxin systems. This method was then applied to study ncRNAs in *Pseudomonas syringae* pv. *actinidiae* (*Psa*), the causal agent of kiwifruit canker disease. We have generated transcriptomes of a pandemic strain of *Psa* in multiple growth conditions *in vitro*, and analysed gene expression changes and identified novel non-coding transcripts. We then studied the expression and conservation of these candidate ncRNAs, and identified several with predicted secondary structure motifs characteristic to known functional ncRNAs.

This thesis also includes a summary of two genome assembly projects of *Gemmata* and *Legionella* isolates, as part of larger collaborations. All diagrams in this thesis are my own work, unless otherwise stated.

Acknowledgements

Thanks to Amy, Andrew, Ant, Astra, Camille, Ciarán, Chris, Dorien, Elsie, Hannah, Harry, Jade, Jen, Jenny, Lars, Lim, Luna the cat, Mackenzie, Maud, Matt, Megan, Moizle, Paul, Rima, Ruaire, Rudolf, and many others for their love, help and support.

Abbreviations used in this thesis

BLAST	Basic local alignment search tool
BWA	Burrows-wheeler aligner
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DEG	Differentially expressed gene
dNTP	Dye Nucleotide Triphosphate
HGT	Horizontal gene transfer
HMM	Hidden Markov model
HR	Hypersensitive response
IS	Insertion sequence
l_2FC	Log$_2$ fold-change
LB	Lysogeny broth
MFE	Minimum free energy
MGE	Mobile genetic element
mRNA	Messenger RNA
ncRNA	Non-coding RNA
NGS	Next-generation Sequencing
nt	Nucleotide(s)
OD A_{600}	Optical density at 600 nm
ORF	Open Reading Frame
Psa	*Pseudomonas syringae* pv. *actinidiae*
Pta	*Pseudomonas syringae* pv. *tabaci*
Pto	*Pseudomonas syringae* pv. *tomato*
pv.	Pathovar
rRNA	Ribosomal RNA
RT-PCR	Reverse transcription PCR
SMRT	Single molecule real-time
SRA	Sequence read archive
sRNA	Small non-coding RNA
SRP	Signal recognition particle
str.	Strain
sv.	Serovar
T3SS	Type III secretion system
TA system	Toxin-antitoxin system

TCA cycle	Tricarboxylic acid cycle
tRNA	transfer RNA
tmRNA	Transfer-messenger RNA
UTR	Un-translated Region

Contents

List of Tables

List of Figures

Chapter 1

An introduction to RNA bioinformatics and pathogen genomics

1.1 Preface

Bacteria are the most diverse and abundant type of organism on the planet (Hug *et al.*, 2016). Among the multitude of possible evolutionary niches, only a small subset of bacteria occupy pathogenic roles (Balloux and van Dorp, 2017, "Microbiology by numbers", 2011). Bacterial pathogens exhibit a remarkable range of adaptions to survive stresses encountered both in the environment and within the host. These include the acquisition of new genes *via* horizontal transfer (Groisman and Ochman, 1996, Kado, 2009), as well as changes in both gene function (Moxon *et al.*, 1994) and gene expression (Thomas and Wigneshweraraj, 2014). Despite having far fewer genes than eukaryotes, sophisticated regulatory networks allow bacteria to sense and respond to their environment (Dorman *et al.*, 2018). The regulatory function and evolution of many non-coding RNAs (ncRNAs) have been identified as major contributors to the virulence and survival of many bacterial pathogens (Gottesman, 2005, Gripenland *et al.*, 2010). Of particular interest are the small non-coding RNAs (sRNAs) that participate in regulatory networks *via* interactions with mRNA and proteins (Gottesman and Storz, 2011a).

This thesis will primarily focus on the challenges of identifying and studying non-coding RNA sequences in bacterial pathogens. Specifically, within the genus *Salmonella* and other pathogens of the Enterobacteriaceae that infect animals and plants, and the kiwifruit pathogen *Pseudomonas syringae* pv. *actinidiae*, also known as *Psa*.

Section 1.5 contains material from a published paper (Jose *et al.*, 2019) associated with this thesis. This is material researched and drafted by myself for the paper, and it has been rewritten to fit in with the context of this introduction.

1.2 A brief introduction to genomics

Pathogen genomics is largely the analysis of biological sequences to identify features which directly or indirectly influence the virulence and fitness of pathogens. This may involve the identification of antibiotic resistance genes, virulence factors such as toxins, or track the evolution and horizontal gene transfer of such elements. The large scale study of such information has been facilitated by the introduction of next-generation sequencing (NGS) technologies, which are both increasingly affordable and capable of producing vast quantities of sequence data (Goodwin *et al.*, 2016). As a result, the number of bacterial genome sequences available for comparison has increased exponentially (Figure 1.1).

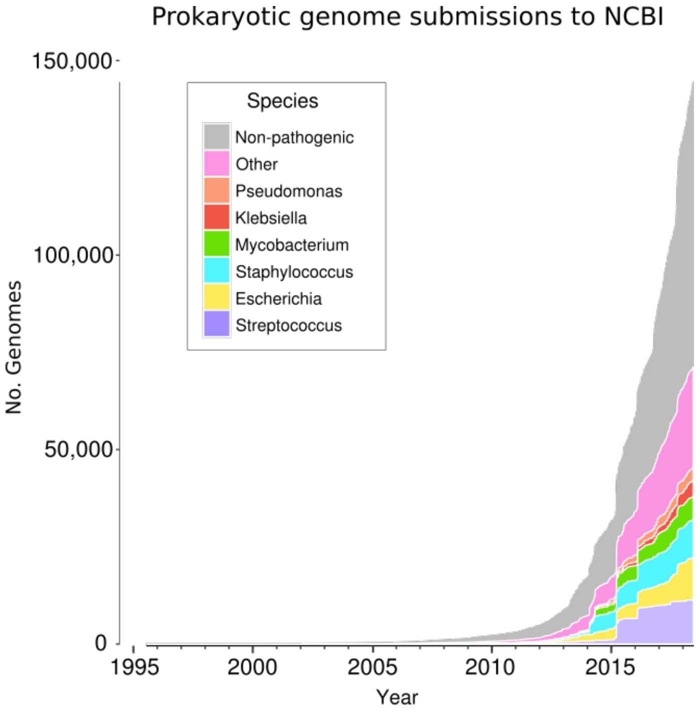


Figure 1.1: Total prokaryotic genome submissions to NCBI over time, showing the exponential growth in prokaryotic genome submissions. Data taken from the NCBI prokaryote genome assembly summary (date range 1995-2018). Relative proportions of pathogenic species and strains identified as pathogenic by Ecker *et al.* (2005) are shown as coloured stacks grouped by genera, showing that a large proportion of prokaryotic genomes in NCBI are from pathogenic species. The top 6 genera are ordered by proportion from smallest to largest (*Pseudomonas*, *Klebsiella*, *Mycobacterium*, *Staphylococcus*, *Escherichia* and *Streptococcus*), with "Other" representing the combined sum of remaining pathogen genomes.

Despite representing a tiny fraction of microbial diversity (Balloux and van Dorp, 2017), prokaryotic pathogens are over-represented within genome databases (Figure 1.1). Genera such as *Streptococcus*, *Escherichia*, and *Salmonella* that contain multiple pathogenic species have been densely sampled, facilitating increasingly detailed and granular genomic studies of pathogenic bacteria.

Genomics, the study of genome sequences, is intertwined with the field of bioinformatics, which is tasked with both managing and interpreting these increasingly large amounts of sequence data. Early bioinformatics tools for biological sequence analysis were designed to assemble (Dayhoff and Ledley, 1962, Staden, 1979) and compare (Felsenstein, 1981, Needleman and Wunsch, 1970, Smith and Waterman, 1981, Thompson *et al.*, 1994) amino acid and nucleotide sequences, which may be simply represented as character strings in a computing context (Derome *et al.*, 2018).

Much of modern bioinformatics remains focused on these same problems; studying both the information content within, and comparisons between biological sequences. Detailed introductions to these methods can be found in numerous textbooks covering bioinformatics methods and algorithms, e.g Durbin *et al.* (1998), Mount (2004) and Waterman (1995).

1.2.1 Next Generation Sequencing technologies

The introduction of NGS technologies with massively increased capacity in the mid-2000's has revolutionised the field of genomics. Current sequencing methods can be broadly split into 'short' and 'long' read technologies, depending on the size of fragments that can be sequenced. Most modern sequencing platforms use a 'sequencing-by-synthesis' approach, using nucleotides modified with dye molecules (dye dNTPs) which fluoresce at wavelengths specific to each nucleotide base (Bentley *et al.*, 2008, Eid *et al.*, 2009).

Short read sequencing technologies typically utilise dye dNTPs with reversible dye termination chemistry (Bentley *et al.*, 2008, Croucher *et al.*, 2009), which terminate strand extension upon incorporation. In Illumina short-read sequencing, individual fragments are first ligated and amplified on a substrate, generating clonal clusters of forward and reverse strands from the same sequence. During sequencing-by-synthesis, single dye dNTPs are incorporated, and the base is determined by excitation of the dye dNTP's fluorophore over a cluster of fragments, which is then cleaved prior the next cycle. This method of sequencing has a low error rate during base calling, but is limited to smaller fragments (50–500 nt).

Long-read technologies have no upper limit on molecule size and excel at resolving repetitive sequence regions, but require large amounts of DNA to produce enough data to compensate for high error rates. The SMRT$^{\text{TM}}$ sequencing approach used by Pacific Biosciences uses a specialised DNA polymerase fixed to a substrate, and dye dNTPs with a fluorophore linked to the terminal phosphate that are cleaved during synthesis. This approach enables reactions to proceed rapidly, and produces a natural double-stranded DNA product (Korlach *et al.*, 2010). The dye dNTPs fluoresce as they approach the illuminated substrate and are incorporated into the new strand, enabling bases to be called (Eid *et al.*, 2009).

Another emerging class of long-read sequencing technologies developed by Oxford Nanopore detects nucleotides by measuring changes in electrical current as samples pass through protein nanopores embedded in a polymer membrane. Voltage is applied to the membrane, causing a current to pass through the nanopore (Feng *et al.*, 2015, Jain *et al.*, 2016). An affixed φ29 DNA polymerase acts as a processive enzyme, ratcheting nucleotides through the pore one base at a time in a controllable reaction (Cherf *et al.*, 2012). Bases are determined by characteristic fluctuations in membrane potential with the passing of each nucleotide (Stoddart *et al.*, 2009). Each nanopore is part of a single circuit, allowing the current across individual nanopores to be controlled and measured.

Nanopore technologies such as the MinION sequencer have been an exciting development due to the small size and portability of the instrument, and its ability to produce extremely long (>140 kb) sequencing reads (Jain *et al.*, 2016). Improvements in both nanopore technology and assembly programs are generating high quality bacterial (Quick *et al.*, 2014), human (Jain *et al.*, 2018) and even native RNA viral genomes (Keller *et al.*, 2018).

1.2.2 Standard Bioinformatic analysis of NGS sequencing data

Genomics is a rapidly changing field, both in terms of sequencing technologies and the bioinformatics tools used to analyse sequence data. While there are numerous tools and workflows for processing sequencing data, bioinformatics pipelines generally include quality control of raw data, an alignment step for assembling reads or mapping reads to a reference sequence, and downstream quantitative or qualitative analysis.

Quality control of data using tools such as FastQC (Andrews, 2010) provide summary statistics that can identify problems with the data-set that originate from library composition or from the sequencing process. Typically some correction must be performed on the sequencing reads to remove any remaining sequencing adapters, or remove reads that are too short or that have missing sequencing pairs. For short-read sequencing data, trimming tools such as trimmomatic (Bolger *et al.*, 2014) or cutAdapt (Martin, 2011) can remove regions of sequence below a certain quality score threshold. For long-read sequencing data, which contain semi-randomly introduced errors, reads can be corrected by alignment of sequencing reads. This may be achieved by consensus overlay correction if there is sufficient read depth, or using additional short-read sequencing data, which have lower error rates.

Many sequencing projects involve an assembly step, which aims to reconstruct genomes or transcripts from sequencing reads. Assembly tools are a particularly active area within bioinformatics, as new software is rapidly developed, aiming to both reduce computational time and improve assembly quality (Simpson and Pop, 2015, Sohn and Nam, 2018). Common algorithmic approaches for assembly include de Bruijn graph-based methods for short read data, and alignment-based methods for long-read data (Simpson and Pop, 2015).

For transcriptome analysis, reads can be assigned to their originating genome sequence by a process called mapping. The number of reads originating from a particular gene or genomic feature is then used to give a quantitative estimate of gene expression. Alignment tools such as BWA (Li and Durbin, 2009) and Bowtie (Langmead and Salzberg, 2012) can align sequence reads to a reference genome. Downstream quantification tools can then be used to provide 'counts' of reads mapped for individual genes (Anders *et al.*, 2015, Li and Dewey, 2011, Trapnell *et al.*, 2012). Alternatively, recent alignment-free tools such as Kallisto (Bray *et al.*, 2016) and Salmon (Patro *et al.*, 2017) use k-mer based methods to rapidly classify and quantify reads, using reference transcripts (Wu *et al.*, 2018). A variety of statistical analysis packages can then be used to interpret gene expression data, including DESeq2 (Love *et al.*, 2014), EdgeR (Robinson *et al.*, 2010) and Limma (Ritchie *et al.*, 2015).

High-throughput pipelines for assembly and annotation are commonly used in bacterial genome sequencing projects (Aziz *et al.*, 2008, Seemann, 2014, Tatusova *et al.*, 2016). These workflows implement a variety of tools to identify open reading frames (Delcher *et al.*, 2007, Hyatt *et al.*, 2010) and annotate genes based on homology to reference proteins or non-coding RNAs (Altschul *et al.*, 1990, Huerta-Cepas *et al.*, 2017, Nawrocki *et al.*, 2015a, Nawrocki and Eddy, 2013, Punta *et al.*, 2012).

1.2.3 RNA-seq

A revolutionary application of NGS technologies is the sequencing of RNA transcripts (RNA-seq). RNA-seq allows the transcription rate of a gene to be estimated by providing a snapshot of the population of transcribed RNAs present in a sample (transcriptome). Protocols for RNA-seq employ similar methods as short-read DNA sequencing, with an additional step where complementary DNA (cDNA) is synthesised from RNA transcripts prior to sequencing (Wang *et al.*, 2009).

As with DNA sequencing, RNA-seq technologies are rapidly developing, producing increasingly high-resolution transcriptomes (Creecy and Conway, 2015). Increases in sequencing capacity and depth allows low abundance transcripts to be detected, and stranded sequencing methods have been developed to determine the originating strand and orientation of a sequencing read (Fullwood *et al.*, 2009, Parkhomchuk *et al.*, 2009). Almost all current RNA-seq protocols are based on Illumina short-read sequencing, however, nanopore-based high-throughput RNA-seq technology is emerging (Workman *et al.*, 2018).

In the past decade, a variety of novel technologies have emerged built on RNA-seq methods (Hrdlickova *et al.*, 2017, Saliba *et al.*, 2017). Improvements in library preparation can determine the transcriptomes of single cells (single-cell RNA-seq, or scRNASeq) (Islam *et al.*, 2014), enabling transcriptomic studies across whole tissues or populations of bacteria.

Other technologies can shed light on the functional roles and fate of RNA molecules. Interactions between transcripts (RNA-RNA interactions) can be identified by cross-linking, ligating and sequencing RNA molecules participating in RNA duplexes (Kudla *et al.*, 2011). Alternatively, RNA duplexes can be purified by immunoprecipitation of enzymes which preferentially bind double-stranded RNA (Lioliou *et al.*, 2013).

RNAs participating in complexes with RNA-binding proteins can be studied by gradient separation (GRAD-seq) (Rederstorff *et al.*, 2010, Smirnov *et al.*, 2016) or UV-RNA-protein crosslinking (Licatalosi *et al.*, 2008), which can be combined with immunoprecipitation of specific RNPs of interest (Lu *et al.*, 2014, Melamed *et al.*, 2016, Zhao *et al.*, 2010). The dynamics of transcription initiation and termination can also be studied by selectively sequencing the $5'$ (Sharma *et al.*, 2010) and $3'$ end of transcripts (Dar *et al.*, 2016), called TSS-seq and Term-seq respectively.

Increased depth of sequencing and improved mapping software has also enabled simultaneous studies of gene expression in bacterial pathogens and their host without the prior separation of samples (Westermann *et al.*, 2012). As the rapid pace of development of such technologies continues, we can expect RNA-seq to provide an increasingly nuanced and comprehensive view of gene expression.

1.3 Bacterial ncRNAs in pathogens

Much of the focus of early molecular biology concerned proteins and their functions. It has since been found, however, that some RNA molecules directly enact specific biological functions both as functional components and regulators of transcription and translation

(Eddy, 2001, Morris and Mattick, 2014) (Figure 1.3). These RNA molecules, termed 'non-coding RNAs' or ncRNAs, are a heterogeneous class of genes which enact their function as RNA transcripts, and are found in all domains of life (Figure 1.3).

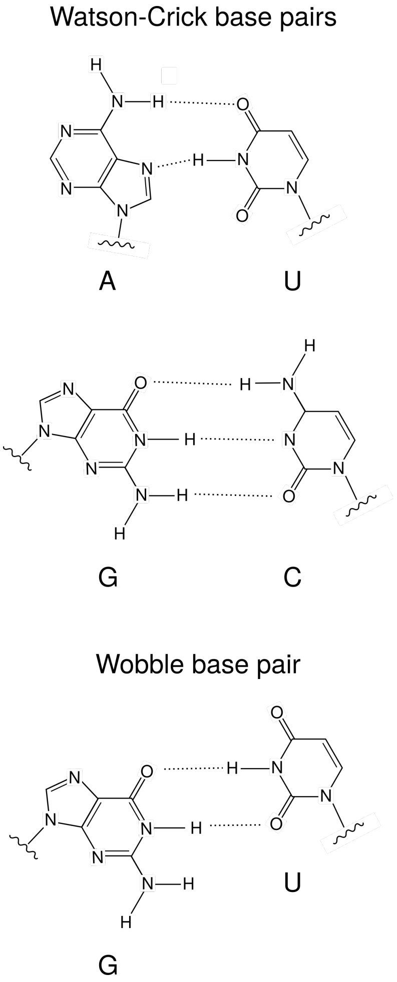


Figure 1.2: Nucleotide base pairs found in RNA secondary structures. **Top:** Watson-crick base pairs between Adenosine (A) and Uracil (U), and Guanine (G) and Cytosine (C). **Bottom:** Wobble base pair formed between Guanine and Uracil. Hydrogen bonds shown as dashed lines.

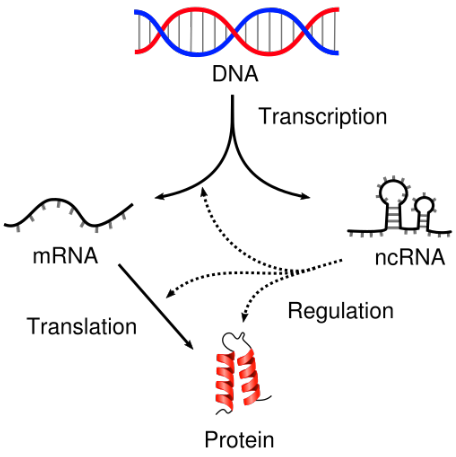


Figure 1.3: Roles of ncRNA regulation in different processes within the central dogma of molecular biology. Adapted from Figure 1, Wahlestedt (2013).

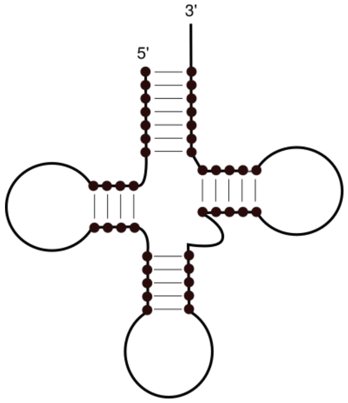


Figure 1.4: Example of an RNA secondary structure: Representation of base-pairing secondary structure of a tRNA, based on Figure 2 by Rich and RajBhandary (1976).

Much like proteins, ncRNAs possess a diverse array of sizes, structure, function and modes of action. Many ncRNA transcripts form complex secondary and tertiary structures *via* intramolecular base pairing (see Figure 1.2 and Figure 1.4), which is the functional determinant of many ncRNAs.

Non-coding RNAs can be broadly grouped into three functional categories: catalytic RNA (ribozymes), RNAs which participate in ribonucleoprotein complexes (RNPs), and regulatory ncRNA (Cech and Steitz, 2014). Famous examples of ncRNAs include ubiquitously conserved elements of the translation machinery, such as the ribosome, tRNAs (Figure 1.4), and the RNase P ribonuclease involved in tRNA processing. Other ribozymes found in bacteria include Group I and II introns (Hausner *et al.*, 2014, Lambowitz and Zimmerly, 2011), which are self-splicing ribozymes that self-propagate in many lineages.

Aside from ribosomes, few RNPs are known in bacteria. Notable examples include the highly conserved 6S RNA, which binds with σ70-RNA polymerase with widespread effects on transcriptional regulation (Wassarman, 2007), the transfer-messenger RNA (tmRNA) that recycles stalled ribosomes and their associated mRNAs and protein products (Moore and Sauer, 2007), the signal recognition particle (SRP) RNA involved in protein transport and localisation (Akopian *et al.*, 2013), and the CRISPR-Cas protein-RNA complex which recognises and guides endonucleases to foreign DNA (Barrangou, 2015).

1.3.1 Gene regulation in pathogens

As single-celled organisms, bacteria must dynamically control internal concentrations of macromolecules in order to respond to changes in the local environment (Dorman *et al.*, 2018). While a handful of genes involved in processes essential to life are constitutively expressed, bacterial genomes also contain cohorts of genes that contribute most to fitness when expressed under specific growth conditions (Price *et al.*, 2016). Effective gene regulation increases fitness by preventing detrimental epistasis, and by allocating resources appropriately.

Bacterial pathogens face additional stresses during infection as they must traverse through radically different environments both outside and inside the host, each with its own set of stressors (Cotter and DiRita, 2000). Virulence genes, such as those involved in toxin secretion or adhesion to host cells, must also be tightly controlled for maximum effect. Gene regulation therefore directly influences pathogenicity.

In addition to gene regulation by transcription factors and changes in DNA and RNA conformation or topology, advances in RNA biology have revealed complex and sophisticated regulatory roles of ncRNA in gene regulation in bacteria (Hör *et al.*, 2018).

1.3.2 Regulatory ncRNA

Regulatory RNAs are a diverse class of ncRNA whose roles as regulators at transcription, translation and at the protein level are being increasingly appreciated (Cech and Steitz, 2014). Regulatory RNAs were initially discovered in the 1980's, mostly as antisense regulators of plasmid replication proteins in *Escherichia coli* (Conrad and Campbell, 1979, Inouye and Delihas, 1988, Stougaard *et al.*, 1981). Although the 6S RNA was technically the first regulatory ncRNA discovered (Brownlee, 1971), its exact function remained unknown until decades later (Barrick *et al.*, 2005). Few regulatory ncRNAs were discovered until the early 2000's, but since then a wide variety of bacterial ncRNAs have been found that contribute to pathogenicity through versatile mechanisms (Papenfort and Vogel, 2010).

Interesting examples of bacterial ncRNAs include *cis*-regulatory elements such as riboswitches, which are structured elements that change conformation to allow or deny the translation of the downstream coding sequence. Most known examples of riboswitches act as sensors, detecting metabolite ligands to self-regulate biosynthetic pathways (Serganov and Nudler, 2013). A particularly elegant example of a riboswitch involved in pathogenicity is the *prfA* thermoswitch of *Listeria monocytogenes*, which unfolds at 37°C to allow the translation of Prfa, a transcriptional activator of virulence genes (Johansson *et al.*, 2002).

Bacterial small RNAs

Bacterial sRNAs are a broad and heterogeneous class of ncRNAs that regulate gene expression by interacting with both mRNA and proteins (Figure 1.5). These small (~50–500 nt) transcripts are often expressed only in specific growth conditions, which gives bacteria an effective toolkit to fine-tune protein expression to their needs.

The majority of characterised sRNAs act by an antisense binding mechanism. These transcripts can base-pair with their targets to various extents in order to degrade, stabilise or enhance transcription of mRNAs, or sequester or inhibit proteins by targeting RNA-binding sites (Barquist and Vogel, 2015a, Storz *et al.*, 2011) (Figure 1.6). Antisense sRNAs may be broadly classified into two categories: firstly, '*cis*-acting' or '*cis*-antisense' sRNAs, that are expressed from the opposite strand overlapping with their target protein, giving extensive complementarity for mRNA-sRNA interactions. Few *cis*-antisense sRNAs have been extensively characterised, in part due to technical difficulties in discerning antisense transcription (Georg and Hess, 2011). Many have been discovered as part of Type I toxin-antitoxin addiction modules, where they are transcribed opposite to toxin protein and prevent host death by sRNA-toxin mRNA inhibition (Brantl, 2012).

A second type of antisense sRNAs, the ‘*trans*-acting’ or ‘intergenic antisense’ sRNAs, are transcribed from intergenic sequences and tend to have smaller regions of complementarity (6–25 nt) with their targets. Many *trans*-acting sRNAs interact with RNA chaperones to facilitate or stabilise sRNA-mRNA base-pairing interactions which would otherwise be weak or transient. The most well studied sRNA chaperone is the Hfq protein found in many gram-negative bacteria, but similar regulatory roles have recently been discovered in ProQ and FinO (Attaiech *et al.*, 2017). A handful of sRNAs have also been found to interact directly with proteins, either by competitive inhibition (Babitzke and Romeo, 2007) or by encouraging dimerisation (Duss *et al.*, 2014).

Bacterial sRNAs may contribute to virulence by regulating the timing of virulence gene expression; either by direct activation/repression, as part of a larger regulatory circuit, or through multiple routes, such as the *Salmonella* sRNA IsrM which attenuates an effector and its corresponding virulence regulator (Gong *et al.*, 2011).

Small RNAs involved in virulence have been found in diverse roles such as adhesion, motility (Vannini *et al.*, 2016), toxin secretion and biofilm formation (Bradley *et al.*, 2011). Many sRNAs involved in virulence are found in horizontally acquired mobile elements and pathogenicity islands (Papenfort and Vogel,

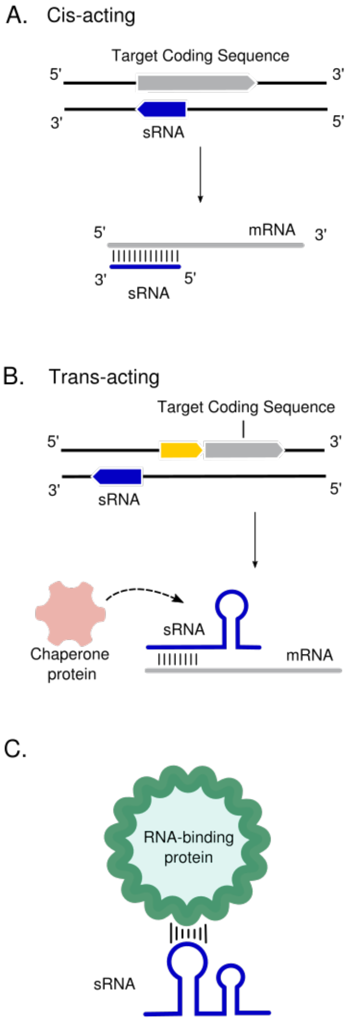


Figure 1.5: Examples of antisense sRNA types and functions. **A.** A *cis*-acting sRNA is expressed antisense to a protein coding sequence. The transcribed sRNA then base-pairs extensively with the mRNA of the target gene. **B.** A *trans*-acting sRNA gene is transcribed from a region distal to its target. The transcribed sRNA then interacts with the target mRNA, utilising a chaperone protein to stabilise the small antisense binding region. **C.** A *trans*-acting sRNA interacts with the RNA-binding site of an RNA-binding protein to inhibit or sequester the protein. Figure based on Figure 1 from Pernitzsch and Sharma (2012).

2010). These sRNAs may control genes within the same element, or interact with the core genome. An example of the latter is *Salmonella* pathogenicity island sRNA InvR, which controls membrane protein expression during cell invasion (Pfeiffer *et al.*, 2007).

Many sRNAs are also involved in the regulation of stress responses and cell survival either through a global stress response or interactions with individual genes (Bhatt *et al.*, 2016, Hoe *et al.*, 2013). Some sRNAs involved in stress responses also contribute to virulence. An example of this can be seen with the *Salmonella* RaoN and *Pseudomonas aeruginosa* PesA sRNAs. Both of these transcripts function generally within the oxidative stress response, that also benefit the bacterium's survival within the host. RaoN is required for replication within macrophages (Lee *et al.*, 2013); PesA is induced in low oxygen conditions and temperatures that may be encountered in the host (Ferrara *et al.*, 2017).

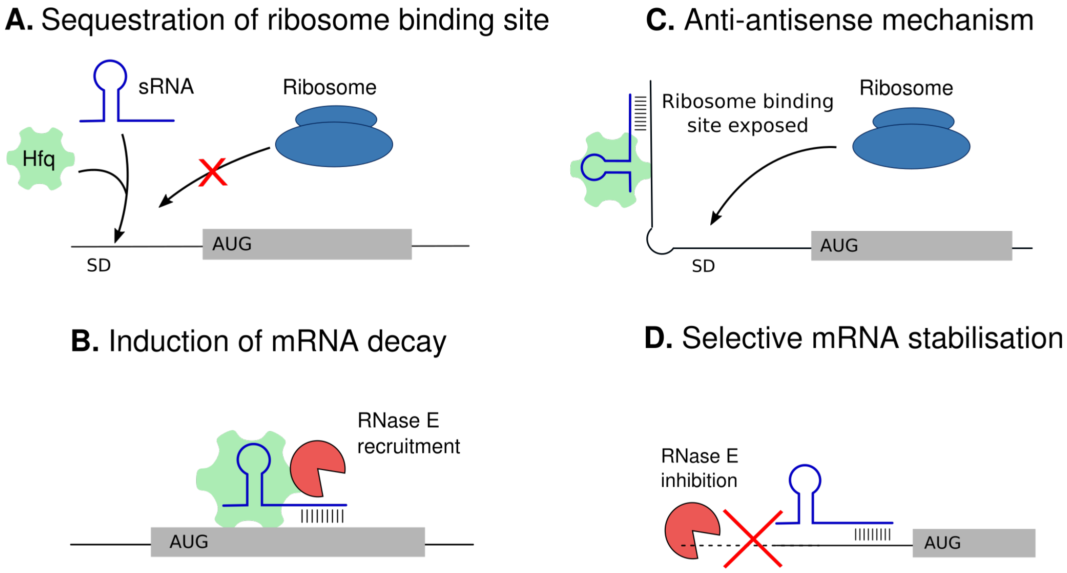

A. Sequestration of ribosome binding site
sRNA
Ribosome
Hfq
AUG
SD
C. Anti-antisense mechanism
Ribosome binding
site exposed
Ribosome
AUG
SD
B. Induction of mRNA decay
RNase E
recruitment
AUG
D. Selective mRNA stabilisation
RNase E
inhibition
AUG

Figure 1.6: Mechanisms of antisense sRNA regulation of mRNA. Antisense sRNAs may decrease protein expression by **(A)** binding to the Shine-dalgarno (SD) region of the target mRNA, preventing the mRNA from being loaded into the ribosome, or **(B)** inducing mRNA decay by recruiting RNase E. Conversely, sRNA binding may increase protein expression by **(C)** causing a conformational change in the mRNA that exposes a ribosome binding site or **(D)** increasing mRNA stability by preventing degradation. These interactions may be stabilised by the RNA chaperone protein Hfq. Based on Figure 1 from Barquist and Vogel (2015a).

The small binding sites of many sRNAs adds flexibility to regulatory networks, as relatively few changes are required to alter the sRNA target range (regulon). An example of regulon reshaping in virulence can be seen in the conserved paralogues *glmY* and *glmZ*. In *E. coli* these sRNAs are part of the core genome in both pathogenic and non-pathogenic strains, however, the GlmY/Z regulon includes virulence genes when these sRNAs are present (Bhatt *et al.*, 2016, Gruber and Sperandio, 2015).

Finally, an example has been found of an sRNA which affects host cells. The *Salmonella* Typhimurium sRNA PinT is expressed within the host cell and both controls the expression of *Salmonella* virulence factors and affects JAK-STAT signalling in the host transcriptome (Westermann *et al.*, 2016).

1.4 ncRNA discovery and annotation

The rate of discovery of ncRNAs has lagged behind that of proteins for several reasons. Most ncRNAs are small molecules, typically shorter than the mRNAs of protein-coding genes. Early ncRNA discovery was a laborious process, where ncRNAs were isolated by size-separation of RNA fractions using sucrose gradients, gels and radioactive labelling (Andersen *et al.*, 1987). Many ncRNAs are also difficult targets for classical mutagenesis due to their size and tolerance of mutations (Gottesman, 2004), the presence of paralogues, and functional redundancy within complex regulatory networks. Bioinformatics approaches are also used to identify and study ncRNA genes. However, unlike proteins, several properties of ncRNAs make them difficult to identify from genome sequence alone. Firstly, ncRNA genes lack shared, recognisable sequence motifs, such as stop/start codons or long open-reading frames found in protein-coding genes. Secondly, many ncRNAs exhibit poor conservation, and instead tend to preserve base-pairs that contribute to structure (Rivas and Eddy, 2001).

1.4.1 Experimental techniques

Historically, regulatory ncRNAs were discovered by chance (Thomason and Storz, 2010). Large numbers of ncRNAs were not discovered until the development of tiling microarrays that could screen for expressed intergenic regions (Gottesman, 2004). Phenotype assays of sRNA deletion mutants have also been used to identify the importance of sRNAs to organismal fitness in different growth conditions (Hobbs *et al.*, 2010, Santiviago *et al.*, 2009). Most of these techniques have now been largely superseded by RNA-seq. Multi-condition transcriptomics are becoming standard for identifying sRNAs. The most notable of these to date have used over a dozen separate growth conditions to comprehensively screen for sRNAs in *E. coli* (Rau *et al.*, 2015) and *Salmonella* Typhimurium (Kröger *et al.*, 2013a).

Several RNA-seq based methods for studying gene fitness and RNA interactions have been applied to bacterial sRNAs. Some sRNA-specific methods have also recently been developed which utilise Hfq-based assays to detect Hfq-sRNA interactions (Holmqvist *et al.*, 2016), or to enrich sRNAs ligated to target mRNA (Melamed *et al.*, 2016). Others can identify novel sRNA chaperones by crosslinking known sRNAs attached to proteins (Smirnov *et al.*, 2016). Transposon insertion sequencing (Tn-Seq or TraDIS), which probes the genome for genes that are essential for growth, is also a useful way to determine functional intergenic regions and validate the contribution of sRNAs to fitness in a specific growth condition (Barquist *et al.*, 2013, van Opijnen *et al.*, 2009).

RNA structure determination is another important step in understanding transcript stability and function. Several methods use structure-dependent covalent modifications of nucleotides. An example of this is SHAPE-seq, which modifies bases that are structurally flexible (and hence less likely to participate in secondary structure) which are then detected by RT-PCR as the modified base inhibits primer extension (Tyrrell, McGinnis, Weeks, and Pielak, Tyrrell *et al.*, Watters and Lucks, 2016). PARIS, another technique, can globally determine RNA-RNA interactions and local regions of secondary structure by cross-linking double-stranded RNA, which is then purified and sequenced (Lu *et al.*, 2016). INTERFACE provides an alternative approach, by surveying regions of ncRNAs that are not participating in secondary structure. This technique uses a reporter system that attenuates transcript elongation when RNA probes are not hybridised (Mihailovic *et al.*, 2018).

1.4.2 Prediction from genomic sequence

Although ncRNAs have few distinguishable features at the sequence level, several approaches have been attempted to predict ncRNAs *de-novo* by comparative genomics. Several ncRNAs were discovered by early comparative genomics studies, which looked for conserved intergenic regions with signals of transcription, such as promoters and terminators (Argaman *et al.*, 2001, Wassarman *et al.*, 2001). Subsequent tools analysed nucleotide substitution patterns within alignments to predict protein-coding potential and look for signals of RNA secondary structure conservation, or the maintenance of thermodynamic stability (Gruber *et al.*, 2010, Pedersen *et al.*, 2006, Rivas and Eddy, 2001).

Conversely, comparative genomics has focused on identifying ncRNA-like sequences in intergenic regions within poorly conserved genomic islands, which are often sources of species or genus-specific genes involved in pathogenicity (Padalon-Brauch *et al.*, 2008a). Promoter and terminator predictions within individual genomes can be used to predict candidate intergenic ncRNAs, where confident promoter models are available (Herbig and Nieselt, 2011, Sridhar *et al.*, 2010). A study by Klein *et al.* (2002) used sequence composition to identify structured ncRNAs, which often have GC-rich sequence compositions, against a background genomic AT bias in thermophilic bacteria.

1.4.3 ncRNA homology search

Non-coding RNAs are generally much more poorly conserved than proteins at the sequence level (Lindgreen *et al.*, 2014). As homology search relies on sequence alignment, homology search techniques for ncRNA are generally less effective than for proteins (McGimpsey, 2019, Rivas and Eddy, 2001, Sun *et al.*, 2012). BLAST is the most widely used tool for nucleotide homology search, and one of the most highly cited scientific papers (Altschul *et al.*, 1990). BLAST is based on a modified Smith-Waterman heuristic pairwise alignment algorithm — homology is determined by finding the optimal alignment of pairs of sequences, factoring in penalties for substitutions, gaps and indels.

More recently developed homology search tools incorporate multiple sequence alignments, and are more effective at determining homologous but divergent sequences (Lindahl and Elofsson, 2000, Park *et al.*, 1998). One such approach is profile hidden Markov model (profile HMM) homology search. Profile HMMs are built from a trusted multiple sequence alignment, and are statistical models that incorporate at each point in a sequence the probability that there is a specific residue, a deletion, or an insertion. By aligning sequences against profile HMMs, more divergent sequences can be detected. Profile HMMs have been shown to be effective for investigating deeply conserved protein (Madera and Gough, 2002) and nucleotide sequences across large phylogenetic distances (Freyhult *et al.*, 2007), and tools like the HMMER (Eddy, 2009) suite of programs can utilise profile HMMs for effective protein and nucleotide homology search (Eddy, 2011).

RNA structure is important for the function of many ncRNAs, and base-pairs which contribute to functional components of secondary structure may be more conserved than the underlying nucleotide sequence. Multiple sequence alignment tools have low efficacy for structured RNAs with low sequence similarity (50–60%) (Gardner *et al.*, 2005), which is a limiting factor for homology search.

The incorporation of structural predictions by RNA-folding software can be used to improve homology search. Early tools to predict RNA secondary structure aimed to maximise the number of base pairs from single primary sequences, and hence identify the most thermodynamically stable secondary structure (Nussinov and Jacobson, 1980, Tinoco *et al.*, 1971, 1973). This approach was subsequently modified to use a nearest-neighbour energy model that captures additional information from stacked base-pairs (Zuker and Stiegler, 1981). Modern RNA folding tools use multiple sequence alignments to incorporate nucleotide or base-pair conservation as a predictor of structure (Puton *et al.*, 2013). This may be done by two approaches: folding a multiple sequence alignment, or folding individual sequences and aligning secondary structures (Gardner and Giegerich, 2004).

Covarying base pairs, where complementary mutations occur to maintain structurally important base pairs, can be incorporated into covariance models as a measure of conserved secondary structure. Covariance models can compensate for large differences in sequences that may occur over large divergence times or due to loose constraints on sequence. While covariance models are the most effective method for ncRNA homology search (Freyhult *et al.*, 2007), they require the query ncRNA to have a conserved secondary structure, excluding ncRNAs with weak associations between structure and function (Peterman *et al.*, 2014, Rivas *et al.*, 2017).

1.5 Possible evolutionary origins of bacterial sRNAs

1.5.1 Are sRNAs poorly conserved or just hard to find?

Most bacterial ncRNA genes appear to have relatively recent origins, as only a handful that are involved in translation, transcription, and translocation exhibit deep conservation (Hoeppner *et al.*, 2012) (Figure 1.7).

Bacterial sRNA gene evolution appears to be exceptionally dynamic over small evolutionary timescales. The development of RNA-seq-based techniques to identify transcripts has led to an explosion of uncharacterised putative sRNAs (Barquist and Vogel, 2015b), with whole-genome screens regularly reporting hundreds of novel sRNAs per genome (Gómez-Lozano *et al.*, 2015, Kröger *et al.*, 2012, Miotto *et al.*, 2012, Rau *et al.*, 2015). Despite their prevalence in bacterial genomes, most known sRNA genes can only be identified across a narrow phylogenetic range. BLAST-based analyses of *Escherichia* (Peer and Margalit, 2014), *Salmonella* (Kröger *et al.*, 2013b, 2012), *Listeria* (Cerutti *et al.*, 2017) and *Campylobacter* (Dugar *et al.*, 2013) sRNAs indicate that the majority are order or genus-specific, with some specific to serovars or even strains.

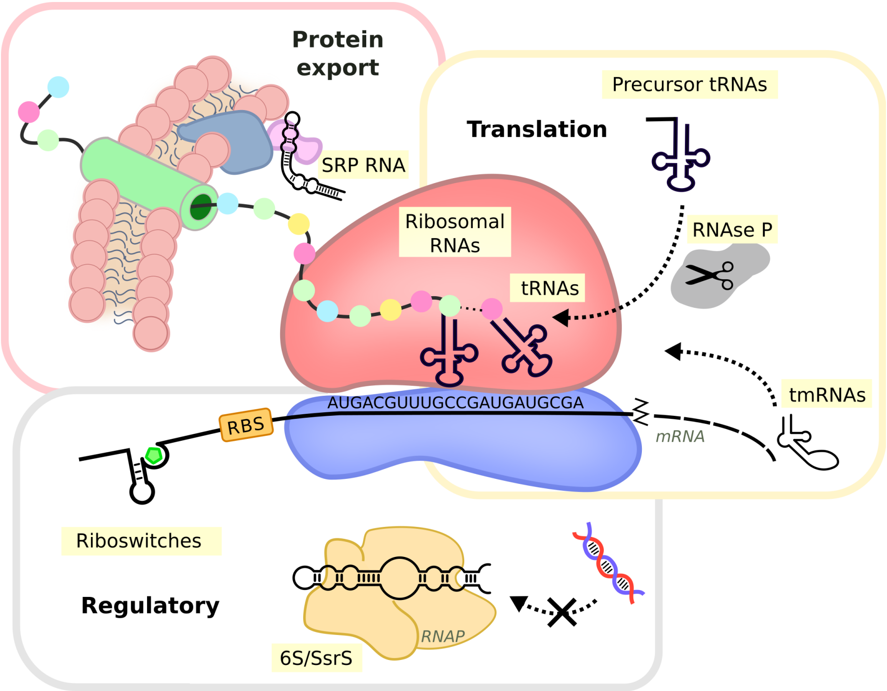


Figure 1.7: Broadly conserved ncRNAs in bacteria. These ncRNAs are involved in core cellular processes: Translation (ribosomal RNAs, tRNAs, RNAse P, and tmRNAs), regulation of transcription (the 6S RNA, aka SsrS) and translation (TPP riboswitch, aka Thi-element), and protein export (the Signal Recognition Particle (SRP) RNA, aka 4.5S). Conserved ncRNAs were identified by phylogenetic distribution of ncRNA annotations in the Rfam database (Hoeppner *et al.*, 2012)

Many bacterial sRNAs appear to present a particularly extreme example of rapid evolution, positing two interesting questions. Firstly, the fact that sRNAs that are known to be homologous through experimental evidence can diverge in nucleotide sequence at a rate near or beyond the limits of our ability to detect genes via homology search, makes them an interesting edge case for testing and improving homology search methods. Tackling this problem may also provide insights into methods used for studying deep-time divergences, but on a timescale that can be observed. Secondly, the presence of many sRNAs with no apparent homologues raises the possibility that these genes have formed *de novo*, rather than through duplication and diversification.

1.5.2 Homology search

One explanation for the apparent lack of conservation of sRNAs is that the rapid evolution of sRNA sequences escapes the sensitivity of homology search tools, as is the case for the Enterobacteriaceae sRNA *micF* (Delihas, 2003). Bacterial sRNAs may be relatively robust to mutation, particularly where function is primarily attributed to "seed regions" of complementarity to their targets that may be as short as 6–8 nt, as is the case for many trans-acting sRNAs (Papenfort *et al.*, 2010, Richter and Backofen, 2012). However, phylogeny-wide analysis of bacterial ncRNA conservation utilising more sensitive profile HMMs and curated CMs from the Rfam database (Nawrocki *et al.*, 2015b), suggests that ncRNAs are overall poorly conserved, with a sharp drop in the number of conserved ncRNAs at the family level and above (Lindgreen *et al.*, 2014) (Figure 1.8). However, as Rfam families are overwhelmingly constructed from conserved sequences it is probable that this is an overestimate of conservation. It has been suggested that the limits of homology search also highlights a "Goldilocks Zone" - the phylogenetic range above which comparative transcriptomics studies would be most useful for identifying conserved transcripts (Lindgreen *et al.*, 2014).

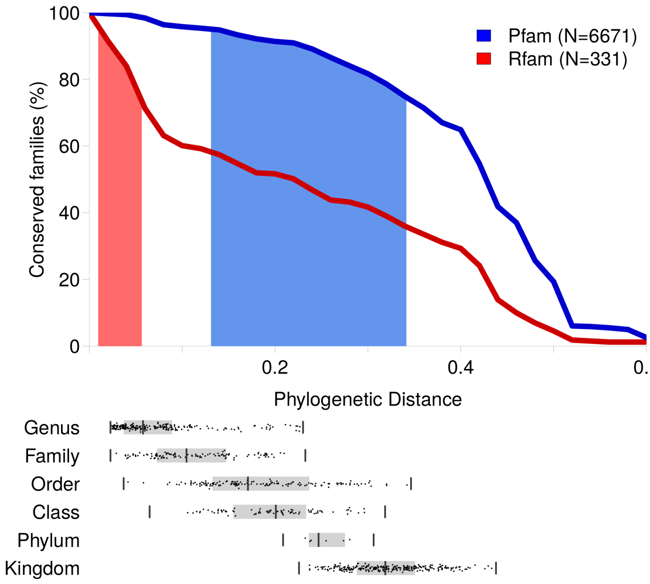


Figure 1.8: Plot showing a comparison of conservation of RNA and protein families. Rfam and Pfam families were identified in 2,562 bacterial genomes using the Infernal and HMMER packages, respectively. The maximum phylogenetic distance spanned by each family was estimated using pairwise phylogenetic distances estimated from 16S rRNA sequences. Top: The percentage of Rfam (N=331) or Pfam (N=6,671) families that are conserved over phylogenetic distances ranging from 0 (closely related) to 0.6 (very divergent). The shaded regions under the curves indicate the "Goldilocks" phylogenetic ranges where between 95 and 75% of RNA (light red) or protein (light blue) families are conserved. Bottom: The distribution of phylogenetic distances between randomly sampled pairs of genomes at 6 taxonomic ranks: genus, family, order, class, phylum and kingdom. At each taxonomic level, pairs were chosen from e.g. the same genus but different species, and so on up the taxonomic hierarchy until pairs of species from the same kingdom (bacteria) and different phyla are shown. Figure adapted from freely available data and code from (Lindgreen *et al.*, 2014).

1.5.3 Potential routes to sRNA evolution

***De novo* evolution**

Bacteria present unique opportunities for both the natural history and experimental study of gene origins, due to their large effective population sizes, rapid generation times, and ease of genetic manipulation. Assuming that a simple sRNA gene requires three components: promotion, termination, and transcript stabilisation, it appears possibile that transcriptional units that function as sRNAs could form "*de novo*" through a variety of mechanisms.

Step-wise generation of genes Theoretically it is possible, albeit unlikely, that selectively advantageous transcript could form through the spontaneous acquisition of promoters and terminators in close proximity by random mutation or deposition by genomic rearrangements. Many bacterial promoter sequences are degenerate, and transcription factor binding sites may arise from just a few point mutations (Stone and Wray, 2001, Yona *et al.*, 2018). Classical Rho-independent terminators have complex sequence requirements to form a stable stem loop and poly-uridine stretch (Ray-Soni *et al.*, 2016). The frequent occurrence of inverted repeat structures in intergenic sequences suggests it may be possible for a terminator-like sequence to be generated *de novo* (Ladoukakis and Eyre-Walker, 2008, Lillo *et al.*, 2002), however it is much more likely that they would be deposited by mobile elements (Naville and Gautheret, 2010), or through genomic rearrangements. Rho-dependent termination appears to occur in a number of intergenic transcripts, however the sequence requirements for these are unclear (Peters *et al.*, 2009). While it is difficult to identify these events without densely sampling closely-related organisms, a recent study has indicated that a genus-specific non-coding transcript has formed due to a rearrangement in *Salmonella* depositing a sequence containing a promoter in a new intergenic region (Raghavan *et al.*, 2015).

Transcriptional noise One possible route for gene formation is through the capture of "transcriptional noise", a term used here to describe transcripts generated through interactions of promiscuous DNA-binding proteins, or through the generation of promoter-like sequences through random mutations in the genome (Eddy, 2013, Struhl, 2007). This shifting pool of diverse transcripts can eventually generate transcripts that interact with other biomolecules and will be captured and stabilised by selection, providing a potential route to *de novo* gene formation through a semi-random process of genomic "sampling".

This sampling process has also been proposed as a source of *de novo* genes in eukaryotes. A comparative transcriptomic study across the mouse taxa showed a high turnover rate of non-genic transcription sites between taxa, and identified lineage-specific transcripts that were weakly but stably expressed (Neme and Tautz, 2016). Bacterial sRNAs may provide a more tractable system in which to observe these processes.

Selection on novel transcripts As has been suggested for eukaryotic microRNAs (Chen and Rajewsky, 2007), bacterial noise transcripts that survive initial purification by selection are likely to have weak initial interactions with target genes, and low-level, condition-specific expression, as has been observed for young non-coding transcripts in *Escherichia coli* and *Salmonella enterica* (Kacharia *et al.*, 2017).

It has also been argued that most antisense transcripts are expressed at levels too low to have major impacts on fitness (Lloréns-Rico *et al.*, 2016). This line of argument also suggests affinity for RNA-binding proteins should follow a similar pattern.

The probabilities of any of these events occurring will vary between due to the differences in genome structure and selection pressure between organisms. For example, genome-wide A/T content may make promoters and terminators more or less likely to occur (Lloréns-Rico *et al.*, 2016), or that large effective population sizes may make selection more effective in some bacteria than others (Novichkov *et al.*, 2009), or that the presence of RNA-binding proteins may make sRNA-based regulation more effective in some lineages than others (Peer and Margalit, 2014).

1.5.4 Exaptation

Exaptation is the process through which existing genomic features are co-opted to provide a new function not previously selected for (Gould and Vrba, 1982). These events can increase the probability of novel sRNA generation, as the prerequisites for stable transcription are provided, and the sequences are already adapted to the host and hence less likely to be deleterious than a *de novo* transcript.

Gene duplication Gene duplication of sRNAs appears to be fairly common, allowing sRNAs to diverge and gain new functions through a process of subfunctionalisation and neofunctionalisation (Caswell *et al.*, 2014, Rastogi and Liberles, 2005). Tandem duplications appear to be an effective way to fine-tune transcript dosage, as has been proposed for the

Pseudomonas RsmX sRNAs (Kay *et al.*, 2005), or to provide a self-regulatory mechanism through subfunctionalisation, such as the Qrr sRNAs involved in quorum sensing in *Vibrio* species (Papenfort and Bassler, 2016). Duplication and neofunctionalisation of sRNAs has also been observed in the PrrF sRNAs involved in regulating iron homeostasis in *Pseudomonas* (Wilderman *et al.*, 2004), and the paragolous OmrA/B sRNAs in *Escherichia coli* (Holmqvist *et al.*, 2010). Horizontally-acquired elements and genomic islands are common in bacterial genomes (Dobrindt *et al.*, 2004), and are frequent sources of novel sRNAs (Padalon-Brauch *et al.*, 2008b, Pichon and Felden, 2005, Tree *et al.*, 2014). Many mobile elements utilise ncRNA regulation to control their replication and integration (Wagner and Simons, 1994), or to ensure their maintenance, such as the sRNA antitoxins in Type I toxin-antitoxin systems that prevent the translation of toxin mRNAs by base-pairing interactions (Brantl, 2012). One recent example of an apparent exaptation event from a mobile genetic element is the *Salmonella* Typhimurium *art200* antisense RNA derived from *tnpA* transposase loci, which regulates pathogenicity responses(Ellis *et al.*, 2017).

UTRs Alternatively, novel transcripts may be generated from existing genes, which provide some protection from selection pressures. Untranslated regions (UTRs) are protected from large-scale deletions by the selective pressures maintaining the associated coding sequences, eliminating the need for *de novo* evolution of transcription and lowering the probability of transcript removal. Recent work has suggested ncRNAs may be generated from within coding sequences themselves (Dar and Sorek, 2018), though no specific function has been proposed for these fragments as of yet.

Several instances of sRNAs derived from the 3′of mRNA transcripts have been characterised (Miyakoshi *et al.*, 2015). This may be in part due to the presence of Rho-independent terminators at many gene ends, which provide several adaptive features for a novel transcript such as a natural termination point, stable secondary structure, and possibly a precursor to an Hfq binding site. The degradation of mRNA transcripts by RNAses can provide a stable fragment which is then free to acquire a secondary function, as appears to be the case for the 3′ UTR derived sRNA CpxQ (Chao and Vogel, 2016) (Figure 1.9A).

Another possibility is the evolution of independent cryptic promoters that allow a UTR-derived sRNA to be transcribed independently from parent gene, as is the case for the sRNA MicL expressed from the 3′ UTR of the *cutC* locus (Figure 1.9B)(Guo *et al.*, 2014). Interestingly, MicL has retained a processing site that is required for function, hinting that other sRNAs such as ArcZ that are similarly processed (Papenfort *et al.*, 2009) may have a 3′ UTR origin.

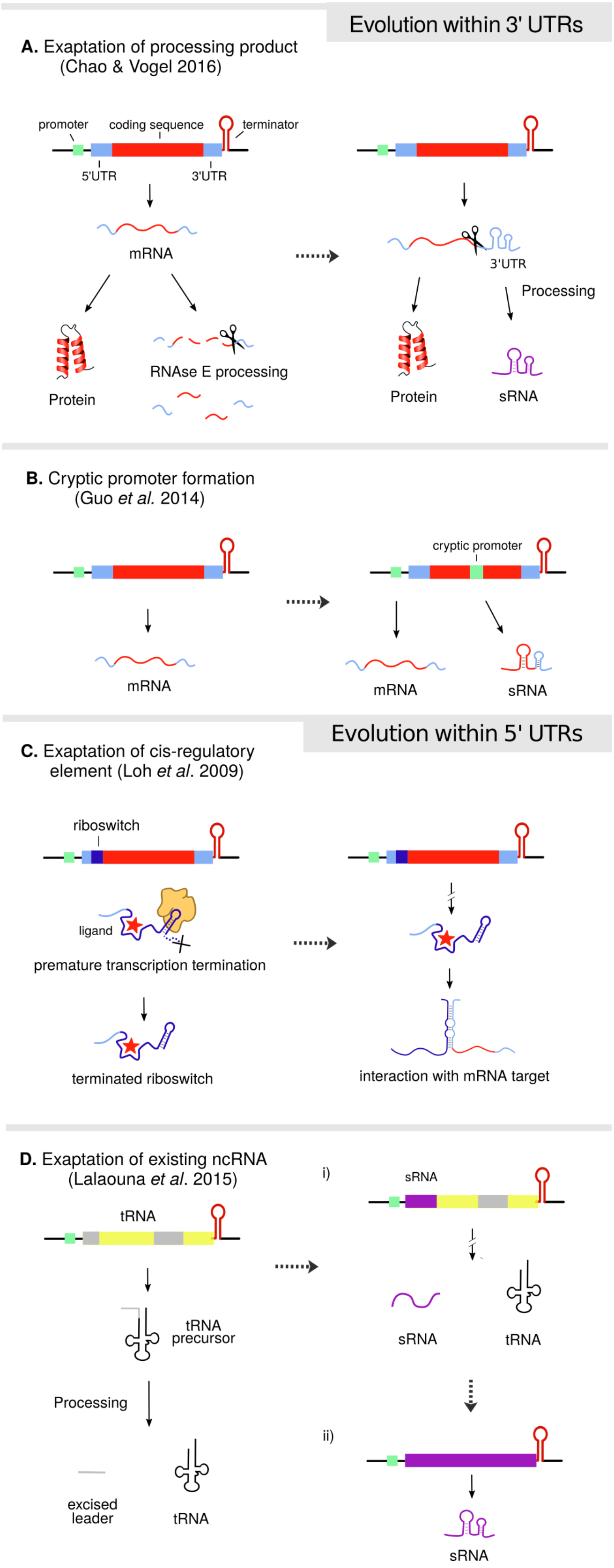


Figure 1.9: Routes for sRNA evolution within "protected spaces" - Functional genomic elements sheltered from deletion by selection on neighbouring genes, that may increase the probability of sRNA evolution. These include known examples from within 3′ UTRs, either derived from processing products (A), or cryptic promoters (B); from existing cis-regulatory elements in 5′ UTRs (C); or from existing ncRNAs, such as tRNAs (D).

Some transcriptional attenuators in 5$'$ UTRs generate stable, or at least abundant, short transcripts that can form the basis for a novel sRNA (Figure 1.9C). Functionally characterised examples of such dual-purpose cis/trans regulators have been found, such as the riboswitch-derived SreA in *Listeria monocytogenes* (Loh *et al.*, 2009), and the leader sequence of the *Salmonella* mgtCBR operon (Choi *et al.*, 2017).

The processing products and intermediaries of core ncRNAs such as ribosomal and tRNAs also have the potential to be exapted. A number of tRNAs and their intermediates have been suggested to interact with Hfq (Lee and Feig, 2008, Zhang *et al.*, 2003), making them especially promising candidates for exaptation events (Gottesman and Storz, 2011b). Spacers excised from pre-tRNA transcripts during tRNA processing have been found to act as sRNA sponges (Figure 1.9D) (Lalaouna *et al.*, 2015).

1.6 Thesis Outline

Many factors that make ncRNA molecules effective and adaptable regulators also make them difficult to study. Bacterial sRNAs are diverse, dynamic regulators of gene expression, however their poor conservation makes them difficult to identify and study using comparative sequence methods. It is difficult to ascertain whether this is due to rapid generation and loss of sRNA genes or due to the limitations of nucleotide homology search. This thesis presents several studies aiming to identify ncRNAs, in particular sRNAs, and better understand their evolution and the tools we use to study them.

Chapter 2 describes an approach to measure the conservation of *Salmonella* sRNAs across the Enterobacteriaceae, and compares the phylogenetic distribution of sRNAs with their predicted evolutionary origins, which are estimated by conservation patterns and flanking proteins. This work demonstrates that many poorly conserved sRNAs are ‘new’ to the *Salmonella* lineage, and have been introduced by mobile genetic elements. A complementary published review (Jose *et al.*, 2019) is also included in the appendix, which further discusses the mechanisms by which new sRNA genes may arise, in particular *via* the capture of transcriptional noise.

Many sRNAs are only expressed under specific growth conditions, enabling careful timing of environmental responses in bacteria. *Pseudomonas syringae* pv. *actinidiae* (*Psa*) is a plant pathogen which causes kiwifruit canker disease, and has had a significant impact on kiwifruit agriculture worldwide. Transcriptomes of *Psa* grown *in vitro* in multiple growth mediums were generated for this thesis by collaborators at Plant and Food Research, Auckland, NZ. In Chapter 3, these transcriptomes are used to study changes in gene expression relevant to pathogenicity between *in vitro* samples, and also compared to another data-set of *Psa* grown *in planta*. In Chapter 4, these *in vitro* and *in planta* transcriptomes are also used to annotate potential ncRNAs in *Psa*. Methods developed for sRNA homology search in Chapter 3 are used to measure candidate ncRNA conservation across *Pseudomonas* genomes. Conservation, structural predictions and expression patterns are used to rank candidate ncRNAs, aiming to find genes that function in virulence.

Comparative transcriptomics between more distantly related organisms (for example, at the phyla level) has been shown to be an effective way to identify novel conserved ncRNAs. Chapter 5 presents initial work two bacterial genome assembly and comparison projects, which aims to form the basis of a larger project of ncRNA identification by comparative transcriptomics.

References

Akopian, D., Shen, K., Zhang, X., and Shan, S.-O. (2013). Signal recognition particle: an essential protein-targeting machine. *Annu. Rev. Biochem.*, *82*, 693–721.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, *215*(3), 403–410.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169.

Andersen, J., Delihas, N., Ikenaka, K., Green, P. J., Pines, O., Ilercil, O., and Inouye, M. (1987). The isolation and characterization of RNA coded by the *micF* gene in *Escherichia coli*. *Nucleic Acids Res.*, *15*(5), 2089–2101.

Andrews, S. (2010). FastQC A Quality Control tool for High Throughput Sequence Data.

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E. G., Margalit, H., and Altuvia, S. (2001). Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.*, *11*(12), 941–950.

Attaiech, L., Glover, J. N. M., and Charpentier, X. (2017). RNA Chaperones Step Out of Hfq's Shadow. *Trends Microbiol.*, *25*(4), 247–249.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, *9*, 75.

Babitzke, P. and Romeo, T. (2007). CsrB sRNA family: sequestration of RNA-binding regulatory proteins. *Curr. Opin. Microbiol.*, *10*(2), 156–163.

Balloux, F. and van Dorp, L. (2017). Q&A: What are pathogens, and what have they done to and for us? *BMC Biology*, *15*(1), 91.

Barquist, L., Langridge, G. C., Turner, D. J., Phan, M.-D., Turner, A. K., Bateman, A., Parkhill, J., Wain, J., and Gardner, P. P. (2013). A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium. *Nucleic Acids Res.*, *41*(8), 4549–4564.

Barquist, L. and Vogel, J. (2015a). Accelerating Discovery and Functional Analysis of Small RNAs with New Technologies. *Annu. Rev. Genet.*, *49*, 367–394.

Barquist, L. and Vogel, J. (2015b). Accelerating Discovery and Functional Analysis of Small RNAs with New Technologies. *Annu. Rev. Genet.*, *49*, 367–394.

Barrangou, R. (2015). The roles of CRISPR-Cas systems in adaptive immunity and beyond. *Curr. Opin. Immunol.*, *32*, 36–41.

Barrick, J. E., Sudarsan, N., Weinberg, Z., Ruzzo, W. L., and Breaker, R. R. (2005). 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *RNA*, *11*(5), 774–784.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E. Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Furey, W. S., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Jones, T. A. H., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ng, B. L., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Pinkard, D. C., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Rodriguez, A. C., Roe, P. M., Rogers, J., Bacigalupo, M. C. R., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Sohna, J. E., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., van de Vondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*, 53.

Bhatt, S., Egan, M., Jenkins, V., Muche, S., and El-Fenej, J. (2016). The Tip of the Iceberg: On the Roles of Regulatory Small RNAs in the Virulence of Enterohemorrhagic and Enteropathogenic *Escherichia coli*. *Front. Cell. Infect. Microbiol.*, *6*, 105.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Bradley, E. S., Bodi, K., Ismail, A. M., and Camilli, A. (2011). A genome-wide approach to discovery of small RNAs involved in regulation of virulence in *Vibrio cholerae*. *PLoS Pathog.*, *7*(7), e1002126.

Brantl, S. (2012). Bacterial type I toxin-antitoxin systems. *RNA Biol.*, *9*(12), 1488–1490.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, *34*(5), 525–527.

Brownlee, G. G. (1971). Sequence of 6S RNA of *E. coli*. *Nat. New Biol.*, *229*(5), 147–149.

Caswell, C. C., Oglesby-Sherrouse, A. G., and Murphy, E. R. (2014). Sibling rivalry: related bacterial small RNAs and their redundant and non-redundant roles. *Front. Cell. Infect. Microbiol.*, *4*, 151.

Cech, T. R. and Steitz, J. A. (2014). The noncoding RNA revolution – trashing old rules to forge new ones. *Cell*, *157*(1), 77–94.

Cerutti, F., Mallet, L., Painset, A., Hoede, C., Moisan, A., Bécavin, C., Duval, M., Dussurget, O., Cossart, P., Gaspin, C., and Chiapello, H. (2017). Unraveling the evolution and co-evolution of small regulatory RNAs and coding genes in Listeria. *BMC Genomics*, *18*(1), 882.

Chao, Y. and Vogel, J. (2016). A 3′ UTR-Derived Small RNA Provides the Regulatory Noncoding Arm of the Inner Membrane Stress Response. *Mol. Cell*, *61*, 352–363.

Chen, K. and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.*, *8*(2), 93–103.

Cherf, G. M., Lieberman, K. R., Rashid, H., Lam, C. E., Karplus, K., and Akeson, M. (2012). Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nat. Biotechnol.*, *30*(4), 344–348.

Choi, E., Han, Y., Cho, Y.-J., Nam, D., and Lee, E.-J. (2017). A trans-acting leader RNA from a Salmonella virulence gene. *PNAS*, *114*(38), 10232–10237.

Conrad, S. E. and Campbell, J. L. (1979). Role of plasmid-coded RNA and ribonuclease III in plasmid DNA replication. *Cell*, *18*(1), 61–71.

Cotter, P. A. and DiRita, V. J. (2000). Bacterial virulence gene regulation: an evolutionary perspective. *Annu. Rev. Microbiol.*, *54*, 519–565.

Creecy, J. P. and Conway, T. (2015). Quantitative bacterial transcriptomics with RNA-seq. *Curr. Opin. Microbiol.*, *23*, 133–140.

Croucher, N. J., Fookes, M. C., Perkins, T. T., Turner, D. J., Marguerat, S. B., Keane, T., Quail, M. A., He, M., Assefa, S., Bähler, J., Kingsley, R. A., Parkhill, J., Bentley, S. D., Dougan, G., and Thomson, N. R. (2009). A simple method for directional transcriptome sequencing using Illumina technology. *Nucleic Acids Res.*, *37*(22), e148.

Dar, D., Shamir, M., Mellin, J. R., Koutero, M., Stern-Ginossar, N., Cossart, P., and Sorek, R. (2016). Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science*, *352*(6282), aad9822.

Dar, D. and Sorek, R. (2018). Bacterial Noncoding RNAs Excised from within Protein-Coding Transcripts. *MBio*, *9*(5), e01730–18.

Dayhoff, M. O. and Ledley, R. S. (1962). Comprotein: A Computer Program to Aid Primary Protein Structure Determination. In *Proceedings of the December 4-6, 1962, Fall Joint Computer Conference*, AFIPS '62 (Fall), New York, NY, USA, 262–274. ACM.

Delcher, A. L., Bratke, K. A., Powers, E. C., and Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, *23*(6), 673–679.

Delihas, N. (2003). Annotation and evolutionary relationships of a small regulatory RNA gene *micF* and its target *ompF* in *Yersinia* species. *BMC Microbiol.*, *3*, 13.

Derome, N., Gauthier, J., Vincent, A. T., and Charette, S. J. (2018). A brief history of bioinformatics. *Briefings in Bioinformatics*, *online*(eprint), bby063.

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

Dorman, C. J., Bhriain, N. N., and Dorman, M. J. (2018). The Evolution of Gene Regulatory Mechanisms in Bacteria. In P. H. Rampelotto (Ed.), *Molecular Mechanisms of Microbial Evolution*, 125–152. Cham: Springer International Publishing.

Dugar, G., Herbig, A., Förstner, K. U., Heidrich, N., Reinhardt, R., Nieselt, K., and Sharma, C. M. (2013). High-resolution transcriptome maps reveal strain-specific regulatory features of multiple Campylobacter jejuni isolates. *PLoS Genet.*, *9*(5), e1003495.

Durbin, R., Eddy, S. R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Duss, O., Michel, E., Yulikov, M., Schubert, M., Jeschke, G., and Allain, F. H.-T. (2014). Structural basis of the non-coding RNA RsmZ acting as a protein sponge. *Nature*, *509*(7502), 588–592.

Ecker, D. J., Sampath, R., Willett, P., Wyatt, J. R., Samant, V., Massire, C., Hall, T. A., Hari, K., McNeil, J. A., Büchen-Osmond, C., and Budowle, B. (2005). The Microbial Rosetta Stone Database: A compilation of global and emerging infectious microorganisms and bioterrorist threat agents. *BMC Microbiol.*, *5*(1), 19.

Eddy, S. R. (2001). Non–coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, *2*(12), 919–929.

Eddy, S. R. (2009). A New Generation of Homology Search Tools Based on Probabilistic Inference. *Genome Inform.*, *23*(1), 205–11.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, *7*(10), e1002195.

Eddy, S. R. (2013). The ENCODE project: missteps overshadowing a success. *Curr. Biol.*, *23*(7), R259–61.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138.

Ellis, M. J., Trussler, R. S., Charles, O., and Haniford, D. B. (2017). A transposon-derived small RNA regulates gene expression in Salmonella Typhimurium. *Nucleic Acids Res.*, *45*(9), 5470–5486.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evo.*, *17*(6), 368–376.

Feng, Y., Zhang, Y., Ying, C., Wang, D., and Du, C. (2015). Nanopore-based fourth-generation DNA sequencing technology. *Genomics Proteomics Bioinformatics*, *13*(1), 4–16.

Ferrara, S., Falcone, M., Macchi, R., Bragonzi, A., Girelli, D., Cariani, L., Cigana, C., and Bertoni, G. (2017). The PAPI-1 pathogenicity island-encoded small RNA PesA influences *Pseudomonas aeruginosa* virulence and modulates pyocin S3 production. *PLoS One*, *12*(6), e0180386.

Freyhult, E. K., Bollback, J. P., and Gardner, P. P. (2007). Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.*, *17*(1), 117–125.

Fullwood, M. J., Wei, C.-L., Liu, E. T., and Ruan, Y. (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.*, *19*(4), 521–532.

Gardner, P. P. and Giegerich, R. (2004). A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, *5*, 140.

Gardner, P. P., Wilm, A., and Washietl, S. (2005). A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, *33*(8), 2433–2439.

Georg, J. and Hess, W. R. (2011). *cis*-antisense RNA, another level of gene regulation in bacteria. *Microbiol. Mol. Biol. Rev.*, *75*(2), 286–300.

Gómez-Lozano, M., Marvig, R. L., Molina-Santiago, C., Tribelli, P. M., Ramos, J.-L., and Molin, S. (2015). Diversity of small RNAs expressed in Pseudomonas species. *Environ. Microbiol. Rep.*, *7*(2), 227–236.

Gong, H., Vu, G.-P., Bai, Y., Chan, E., Wu, R., Yang, E., Liu, F., and Lu, S. (2011). A *Salmonella* small non-coding RNA facilitates bacterial invasion and intracellular replication by modulating the expression of virulence factors. *PLoS Pathog.*, *7*(9), e1002120.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, *17*(6), 333–351.

Gottesman, S. (2004). The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annu. Rev. Microbiol.*, *58*, 303–328.

Gottesman, S. (2005). Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet.*, *21*(7), 399–404.

Gottesman, S. and Storz, G. (2011a). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb. Perspect. Biol.*, *3*(12).

Gottesman, S. and Storz, G. (2011b). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb. Perspect. Biol.*, *3*(12), a003798.

Gould, S. J. and Vrba, E. S. (1982). Exaptation—a Missing Term in the Science of Form. *Paleobiology*, *8*(1), 4–15.

Gripenland, J., Netterling, S., Loh, E., Tiensuu, T., Toledo-Arana, A., and Johansson, J. (2010). RNAs: regulators of bacterial virulence. *Nat. Rev. Microbiol.*, *8*(12), 857–866.

Groisman, E. A. and Ochman, H. (1996). Pathogenicity islands: bacterial evolution in quantum leaps. *Cell*, *87*(5), 791–794.

Gruber, A. R., Findeiß, S., Washietl, S., Hofacker, I. L., and Stadler, P. F. (2010). RNAz 2.0: improved noncoding RNA detection. *Pacific Symposium on Biocomputing*, 69–79.

Gruber, C. C. and Sperandio, V. (2015). Global analysis of posttranscriptional regulation by GlmY and GlmZ in enterohemorrhagic *Escherichia coli* O157:H7. *Infect. Immun.*, *83*(4), 1286–1295.

Guo, M. S., Updegrove, T. B., Gogol, E. B., Shabalina, S. A., Gross, C. A., and Storz, G. (2014). MicL, a new σE-dependent sRNA, combats envelope stress by repressing synthesis of Lpp, the major outer membrane lipoprotein. *Genes Dev.*, *28*(14), 1620–1634.

Hausner, G., Hafez, M., and Edgell, D. R. (2014). Bacterial group I introns: mobile RNA catalysts. *Mob. DNA*, *5*(1), 8.

Herbig, A. and Nieselt, K. (2011). nocoRNAc: characterization of non-coding RNAs in prokaryotes. *BMC Bioinformatics*, *12*, 40.

Hobbs, E. C., Astarita, J. L., and Storz, G. (2010). Small RNAs and small proteins involved in resistance to cell envelope stress and acid shock in *Escherichia coli*: analysis of a barcoded mutant collection. *J. Bacteriol.*, *192*(1), 59–67.

Hoe, C.-H., Raabe, C. A., Rozhdestvensky, T. S., and Tang, T.-H. (2013). Bacterial sRNAs: regulation in stress. *Int. J. Med. Microbiol.*, *303*(5), 217–229.

Hoeppner, M. P., Gardner, P. P., and Poole, A. M. (2012). Comparative analysis of RNA families reveals distinct repertoires for each domain of life. *PLoS Comput. Biol.*, *8*(11), e1002752.

Holmqvist, E., Reimegård, J., Sterk, M., Grantcharova, N., Römling, U., and Wagner, E. G. H. (2010). Two antisense RNAs target the transcriptional regulator CsgD to inhibit curli synthesis. *EMBO J.*, *29*(11), 1840–1850.

Holmqvist, E., Wright, P. R., Li, L., Bischler, T., Barquist, L., Reinhardt, R., Backofen, R., and Vogel, J. (2016). Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo. *EMBO J.*, *35*(9), 991–1011.

Hör, J., Gorski, S. A., and Vogel, J. (2018). Bacterial RNA Biology on a Genome Scale. *Mol. Cell*, *70*(5), 785–799.

Hrdlickova, R., Toloue, M., and Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip. Rev. RNA*, *8*(1), ePub.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., and Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.*, *34*(8), 2115–2122.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., and Banfield, J. F. (2016). A new view of the tree of life. *Nat. Microbiol*, *1*, 16048.

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 119.

Inouye, M. and Delihas, N. (1988). Small RNAs in the prokaryotes: a growing list of diverse roles. *Cell*, *53*(1), 5–7.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, *11*(2), 163–166.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., Snutch, T. P., Tee, L., Paten, B., Phillippy, A. M., Simpson, J. T., Loman, N. J., and Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, *36*(4), 338–345.

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.*, *17*(1), 239.

Johansson, J., Mandin, P., Renzoni, A., Chiaruttini, C., Springer, M., and Cossart, P. (2002). An RNA thermosensor controls expression of virulence genes in *Listeria monocytogenes*. *Cell*, *110*(5), 551–561.

Jose, B. R., Gardner, P. P., and Barquist, L. (2019). Transcriptional noise and exaptation as sources for bacterial sRNAs. *Biochem. Soc. Trans.*.

Kacharia, F. R., Millar, J. A., and Raghavan, R. (2017). Emergence of New sRNAs in Enteric Bacteria is Associated with Low Expression and Rapid Evolution. *J. Mol. Evol.*, *84*(4), 204–213.

Kado, C. I. (2009). Horizontal gene transfer: sustaining pathogenicity and optimizing host-pathogen interactions. *Mol. Plant Pathol.*, *10*(1), 143–150.

Kay, E., Dubuis, C., and Haas, D. (2005). Three small RNAs jointly ensure secondary metabolism and biocontrol in *Pseudomonas fluorescens* CHA0. *PNAS*, *102*(47), 17136–17141.

Keller, M. W., Rambo-Martin, B. L., Wilson, M. M., Ridenour, C. A., Shepard, S. S., Stark, T. J., Neuhaus, E. B., Dugan, V. G., Wentworth, D. E., and Barnes, J. R. (2018). Direct RNA Sequencing of the Coding Complete Influenza A Virus Genome. *Sci. Rep.*, *8*(1), 14408.

Klein, R. J., Misulovin, Z., and Eddy, S. R. (2002). Noncoding RNA genes identified in AT-rich hyperthermophiles. *PNAS*, *99*(11), 7542–7547.

Korlach, J., Bjornson, K. P., Chaudhuri, B. P., Cicero, R. L., Flusberg, B. A., Gray, J. J., Holden, D., Saxena, R., Wegener, J., and Turner, S. W. (2010). Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol.*, *472*, 431–455.

Kröger, C., Colgan, A., Srikumar, S., Händler, K., Sivasankaran, S. K., Hammarlöf, D. L., Canals, R., Grissom, J. E., Conway, T., Hokamp, K., and Hinton, J. C. D. (2013a).

An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium. *Cell Host Microbe*, *14*(6), 683–695.

Kröger, C., Colgan, A., Srikumar, S., Händler, K., Sivasankaran, S. K., Hammarlöf, D. L., Canals, R., Grissom, J. E., Conway, T., Hokamp, K., and Hinton, J. C. D. (2013b). An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium. *Cell Host Microbe*, *14*(6), 683–695.

Kröger, C., Dillon, S. C., Cameron, A. D. S., Papenfort, K., Sivasankaran, S. K., Hokamp, K., Chao, Y., Sittka, A., Hébrard, M., Händler, K., Colgan, A., Leekitcharoenphon, P., Langridge, G. C., Lohan, A. J., Loftus, B., Lucchini, S., Ussery, D. W., Dorman, C. J., Thomson, N. R., Vogel, J., and Hinton, J. C. D. (2012). The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. *PNAS*, *109*(20), E1277–86.

Kudla, G., Granneman, S., Hahn, D., Beggs, J. D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *PNAS*, *108*(24), 10010–10015.

Ladoukakis, E. D. and Eyre-Walker, A. (2008). The excess of small inverted repeats in prokaryotes. *J. Mol. Evol.*, *67*(3), 291–300.

Lalaouna, D., Carrier, M.-C., Semsey, S., Brouard, J.-S., Wang, J., Wade, J. T., and Massé, E. (2015). A $3'$ External Transcribed Spacer in a tRNA Transcript Acts as a Sponge for Small RNAs to Prevent Transcriptional Noise. *Mol. Cell*, *58*(3), 393–405.

Lambowitz, A. M. and Zimmerly, S. (2011). Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb. Perspect. Biol.*, *3*(8), a003616.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, *9*(4), 357–359.

Lee, T. and Feig, A. L. (2008). The RNA binding protein Hfq interacts specifically with tRNAs. *RNA*, *14*(3), 514–523.

Lee, Y. H., Kim, S., Helmann, J. D., Kim, B.-H., and Park, Y. K. (2013). RaoN, a small RNA encoded within *Salmonella* pathogenicity island-11, confers resistance to macrophage-induced stress. *Microbiology*, *159*(Pt 7), 1366–1378.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C., and Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, *456*(7221), 464–469.

Lillo, F., Basile, S., and Mantegna, R. N. (2002). Comparative genomics study of inverted repeats in bacteria. *Bioinformatics*, *18*(7), 971–979.

Lindahl, E. and Elofsson, A. (2000). Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.*, *295*(3), 613–625.

Lindgreen, S., Umu, S. U., Lai, A. S.-W., Eldai, H., Liu, W., McGimpsey, S., Wheeler, N. E., Biggs, P. J., Thomson, N. R., Barquist, L., Poole, A. M., and Gardner, P. P. (2014). Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Comput. Biol.*, *10*(10), e1003907.

Lioliou, E., Sharma, C. M., Altuvia, Y., Caldelari, I., Romilly, C., Helfer, A.-C., Margalit, H., and Romby, P. (2013). In vivo mapping of RNA-RNA interactions in *Staphylococcus aureus* using the endoribonuclease III. *Methods*, *63*(2), 135–143.

Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.-H., Bork, P., Glass, J. I., Serrano, L., and Lluch-Senar, M. (2016). Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci Adv*, *2*(3), e1501363.

Loh, E., Dussurget, O., Gripenland, J., Vaitkevicius, K., Tiensuu, T., Mandin, P., Repoila, F., Buchrieser, C., Cossart, P., and Johansson, J. (2009). A trans-acting riboswitch controls expression of the virulence regulator PrfA in Listeria monocytogenes. *Cell*, *139*(4), 770–779.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, *15*(12), 550.

Lu, Z., Guan, X., Schmidt, C. A., and Matera, A. G. (2014). RIP-seq analysis of eukaryotic Sm proteins identifies three major categories of Sm-containing ribonucleoproteins. *Genome Biol.*, *15*(1), R7.

Lu, Z., Zhang, Q. C., Lee, B., Flynn, R. A., Smith, M. A., Robinson, J. T., Davidovich, C., Gooding, A. R., Goodrich, K. J., Mattick, J. S., Mesirov, J. P., Cech, T. R., and Chang, H. Y. (2016). RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell*, *165*(5), 1267–1279.

Madera, M. and Gough, J. (2002). A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.*, *30*(19), 4321–4328.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12.

McGimpsey, S. (2019). The Twilight Zone of Nucleotide Homology. Master's thesis, University of Otago.

Melamed, S., Peer, A., Faigenbaum-Romm, R., Gatt, Y. E., Reiss, N., Bar, A., Altuvia, Y., Argaman, L., and Margalit, H. (2016). Global Mapping of Small RNA-Target Interactions in Bacteria. *Mol. Cell*, *63*(5), 884–897.

"Microbiology by numbers" (2011). Microbiology by numbers. *Nat. Rev. Microbiol.*, *9*, 628.

Mihailovic, M. K., Vazquez-Anderson, J., Li, Y., Fry, V., Vimalathas, P., Herrera, D., Lease, R. A., Powell, W. B., and Contreras, L. M. (2018). High-throughput *in vivo* mapping of RNA accessible interfaces to identify functional sRNA binding sites. *Nat. Commun.*, *9*(1), 4084.

Miotto, P., Forti, F., Ambrosi, A., Pellin, D., Veiga, D. F., Balazsi, G., Gennaro, M. L., Di Serio, C., Ghisotti, D., and Cirillo, D. M. (2012). Genome-wide discovery of small RNAs in Mycobacterium tuberculosis. *PLoS One*, *7*(12), e51950.

Miyakoshi, M., Chao, Y., and Vogel, J. (2015). Regulatory small RNAs from the $3'$ regions of bacterial mRNAs. *Curr. Opin. Microbiol.*, *24*(0), 132–139.

Moore, S. D. and Sauer, R. T. (2007). The tmRNA system for translational surveillance and ribosome rescue. *Annu. Rev. Biochem.*, *76*, 101–124.

Morris, K. V. and Mattick, J. S. (2014). The rise of regulatory RNA. *Nat. Rev. Genet.*, *15*(6), 423–437.

Mount, D. W. (2004). *Bioinformatics: Sequence and Genome Analysis*. CSHL Press.

Moxon, E. R., Rainey, P. B., Nowak, M. A., and Lenski, R. E. (1994). Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.*, *4*(1), 24–33.

Naville, M. and Gautheret, D. (2010). Transcription attenuation in bacteria: theme and variations. *Brief. Funct. Genomics*, *9*(2), 178–189.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D. (2015a). Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, *43*(Database issue), D130–7.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D. (2015b). Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, *43*(Database issue), D130–7.

Nawrocki, E. P. and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933–2935.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, *48*(3), 443 – 453.

Neme, R. and Tautz, D. (2016). Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *Elife*, *5*, e09977.

Novichkov, P. S., Wolf, Y. I., Dubchak, I., and Koonin, E. V. (2009). Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. *J. Bacteriol.*, *191*(1), 65–73.

Nussinov, R. and Jacobson, A. B. (1980). Fast algorithm for predicting the secondary structure of single-stranded RNA. *PNAS*, *77*(11), 6309–6313.

Padalon-Brauch, G., Hershberg, R., Elgrably-Weiss, M., Baruch, K., Rosenshine, I., Margalit, H., and Altuvia, S. (2008b). Small RNAs encoded within genetic islands of Salmonella typhimurium show host-induced expression and role in virulence. *Nucleic Acids Res.*, *36*(6), 1913–1927.

Padalon-Brauch, G., Hershberg, R., Elgrably-Weiss, M., Baruch, K., Rosenshine, I., Margalit, H., and Altuvia, S. (2008a). Small RNAs encoded within genetic islands of *Salmonella* Typhimurium show host-induced expression and role in virulence. *Nucleic Acids Res.*, *36*(6), 1913–1927.

Papenfort, K. and Bassler, B. L. (2016). Quorum sensing signal-response systems in Gram-negative bacteria. *Nat. Rev. Microbiol.*, *14*(9), 576–588.

Papenfort, K., Bouvier, M., Mika, F., Sharma, C. M., and Vogel, J. (2010). Evidence for an autonomous 5′ target recognition domain in an Hfq-associated small RNA. *PNAS*, *107*(47), 20435–20440.

Papenfort, K., Said, N., Welsink, T., Lucchini, S., Hinton, J. C. D., and Vogel, J. (2009). Specific and pleiotropic patterns of mRNA regulation by ArcZ, a conserved, Hfq-dependent small RNA. *Mol. Microbiol.*, *74*(1), 139–158.

Papenfort, K. and Vogel, J. (2010). Regulatory RNA in bacterial pathogens. *Cell Host Microbe*, *8*(1), 116–127.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, *284*(4), 1201–1210.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.*, *37*(18), e123.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, *14*(4), 417–419.

Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D. (2006). Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, *2*(4), e33.

Peer, A. and Margalit, H. (2014). Evolutionary patterns of Escherichia coli small RNAs and their regulatory interactions. *RNA*, *20*(7), 994–1003.

Pernitzsch, S. R. and Sharma, C. M. (2012). Transcriptome complexity and riboregulation in the human pathogen *Helicobacter pylori*. *Front. Cell. Infect. Microbiol.*, *2*, 14.

Peterman, N., Lavi-Itzkovitz, A., and Levine, E. (2014). Large-scale mapping of sequence-function relations in small regulatory RNAs reveals plasticity and modularity. *Nucleic Acids Res.*, *42*(19), 12177–12188.

Peters, J. M., Mooney, R. A., Kuan, P. F., Rowland, J. L., Keles, S., and Landick, R. (2009). Rho directs widespread termination of intragenic and stable RNA transcription. *PNAS*, *106*(36), 15406–15411.

Pfeiffer, V., Sittka, A., Tomer, R., Tedin, K., Brinkmann, V., and Vogel, J. (2007). A small non-coding RNA of the invasion gene island (SPI-1) represses outer membrane protein synthesis from the *Salmonella* core genome. *Mol. Microbiol.*, *66*(5), 1174–1191.

Pichon, C. and Felden, B. (2005). Small RNA genes expressed from Staphylococcus aureus genomic and pathogenicity islands with specific expression among pathogenic strains. *PNAS*, *102*(40), 14249–14254.

Price, M. N., Wetmore, K. M., Deutschbauer, A. M., and Arkin, A. P. (2016). A Comparison of the Costs and Benefits of Bacterial Gene Expression. *PLoS One*, *11*(10), e0164314.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Res.*, *40*(Database issue), D290–301.

Puton, T., Kozlowski, L. P., Rother, K. M., and Bujnicki, J. M. (2013). CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res.*, *41*(7), 4307–4323.

Quick, J., Quinlan, A. R., and Loman, N. J. (2014). A reference bacterial genome dataset generated on the MinION^{TM} portable single-molecule nanopore sequencer. *Gigascience*, *3*, 22.

Raghavan, R., Kacharia, F. R., Millar, J. A., Sislak, C. D., and Ochman, H. (2015). Genome rearrangements can make and break small RNA genes. *Genome Biol. Evol.*, *7*(2), 557–566.

Rastogi, S. and Liberles, D. A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol. Biol.*, *5*, 28.

Rau, M. H., Bojanovič, K., Nielsen, A. T., and Long, K. S. (2015). Differential expression of small RNAs under chemical stress and fed-batch fermentation in *E. coli*. *BMC Genomics*, *16*, 1051.

Ray-Soni, A., Bellecourt, M. J., and Landick, R. (2016). Mechanisms of Bacterial Transcription Termination: All Good Things Must End. *Annu. Rev. Biochem.*, *85*, 319–347.

Rederstorff, M., Bernhart, S. H., Tanzer, A., Zywicki, M., Perfler, K., Lukasser, M., Hofacker, I. L., and Hüttenhofer, A. (2010). RNPomics: defining the ncRNA transcriptome by cDNA library generation from ribonucleo-protein particles. *Nucleic Acids Res.*, *38*(10), e113.

Rich, A. and RajBhandary, U. L. (1976). Transfer RNA: molecular structure, sequence, and properties. *Annu. Rev. Biochem.*, *45*, 805–860.

Richter, A. S. and Backofen, R. (2012). Accessibility and conservation: general features of bacterial small RNA-mRNA interactions? *RNA Biol.*, *9*(7), 954–965.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, *43*(7), e47.

Rivas, E., Clements, J., and Eddy, S. R. (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods*, *14*(1), 45–48.

Rivas, E. and Eddy, S. R. (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, *2*, 8.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Saliba, A.-E., C Santos, S., and Vogel, J. (2017). New RNA-seq approaches for the study of bacterial pathogens. *Curr. Opin. Microbiol.*, *35*, 78–87.

Santiviago, C. A., Reynolds, M. M., Porwollik, S., Choi, S.-H., Long, F., Andrews-Polymenis, H. L., and McClelland, M. (2009). Analysis of pools of targeted *Salmonella* deletion mutants identifies novel genes affecting fitness during competitive infection in mice. *PLoS Pathog.*, *5*(7), e1000477.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069.

Serganov, A. and Nudler, E. (2013). A decade of riboswitches. *Cell*, *152*(1-2), 17–24.

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F., and Vogel, J. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, *464*(7286), 250–255.

Simpson, J. T. and Pop, M. (2015). The Theory and Practice of Genome Sequence Assembly. *Annual Review of Genomics and Human Genetics*, *16*, 153–172.

Smirnov, A., Förstner, K. U., Holmqvist, E., Otto, A., Günster, R., Becher, D., Reinhardt, R., and Vogel, J. (2016). Grad-seq guides the discovery of ProQ as a major small RNA-binding protein. *PNAS*, *113*(41), 11591–11596.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, *147*(1), 195–197.

Sohn, J.-I. and Nam, J.-W. (2018). The present and future of *de novo* whole-genome assembly. *Briefings in Bioinformatics*, *19*(1), 23–40.

Sridhar, J., Sambaturu, N., Sabarinathan, R., Ou, H.-Y., Deng, Z., Sekar, K., Rafi, Z. A., and Rajakumar, K. (2010). sRNAscanner: a computational tool for intergenic small RNA detection in bacterial genomes. *PloS One*, *5*(8), e11970.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.*, *6*(7), 2601–2610.

Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G., and Bayley, H. (2009). Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *PNAS*, *106*(19), 7702–7707.

Stone, J. R. and Wray, G. A. (2001). Rapid evolution of cis-regulatory sequences via local point mutations. *Mol. Biol. Evol.*, *18*(9), 1764–1770.

Storz, G., Vogel, J., and Wassarman, K. M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. *Mol. Cell*, *43*(6), 880–891.

Stougaard, P., Molin, S., and Nordström, K. (1981). RNAs involved in copy-number control and incompatibility of plasmid R1. *PNAS*, *78*(10), 6008–6012.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.*, *14*(2), 103–105.

Sun, Y., Aljawad, O., Lei, J., and Liu, A. (2012). Genome-scale NCRNA homology search using a Hamming distance-based filtration strategy. *BMC Bioinformatics*, *13 (Suppl 3)*, S12.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., and Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.*, *44*(14), 6614–6624.

Thomas, M. S. and Wigneshweraraj, S. (2014). Regulation of virulence gene expression. *Virulence*, *5*(8), 832–834.

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

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, *22*(22), 4673–4680.

Tinoco, I., Uhlenbeck, O. C., and Levine, M. D. (1971). Estimation of Secondary Structure in Ribonucleic Acids. *Nature*, *230*(5293), 362–367.

Tinoco, I. J., Borer, P. N., Dengler, B., Levin, M. D., Uhlenbeck, O. C., Crothers, D. M., and Bralla, J. (1973). Improved estimation of secondary structure in ribonucleic acids. *Nat New Biol.*, *246*(150), 40–1.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols*, *7*(3), 562–578.

Tree, J. J., Granneman, S., McAteer, S. P., Tollervey, D., and Gally, D. L. (2014). Identification of bacteriophage-encoded anti-sRNAs in pathogenic Escherichia coli. *Mol. Cell*, *55*(2), 199–213.

Tyrrell, J., McGinnis, J. L., Weeks, K. M., and Pielak, G. J. The cellular environment stabilizes adenine riboswitch RNA structure.

van Opijnen, T., Bodi, K. L., and Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods*, *6*(10), 767–772.

Vannini, A., Roncarati, D., and Danielli, A. (2016). The *cag*-pathogenicity island encoded CncR1 sRNA oppositely modulates *Helicobacter pylori*f motility and adhesion to host cells. *Cell. Mol. Life Sci.*, *73*(16), 3151–3168.

Wagner, E. G. and Simons, R. W. (1994). Antisense RNA control in bacteria, phages, and plasmids. *Annu. Rev. Microbiol.*, *48*, 713–742.

Wahlestedt, C. (2013). Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nat. Rev. Drug Discov.*, *12*(6), 433–446.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, *10*(1), 57–63.

Wassarman, K. M. (2007). 6S RNA: a regulator of transcription. *Mol. Microbiol.*, *65*(6), 1425–1431.

Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G., and Gottesman, S. (2001). Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.*, *15*(13), 1637–1651.

Waterman, M. S. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes*. CRC Press.

Watters, K. E. and Lucks, J. B. (2016). Mapping RNA Structure *In Vitro* with SHAPE Chemistry and Next-Generation Sequencing (SHAPE-Seq). In D. H. Turner and D. H. Mathews (Eds.), *RNA Structure Determination: Methods and Protocols*, 135–162. New York, NY: Springer New York.

Westermann, A. J., Förstner, K. U., Amman, F., Barquist, L., Chao, Y., Schulte, L. N., Müller, L., Reinhardt, R., Stadler, P. F., and Vogel, J. (2016). Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. *Nature*, *529*(7587), 496–501.

Westermann, A. J., Gorski, S. A., and Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.*, *10*(9), 618–630.

Wilderman, P. J., Sowa, N. A., FitzGerald, D. J., FitzGerald, P. C., Gottesman, S., Ochsner, U. A., and Vasil, M. L. (2004). Identification of tandem duplicate regulatory small RNAs in Pseudomonas aeruginosa involved in iron homeostasis. *PNAS*, *101*(26), 9792–9797.

Workman, R. E., Tang, A., Tang, P. S., Jain, M., Tyson, J. R., Zuzarte, P. C., Gilpatrick, T., Razaghi, R., Quick, J., Sadowski, N., Holmes, N., de Jesus, J. G., Jones, K., Snutch, T. P., Loman, N. J., Paten, B., Loose, M. W., Simpson, J. T., Olsen, H. E., Brooks, A. N., Akeson, M., and Timp, W. (2018). Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv*, *Online*(Preprint), 459529.

Wu, D. C., Yao, J., Ho, K. S., Lambowitz, A. M., and Wilke, C. O. (2018). Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics*, *19*(1), 510.

Yona, A. H., Alm, E. J., and Gore, J. (2018). Random sequences rapidly evolve into de novo promoters. *Nat. Commun.*, *9*(1), 1530.

Zhang, A., Wassarman, K. M., Rosenow, C., Tjaden, B. C., Storz, G., and Gottesman, S. (2003). Global analysis of small RNA and mRNA targets of Hfq. *Mol. Microbiol.*, *50*(4), 1111–1124.

Zhao, J., Ohsumi, T. K., Kung, J. T., Ogawa, Y., Grau, D. J., Sarma, K., Song, J. J., Kingston, R. E., Borowsky, M., and Lee, J. T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell*, *40*(6), 939–953.

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, *9*(1), 133–148.

Chapter 2

Horizontal-gene transfer as a source of novel sRNAs in *Salmonella* Typhimurium

2.1 Preface

Despite their important role in gene regulation, studies of the evolutionary dynamics and origins of bacterial sRNA genes have been hindered by their poor sequence conservation, which makes annotation *via* sequence homology challenging. We implemented a profile HMM-based homology search pipeline to explore the conservation of *Salmonella* Typhimurium sRNA sequences, which supports previous observations of high sequence turnover and rapid gain and loss of genes across lineages. Synteny and genomic context were used to infer the evolutionary origins of these sRNAs, which shows that many *Salmonella*-specific sRNAs have been recently acquired through horizontal gene transfer. Examples of potential exaptation from mobile genetic element and toxin-antitoxin system sRNAs are described, and potential pitfalls when performing sRNA homology search are highlighted.

2.1.1 Contributions

Paul Gardner and Lars Barquist developed the idea for the project and provided supervision and feedback. I implemented all data analysis pipelines and developed methods to predict gene origins.

2.2 Introduction

Bacterial sRNAs are a heterogeneous class of ncRNAs that can regulate gene expression *via* interactions with mRNAs and RNA-binding proteins, granting bacteria extensive transcriptional flexibility. These short (< ~200 nt), often highly structured molecules can base-pair with mRNA targets in order to reduce, stabilise, or enhance transcriptional output, or to sequester, inhibit, or guide protein targets through interactions with RNA-binding sites (Barquist and Vogel, 2015, Storz *et al.*, 2011).

The majority of characterised sRNAs act through antisense base-pairing, similar to eukaryotic miRNAs. These are further classified into *cis*-encoded antisense sRNAs, transcribed opposite their target gene, and *trans*-encoded sRNAs, transcribed distal to their targets. *Trans*-encoded sRNAs generally have short regions of complementarity to their targets, as small as 5–8 nt, that seed RNA:RNA interactions, which are often stabilised by association with the RNA chaperone Hfq (Santiago-Frangos and Woodson, 2018). These sRNAs tend to be expressed under specific growth conditions, allowing bacteria to adapt and react to different environmental stimuli (Gottesman and Storz, 2011). A small number of sRNAs also regulate RNA-binding proteins, such as the global regulator CsrA, by competitive inhibition (Babitzke and Romeo, 2007, Valverde *et al.*, 2004, Weilbacher *et al.*, 2003).

While the vast majority of sRNAs have unknown functions, those which have been functionally characterised act as regulators in a wide range of biological processes. Examples of recently characterised sRNA functions include responses to physiological (Chao and Vogel, 2016, Guo *et al.*, 2014) and nutritional stress (Yin *et al.*, 2018), quorum sensing (Zhao *et al.*, 2018) and biofilm formation (Papenfort *et al.*, 2015, 2017).

Difficulties in identifying sRNAs by experimental and computational approaches has limited comprehensive studies to model organisms such as *E. coli* and *Salmonella enterica* serovar Typhimurium. These studies indicate that bacterial genomes may harbour several hundred sRNAs, many of which appear to be unique to a genus or species (Barquist and Vogel, 2015). Previous studies by Kröger *et al.* (2013) identified 281 sRNAs in *S.* Typhimurium from RNA-seq data generated in 27 infection-relevant growth conditions, and estimated that as many as 60% of these may be genus-specific (Kröger *et al.*, 2013, 2012).

Studies of sRNA evolution remain limited: aside from a handful of well-conserved sRNAs involved in essential biological processes, most sRNAs studied to date appear to have a limited phylogenetic distribution, and undergo rapid sequence turnover (Gómez-Lozano *et al.*, 2012, Gottesman and Storz, 2011, Skippington and Ragan, 2012). This poor conservation frustrates annotation by homology search and leaves many questions about their evolutionary origins unanswered (Dutcher and Raghavan, 2018, Jose *et al.*, 2019, Lindgreen *et al.*, 2014, Updegrove *et al.*, 2015).

The discovery of poorly conserved sRNAs with niche-specific functions, such as the *Salmonella* PinT sRNA that acts to regulate genes during macrophage infection (Westermann *et al.*, 2016), suggests that sRNAs may be gained and rapidly acquire a beneficial function. However, it is difficult to establish whether many sRNAs are 'new' genes unique to a lineage, or vertically inherited and rapidly diversifying.

Profile hidden Markov model-based tools such as HMMER (Eddy, 2009), which build probabilistic models of sequence variation from a trusted multiple sequence alignment, are becoming increasingly popular approaches to study genes that are deeply conserved but have diverged significantly at the sequence level (Eddy, 2011). Other tools such as Infernal (Nawrocki and Eddy, 2013) can incorporate annotations of structural conservation into covariance models, but are only applicable to ncRNAs with conserved secondary structure. When alignments of trusted orthologues are not available, an iterative approach to building HMMs has been effective for studying highly divergent ncRNA (Barquist *et al.*, 2016) and protein (Johnson *et al.*, 2010) sequences.

We studied the sequence conservation of 200 intergenic *Salmonella enterica* serovar Typhimurium sRNAs by using an iterative profile HMM-based homology search pipeline, and used functional annotations of nearby proteins to infer their evolutionary origins. Our results show that many sRNAs are new to the *Salmonella* lineage, and exhibit rapid sequence turnover. We found that horizontal gene transfer is the main driver of sRNA acquisition in *Salmonella*, and identified *Salmonella*-specific sRNAs that appear to be derived from phage control systems, and other mobile genetic elements, as well as type I toxin-antitoxin systems.

2.3 Methods

2.3.1 Starting data-set

A set of 200 experimentally verified intergenic sRNA genes from *Salmonella enterica* Typhimurium str. ST4/74 (NCBI accession: CP002487) identified by Kröger *et al.* (2013) (http://bioinf.gen.tcd.ie/cgi-bin/salcom.pl) were used as starting sequences for homology search. Locus ID names from this dataset were used as gene names in this study (gene name synonyms can be found in Table B.2). A set of 3284 complete bacterial genomes from EMBL, including a subset of 128 *Salmonella* genomes, were used as reference databases for the homology search.

2.3.2 Initial models and quality control

For each sRNA, `hmmbuild` (Wheeler and Eddy, 2013) was used to build three separate single-sequence HMMs, consisting of the sRNA sequence and two 150 nt adjacent flanking sequences. These HMMs were used to identify homologous sequences in *Salmonella* genomes from the bacterial genome database with nhmmer (E-value threshold = 0.001). A custom Perl script was used to perform two filtering functions. First, flanking sequences were used as synteny anchors to improve annotation accuracy for sRNAs with palindromic sequences, and to separate duplicate genes and recent paralogues. Annotation overlaps and multi-copy results were then resolved by selecting annotations with the highest combined bit-score of the sRNA and correctly orientated synteny anchors within 150 nt (Figure 2.1).

Homologous *Salmonella* sequences from both the sRNA and flanking sequences were then aligned using the ncRNA sequence alignment tool `mafft-qinsi` (Katoh and Standley, 2013), and new HMMs were built from each alignment. These HMMs were then used in a second `nhmmer` (Wheeler and Eddy, 2013) homology search over the entire bacterial genome database (E-value threshold = 0.01). Sequences from this step were filtered and used to generate final multiple sequence alignments for each sRNA.

Final alignments were analysed to look for coding potential (`RNAcode`, Washietl *et al.* (2011)), presence of covariation to maintain secondary structure (`RNAz`, Gruber *et al.* (2010) and `R-scape`, Rivas *et al.* (2017)). Consensus secondary structure predictions were generated using `RNAalifold` and visualised using `RALEE` (Griffiths-Jones, 2005).

2.3.3 Estimating conservation

Two measurements of conservation were performed for each sRNA to generate a heatmap (Figures 2.8 and 2.7). Pairwise sequence identity was calculated by aligning each sequence annotated from the homology search against the single-sequence model for that sRNA with `hmmalign`. Overall conservation across the Enterobacteriaceae was also measured by counting the presence or absence of an annotation in each genome. For each genus, the proportion of the genomes annotated was also used as an estimate of conservation.

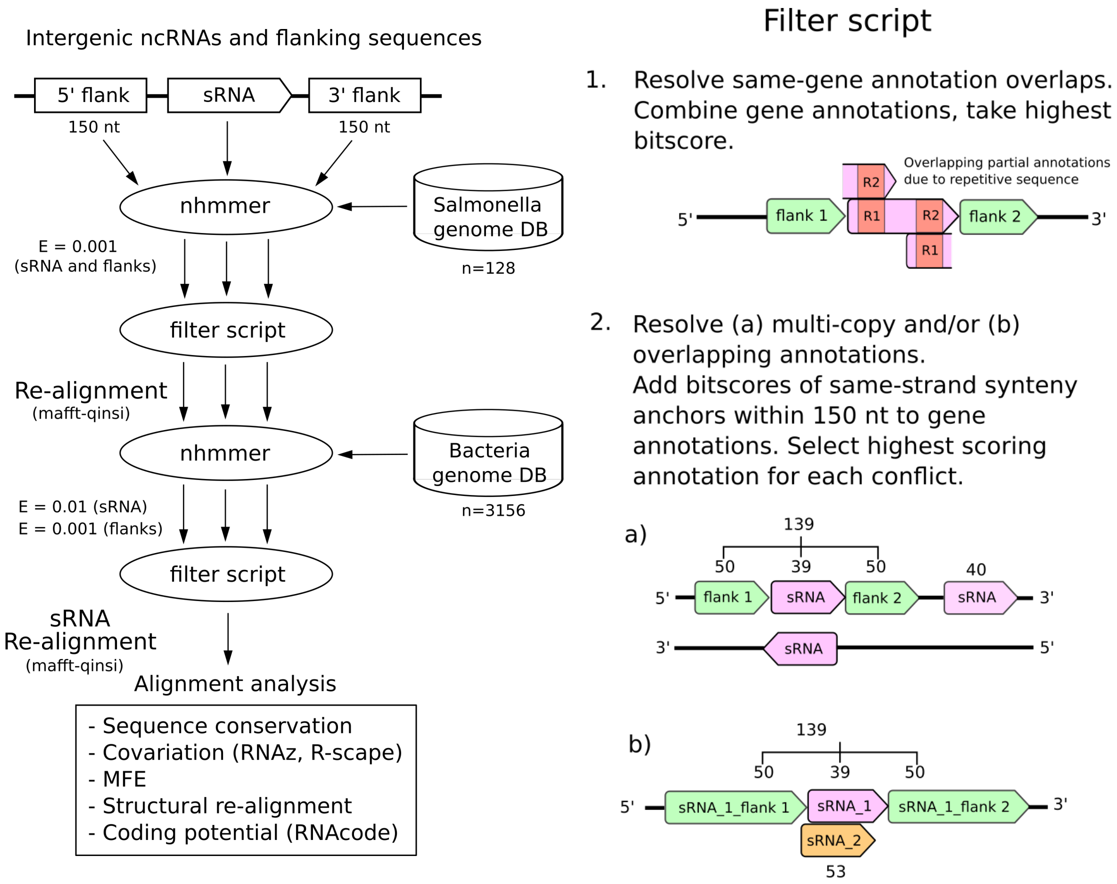

Intergenic ncRNAs and flanking sequences
5' flank
sRNA
3' flank
150 nt
150 nt
nhmmer
Salmonella genome DB
n=128
E = 0.001
(sRNA and flanks)
filter script
Re-alignment
(mafft-qinsi)
nhmmer
Bacteria genome DB
n=3156
E = 0.01 (sRNA)
E = 0.001 (flanks)
filter script
sRNA
Re-alignment
(mafft-qinsi)
Alignment analysis
- Sequence conservation
- Covariation (RNAz, R-scape)
- MFE
- Structural re-alignment
- Coding potential (RNAcode)
Filter script
1. Resolve same-gene annotation overlaps. Combine gene annotations, take highest bitscore.
Overlapping partial annotations due to repetitive sequence
5'
flank 1
R2
R1
R2
R1
flank 2
3'
2. Resolve (a) multi-copy and/or (b) overlapping annotations. Add bitscores of same-strand synteny anchors within 150 nt to gene annotations. Select highest scoring annotation for each conflict.
a)
139
50
39
50
40
5'
flank 1
sRNA
flank 2
sRNA
3'
3'
sRNA
5'
b)
139
50
39
50
5'
sRNA_1_flank 1
sRNA_1
sRNA_1_flank 2
3'
sRNA_2
53

Figure 2.1: Left: Flow-chart describing the homology search pipeline used. Single-sequence HMMs are used to search for homologous sequences within a genus restricted database (128 *Salmonella* genomes) using nhmmer (Wheeler and Eddy, 2013). A stringent E-value threshold of 0.001 is used for this step to improve annotation accuracy and generate a trusted multiple sequence alignment. A filtering script is then used to resolve multi-copy and overlapping annotations. Multiple sequence alignments for both the sRNA and flanking sequences are generated using mafft-qinsi (Katoh and Standley, 2013). New HMMs were built from each alignment and used in a second nhmmer homology search over the entire bacterial genome database. For this search, the E-value threshold is increased to 0.01 for sRNA genes to capture more divergent sequences. Results are filtered and used to generate final alignments. Alignments are ranked by conservation, structural stability (RNAalifold minimum free energy prediction), and covarying base pairs. **Right:** Filtering strategy for resolving multi-copy and overlapping ncRNA annotations. **(1)** Partial overlapping annotations of the same gene are combined. An example shows an sRNA with partial annotations around repeat regions. **(2)** For **(a)** multi-copy and **(b)** overlapping annotations, the bit-scores of each annotation and any correctly oriented synteny anchors within 150 nt are summed. The annotation with the highest bit-score is then chosen.

2.3.4 Predicting sRNA evolutionary origins

To predict the evolutionary origin of the genes, a database of 317,498 proteins was created from three upstream and three downstream proteins for each sRNA annotation, based on EMBL gene annotations. These proteins were functionally annotated with `eggnog-mapper` (Huerta-Cepas *et al.*, 2017) using the eggNOG bactNOG HMMs (Powell *et al.*, 2012). The resulting 1,444 eggNOG functional descriptions were used to estimate if proteins were part of a mobile genetic element (MGE), such as transposons (e.g transposases and integrases) or phages (e.g integrases, capsid and tail proteins). Protein descriptions from the eggnog annotations were manually binned into 'MGE' (n=55, list provided in Table B.3) and 'non-MGE' (n=1089).

Individual sRNAs were manually classified as "horizontally" or "vertically" inherited based on several lines of evidence. First, sRNAs were classed as candidate horizontally-acquired genes based on their proximity to annotated pathogenicity islands and prophages (annotations by Kröger *et al.* (2013, 2012)). EggNOG protein descriptions were also used to indicate whether an sRNA was near, or within, a cryptic or undiscovered mobile genetic element. If any protein annotations flanking an sRNA were consistently annotated as an MGE, this was taken as a signal of horizontal gene transfer.

A sub-set of vertically-inherited sRNAs were classed as "divergence". Vertically-inherited sRNAs with inconsistent annotation across a lineage, for example otherwise conserved genes missing in a specific genus, were investigated to identify if the sRNA was missing or not annotated due to sequence divergence. The presence of closely-spaced sRNA flanking sequences was used to identify if an intergenic region known to contain an sRNA was present but not annotated due to sequence divergence.

Conservation was then used to resolve ambiguities. For example, if an sRNA was inconsistently annotated and conserved at the sequence level across the Enterobacteriaceae, this was taken as a signal of horizontal gene transfer (Figure 2.2A). Unfiltered homology search results were also used to identify sRNAs with highly multi-copy annotations or partial matches to MGEs (Figure 2.2B). The combination of these factors was used to predict whether sRNAs were part of mobile genetic elements, or merely frequent insertion sites for MGEs.

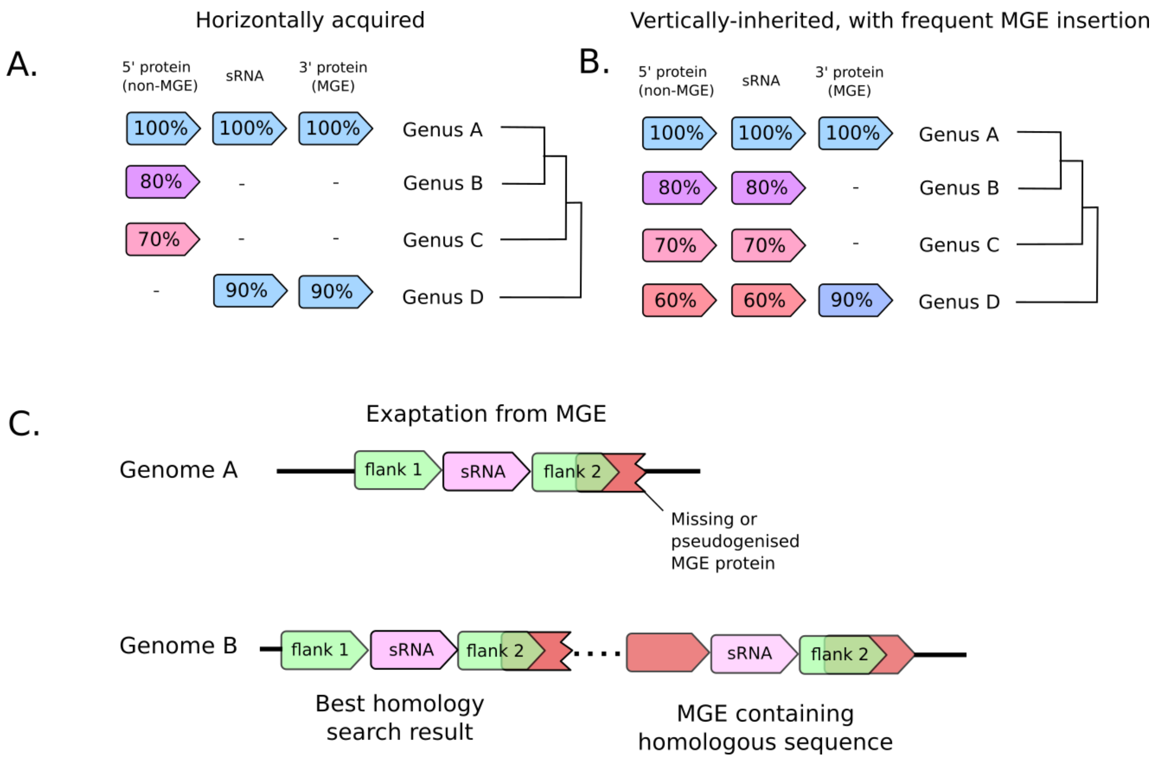


Figure 2.2: Diagram illustrating how homology search results were used to identify sRNAs associated with MGEs. Two examples (A and B) illustrate how context can be used to identify if an sRNA associated with MGE proteins is horizontally acquired or vertically inherited. In this example, Genus A and Genus D are in the same family, but have diverged significantly. **(A)** A horizontally-acquired sRNA is present only in two distantly-related genera, and associated with an MGE protein. The sRNA also shows high sequence conservation over a large phylogenetic distance. **(B)** A vertically-inherited sRNA is consistently annotated throughout the phylogeny, and shows high sequence turnover. An MGE protein is often found near the sRNA, but is due to independent insertions within Genus A and Genus D. **(C)** An illustration showing how multi-copy homology search results can be used to identify sRNAs with homology to an sRNA contained within an MGE. An example is shown of an sRNA deposited by a previous MGE in the ancestor of Genome A and Genome B, that has been maintained after the rest of the MGE has been deleted or pseudogenised. Genome B contains both the exapted MGE sRNA locus, which is annotated by homology search, as well as an intact MGE containing a homologous sequence. Homology with this MGE can be used to infer a horizontally-acquired origin for this sRNA.

2.4 Results and Discussion

2.4.1 Filtering is required to reduce false positives

Prior to filtering many sRNAs were annotated as multi-copy (Figure 2.3), due to shared sequence features with each other, and with existing genomic elements. Some HMMs built from short sRNA sequences generated highly multi-copy homology search results, producing hundreds of annotations per genome. An extreme example of this was *c0664*, which was annotated at the end of almost all coding sequences in each *Salmonella* genome prior to filtering.

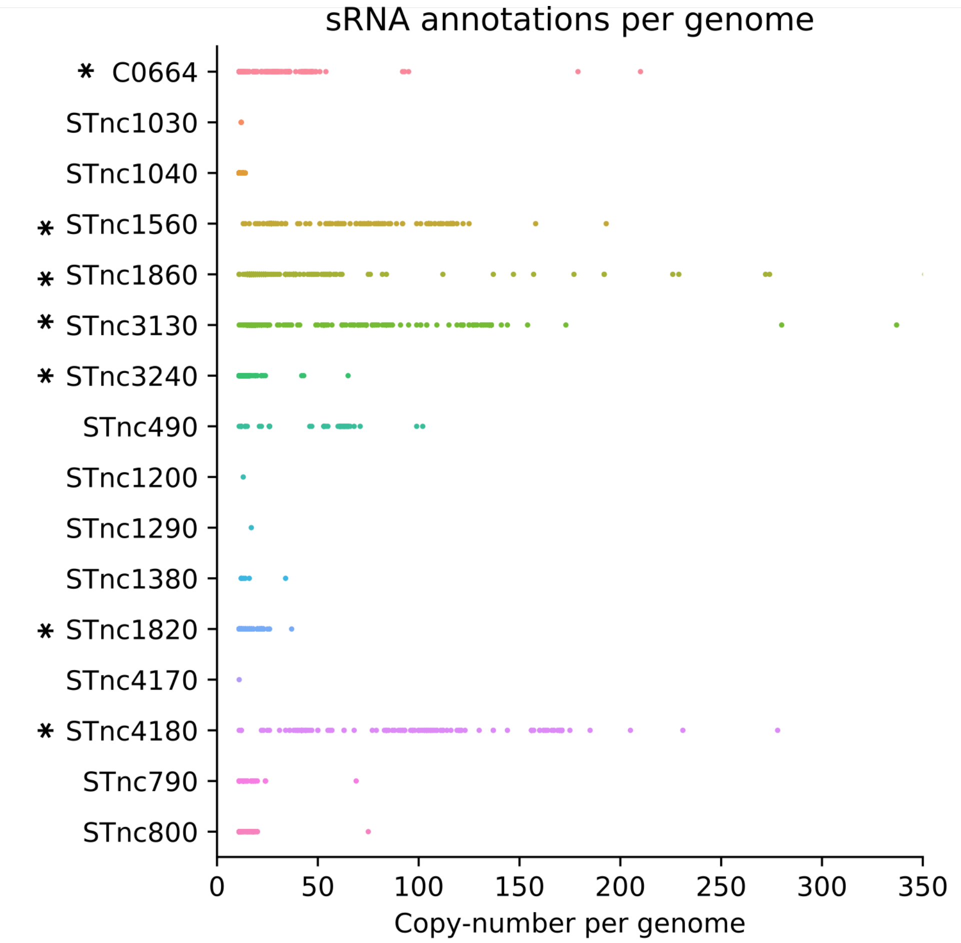


Figure 2.3: Copy number per genome for highly multi-copy sRNAs. Number of homology search results per genome prior to filtering are plotted for genes with >10 copies per genome. * denotes genes that were found to overlap with annotated repetitive extragenic palindromes (REPs) in *E. coli* str. K12 substr. MG1655 (genome accession U00096.3) (shown in Figure 2.4).

Many of the most highly multi-copy sRNAs (*c0664*, *STnc1560*, *STnc1820*, *STnc1860*, *STnc3130*, *STnc3240* and *STnc4180*) had annotations that overlapped with repetitive extragenic palindromic (REP) elements in *E. coli* str. K12 substr. MG1655 (genome accession U00096.3). REP elements are 30-40 nt short intergenic repeats found downstream of stop codons in gammaproteobacteria, which make up ~0.5-1% of the *E. coli* genome (Dimri *et al.*, 1992, Stern *et al.*, 1984).

REPs can help form and stabilise hairpin loops within transcripts, and can function to regulate mRNA transcription (Espéli *et al.*, 2001), polycistronic transcript stability and decay (Khemici and Carpousis, 2004), and translation (Liang *et al.*, 2015). *c0664* is known to contain an REP element (Hershberg *et al.*, 2003), and other sRNAs which overlapped with REPs had conserved REP-like imperfect palindromic sequences in predicted stem-loop secondary structures (Figure 2.3).

Short sRNAs with sequence similarity to REP elements are likely to produce false positive annotations where the REP element forms a substantial proportion of the sRNA gene. For example, *c0664* has previously been reported at over 40 copies per genome in a BLAST-based analysis (Skippington and Ragan, 2012). Further work is needed to confirm if these sRNAs are REP elements, and if highly multi-copy sRNAs that are annotated at the $3'$ end of protein-coding sequences may contain novel REP motifs.

Other structured sRNAs containing palindromic sequences produced multiple partial intra-gene annotations on each strand, and models for recent paralogues such as *glmZ* and *glmY* converged on each other prior to the implementation of synteny anchor-based filtering. Filtered annotations generated in this project closely matched those from the original *Salmonella* Typhimurium ST4/74 sRNAs, and closely overlapped (0-64 nt difference in gene boundaries) with known homologues annotated in *E. coli* K12, and a set of sRNA annotations collected from literature by Richter and Backofen (2012) (Figure 2.5), indicating that filtering with synteny anchors removed the majority of false positives for intergenic sRNAs.

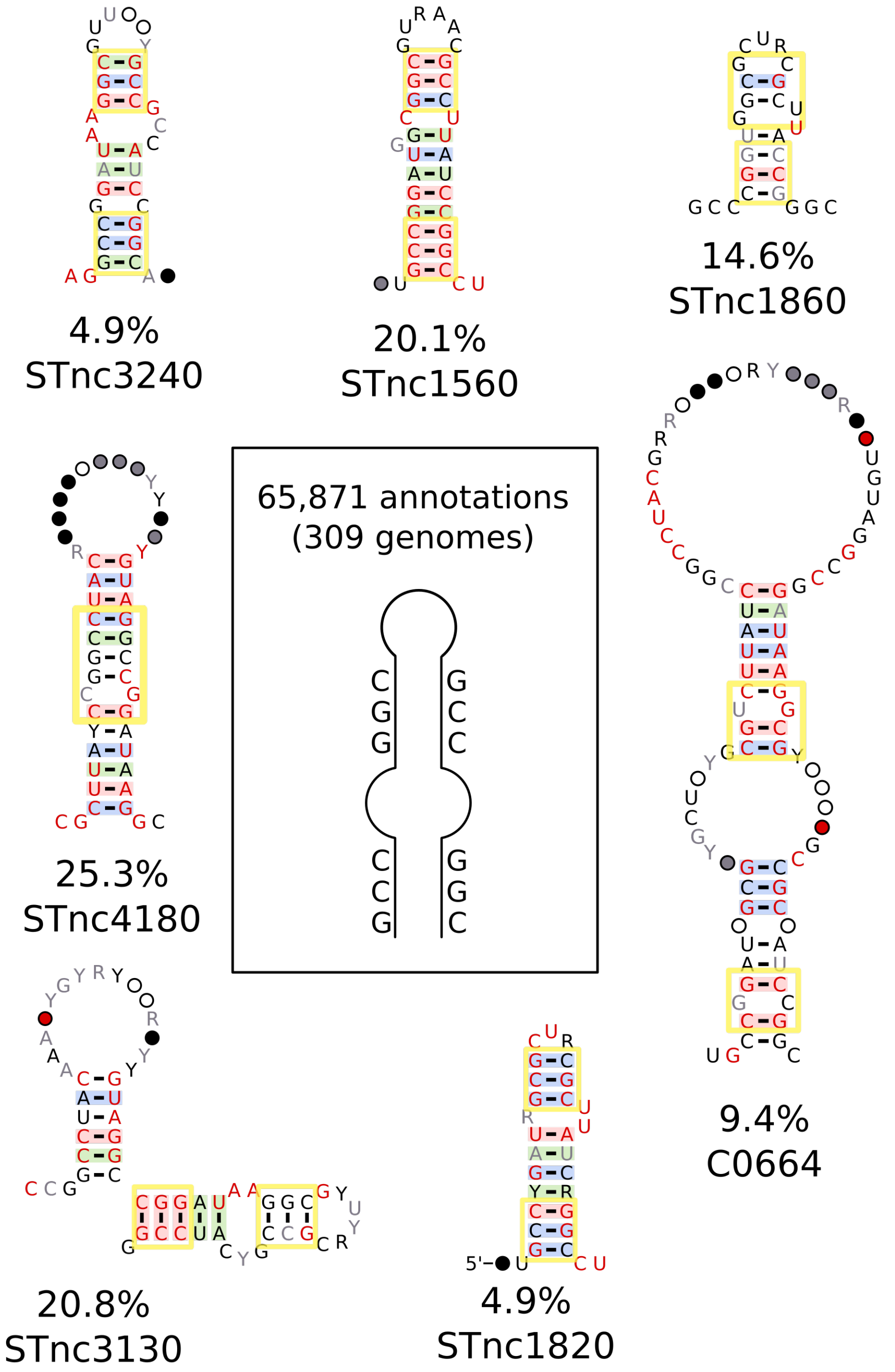


Figure 2.4: Predicted consensus secondary structures for sRNAs with sequence similarity to annotated REPs in *E. coli* str. K12 substr. MG1655 (based on annotation overlap with REP annotations for U00096.3 on NCBI). A total of 65,871 annotations of these seven genes were generated during homology search, the proportions of which are shown for each sRNA. A representative *E. coli* REP motif (Tobes and Pareja, 2006) within a stem-loop structure is shown. REP-like motifs are highlighted on the structures in yellow.

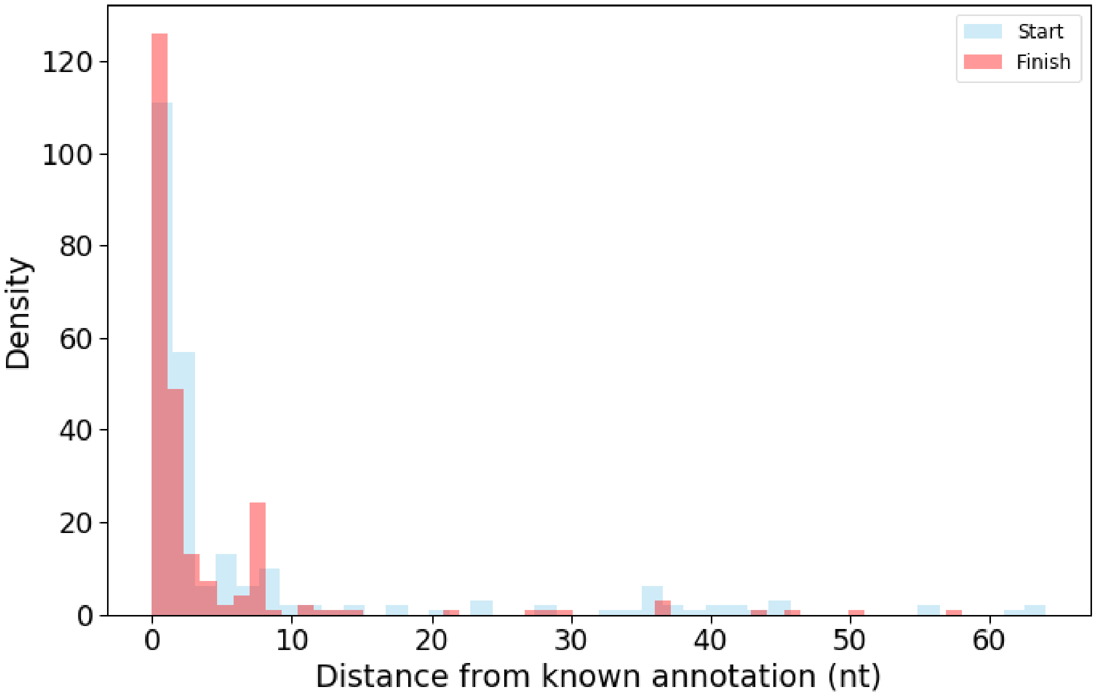


Figure 2.5: Comparison of sRNA annotations by Richter and Backofen (2012) and from this study. Gene boundary differences (distance from start and finish) in genomes and genes present in both data-sets (n=242) are plotted as a histogram, showing that these annotations closely overlap. Some annotations of longer sRNAs from this study are truncated, but represent minimum 70% coverage of the gene. The annotations had no differences in orientation.

2.4.2 Recent acquisition and sequence divergence of *Salmonella* sRNAs

The sequence conservation of homology search results across the Enterobacteriaceae, visualised as a heatmap in Figures 2.8B and 2.7B, broadly mirrors previous observations that the majority of sRNAs both evolve rapidly and are restricted to a genus or species (Gómez-Lozano *et al.*, 2015, Gottesman and Storz, 2011, Skippington and Ragan, 2012). The conservation of sequences flanking sRNAs, most of which overlapped with coding sequences, were used to estimate if sRNAs were limited to a lineage and not approaching the lower bound of sequence alignment due to sequence variation. The majority of vertically-inherited sRNAs had similar phylogenetic distributions to at least one flanking sequences, which were limited to a sub-lineage of the Enterobacteriaceae.

The rate of sequence turnover, defined here as observable change in nucleotide sequence over time, appears to be proportional to the age of sRNA genes investigated in this study. High rates of sequence divergence (a change from 100% to ~50% sequence identity) could be seen between sRNAs limited to the *Salmonella-Escherichia* lineage, as between larger phylogenetic distances such as between *Salmonella* and *Yersinia* sequences (Figure 2.6).

Only 9 intergenic *Salmonella* sRNAs were found to be conserved across the Enterobacteriaceae, all of which except *sdsR*, *spf*, *sroG* and *STnc700* were family-specific. Several other sRNAs were highly conserved in all genera except in the obligate plant pathogens *Edwardsiella*, *Xehnorhabdus* and *Pectobacterium*. With the exception of *spf*, all highly conserved sRNA genes exhibited high sequence divergence, with homologous sequences in the genera most distantly related to *Salmonella* having only ~40% sequence similarity to their *Salmonella* counterparts. *spf* was found to be highly conserved throughout the Enterobacteriaceae, and had the least amount of sequence variation, with the lowest average percentage sequence identity of 93% in *Dickeya*.

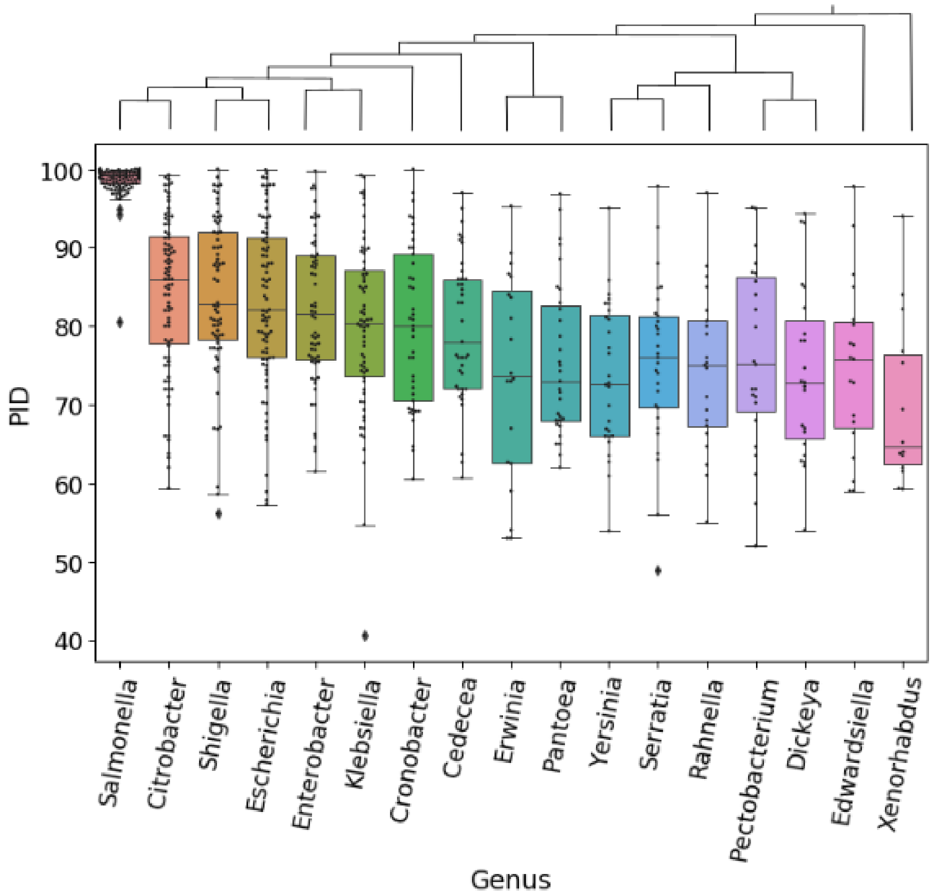


Figure 2.6: Boxplots showing sRNA sequence variation within different genera. PID (percentage sequence identity) relative to the *S.* Typhimurium ST4/74 sequence measurements for vertically-inherited sRNA annotations from the homology search are shown, separated by genus. Genera are ordered by phylogenetic distance from *Salmonella*, and approximate phylogenetic relationships are shown above the plot.

In total 74/200 sRNAs were only found within *Salmonella* (shown in Figure 2.8), and overall the majority of sRNAs were restricted to the *Salmonella-Escherichia* lineage, and showed rapid within-lineage divergence. Although many sRNAs in this data-set have yet to be functionally characterised, there does appear to be a correlation between the specificity of function and conservation of an sRNA gene.

Highly conserved sRNAs are known to be involved in essential cellular processes, such as the 6S RNA, which is involved in the regulation of transcription (Wassarman, 2007), or well integrated into large regulons, such as the CsrB sRNA, which sequesters the highly conserved carbon storage regulator CsrA (Babitzke and Romeo, 2007). Conversely, sequence and gene conservation was reduced in *Erwinia, Pectobacterium, Pantoea, Dickeya*, and *Xenorhabdus*, which are obligate plant pathogens that have substantial genomic and phenotypic differences other genera in the family, which are animal gut pathogens and commensals (Toth *et al.*, 2006).

The variation of function and sequence diversity decreases for sRNA genes are more widely conserved throughout the Enterobacteriaceae. Genus or lineage-specific sRNAs often contribute to niche adaption, and show more sequence variation across small phylogenetic distances, indicating that selection pressures and essentiality of the sRNA regulon restricts sequence diversity.

S. Typhimurium sRNAs conserved in the *Salmonella-Escherichia* lineage are often involved in niche-specific stress responses that promote survival *in vivo*. Two characterised examples are RydC, which suppresses biofilm formation and cell adhesion during nutrient availability (Bordeau and Felden, 2014), and MgrR, which is expressed in response to low Mg^{2+} concentrations conditions found *in vivo*, and regulates the formation of external lipopolysaccharides as part of a strategy to avoid recognition by the host immune system (Moon and Gottesman, 2009, Moon *et al.*, 2013). Both of these sRNAs show sequence divergence (~80-85% sequence identity in *E. coli* relative to *Salmonella*) and have been observed to have additional specific functions unique to *Salmonella* or *Escherichia*. Similarly, many *Salmonella enterica*-specific sRNAs have been found to function in survival and virulence in highly specific infection conditions (Barquist *et al.*, 2013, Colgan *et al.*, 2016). Two *Salmonella*-specific sRNAs, *STnc2050* and *STnc3750*, also have varied expression across different strains of *S.* Typhimurium (Canals *et al.*, 2019).

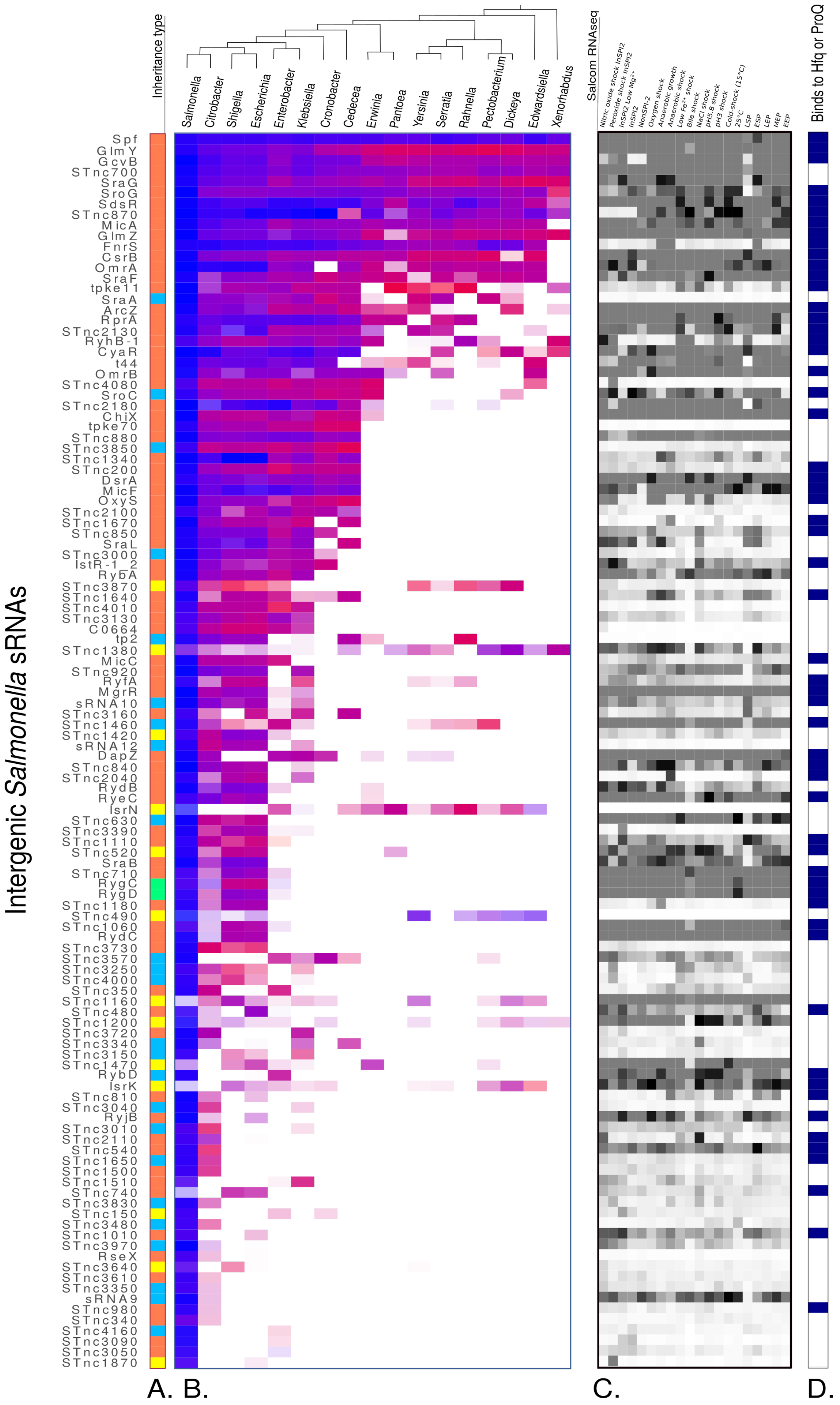

Intergenic *Salmonella* sRNAs
Inheritance type
Salmonella
Citrobacter
Shigella
Escherichia
Enterobacter
Klebsiella
Cronobacter
Cedecea
Erwinia
Pantoea
Yersinia
Serratia
Rahnella
Pectobacterium
Dickeya
Edwardsiella
Xenorhabdus
Salcom RNAseq
Nitric oxide shock InSPI2
Peroxide shock InSPI2
InSPI2 Low Mg2+
InSPI2
NonSPI-2
Oxygen shock
Anaerobic growth
Anaerobic shock
Low Fe2+ shock
Bile shock
NaCl shock
pH5.8 shock
pH3 shock
Cold-shock (15°C)
25°C
LSP
ESP
LEP
MEP
EEP
Binds to Hfq or ProQ
Spf
GlmY
GcvB
STnc700
SraG
SroG
SdsR
STnc870
MicA
GlmZ
FnrS
CsrB
OmrA
SraF
tpke11
SraA
ArcZ
RprA
STnc2130
RyhB-1
CyaR
t44
OmrB
STnc4080
SroC
STnc2180
ChiX
tpke70
STnc880
STnc3850
STnc1340
STnc200
DsrA
MicF
OxyS
STnc2100
STnc1670
STnc850
SraL
STnc3000
IstR-1_2
RybA
STnc3870
STnc1640
STnc4010
STnc3130
C0664
tp2
STnc1380
MicC
STnc920
RyfA
MgrR
sRNA10
STnc3160
STnc1460
STnc1420
sRNA12
DapZ
STnc840
STnc2040
RydB
RyeC
IsrN
STnc630
STnc3390
STnc1110
STnc520
SraB
STnc710
RygC
RygD
STnc1180
STnc490
STnc1060
RydC
STnc3730
STnc3570
STnc3250
STnc4000
STnc350
STnc1160
STnc480
STnc1200
STnc3720
STnc3340
STnc3150
STnc1470
RybD
IsrK
STnc810
STnc3040
RyjB
STnc3010
STnc2110
STnc540
STnc1650
STnc1500
STnc1510
STnc740
STnc3830
STnc150
STnc3480
STnc1010
STnc3970
RseX
STnc3640
STnc3610
STnc3350
sRNA9
STnc980
STnc340
STnc4160
STnc3090
STnc3050
STnc1870
A.
B.
C.
D.

Figure 2.7

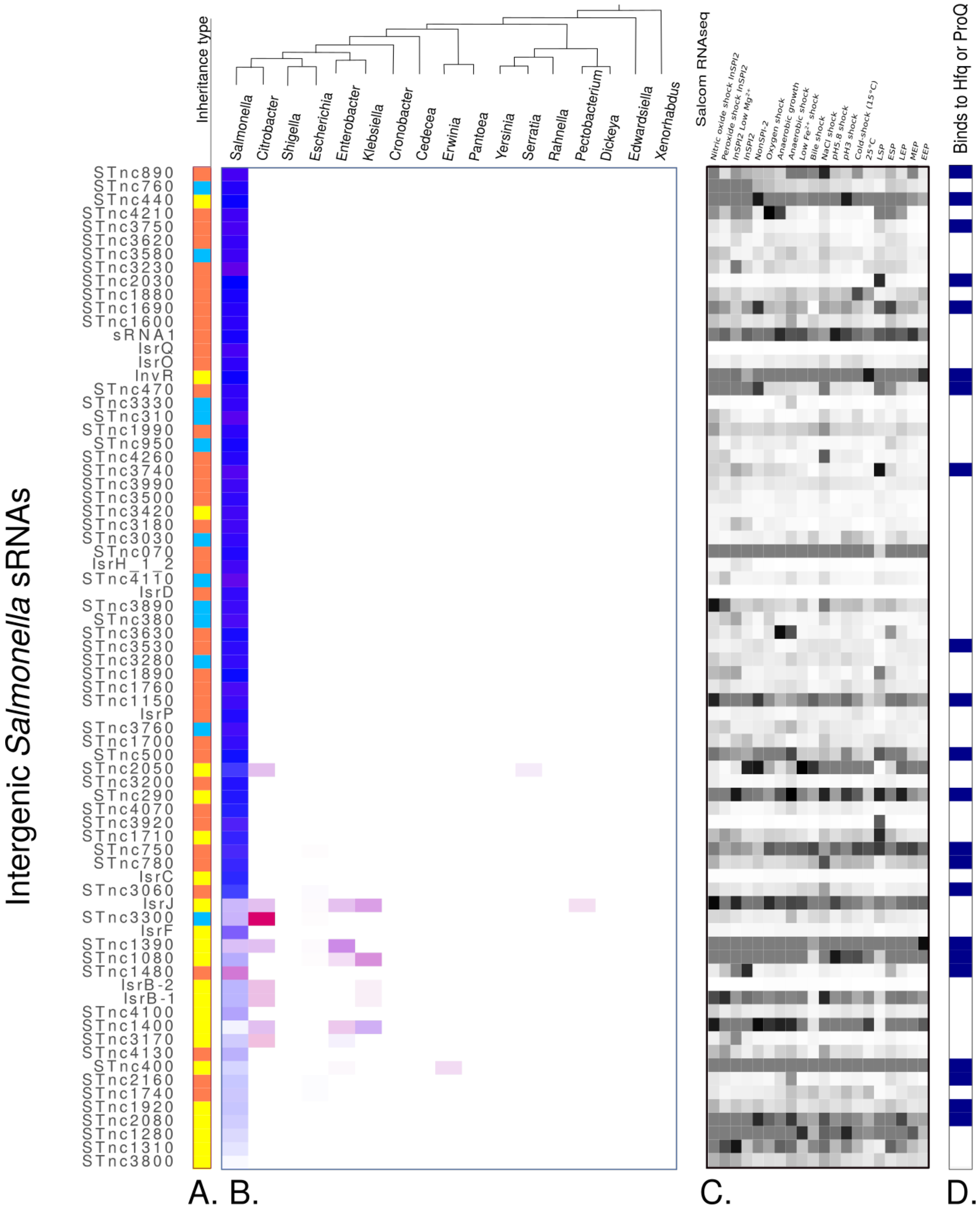

Intergenic *Salmonella* sRNAs
Inheritance type
Salmonella
Citrobacter
Shigella
Escherichia
Enterobacter
Klebsiella
Cronobacter
Cedecea
Erwinia
Pantoea
Yersinia
Serratia
Rahnella
Pectobacterium
Dickeya
Edwardsiella
Xenorhabdus
Salcom RNAseq
Nitric oxide shock InSPI2
Peroxide shock InSPI2
InSPI2 Low Mg²⁺
InSPI2
NonSPI-2
Oxygen shock
Anaerobic growth
Anaerobic shock
Low Fe²⁺ shock
Bile shock
NaCl shock
pH5.8 shock
pH3 shock
Cold-shock (15°C)
25°C
LSP
ESP
LEP
MEP
EEP
Binds to Hfq or ProQ
STnc890
STnc760
STnc440
STnc4210
STnc3750
STnc3620
STnc3580
STnc3230
STnc2030
STnc1880
STnc1690
STnc1600
sRNA1
IsrQ
IsrO
InvR
STnc470
STnc3330
STnc310
STnc1990
STnc950
STnc4260
STnc3740
STnc3990
STnc3500
STnc3420
STnc3180
STnc3030
STnc070
IsrH_1_2
STnc4110
IsrD
STnc3890
STnc380
STnc3630
STnc3530
STnc3280
STnc1890
STnc1760
STnc1150
IsrP
STnc3760
STnc1700
STnc500
STnc2050
STnc3200
STnc290
STnc4070
STnc3920
STnc1710
STnc750
STnc780
IsrC
STnc3060
IsrJ
STnc3300
IsrF
STnc1390
STnc1080
STnc1480
IsrB-2
IsrB-1
STnc4100
STnc1400
STnc3170
STnc4130
STnc400
STnc2160
STnc1740
STnc1920
STnc2080
STnc1280
STnc1310
STnc3800
A.
B.
C.
D.

Figure 2.8

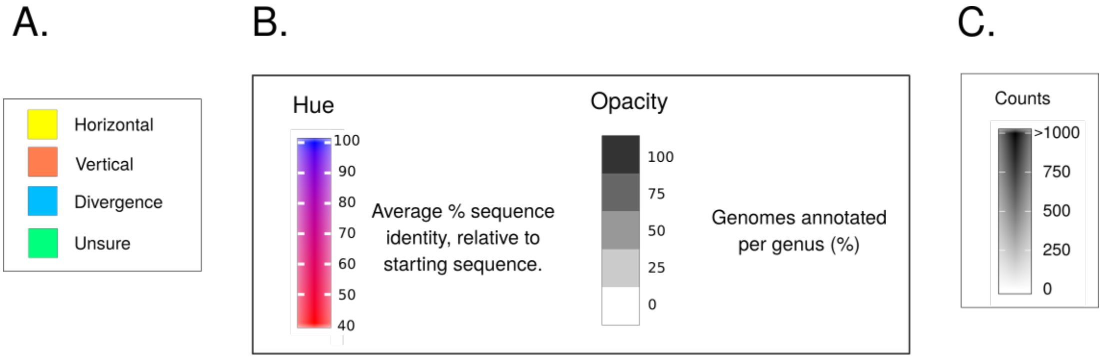


Figure 2.9: Conservation, expression and predicted evolutionary origins of *Salmonella* sRNA genes. **(A)** Predicted inheritance type. **Horizontal**: Acquired through horizontal gene transfer. **Divergence**: The locus containing an sRNA is conserved, but sRNA sequence similarity is too low for it to be annotated *via* homology search. **Vertical**: No signals of horizontal gene transfer. **(B)** Heatmap showing the annotation range and sequence conservation of intergenic *Salmonella* sRNA genes in Enterobacteriaceae genomes. Sequence conservation is shown as a colour gradient from blue (100%) to red (40%), representing genus average percent sequence identity based on alignment of homology search results to the corresponding *Salmonella* Typhimurium ST4/74 sequence. Gene conservation is shown as a change in opacity, represented by the percentage of genomes with an annotation within that genus. Genes are ordered based on overall conservation. Additional panels show information for the same genes from different data-sets. **(C)** RNA-seq counts (Transcripts per million) across 22 infection-relevant growth conditions (capped at 1000 TPM) from Kröger *et al.* (2013). **(D)** sRNA binds to Hfq (Holmqvist *et al.*, 2016) or ProQ (Smirnov *et al.*, 2016).

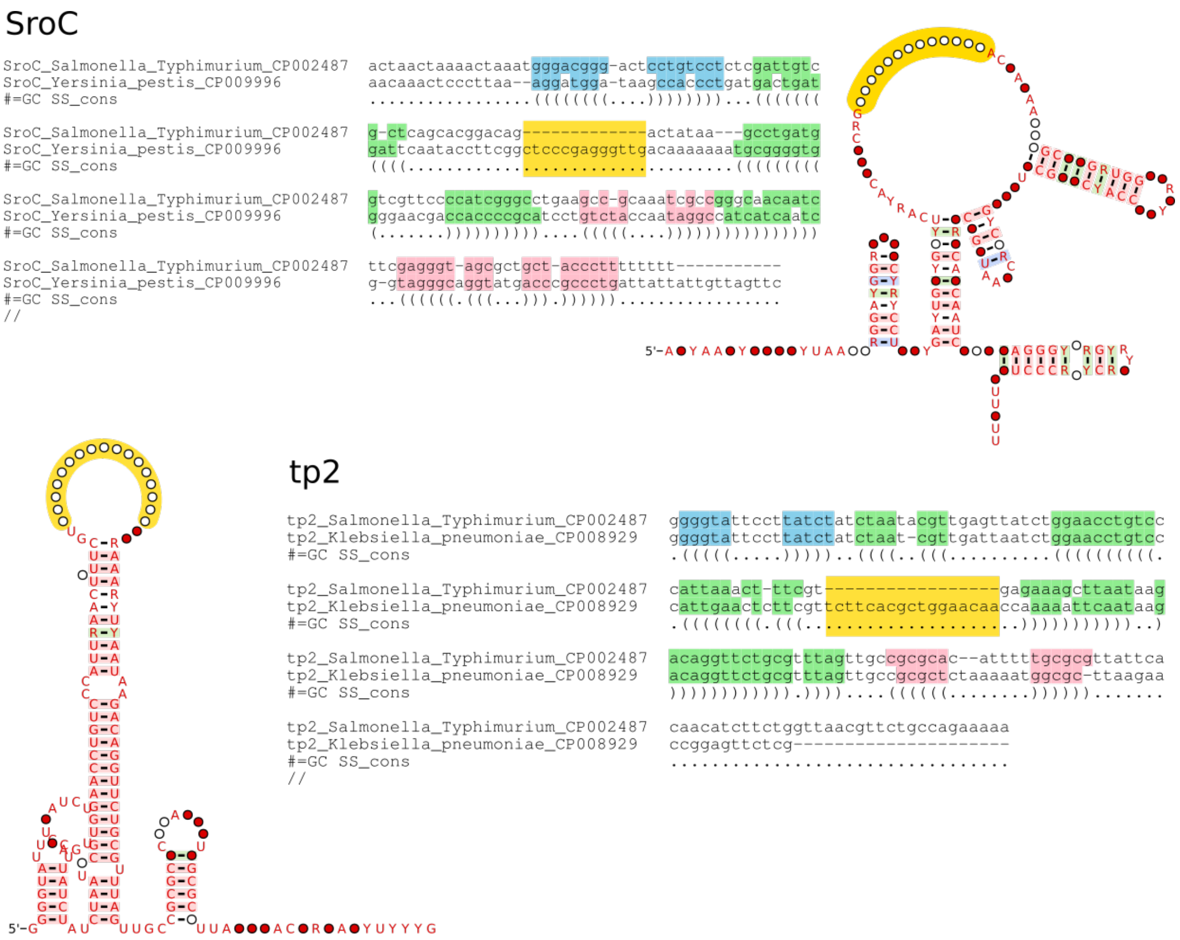
2.4.3 Sequence composition and turnover can hamper annotation

Thirty-six sRNAs were undetectable by homology search due to sequence divergence, large insertions or deletions. An example of this is the *tp2* gene located within a pyruvate metabolic locus, which is highly conserved across the Enterobacteriaceae. The *tp2* synteny anchors, consisting of the $3'$ and $5'$ ends of the proteins adjacent to *tp2* (a pyruvate dehydrogenase sub-unit, and a pyruvate dehydrogenase regulator) were found in same orientation and approximate distance apart throughout the family. Alignment of the intergenic region between the *tp2* synteny anchors in *Klebsiella pneumoniae* (NCBI accession: CP008929) to the *Salmonella* Typhimurium ST4/74 *tp2* sequence identified an 18 nt insertion in *Klebsiella pneumoniae*, accounting for a large proportion of the sequence (Figure 2.10).

Large changes in sequence length have been previously observed in Enterobacteriaceae sRNAs. The MicF RNA, which is also in a conserved locus, has been previously annotated in *Yersinia* using synteny information and BLAST searches for *micF* flanking sequences from *E. coli*. (Delihas, 2003). Comparison of the *micF* sequences from *E. coli*, *Yersinia* and *Serratia* sp. found indels of 6–12 nt across the RNA and promoter region, in addition to high sequence variation, despite the MicF RNAs from these species having conserved functions.

Synteny has also been used to annotate SgrS, an sRNA containing a protein-coding region, which can vary from in length by over 300 nt, by using the presence of a conserved flanking gene *sgrR* to identify the correct intergenic region (Horler and Vanderpool, 2009).

The sequence composition of certain sRNAs made annotation with HMMs difficult. The RseX sRNA, which has been experimentally confirmed in *E. coli* (Douchin *et al.*, 2006), was not found outside of *Salmonella* in this study. Although the *Salmonella* and *E. coli rseX* sequences share 70% sequence identity, the short length and biased composition of single-nucleotide repeats in both sequences led to the results being filtered by the composition bias filter in `nhmmer`. Results returned for *E. coli rseX* with and without (`-nobias`) this filter generated annotations with bit-scores below the detection threshold of the pipeline.



SroC

```
SroC_Salmonella_Typhimurium_CP002487  actaactaaaactaaatgggacggg-actcctgtcctctcgattgtc
SroC_Yersinia_pestis_CP009996         aacaaactcccttaa--aggatgga-taagccaccctgatgactgat
#=GC SS_cons                          .................((((((((....))))))))...(((((((

SroC_Salmonella_Typhimurium_CP002487  g-ctcagcacggacag-------------actataa---gcctgatg
SroC_Yersinia_pestis_CP009996         gattcaataccttcggctcccgagggttgacaaaaaaatgcggggtg
#=GC SS_cons                          ((((..................................(((((((((

SroC_Salmonella_Typhimurium_CP002487  gtcgttccccatcgggcctgaagcc-gcaaatcgccgggcaacaatc
SroC_Yersinia_pestis_CP009996         gggaacgaccaccccgcatcctgtctaccaataggccatcatcaatc
#=GC SS_cons                          (.......))))))))))....(((((....))))))))))))))))

SroC_Salmonella_Typhimurium_CP002487  ttcgagggt-agcgctgct-acccttttttt-----------
SroC_Yersinia_pestis_CP009996         g-gtagggcaggtatgacccgccctgattattattgttagttc
#=GC SS_cons                          ...((((((.(((...))).)))))).................
//
```

tp2

```
tp2_Salmonella_Typhimurium_CP002487   ggggtattccttatctatctaatacgttgagttatctggaacctgtcc
tp2_Klebsiella_pneumoniae_CP008929    ggggtattccttatctatctaat-cgttgattaatctggaacctgtcc
#=GC SS_cons                          .(((((.....)))))..((((..(((..........((((((((((.

tp2_Salmonella_Typhimurium_CP002487   cattaaact-ttcgt-----------------gagaaagcttaataag
tp2_Klebsiella_pneumoniae_CP008929    cattgaactcttcgttcttcacgctggaacaaccaaaaattcaataag
#=GC SS_cons                          .((((((((.(((....................)))))))))))..)

tp2_Salmonella_Typhimurium_CP002487   acaggttctgcgtttagttgccgcgcac--atttttgcgcgttattca
tp2_Klebsiella_pneumoniae_CP008929    acaggttctgcgtttagttgccgcgctctaaaaatggcgc-ttaagaa
#=GC SS_cons                          ))))))))))))).))))....((((((........)))))).......

tp2_Salmonella_Typhimurium_CP002487   caacatcttctggttaacgttctgccagaaaaa
tp2_Klebsiella_pneumoniae_CP008929    ccggagttctcg---------------------
#=GC SS_cons                          .................................
//
```

Figure 2.10: Examples of difficult to annotate ncRNAs. These sequences contain large indels which make sequence alignment challenging. "Missing" intergenic sequences, consisting of the sequence located between flanking region annotations, and sequences annotated by the homology search were aligned using `mafft-qinsi`, and consensus secondary structure predictions generated with `RNAalifold`. **Top:** Alignment of *SroC* from *S.* Typhimurium and intergenic sequence from same locus in *Yersinia pestis*. **Bottom:** Alignment of *tp2* from *Salmonella* Typhimurium and intergenic sequence from the same locus in *Klebsiella pneumoniae*. Bases are coloured by consensus structure using `RALEE` (Griffiths-Jones, 2005).

2.4.4 Poorly conserved sRNAs are associated with horizontal gene transfer within *Salmonella*

Homology search results show that a large proportion of *S.* Typhimurium ST4/74 sRNAs are *Salmonella*-specific. Sixty sRNAs were found only in *Salmonella*, 39 of which were annotated and highly conserved across all *Salmonella* genomes. Twenty-one sRNAs were limited to *S. enterica* subsp. *enterica*, with *STnc3800* being the only sRNA that was specific to *S.* Typhimurium.

Eighteen sRNAs were present only in *S. enterica* subsp. *enterica* within *Salmonella*, but were also annotated in other genera. Annotations of surrounding proteins and genome comparisons found these sRNAs are associated with horizontal gene transfer events into *Salmonella*, or are homologous to mobile genetic elements (Figures 2.11 and 2.2). These include sRNAs found in large horizontally-acquired regions, such as prophages and pathogenicity islands, and sRNAs associated with transposons, insertion elements, and cryptic phages.

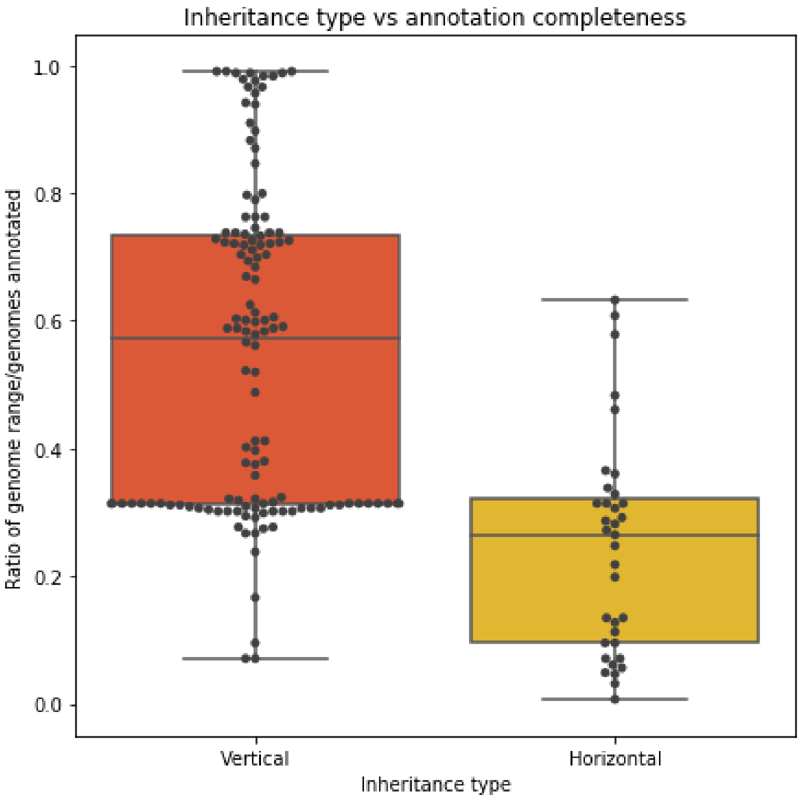


Figure 2.11: Box-plot showing annotation density for sRNAs for horizontally-acquired and vertically inherited sRNAs. The y axis shows a ratio of annotated vs non-annotated genomes for an sRNA (all genomes for the genera containing at least one sRNA annotation). This shows that the proportion of genomes annotated is lower for horizontally-acquired sRNAs, as the mobile genetic elements are not consistently present in any lineage.

Many *Salmonella* pathogenicity islands are horizontally transferred between strains or closely related species within *Salmonella*. These islands are thought to be unique to the *Salmonella* lineage, acquired after the *Salmonella/E. coli* split several million years ago (Bäumler, 1997). Many pathogenicity islands appear to have a bacteriophage origin, and may have formed from phage sequence, or sequence laterally transferred by phage machinery (Schmidt and Hensel, 2004). Several sRNAs located in pathogenicity islands appear to also have a phage origin, as they had homologous sequences in annotated prophages across the Enterobacteriaceae.

Other *Salmonella*-specific sRNAs did not have high sequence identity to MGE ncRNAs in other genera, but context such as surrounding protein-coding genes and analogous secondary structure (Hershko-Shalev *et al.*, 2016) are indicative of a MGE origin. Many horizontally-acquired sRNAs were located within intact and cryptic transposons, such as *isrF*, which is located next to a frame-shifted transposase in *S.* Typhimurium. The *isrN* gene is located in between putative transposases; BLAST and Pfam annotations of the region encompassing *isrN* and both flanking ORFs indicate that this is a frame-shifted IS3 transposase. The *STnc490* sRNA (renamed *tnpA*, previously *art200*), which is located in the $5'$ UTR of IS200 transposons, has recently been identified as an important regulator of of pathogenicity in *Salmonella* (Ellis *et al.*, 2018, 2017). Transposon-derived sRNAs have also been introduced by *Tn10/IS10* and *Tn5/IS50* in *E. coli* (Ellis *et al.*, 2015, Ross *et al.*, 2013, 2014) suggesting that transposons may be a frequent source of ncRNA genes in these genera.

2.4.5 Expression of conserved and horizontally-acquired sRNAs

Highly conserved sRNAs were also both highly and constitutively expressed across multiple growth conditions in experiments by Kröger *et al.* (2012) and Kröger *et al.* (2013). Almost all of these sRNAs have verified interactions with the RNA-binding chaperones Hfq and ProQ (Holmqvist *et al.*, 2016, Smirnov *et al.*, 2016) (shown in Figures 2.8D and 2.7D), and are known to participate in complex regulons which influence essential stress responses and metabolic processes (Updegrove *et al.*, 2015).

The less well conserved vertically inherited *Salmonella*-specific sRNAs were often poorly expressed, and often expressed in specific conditions. Few of these sRNAs have been experimentally verified or functionally characterised, however those exhibiting condition-specific expression presumably participate in or are linked to specific responses to the environment.

Horizontally-acquired *Salmonella*-specific sRNAs were generally highly and constitutively expressed, and few are predicted to bind to Hfq or ProQ. Non-coding RNAs contained within MGEs are often constitutively expressed. This may be to ensure dosage is sufficient when interacting with highly-expressed elements of the core genome, such as the Hfq-interacting sRNAs in *E. coli* prophages (Tree *et al.*, 2014), and the *Salmonella* SPI-1 sRNA InvR (Pfeiffer *et al.*, 2007). These sRNAs promote maintenance of the element by providing a beneficial phenotype to the host, which is best achieved by targeting genes which are themselves highly or constitutively expressed, allowing them to integrate into or form regulons essential for growth and survival (Fröhlich *et al.*, 2016, Pfeiffer *et al.*, 2007). More well-known examples of MGE-encoded bacterial sRNAs are ncRNA repressors, which are highly expressed to suppress toxic effects of MGE genes, such as the phage-encoded IsrK (Hershko-Shalev *et al.*, 2016), or to act as antitoxins in addiction modules (Lobato-Márquez *et al.*, 2015).

Genomic island sRNAs are generally expressed in specific growth conditions. While these sRNAs appear to have an MGE-origin, their relatively ancient association with *Salmonella* may have allowed them to integrate with regulatory circuitry and take on more specific functions (Fröhlich *et al.*, 2016).

2.4.6 Vertically-inherited sRNAs are associated with integration

The adjacent proteins for many highly conserved sRNAs (*micA*, *micC*, *omrA*, *omrB*, *rybA*, *rydC*, *ryeC*, *fnrS*) were sometimes annotated as MGE-like (i.e a transposase or integrase), which was infrequent and typically lineage-specific. The frequency of these annotations made it difficult to confidently assign inheritance type from functional annotations of nearby proteins alone (Figures 2.12 and 2.2).

Several vertically-inherited sRNAs were highly enriched for MGE insertion. The *sdsR* gene has been identified as a phage insertion hot-spot in *Salmonella* and *E. coli* (Balbontín *et al.*, 2008), had 12.9% of flanking protein annotations classified as MGE-like. MGE-like proteins were found near *sdsR* in all enteric genera, indicating that *sdsR* is also a preferred MGE insertion site outside of the *Salmonella-Escherichia* lineage. Pathogenicity island sRNAs had varying proportions of nearby MGE-like protein annotations (*STnc1710, STnc3870* in Figure 2.12), due to the large size of these elements.

The frequency and distribution of these annotations across the phylogeny suggest frequent insertion of transposable elements near sRNAs. An interesting example of a transposase insertion can be seen in *spf*, which is located in a highly conserved ribosomal protein locus. In all *Yersinia pestis* genomes in this study, *spf* was next to an IS1541a transposase insertion. Some signals of past insertions can also be seen, for example the highly conserved iron-storage regulator *fnrS* was located near a cryptic phage integrase in all *Salmonella* and some *E. coli* genomes in the data-set.

Some sRNAs are also located next to tRNAs, which are common insertion sites for phages and MGEs (Reiter *et al.*, 1989, Williams, 2002). The *glmZ* sRNA is located next to a tRNA locus, and is associated with integrases in strains from *Salmonella*, *Shigella* and *Yersinia*. A recent example of this can be seen in *S. enterica* sv. *Montevideo*, where *glmZ* appears to have been an insertion site for IS200. Poorly conserved sRNAs were also associated with insertions: the *Salmonella*-specific *isrO* sRNA, which is located next to a tRNA gene, is next to integrase annotations in *Salmonella* Typhi, which is a recently diverged strain (~50,000 years old) (Kidgell *et al.*, 2002).

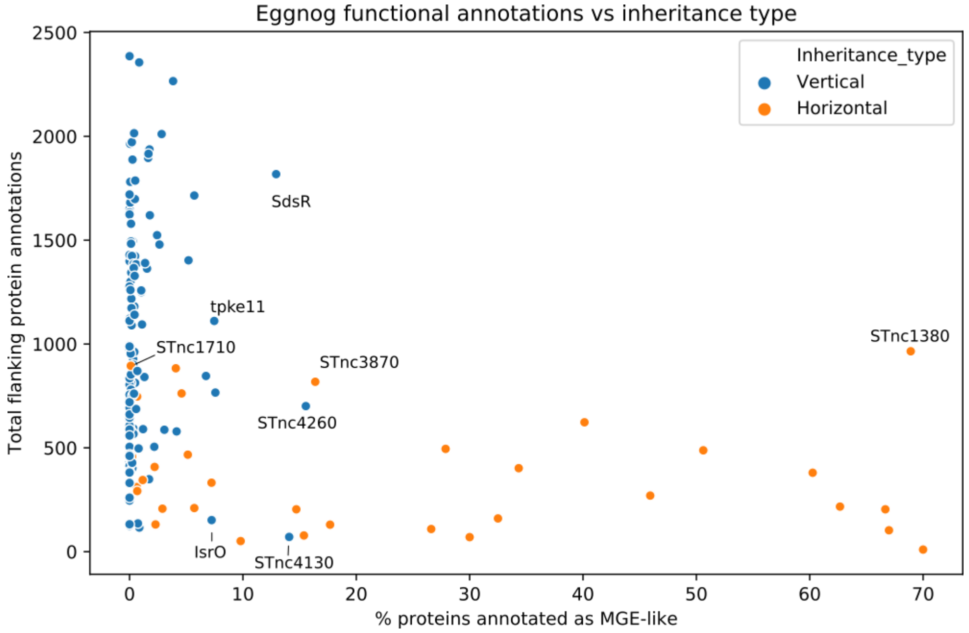


Figure 2.12: Plot showing the association between sRNA flanking protein function and number of total number of annotations. Total annotations are for 6 proteins (+/-3 either side of the sRNA) per genome. Points for each gene are coloured by inheritance type. Most vertically inherited sRNAs have less than 0-5% of flanking proteins annotated as MGE-like, and no vertically-inherited sRNAs have more than 20% MGE-like flanking proteins. Horizontally-acquired sRNAs range from 0% of proteins annotated as MGE-like, for sRNAs that are part of larger MGEs such as pathogenicity islands, to 70% for sRNAs within transposition machinery. Several outliers from each inheritance type are highlighted, and are discussed in the text.

Many mobile genetic elements recognise specific sequence motifs as insertion sites, such as inverted repeats or tRNAs, which share similarities with structural components of sRNAs (Darmon and Leach, 2014, Siguier *et al.*, 2014, Williams, 2002). This has been proposed as the reason for phage insertion hotspots near *cyaR*, which is an insertion site for phage P2 in *E. coli* (De Lay and Gottesman, 2009). The frequent association between sRNAs and mobile genetic elements makes it difficult to untangle if these sRNAs are exapted from sequence motifs that permit MGE insertion, or are formed from remnants of previous HGT events (Jose *et al.*, 2019), as inverted repeats are both common within transposable elements and can be deposited within genomes by transposition (Delihas, 2011, Siguier *et al.*, 2014, Vandecraen *et al.*, 2017).

2.4.7 Exaptation of sRNAs from type I toxin-antitoxin systems

Type I toxin-antitoxin (TA) systems utilise sRNAs as antitoxins, which bind to toxin mRNAs to suppress toxin translation (Brantl, 2012). Many of these are chromosomally encoded, and do not show classic signatures of horizontal gene transfer (Coray *et al.*, 2017, Fozo *et al.*, 2010). Several *Salmonella* sRNAs are derived from Type I TA antitoxins, such as the *rygC* and *rygD* sRNAs (now renamed *sibA* and *sibC*), which are known homologues of the Sib RNA from Ibs-Sib TA systems in *E. coli* K12 (Han *et al.*, 2010, Hébrard *et al.*, 2012). Homology search results found that *STnc1420* also shares sequence similarity with annotated Sib RNAs in *E. coli*.

While the primary function of these sRNAs is to protect the host from the toxin, these regulons may also provide a fitness advantage. The *Salmonella* IstR sRNA (Vogel *et al.*, 2004), which inhibits translation of its cognate TisB toxin, but IstR expression decreases during stationary phase, slowing growth as the toxin accumulates, which is thought to assist DNA repair. Recently, the SraC (aka RyeA) and SdsR sRNAs, which are transcribed from overlapping genes located on opposite strands, have been found to act as a novel type of entirely sRNA-based TA system in *E. coli*, in which SdsR acts as a toxin by suppressing the translation of an inner membrane protein essential for growth (Choi *et al.*, 2018). Interestingly, this role does not appear to be conserved, as SdsR has different target ranges in *E. coli* and *Salmonella* (Fröhlich *et al.*, 2016), and is also more highly conserved than its purported antitoxin SraC (Fröhlich *et al.*, 2012).

The *Salmonella* pathogenicity island sRNA *isrA* was also found to share sequence similarity with the *symE-symR* TA system found in *E. coli*. A previous BLAST-based study of sRNA turnover (Skippington and Ragan, 2012) predicted *isrA* homologues in *E. coli* genomes, however *isrA* was not found outside *Salmonella* in this study after filtering. Examination of unfiltered homology search results showed multiple lower-scoring *isrA* annotations in *S. bongori* and *S. arizonae* genomes.

Of the 6 candidate *isrA* annotations in *S. bongori* (NCBI accession CP006693), the majority were located opposite small hypothetical proteins in a Type VI secretion system. A BLAST search for these proteins returned results with ~85% sequence similarity to the *E. coli symE* toxin. A homology search of the *S*. Typhimurium ST4/74 genome with HMMs of TA toxin proteins from Coray *et al.* (2017) annotated the hypothetical proteins adjacent to *isrA* as *symE*, suggesting that *isrA* may be derived from the sRNA antitoxin *symR*. Previous annotation of TAs in *Salmonella* Typhimurium (Lobato-Márquez *et al.*, 2015) did not identify this *symE* homologue with PSI-BLAST, however as with other sRNAs, the study and detection of Type I TA families is difficult due to the short and diverse RNA and protein-coding components (Coray *et al.*, 2017).

2.4.8 Conclusions and Future Directions

We aimed to examine the rates of turnover of *Salmonella* sRNAs using a sensitive iterative profile-HMM approach, which is effective at re-annotating known sRNAs from single starting sequences over large phylogenetic distances. These results confirm previous observations that sRNAs are both rapidly acquired and exhibit rapid sequence turnover. The incorporation of synteny anchors in this pipeline allowed us to resolve annotation conflicts, and predict that the limited phylogenetic distribution of many sRNAs is primarily due to gene gain and loss rather than sequence divergence.

Many sRNAs contain secondary structure motifs in common with each other (Gardner and Eldai, 2015), and with other structured RNAs. This highlighted an additional challenge for sRNA homology search: reducing specificity is required to detect more distant homologues, but this can generate high rates of false positives, requiring careful filtering. This is further compounded by the short length of many sequences, making it easy for individual structural elements to return highly scoring results during homology search. The issue of false positive rates has also been considered for the recently released tool GLASSgo, which also uses an iterative approach to identify sRNA homologues using BLAST. The authors also acknowledge the risks of high false positive rates using a low-specificity iterative approach, and rely on structure-based clustering to increase the true positive rates (Lott *et al.*, 2018).

With any nucleotide-based sRNA homology search, consideration of structural similarities to existing ncRNAs and evolutionary context, such as the presence of paralogues, is essential for studying rapidly evolving genes. Signals other than nucleotide sequence can be incorporated into homology search, for example 2D and 3D structure motifs for sRNAs containing conserved secondary structure (Barquist *et al.*, 2016, Gardner and Eldai, 2015). Genomic or syntenic context, such as the presence of conserved promoters or nearby proteins, is also useful in identifying highly divergent homologous sequences, however methods to incorporate this information are likely to be bespoke and difficult to scale (Menzel *et al.*, 2009), particularly for heterogenous genes such as bacterial sRNAs.

Synteny has been used as part of a strategy to study the conservation of the *Staphylococcus* sRNA RsaC, whic is involved in the oxidative stress response. While RsaC is in a conserved operon, it was found to have variable sequence length, and study of flanking genes were used to find homologues of RsaC(Lalaouna *et al.*, 2019).

This study indicates that a large proportion of *Salmonella enterica*-specific sRNAs (19/51) have been acquired *via* horizontal gene transfer. While the classification of sRNAs into horizontally and vertically inherited is limited by our ability to detect signals of horizontal gene transfer over large periods of evolutionary time, it appears to horizontal gene transfer is the main driver of sRNA acquisition in *Salmonella*. As we discussed previously (Jose *et al.*, 2019), exaptation of sRNA sequences introduced by horizontal gene transfer provides an easy way to integrate host and foreign DNA into regulons which benefit the host. Many sRNAs acquired by horizontal gene transfer interact with the core genome, as well as sRNAs which act to regulate HGT-acquired/derived elements (Fröhlich and Papenfort, 2016) where sRNA regulation 'tames' HGT-acquired genes (Papenfort and Vogel, 2010).

The prevalence of 'new' sRNAs that derive from toxin-antitoxin systems indicate another interesting route for acquiring new sRNAs through exaptation. It is thought that chromosomal Type I TA families spread by duplication; having multiple redundant TA systems may free an sRNA from strictly performing an antitoxin role, allowing it to be easily exapted (Jose *et al.*, 2019). The maintenance of orphan antitoxins sequences in the absence of its toxin counterpart may also be a useful defence against further TA integration, or provide redundancy for existing TAs.

We have also previously considered the likelihood of *de novo* sRNA formation (Jose *et al.*, 2019). This study highlights many *Salmonella* and some *S.* Typhimurium-specific sRNAs that are vertically inherited, raising the question as to how these genes arose. Exploring the function and expression of these genes across *Salmonella* sp. may clarify if these transcripts are conserved *de novo* genes, or transcriptional noise in intergenic regions.

Many aspects of genome architecture, ecology and selection pressures that we have discussed previously in the context of *de novo* sRNA gene evolution (Jose *et al.*, 2019) are also likely to affect the rates of sequence evolution and persistence of sRNA genes. Some bacteria have especially dynamic genomes, with high rates of gene influx from HGT that are beneficial for niche adaptation (Vos *et al.*, 2015). In this context, the overall conservation and evolutionary dynamics of sRNAs within a lineage may be quite different depending on the rate of HGT events, highlighting the need for studies in a wide range of organisms.

The pace and scope of sRNA annotation is increasing (Ami *et al.*, 2020, Pobre *et al.*, 2019), particularly with the development of RNA chaperone assays (Shin *et al.*, 2019). These function-based assays also allow for more confident annotation of candidate sRNAs against a background of transcriptional noise, although recent studies in Gram-positive bacteria suggests the presence of such proteins and interaction systems are not ubiquitous (Hör *et al.*, 2020).

Experiments to identify and functionally characterise Enterobacteriaceae sRNAs outside of the *Escherichia-Salmonella* clade are rapidly increasing the number of sequences available for comparison. Studies in *Yersinia* sp. (Han *et al.*, 2019, Martínez-Chavarría and Vadyvaloo, 2015, Yan *et al.*, 2013), and more recently in *Erwinia* (Schachterle *et al.*, 2019), *Edwardsiella* (Gao *et al.*, 2019) and *Pantoea* (Shin *et al.*, 2019) are providing insight into sRNA sequence and functional conservation across phylogenetic distances that are poorly served by ncRNA homology search.

Studies of sRNA conservation in many genera remain limited to a select number of genes that have been functionally characterised, however, available studies broadly reflect the results of this work. A recent BLAST-based study of a large number of *Pantoea ananatis* candidate sRNAs identified by a Hfq pulldown appears to show similar results to this study. Shin *et al.* (2019) found that the majority of *P. anananatis* sRNAs are genus-specific, with only ~15% having some homologue in *S. enterica*.

Fine-grained studies of sRNA evolution remain lacking in other families, however recent efforts to characterise sRNAs in Pseudomonadaceae (Filiatrault *et al.*, 2010, Gómez-Lozano *et al.*, 2012, 2015, Pita *et al.*, 2018, Pobre *et al.*, 2019), a family that is closely related to the Enterobacteriaceae, are likely to make between-family comparisons of sRNA turnover and evolution feasible in the near future. Although comprehensive studies of sequence conservation are not available, most *Pseudomonas* sRNAs discovered to-date appear to be species-specific (Gómez-Lozano *et al.*, 2015, Pobre *et al.*, 2019).

A study of sRNAs in another close relative of the Enterobacteriaceae, *Vibrio parahaemolyticus*, found that 88% of annotated *V. parahaemolyticus* sRNAs are species-specific, and also identified strain-specific sRNAs associated with a mobile genetic element(Plaza *et al.*, 2019). Another niche-specific sRNA has been found to be conserved across *Coxiella burnetti* strains that are mammalian pathogens (Wachter *et al.*, 2019).

Recent work has also found that many regulatory sRNAs are also present in Archaea (Gelsinger and DiRuggiero, 2018), some of which also appear to show poor conservation. A recent BLAST-based study found a genus-specific sRNA conserved across *Haloferax*, that functions in a niche-specific phosphate starvation response (Kliemt *et al.*, 2019). However, studies in Archaea and more obscure branches of Bacteria are limited by a lack of available genomes from closely related taxa for comparison.

While improved homology search methods are crucial for understanding sRNA sequence conservation, this information must be carefully integrated alongside domain-specific knowledge to properly address larger questions about their evolution and origins (Georg *et al.*, 2019). This concept is the basis for the novel aspects of this study, namely the large-scale curation and integration of a variety of contextual information to predict the origins of individual sRNA genes, which appears to be a major factor in the poor conservation of many *Salmonella* sRNAs.

Work from this Chapter forms the basis of an upcoming publication, "Horizontal-gene transfer is a major source of novel sRNAs in *Salmonella* Typhimurium".

References

Ami, V. K. G., Balasubramanian, R., and Hegde, S. R. (2020). Genome-wide identification of the context-dependent sRNA expression in Mycobacterium tuberculosis. *BMC Genomics*, *21*(1), 167.

Babitzke, P. and Romeo, T. (2007). CsrB sRNA family: sequestration of RNA-binding regulatory proteins. *Curr. Opin. Microbiol.*, *10*(2), 156–163.

Balbontín, R., Figueroa-Bossi, N., Casadesús, J., and Bossi, L. (2008). Insertion hot spot for horizontally acquired DNA within a bidirectional small-RNA locus in *Salmonella enterica*. *J. Bacteriol.*, *190*(11), 4075–4078.

Barquist, L., Burge, S. W., and Gardner, P. P. (2016). Studying RNA Homology and Conservation with Infernal: From Single Sequences to RNA Families. *Curr. Protoc. Bioinformatics*, *54*, 12.13.1–12.13.25.

Barquist, L., Langridge, G. C., Turner, D. J., Phan, M.-D., Turner, A. K., Bateman, A., Parkhill, J., Wain, J., and Gardner, P. P. (2013). A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium. *Nucleic Acids Res.*, *41*(8), 4549–4564.

Barquist, L. and Vogel, J. (2015). Accelerating Discovery and Functional Analysis of Small RNAs with New Technologies. *Annu. Rev. Genet.*, *49*, 367–394.

Bäumler, A. J. (1997). The record of horizontal gene transfer in *Salmonella*. *Trends Microbiol.*, *5*(8), 318–322.

Bordeau, V. and Felden, B. (2014). Curli synthesis and biofilm formation in enteric bacteria are controlled by a dynamic small RNA module made up of a pseudoknot assisted by an RNA chaperone. *Nucleic Acids Res.*, *42*(7), 4682–4696.

Brantl, S. (2012). Bacterial type I toxin-antitoxin systems. *RNA Biol.*, *9*(12), 1488–1490.

Canals, R., Hammarlöf, D. L., Kröger, C., Owen, S. V., Fong, W. Y., Lacharme-Lora, L., Zhu, X., Wenner, N., Carden, S. E., Honeycutt, J., Monack, D. M., Kingsley, R. A., Brownridge, P., Chaudhuri, R. R., Rowe, W. P. M., Predeus, A. V., Hokamp, K., Gordon, M. A., and Hinton, J. C. D. (2019). Adding function to the genome of African *Salmonella* Typhimurium ST313 strain D23580. *PLoS Biol.*, *17*(1), e3000059.

Chao, Y. and Vogel, J. (2016). A $3'$ UTR-Derived Small RNA Provides the Regulatory Noncoding Arm of the Inner Membrane Stress Response. *Mol. Cell*.

Choi, J. S., Kim, W., Suk, S., Park, H., Bak, G., Yoon, J., and Lee, Y. (2018). The small RNA, SdsR, acts as a novel type of toxin in *Escherichia coli*. *RNA Biol.*, *15*(10), 1319–1335.

Colgan, A. M., Kröger, C., Diard, M., Hardt, W.-D., Puente, J. L., Sivasankaran, S. K., Hokamp, K., and Hinton, J. C. D. (2016). The Impact of 18 Ancestral and Horizontally-Acquired Regulatory Proteins upon the Transcriptome and sRNA Landscape of *Salmonella enterica* serovar Typhimurium. *PLoS Genet.*, *12*(8), e1006258.

Coray, D. S., Wheeler, N. E., Heinemann, J. A., and Gardner, P. P. (2017). Why so narrow: Distribution of anti-sense regulated, type I toxin-antitoxin systems compared with type II and type III systems. *RNA Biol.*, *14*(3), 275–280.

Darmon, E. and Leach, D. R. F. (2014). Bacterial genome instability. *Microbiol. Mol. Biol. Rev.*, *78*(1), 1–39.

De Lay, N. and Gottesman, S. (2009). The Crp-activated small noncoding regulatory RNA CyaR (RyeE) links nutritional status to group behavior. *J. Bacteriol.*, *191*(2), 461–476.

Delihas, N. (2003). Annotation and evolutionary relationships of a small regulatory RNA gene *micF* and its target *ompF* in *Yersinia* species. *BMC Microbiol.*, *3*, 13.

Delihas, N. (2011). Impact of small repeat sequences on bacterial genome evolution. *Genome Biol. Evol.*, *3*, 959–973.

Dimri, G. P., Rudd, K. E., Morgan, M. K., Bayat, H., and Ames, G. F. (1992). Physical mapping of repetitive extragenic palindromic sequences in *Escherichia coli* and phylogenetic distribution among *Escherichia coli* strains and other enteric bacteria. *J. Bacteriol.*, *174*(14), 4583–4593.

Douchin, V., Bohn, C., and Bouloc, P. (2006). Down-regulation of porins by a small RNA bypasses the essentiality of the regulated intramembrane proteolysis protease RseP in *Escherichia coli*. *J. Biol. Chem.*, *281*(18), 12253–12259.

Dutcher, H. A. and Raghavan, R. (2018). Origin, Evolution, and Loss of Bacterial Small RNAs. *Microbiol. Spectr.*, *6*(2).

Eddy, S. R. (2009). A New Generation of Homology Search Tools Based on Probabilistic Inference. *Genome Inform.*, *23*(1), 205–11.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, *7*(10), e1002195.

Ellis, M. J., Carfrae, L. A., Macnair, C. R., Trussler, R. S., Brown, E. D., and Haniford, D. B. (2018). Silent but deadly: IS200 promotes pathogenicity in *Salmonella* Typhimurium. *RNA Biol.*, *15*(2), 176–181.

Ellis, M. J., Trussler, R. S., Charles, O., and Haniford, D. B. (2017). A transposon-derived small RNA regulates gene expression in *Salmonella* Typhimurium. *Nucleic Acids Res.*.

Ellis, M. J., Trussler, R. S., and Haniford, D. B. (2015). Hfq binds directly to the ribosome-binding site of IS10 transposase mRNA to inhibit translation. *Mol. Microbiol.*, *96*(3), 633–650.

Espéli, O., Moulin, L., and Boccard, F. (2001). Transcription attenuation associated with bacterial repetitive extragenic BIME elements. *J. Mol. Biol.*, *314*(3), 375–386.

Filiatrault, M. J., Stodghill, P. V., Bronstein, P. A., Moll, S., Lindeberg, M., Grills, G., Schweitzer, P., Wang, W., Schroth, G. P., Luo, S., Khrebtukova, I., Yang, Y., Thannhauser, T., Butcher, B. G., Cartinhour, S., and Schneider, D. J. (2010). Transcriptome analysis of *Pseudomonas syringae* identifies new genes, noncoding RNAs, and antisense activity. *J. Bacteriol.*, *192*(9), 2359–2372.

Fozo, E. M., Makarova, K. S., Shabalina, S. A., Yutin, N., Koonin, E. V., and Storz, G. (2010). Abundance of type I toxin-antitoxin systems in bacteria: searches for new candidates and discovery of novel families. *Nucleic Acids Res.*, *38*(11), 3743–3759.

Fröhlich, K. S., Haneke, K., Papenfort, K., and Vogel, J. (2016). The target spectrum of SdsR small RNA in *Salmonella*. *Nucleic Acids Res.*, *44*(21), 10406–10422.

Fröhlich, K. S. and Papenfort, K. (2016). Interplay of regulatory RNAs and mobile genetic elements in enteric pathogens. *Mol. Microbiol.*, *101*(5), 701–713.

Fröhlich, K. S., Papenfort, K., Berger, A. A., and Vogel, J. (2012). A conserved RpoS-dependent small RNA controls the synthesis of major porin OmpD. *Nucleic Acids Res.*, *40*(8), 3623–3640.

Gao, D., Zhang, Y., Liu, R., Fang, Z., and Lu, C. (2019). EsR240, a non-coding sRNA, is required for the resistance of *Edwardsiella tarda* to stresses in macrophages and for virulence. *Vet. Microbiol.*, *231*, 254–263.

Gardner, P. P. and Eldai, H. (2015). Annotating RNA motifs in sequences and alignments. *Nucleic Acids Res.*, *43*(2), 691–698.

Gelsinger, D. R. and DiRuggiero, J. (2018). The Non-Coding Regulatory RNA Revolution in Archaea. *Genes*, *9*(3).

Georg, J., Lalaouna, D., Hou, S., Lott, S. C., Caldelari, I., Marzi, S., Hess, W. R., and Romby, P. (2019). The power of cooperation: Experimental and computational approaches in the functional characterization of bacterial sRNAs. *Mol. Microbiol.*.

Gómez-Lozano, M., Marvig, R. L., Molin, S., and Long, K. S. (2012). Genome-wide identification of novel small RNAs in *Pseudomonas aeruginosa*. *Environ. Microbiol.*, *14*(8), 2006–2016.

Gómez-Lozano, M., Marvig, R. L., Molina-Santiago, C., Tribelli, P. M., Ramos, J.-L., and Molin, S. (2015). Diversity of small RNAs expressed in *Pseudomonas* species. *Environ. Microbiol. Rep.*, *7*(2), 227–236.

Gottesman, S. and Storz, G. (2011). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb. Perspect. Biol.*, *3*(12).

Griffiths-Jones, S. (2005). RALEE–RNA ALignment editor in Emacs. *Bioinformatics*, *21*(2), 257–259.

Gruber, A. R., Findeiß, S., Washietl, S., Hofacker, I. L., and Stadler, P. F. (2010). RNAz 2.0: improved noncoding RNA detection. *Pacific Symposium on Biocomputing*, 69–79.

Guo, M. S., Updegrove, T. B., Gogol, E. B., Shabalina, S. A., Gross, C. A., and Storz, G. (2014). MicL, a new σE-dependent sRNA, combats envelope stress by repressing synthesis of Lpp, the major outer membrane lipoprotein. *Genes Dev.*, *28*(14), 1620–1634.

Han, K., Kim, K.-S., Bak, G., Park, H., and Lee, Y. (2010). Recognition and discrimination of target mRNAs by Sib RNAs, a *cis*-encoded sRNA family. *Nucleic Acids Res.*, *38*(17), 5851–5866.

Han, Y., Chen, D., Yan, Y., Gao, X., Liu, Z., Xue, Y., Zhang, Y., and Yang, R. (2019). Hfq Globally Binds and Destabilizes sRNAs and mRNAs in *Yersinia pestis*. *mSystems*, *4*(4).

Hébrard, M., Kröger, C., Srikumar, S., Colgan, A., Händler, K., and Hinton, J. C. D. (2012). sRNAs and the virulence of *Salmonella enterica* serovar Typhimurium. *RNA Biol.*, *9*(4), 437–445.

Hershberg, R., Altuvia, S., and Margalit, H. (2003). A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Res.*, (7), 1813–1820.

Hershko-Shalev, T., Odenheimer-Bergman, A., Elgrably-Weiss, M., Ben-Zvi, T., Govindarajan, S., Seri, H., Papenfort, K., Vogel, J., and Altuvia, S. (2016). Gifsy-1 Prophage IsrK with Dual Function as Small and Messenger RNA Modulates Vital Bacterial Machineries. *PLoS Genet.*, *12*(4), e1005975.

Holmqvist, E., Wright, P. R., Li, L., Bischler, T., Barquist, L., Reinhardt, R., Backofen, R., and Vogel, J. (2016). Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking *in vivo*. *EMBO J.*, *35*(9), 991–1011.

Hör, J., Garriss, G., Di Giorgio, S., Hack, L.-M., Vanselow, J. T., Förstner, K. U., Schlosser, A., Henriques-Normark, B., and Vogel, J. (2020). Grad-seq in a Gram-positive bacterium reveals exonucleolytic sRNA activation in competence control. *EMBO J.*, *39*(9), e103852.

Horler, R. S. P. and Vanderpool, C. K. (2009). Homologs of the small RNA SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence. *Nucleic Acids Res.*, *37*(16), 5465–5476.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., and Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.*, *34*(8), 2115–2122.

Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, *11*, 431.

Jose, B. R., Gardner, P. P., and Barquist, L. (2019). Transcriptional noise and exaptation as sources for bacterial sRNAs. *Biochem. Soc. Trans.*.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, *30*(4), 772–780.

Khemici, V. and Carpousis, A. J. (2004). The RNA degradosome and poly(A) polymerase of *Escherichia coli* are required in vivo for the degradation of small mRNA decay intermediates containing REP-stabilizers. *Mol. Microbiol.*, *51*(3), 777–790.

Kidgell, C., Reichard, U., Wain, J., Linz, B., Torpdahl, M., Dougan, G., and Achtman, M. (2002). *Salmonella* typhi, the causative agent of typhoid fever, is approximately 50,000 years old. *Infect. Genet. Evol.*, *2*(1), 39–45.

Kliemt, J., Jaschinski, K., and Soppa, J. (2019). A Haloarchaeal Small Regulatory RNA (sRNA) Is Essential for Rapid Adaptation to Phosphate Starvation Conditions. *Front. Microbiol.*, *10*, 1219.

Kröger, C., Colgan, A., Srikumar, S., Händler, K., Sivasankaran, S. K., Hammarlöf, D. L., Canals, R., Grissom, J. E., Conway, T., Hokamp, K., and Hinton, J. C. D. (2013). An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium. *Cell Host Microbe*, *14*(6), 683–695.

Kröger, C., Dillon, S. C., Cameron, A. D. S., Papenfort, K., Sivasankaran, S. K., Hokamp, K., Chao, Y., Sittka, A., Hébrard, M., Händler, K., Colgan, A., Leekitcharoenphon, P., Langridge, G. C., Lohan, A. J., Loftus, B., Lucchini, S., Ussery, D. W., Dorman, C. J., Thomson, N. R., Vogel, J., and Hinton, J. C. D. (2012). The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. *PNAS*, *109*(20), E1277–86.

Lalaouna, D., Baude, J., Wu, Z., Tomasini, A., Chicher, J., Marzi, S., Vandenesch, F., Romby, P., Caldelari, I., and Moreau, K. (2019). RsaC sRNA modulates the oxidative stress response of Staphylococcus aureus during manganese starvation. *Nucleic Acids Res.*, *47*(18), 9871–9887.

Liang, W., Rudd, K. E., and Deutscher, M. P. (2015). A role for REP sequences in regulating translation. *Mol. Cell*, *58*(3), 431–439.

Lindgreen, S., Umu, S. U., Lai, A. S.-W., Eldai, H., Liu, W., McGimpsey, S., Wheeler, N. E., Biggs, P. J., Thomson, N. R., Barquist, L., Poole, A. M., and Gardner, P. P. (2014). Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Comput. Biol.*, *10*(10), e1003907.

Lobato-Márquez, D., Moreno-Córdoba, I., Figueroa, V., Díaz-Orejas, R., and García-del Portillo, F. (2015). Distinct type I and type II toxin-antitoxin modules control *Salmonella* lifestyle inside eukaryotic cells. *Sci. Rep.*, *5*, 9374.

Lott, S. C., Schäfer, R. A., Mann, M., Backofen, R., Hess, W. R., Voß, B., and Georg, J. (2018). GLASSgo - Automated and Reliable Detection of sRNA Homologs From a Single Input Sequence. *Front. Genet.*, *9*, 124.

Martínez-Chavarría, L. C. and Vadyvaloo, V. (2015). *Yersinia pestis* and *Yersinia pseudotuberculosis* infection: a regulatory RNA perspective. *Front. Microbiol.*, *6*, 956.

Menzel, P., Gorodkin, J., and Stadler, P. F. (2009). The tedious task of finding homologous noncoding RNA genes. *RNA*, *15*(12), 2075–2082.

Moon, K. and Gottesman, S. (2009). A PhoQ/P-regulated small RNA regulates sensitivity of *Escherichia coli* to antimicrobial peptides. *Mol. Microbiol.*, *74*(6), 1314–1330.

Moon, K., Six, D. A., Lee, H.-J., Raetz, C. R. H., and Gottesman, S. (2013). Complex transcriptional and post-transcriptional regulation of an enzyme for lipopolysaccharide modification. *Mol. Microbiol.*, *89*(1), 52–64.

Nawrocki, E. P. and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933–2935.

Papenfort, K., Förstner, K. U., Cong, J.-P., Sharma, C. M., and Bassler, B. L. (2015). Differential RNA-seq of *Vibrio cholerae* identifies the VqmR small RNA as a regulator of biofilm formation. *PNAS*.

Papenfort, K., Silpe, J. E., Schramma, K. R., Cong, J.-P., Seyedsayamdost, M. R., and Bassler, B. L. (2017). A *Vibrio cholerae* autoinducer-receptor pair that controls biofilm formation. *Nat. Chem. Biol.*.

Papenfort, K. and Vogel, J. (2010). Regulatory RNA in bacterial pathogens. *Cell Host Microbe*, *8*(1), 116–127.

Pfeiffer, V., Sittka, A., Tomer, R., Tedin, K., Brinkmann, V., and Vogel, J. (2007). A small non-coding RNA of the invasion gene island (SPI-1) represses outer membrane protein synthesis from the *Salmonella* core genome. *Mol. Microbiol.*, *66*(5), 1174–1191.

Pita, T., Feliciano, J. R., and Leitão, J. H. (2018). Small Noncoding Regulatory RNAs from *Pseudomonas aeruginosa* and *Burkholderia cepacia* Complex. *Int. J. Mol. Sci.*, *19*(12).

Plaza, N., Pérez-Reytor, D., Ramírez-Araya, S., Pavón, A., Corsini, G., Loyola, D. E., Jaña, V., Pavéz, L., Navarrete, P., Bastías, R., Castillo, D., and García, K. (2019). Conservation of Small Regulatory RNAs in Vibrio parahaemolyticus: Possible role of RNA-OUT Encoded by the Pathogenicity Island (VPaI-7) of Pandemic Strains. *Int. J. Mol. Sci.*, *20*(11).

Pobre, V., Graça-Lopes, G., Saramago, M., Ankenbauer, A., Takors, R., Arraiano, C. M., and Viegas, S. C. (2019). Prediction of novel non-coding RNAs relevant for the growth of Pseudomonas putida in a bioreactor. *Microbiology*.

Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Muller, J., Arnold, R., Rattei, T., Letunic, I., Doerks, T., Jensen, L. J., von Mering, C., and Bork, P. (2012). eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, *40*(Database issue), D284–9.

Reiter, W. D., Palm, P., and Yeats, S. (1989). Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic Acids Res.*, *17*(5), 1907–1914.

Richter, A. S. and Backofen, R. (2012). Accessibility and conservation: general features of bacterial small RNA-mRNA interactions? *RNA Biol.*, *9*(7), 954–965.

Rivas, E., Clements, J., and Eddy, S. R. (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods*, *14*(1), 45–48.

Ross, J. A., Ellis, M. J., Hossain, S., and Haniford, D. B. (2013). Hfq restructures RNA-IN and RNA-OUT and facilitates antisense pairing in the Tn10/IS10 system. *RNA*, *19*(5), 670–684.

Ross, J. A., Trussler, R. S., Black, M. D., McLellan, C. R., and Haniford, D. B. (2014). Tn5 transposition in *Escherichia coli* is repressed by Hfq and activated by over-expression of the small non-coding RNA SgrS. *Mob. DNA*, *5*(1), 27.

Santiago-Frangos, A. and Woodson, S. A. (2018). Hfq chaperone brings speed dating to bacterial sRNA. *Wiley Interdiscip. Rev. RNA*, *9*(4), e1475.

Schachterle, J., Onsay, D., and Sundin, G. (2019). Small RNA ArcZ regulates oxidative stress response genes and regulons in *Erwinia amylovora*. *Front. Microbiol.*, *10*, 2775.

Schmidt, H. and Hensel, M. (2004). Pathogenicity islands in bacterial pathogenesis. *Clin. Microbiol. Rev.*, *17*(1), 14–56.

Shin, G. Y., Schachterle, J. K., Shyntum, D. Y., Moleleki, L. N., Coutinho, T. A., and Sundin, G. W. (2019). Functional Characterization of a Global Virulence Regulator Hfq and Identification of Hfq-Dependent sRNAs in the Plant Pathogen Pantoea ananatis. *Front. Microbiol.*, *10*, 2075.

Siguier, P., Gourbeyre, E., and Chandler, M. (2014). Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.*, *38*(5), 865–891.

Skippington, E. and Ragan, M. A. (2012). Evolutionary dynamics of small RNAs in 27 *Escherichia coli* and *Shigella* genomes. *Genome Biol. Evol.*, *4*(3), 330–345.

Smirnov, A., Förstner, K. U., Holmqvist, E., Otto, A., Günster, R., Becher, D., Reinhardt, R., and Vogel, J. (2016). Grad-seq guides the discovery of ProQ as a major small RNA-binding protein. *PNAS*, *113*(41), 11591–11596.

Stern, M. J., Ames, G. F., Smith, N. H., Robinson, E. C., and Higgins, C. F. (1984). Repetitive extragenic palindromic sequences: a major component of the bacterial genome. *Cell*, *37*(3), 1015–1026.

Storz, G., Vogel, J., and Wassarman, K. M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. *Mol. Cell*, *43*(6), 880–891.

Tobes, R. and Pareja, E. (2006). Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements. *BMC Genomics*, *7*, 62.

Toth, I. K., Pritchard, L., and Birch, P. R. J. (2006). Comparative genomics reveals what makes an enterobacterial plant pathogen. *Annu. Rev. Phytopathol.*, *44*, 305–336.

Tree, J. J., Granneman, S., McAteer, S. P., Tollervey, D., and Gally, D. L. (2014). Identification of bacteriophage-encoded anti-sRNAs in pathogenic *Escherichia coli*. *Mol. Cell*, *55*(2), 199–213.

Updegrove, T. B., Shabalina, S. A., and Storz, G. (2015). How do base-pairing small RNAs evolve? *FEMS Microbiol. Rev.*, *39*(3), 379–391.

Valverde, C., Lindell, M., Wagner, E. G. H., and Haas, D. (2004). A repeated GGA motif is critical for the activity and stability of the riboregulator RsmY of *Pseudomonas fluorescens*. *J. Biol. Chem.*, *279*(24), 25066–25074.

Vandecraen, J., Chandler, M., Aertsen, A., and Van Houdt, R. (2017). The impact of insertion sequences on bacterial genome plasticity and adaptability. *Crit. Rev. Microbiol.*, *43*(6), 709–730.

Vogel, J., Argaman, L., Wagner, E. G. H., and Altuvia, S. (2004). The small RNA IstR inhibits synthesis of an SOS-induced toxic peptide. *Curr. Biol.*, *14*(24), 2271–2276.

Vos, M., Hesselman, M. C., Te Beek, T. A., van Passel, M. W. J., and Eyre-Walker, A. (2015). Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends Microbiol.*, *23*(10), 598–605.

Wachter, S., Bonazzi, M., Shifflett, K., Moses, A. S., Raghavan, R., and Minnick, M. F. (2019). A CsrA-Binding, trans-Acting sRNA of Coxiella burnetii Is Necessary for Optimal Intracellular Growth and Vacuole Formation during Early Infection of Host Cells. *J. Bacteriol.*, *201*(22).

Washietl, S., Findeiss, S., Müller, S. A., Kalkhof, S., von Bergen, M., Hofacker, I. L., Stadler, P. F., and Goldman, N. (2011). RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*, *17*(4), 578–594.

Wassarman, K. M. (2007). 6S RNA: a regulator of transcription. *Mol. Microbiol.*, *65*(6), 1425–1431.

Weilbacher, T., Suzuki, K., Dubey, A. K., Wang, X., Gudapaty, S., Morozov, I., Baker, C. S., Georgellis, D., Babitzke, P., and Romeo, T. (2003). A novel sRNA component of the carbon storage regulatory system of *Escherichia coli*. *Mol. Microbiol.*, *48*(3), 657–670.

Westermann, A. J., Förstner, K. U., Amman, F., Barquist, L., Chao, Y., Schulte, L. N., Müller, L., Reinhardt, R., Stadler, P. F., and Vogel, J. (2016). Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. *Nature*, *529*(7587), 496–501.

Wheeler, T. J. and Eddy, S. R. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, *29*(19), 2487–2489.

Williams, K. P. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.*, *30*(4), 866–875.

Yan, Y., Su, S., Meng, X., Ji, X., Qu, Y., Liu, Z., Wang, X., Cui, Y., Deng, Z., Zhou, D., Jiang, W., Yang, R., and Han, Y. (2013). Determination of sRNA expressions by RNA-seq in *Yersinia pestis* grown *in vitro* and during infection. *PLoS One*, *8*(9), e74495.

Yin, X., Wu Orr, M., Wang, H., Hobbs, E. C., Shabalina, S. A., and Storz, G. (2018). The small protein MgtS and small RNA MgrR modulate the PitA phosphate symporter to boost intracellular magnesium levels. *Mol. Microbiol.*.

Zhao, X., Liu, R., Tang, H., Osei-Adjei, G., Xu, S., Zhang, Y., and Huang, X. (2018). A 3′UTR-derived non-coding RNA RibS increases expression of *cfa* and promotes biofilm formation of *Salmonella enterica* serovar Typhi. *Res. Microbiol.*, *169*(6), 279–288.

Chapter 3

Comparative transcriptomics of *Pseudomonas syringae* pv. *actinidiae*

3.1 Preface

Pseudomonas syringae pv. *actinidiae* (*Psa*), a plant pathogen causing kiwifruit canker disease, has emerged as a major threat to kiwifruit agriculture during pandemic outbreaks in the past decade. This pathogen has had major impacts on the New Zealand kiwifruit industry, which is a large contributor to the export economy. This project is funded by Bio-Protection Research Centre, NZ, aiming to explore how this pathogen adapts to its environment, and identify novel ncRNAs that may contribute to virulence.

The next two chapters describe the analysis of multi-condition RNA-seq experiments in *Psa* grown *in vitro*. In this chapter, *in vitro* transcriptomes are compared to look for gene expression changes relevant to pathogenicity. These data are also compared to a recent *in planta* study, in which *Psa* transcriptomes were generated over an infection time course of kiwifruit plantlets.

3.1.1 Contributions

Peter McAtee provided *in planta* transcriptome data and R scripts for Pathview analysis. Matt Templeton provided feedback and guidance on this project, and additional genome annotations and expertise. I performed all RNA-seq data analysis.

3.2 Introduction

Pseudomonas syringae pv. *actinidiae* (*Psa*) is a gram-negative bacterial pathogen, which has emerged as a major threat to the kiwifruit industry during a pandemic outbreak in the past decade (Scortichini *et al.*, 2012, Tanner, 2015). *Psa* has been the focus of much study both due to its economic impact on the kiwifruit industry, and as an example of an emergent pathogen. It was first identified in Japan during the 1980s as the cause of kiwifruit canker disease, which is characterised by the development of necrotic lesions (Figure 3.1) and red exudate, leading to vine death or loss of fruit production (Serizawa *et al.*, 1989, Takikawa *et al.*, 1989).

Pseudomonas syringae is a species complex characterised by the rapid emergence and spread of pathogenic strains which cause disease in plants, most isolates of which tend to have highly specific host ranges (Mazzaglia *et al.*, 2012). Significant genomic diversity has been found within *P. syringae* due to large amounts of homologous recombination between strains, as well as the circulation of mobile genetic elements within the population. These factors appear to be major contributors to the rapid development and evolution of virulence in *P. syringae* plant pathogens (Monteil *et al.*, 2016).

Pandemic strains of *Psa* also share this genomic volatility, which has enabled the pathogen to increase in virulence and develop copper resistance (Colombi *et al.*, 2017). Comparative genomics studies of the structure of the *Psa* population identified several clades (also called biovars) that are independent clonal lineages with substantial between-clade diversity (Chapman *et al.*, 2012, Mazzaglia *et al.*, 2012, McCann *et al.*, 2013). Recent studies have determined that the New Zealand pandemic strain *Psa*-V (biovar 3) appears to have emerged from China (Butler *et al.*, 2013, McCann *et al.*, 2017), and found that a significant proportion of genome can be attributed to homologous recombination both within *Psa* clades (McCann *et al.*, 2013), as well as the acquisition of virulence factors from other *P. syringae* pathovars.

3.2.1 Pathogenesis of *Psa* infection

Psa is a gram-negative rod-shaped bacteria, that produces fluorescent siderophore compounds characteristic to many pseudomonads. Host plants may be inoculated with *Psa via* the environment, either by rainfall or splashing, or by plant material such as infected exudate or pollen. Bacteria can then enter the host plant through stomata, lenticels, wounds or flowers (Donati *et al.*, 2014, 2018, Spinelli *et al.*, 2011).

Within the host plant *Psa* multiplies in the apoplastic space and can cause chlorotic lesions and necrosis by secreting phytotoxic compounds such as syringomycin, coronatine and phaseolotoxin (Figure 3.1). During systemic infections, *Psa* enters the xylem in the plant vasculature, forming alginate-containing biofilms which block water transport, causing wilting (Bae *et al.*, 2015). Infiltration of the xylem then allows *Psa* to travel throughout the plant, eventually leading to host death.

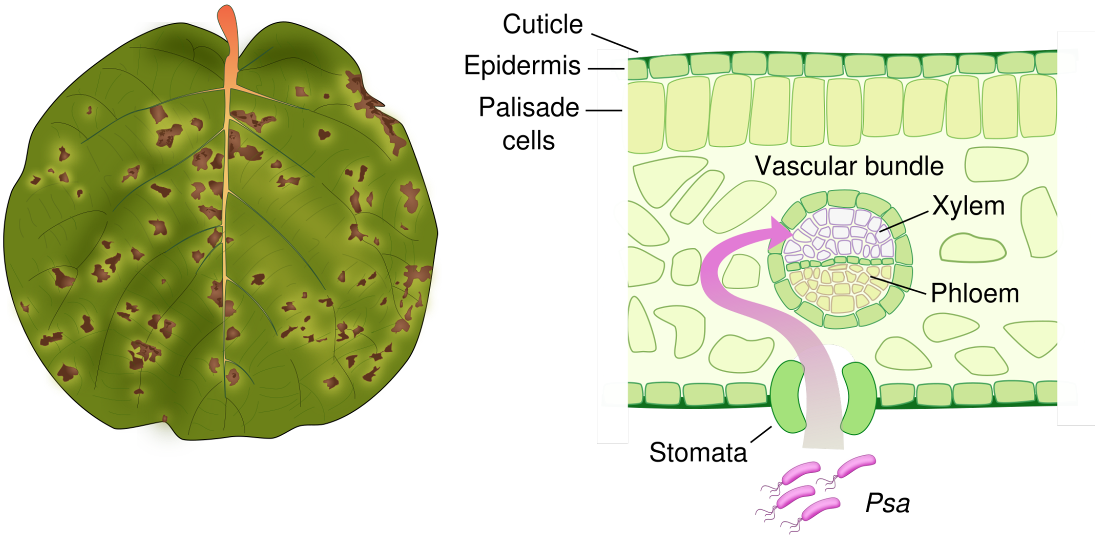


Figure 3.1: Diagram showing *Psa* infection of a kiwifruit leaf. **Left:** Kiwifruit leaf infected by *Psa*. Symptoms of the infection can be seen as brown patches of dead tissue caused by necrosis, as well as yellowing due to a loss of chlorophyll production in affected regions (chlorosis). **Right:** Potential avenue of *Psa* infection of a leaf. *Psa* enters through the stomata, and can reside in the apoplastic space where it triggers the plant necrotic response. *Psa* may also make its way to a vascular bundle and enter the xylem to cause systemic infection.

Pathogenic strains of *Pseudomonas syringae* utilise a type III secretion system (T3SS) to inject virulence factors into plant cells. These include effectors, proteins which interact and module host cells to promote pathogen survival. Like other plant pathogens, the T3SS locus of *P. syringae* includes *hrp* genes, a set of regulators, effectors and T3SS components that are activated in response to environmental stimuli (Lindgren, 1997).

Hrp genes are named after the ability of some *hrp*-gene products to produce a hypersensitive response (HR) in the host plant. The HR is basal plant immune response which is characterised by ion flux and the generation of destructive free radicals, that cause local necrosis which helps to restrict the growth or spread of pathogens (He, 1996). Virulence products produced by pathogens, such as effectors encoded by *hrp* genes in *P. syringae* (He, 1996, He *et al.*, 1993), can trigger the HR response when these products are recognised by a cognate resistance (R) gene in the plant (Keen, 1990). These gene-for-gene responses can be unique to specific interactions between pathogens and hosts (Heath, 2000). Certain T3SS proteins and secreted compounds such as coronatine secreted by plant pathogenic bacteria can also act to suppress the plant HR by interfering with recognition pathways (Nomura *et al.*, 2005).

3.2.2 Aims of this study

Like other bacterial pathogens, *Psa* experiences various levels of nutrient stress during infection and transmission. *Pseudomonas syringae* pathogens famously can participate in the water cycle, enabling their spread to a wide range of aqueous environments and plant hosts by precipitation or splash contamination (Hirano and Upper, 1990, Morris *et al.*, 2013, 2008).

Interestingly, *P. syringae* appears to have evolved complementary adaptions to nutrient stress that optimise its survival and spread both in the environment and within plant hosts. Factors that enable *P. syringae* to grow and spread in oligotrophic freshwater environments, such as metabolic versatility and the ability to sequester nutrients, are also associated with strain virulence and host range (Morris *et al.*, 2013).

For *Psa* growing *in planta*, the epiphytic stage of infection is likely to be nutrient poor with the exception of nectar (Donati *et al.*, 2018). Systemic infection within the xylem also requires adaption to a nutrient poor environment (Vinatzer, 2012) – the uptake and utilisation of sucrose as a carbon source reported in *Psa* may be an adaption to this. Bacteria in these environments are likely to face osmotic changes within different plant tissues, as well as oxidative stress from plant defence mechanisms (Jones *et al.*, 2007, Nomura *et al.*, 2005).

These abilities are made possible with well-tuned gene regulatory mechanisms, including the actions of regulatory ncRNAs which are poorly understood in *P. syringae.* For this study, *Psa* was grown *in vitro* in three culture mediums with different levels of nutrient availability, and transcriptomes generated for these mediums at different parts of the growth cycle. These *in vitro* transcriptomes were also compared with a recent *in planta* study by McAtee *et al.* (2018), in which *Psa* transcriptomes were generated over an infection time course of kiwifruit plantlets.

In this Chapter, the protein-coding genes are compared across these transcriptomes, to identify the broad changes in gene expression as *Psa* adapts to nutrient stress, and identify gene expression changes that may be relevant to survival and pathogenicity. This then provides a background for analysis of condition-specific ncRNA expression in Chapter 4, which describes the annotation, expression and conservation of candidate ncRNAs associated with each growth condition. The analysis in this chapter is important for identifying ncRNAs that may be relevant for pathogenicity, as we can assume that condition-specific ncRNAs expressed concurrently with a stress or pathogenicity response may be involved in its regulation.

3.3 Methods

3.3.1 *In vitro* RNA-seq

Cultures of *Pseudomonas syringae* pv. *actinidiae* isolate ICMP 18884 were grown in three different growth mediums at 20°C. Samples were grown in complete rich media (LB), with four replicates in log phase, and two in log/late log phase. In minimal media (M9 media supplemented with glucose), two samples were generated in log/late log phase, and four samples were generated under starvation conditions (0.1 M9 salts)

Library	Name	Growth medium	Growth phase	OD A_{600}	Ave. fragment size (bp)
2428_01	M9-1	Minimal	log/late log	1.6	306
2428_02	M9-2	Minimal	log/late log	1.6	308
2428_03	LB-3	Rich	late log	4.6	306
2428_04	LB-4	Rich	late log	4.6	309
2428_05	ST-5	Starved	NA	NA	309
2428_06	ST-6	Starved	NA	NA	315
2428_07	LB-7	Rich	log	1.1	314
2428_08	LB-8	Rich	log	1.1	315
2428_09	LB-9	Rich	log	1.6	319
2428_10	LBA-10	Rich	log	1.6	311
2428_11	ST-11	Starved	NA	NA	326
2428_12	ST-12	Starved	NA	NA	318

Table 3.1: Summary of sequencing libraries and growth conditions of the twelve *in vitro Psa* RNA-seq samples. Three growth mediums were used: Rich (LB media), Minimal (M9 media supplemented with glucose) and Starved (0.1 M9 salts). Growth phase was determined by measuring optical density/absorbance at 600 nm (OD A_{600}). Growth phase could not could not be reliably determined for starved samples due to a lack of bacterial growth in this medium. Average fragment size of RNA libraries ranged from 306–326 nt across the samples.

RNA was extracted using an Ambion Pure kit. Ribosomal rRNA depletion was performed by New Zealand Genomics Limited (NZGL) using an Illumina Ribo-Zero rRNA Removal Kit. All libraries for the project were stranded and sequenced on a single Hi-Seq cell.

3.3.2 Differential gene expression

Raw reads were processed with `trimmomatic` (Bolger *et al.*, 2014) to remove adaptors and regions of poor quality sequence. Transcripts for all *Psa* genes were quantified using `Kallisto` (Bray *et al.*, 2016) and count data analysed using `DESeq2` (Love *et al.*, 2014). The full data analysis pipeline is provided as a jupyter notebook in the supplemental materials (see Supplementary Materials).

3.3.3 Pathway analysis

`Pathview` (Luo and Brouwer, 2013) was used to visualise the overall effect of growth condition on gene expression. KEGG annotations for orthologous genes from *Pseudomonas syringae* pv. *tomato* DC3000 (*PSPTO*) were used for pathway analysis. Orthologous genes were annotated using a `blastp` (Altschul *et al.*, 1990) search of *Psa* proteins against *PSPTO* proteins, and taking the highest-scoring significant result for each *Psa* protein (E-value threshold 0.05). Pathways containing more than nine highly differentially expressed genes were visualised with `Pathview` (see Supplementary Materials).

For pathways with multiple genes mapping at the same node, the gene with the largest total change in gene expression across all media comparisons was used. Mapped *Psa* genes were renamed by their IYO gene number on the plots generated by `Pathview` using a custom Perl script. Log_2 fold-change for each comparison (rich vs starved, rich vs minimal, minimal vs starved) were plotted on each node as a heatmap, ranging from -8 (dark blue) to 8 (deep red).

3.3.4 Functional annotation

A set of comprehensive functional annotations was created for predicted protein-coding genes in the *Psa* genome and plasmid (NCBI accessions: CP011792.2 and CP011793.1) (Templeton *et al.*, 2015). Proteins were functionally annotated using the EggNOG web server (Huerta-Cepas *et al.*, 2017). Secondary metabolites were predicted using the anti-SMASH web server(Medema *et al.*, 2011). RAST (Aziz *et al.*, 2008) annotations, and a set of custom annotations were also provided by Matt Templeton.

3.3.5 Comparison to *in planta* data

RNA-seq data from McAtee *et al.* (2018) was obtained from the NCBI Sequence Read Archive (SRA) project SRP148711 (Table 3.2). Raw reads from samples infected with *Psa* at different time-points were downloaded and processed using the same analysis pipeline as the *in vitro* RNA-seq.

Time-point	SRA Accession		
1.5hrs	SRR7204724	SRR7204725	SRR7204726
3hrs	SRR7204727	SRR7204728	SRR7204729
6hrs	SRR7204688	SRR7204694	SRR7204707
12hr	SRR7204687	SRR7204689	SRR7204708
24hrs	SRR7204690	SRR7204691	SRR7204693
48hrs	SRR7204692	SRR7204695	SRR7204706
72hrs	SRR7204685	SRR7204686	SRR7204711
96hrs	SRR7204709	SRR7204710	SRR7204712
120hrs	SRR7204713	SRR7204714	SRR7204715

Table 3.2: Time-points and SRA accessions for *in planta* RNA-seq samples

3.4 Results and Discussion

3.4.1 RNA-seq quality control

***In vitro* samples**

Approximately 60% of all *in vitro* sequencing reads (approximately 10 million counts per sample) were successfully mapped to protein coding genes (Table 3.3). During initial quality control, it was found that ribosomal RNA depletion for Sample 10 had failed. This sample was not included in further analyses.

Principal component analysis (Figure 3.2) showed that *in vitro* samples clustered by media and growth phase, with the majority of variance in PC1 due to growth medium, with some variance between the two distinct growth phases in the rich media samples in PC2. Starved samples showed some within-group variance, which may be due to differences in growth phase.

Hierarchical clustering of sample-to-sample distances (Figure 3.3) showed strong within-group similarity for media type and growth phase, with starved samples showing the same within-group variation observed in the principal component analysis (Figure 3.2. Samples grouped primarily by growth phase instead of media type. Rich media samples formed two separate clusters based on growth phase. Starved samples were closest to rich media samples in log/late log phase, which may be due to similarities in growth limitation in these samples. Minimal media samples were most similar to the rich media samples in log phase.

Sample	Sample name	Total counts	% Reads mapped to genes
Sample_1	M9-1	10560672	60.3
Sample_2	M9-2	10248103	58.9
Sample_3	LB-3	8416904	49.1
Sample_4	LB-4	8707835	49.8
Sample_5	ST-5	9370257	52.8
Sample_6	ST-6	9965784	58.7
Sample_7	LB-7	10435416	64.0
Sample_8	LB-8	10886018	66.8
Sample_9	LB-9	11590564	67.4
Sample_10	LBA-10	8703562	51.6
Sample_11	ST-11	10249949	57.7
Sample_12	ST-12	9256567	58.6

Table 3.3: Number of RNA-seq transcripts for *in vitro* samples. Growth medium and phase associated with sample names are detailed in Table 3.1). The overall proportion of reads mapping to *Psa* transcripts for each sample range from 49.1–67.4%

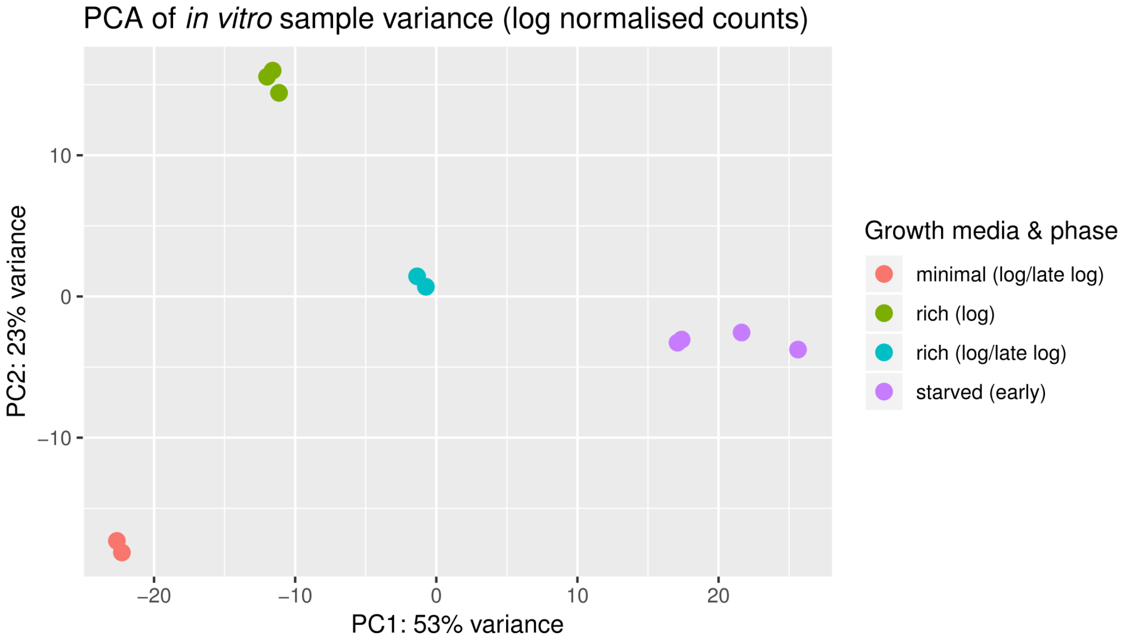


Figure 3.2: Principal component analysis of gene expression data ($\log_2$ normalised counts) from *in vitro* samples. Sample-sample distances are plotted along the first two principal components (PC1 and PC2). Points represent individual samples, which are coloured by experiment group (growth medium and growth phase). Samples clustered by experiment type along PC1 and PC2, showing similarity between replicates. Sample groups were distinct and well separated across both principal components.

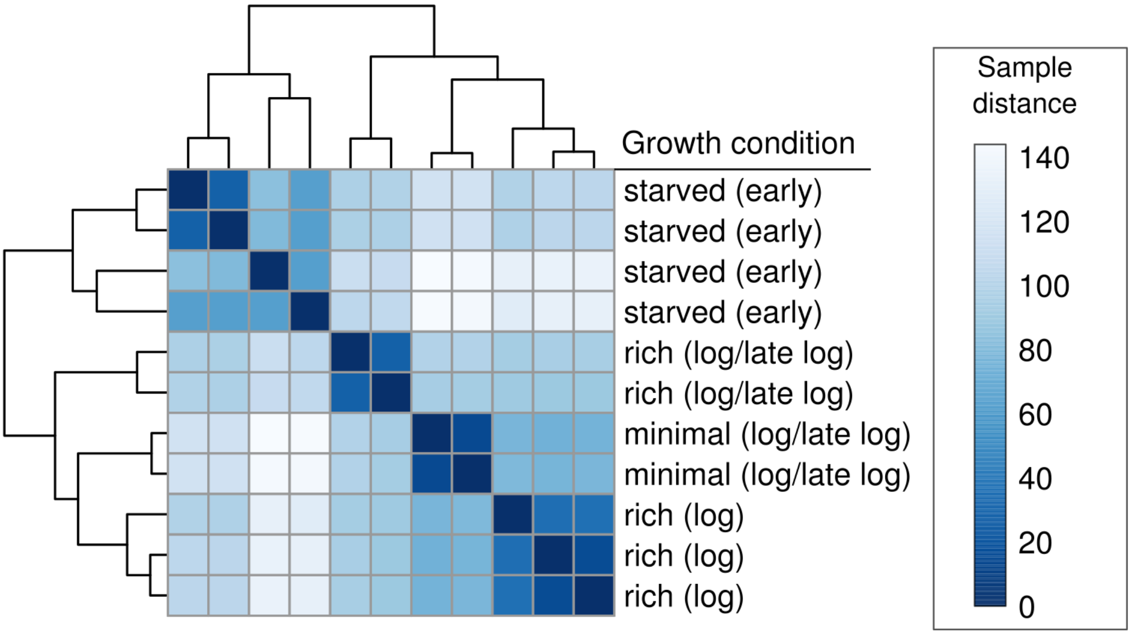


Figure 3.3: Heatmap showing sample-sample distances between *in vitro* samples after hierarchical clustering, based on DESeq2 normalised counts. Replicate samples are clustered into groups showing high similarity to each other. Sample groups are then clustered by growth phase and media type.

An MA-plot visualising the $\log_2$ fold change (l_2FC) compared to the mean of normalised counts for all genes in the *in vitro* samples showed that that few genes had low expression overall, and that the majority of genes were differentially expressed across the data-set. Overall, 2592/5809 genes had a l_2FC with $p < 0.01$, and 2649/5809 genes had a l_2FC with $p < 0.0001$ (Figure 3.4).

Histograms showing the frequency distribution of Benjamini-Hochberg adjusted P values (P_{adj}) for $\log_2$ fold changes for pairwise comparisons of sample types showed 500-1000 differentially expressed genes (DEGs) with a $P_{adj} < 0.01$ for each comparison (Figure 3.5, Figure 3.6). Comparisons between growth phases showed many genes were highly differentially expressed between rich media samples – rich media samples were also more different to each other than to the minimal media sample. Rich media log phase samples were most similar to minimal media log/late log phase samples.

Comparisons between media types showed rich media samples overall were most similar to minimal media samples, with only 522 DEGs $P_{adj} < 0.01$. Approximately 1500 genes were highly differentially expressed ($P_{adj} < 0.0001$) in comparisons between the starved samples and the other media types, whereas only 257 genes were differentially expressed between rich and minimal media samples (Table 3.4).

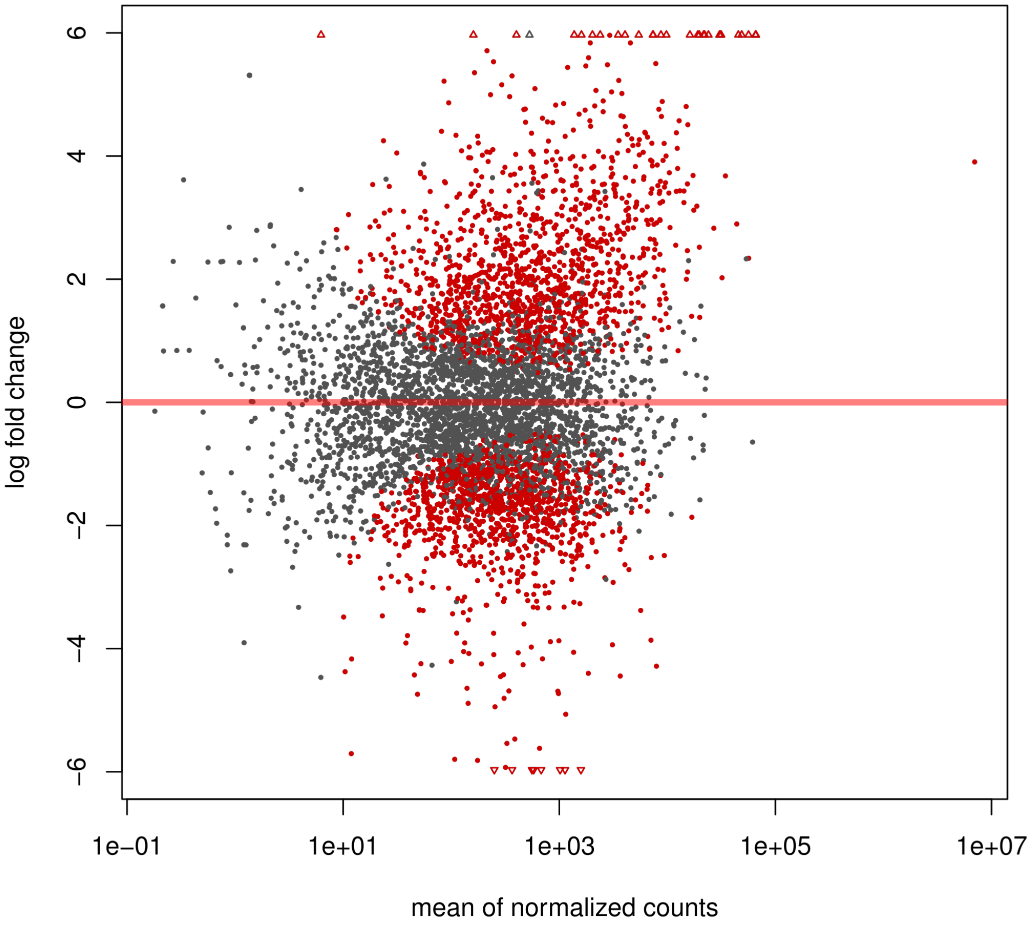
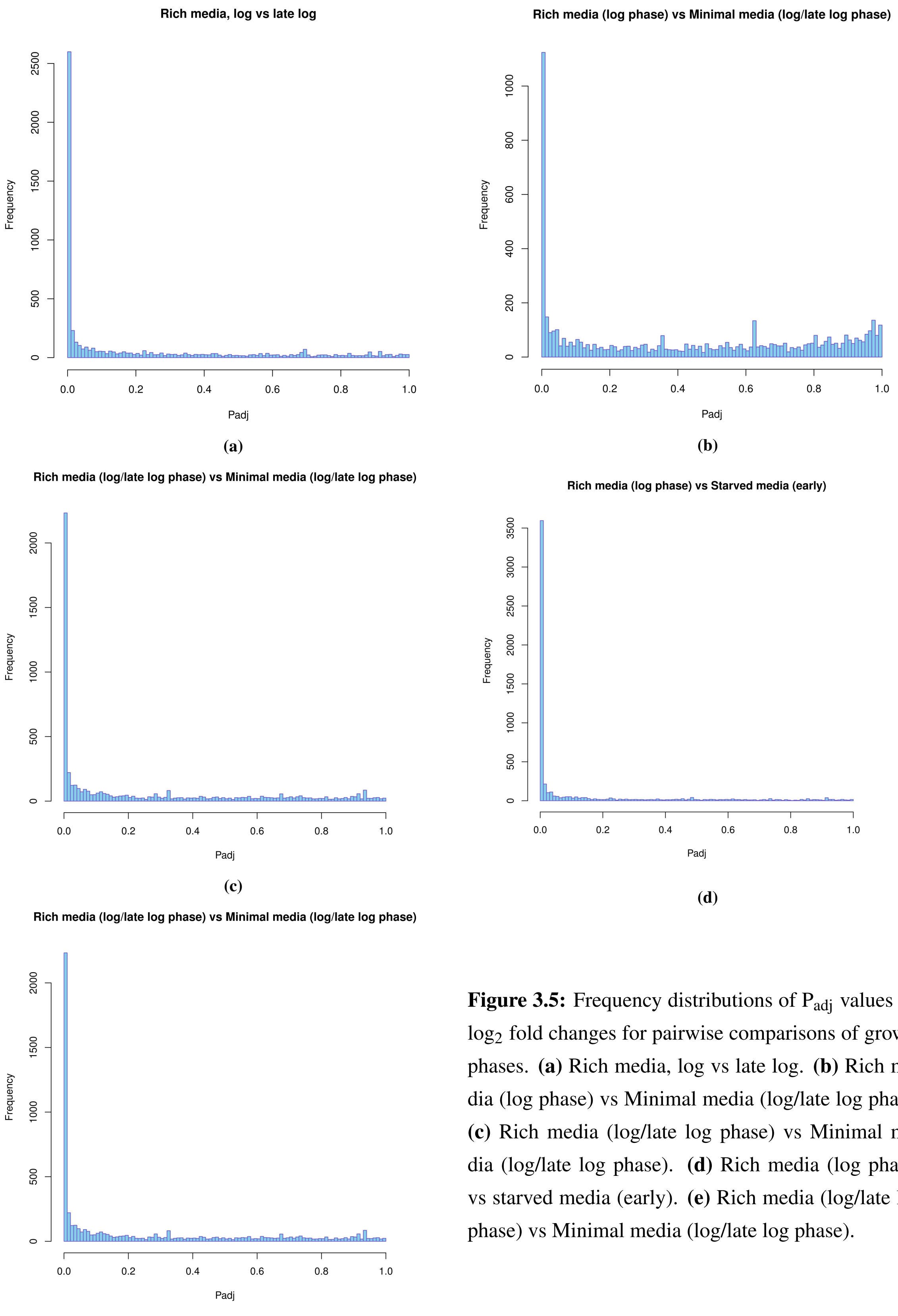


Figure 3.4: MA-plot showing the $\log_2$ fold changes (l_2FC) compared to the mean of normalised counts for all genes in the *in vitro* samples. Genes with a highly significant l_2FC ($p < 0.0001$) are highlighted in red. Plot generated using `DESeq2`.

Sample type	Starved	Minimal	Rich (log)
Starved	-	-	-
Minimal	1561	-	-
Rich (all)	1572	257	-
Rich (log)	2694	651	-
Rich (log/late log)	1878	1368	1590

Table 3.4: Numbers of highly differentially expressed genes for pairwise comparisons of *in vitro* experiment factors ($P_{adj} < 0.0001$)



(a)

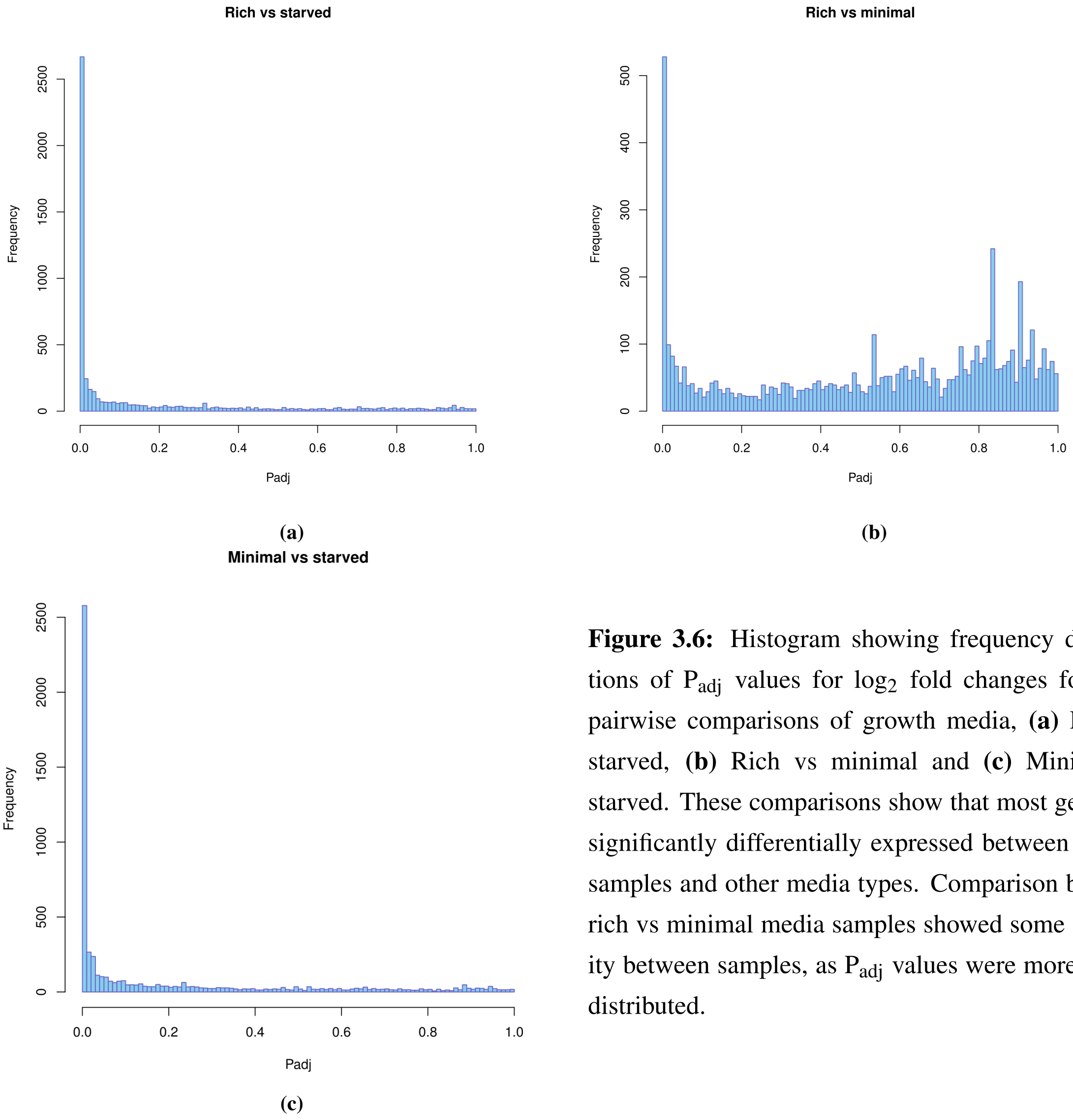
(b)

(c)

(d)

(e)

Figure 3.5: Frequency distributions of P_{adj} values for $\log_2$ fold changes for pairwise comparisons of growth phases. **(a)** Rich media, log vs late log. **(b)** Rich media (log phase) vs Minimal media (log/late log phase) **(c)** Rich media (log/late log phase) vs Minimal media (log/late log phase). **(d)** Rich media (log phase) vs starved media (early). **(e)** Rich media (log/late log phase) vs Minimal media (log/late log phase).



(a)

(b)

(c)

Figure 3.6: Histogram showing frequency distributions of P_{adj} values for $\log_2$ fold changes for three pairwise comparisons of growth media, **(a)** Rich vs starved, **(b)** Rich vs minimal and **(c)** Minimal vs starved. These comparisons show that most genes are significantly differentially expressed between starved samples and other media types. Comparison between rich vs minimal media samples showed some similarity between samples, as P_{adj} values were more evenly distributed.

In planta samples

PCA of all samples showed that the largest contribution to PC1 variance was experiment type (*in planta* and *in vitro*). The top 500 genes with the most variance across the data set (PC1) were inspected to look for major differences in gene expression between the *in planta* and *in vitro* samples. Most of the genes that contribute most to PC1 were effectors and virulence factors that are highly up-regulated in planta. PC2 showed that rich media samples are the most similar to *in planta* samples based on gene expression, and clustered near early infection time-points (1.5hrs, 3hrs, 12hrs and 24hrs post-infection).

Sample	Sample name	Reads	Total counts	% Reads mapped to genes
1.5hr	16	38415111	179234	0.47
1.5hr	17	35889853	175850	0.49
1.5hr	20	37962490	129053	0.34
3hr	25	39563552	190023	0.48
3hr	26	33576040	233502	0.70
3hr	27	40015335	147195	0.37
6hr	28	38141276	259132	0.68
6hr	29	36345675	163657	0.45
6hr	30	39437965	156166	0.40
12hr	34	39065761	349361	0.90
12hr	37	42099962	231399	0.55
12hr	38	36097446	174620	0.48
24hr	3	39371740	99075	0.25
24hr	41	37950153	222691	0.59
24hr	42	35325178	121312	0.34
48hr	10	34148479	93320	0.27
48hr	43	35609939	90506	0.25
48hr	44	38929993	96337	0.25
72hr	12	35007764	61033	0.17
72hr	6	37514121	93645	0.25
72hr	8	38572996	133459	0.35
96hr	55	32227055	97762	0.30
96hr	56	31614369	39471	0.12
96hr	57	28591692	103980	0.36
120hr	49	31642015	262133	0.83
120hr	50	35685140	240136	0.67
120hr	51	31107467	170153	0.55

Table 3.5: Total estimated counts of RNA-seq reads mapping to *Psa* transcripts from *in planta* samples generated by McAtee *et al.* (2018). The overall proportion of reads mapping to *Psa* transcripts for each sample range from 0.12–0.9%

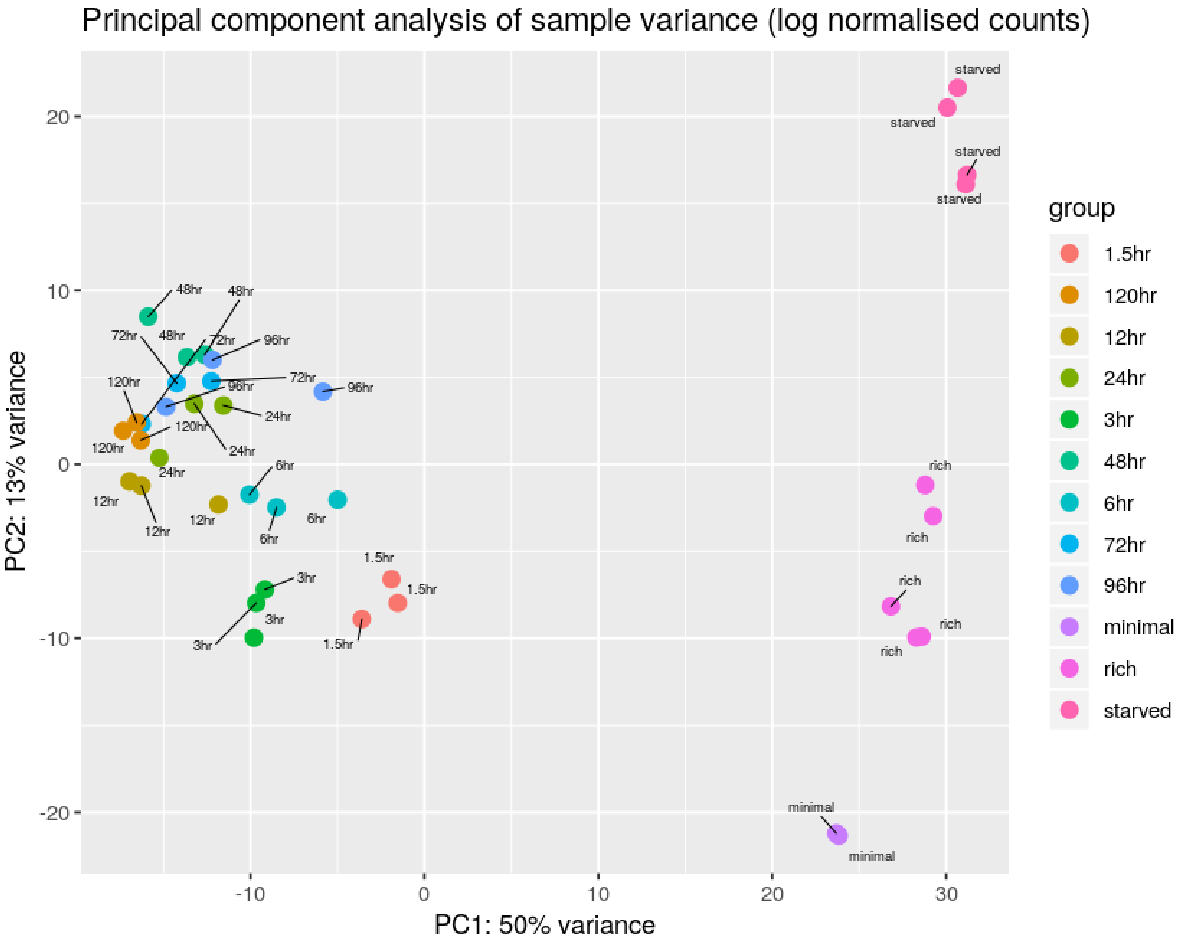


Figure 3.7: Principal component analysis of gene expression data ($\log_2$ normalised counts) from *in vitro* and *in planta* samples. Sample-sample distances are plotted along the first two principal components (PC1 and PC2). Points represent individual samples, which are coloured by experiment group (***In vitro*:** minimal, rich and starved growth mediums. ***In planta*:** Time-points after *Psa* inoculation on kiwifruit plantlets). The largest variance is between *in vitro* and *in planta* samples, which are separated along PC1, indicating that *in vitro* samples do not mimic *in planta* gene expression. Samples clustered by experiment type along PC1 and PC2, showing similarity between replicates. Sample groups were highly distinct among *in vitro* samples, which were separated along PC2. *In planta* samples were more closely grouped along both PC1 and PC2, with early time-points (1.5hrs-12hrs) showing more variance between each other as well as later time-points. Late time-point *in planta* samples were more closely grouped along PC1 and PC2. *In vitro* rich media samples were similar to early *in planta* time-points along PC2, however, other *in vitro* samples groups were well separated from each other as well as *in planta* samples across both principal components.

3.4.2 Differential gene expression analysis

Within *in vitro* comparisons

Functional annotation of the top 250 DEGs from comparisons between in vitro samples grown in different growth mediums showed the biggest transcriptional changes in genes involved in iron metabolism. Pathway analysis found that most significantly differentially expressed genes were involved in energy metabolism, secondary metabolite biosynthesis and virulence factors (Table 3.6).

Iron homeostasis Almost all of the most highly differentially expressed genes in the *in vitro* samples were involved with or dependent on the acquisition and storage of iron. Iron is an essential micro-nutrient for all organisms and is a co-factor in many enzymes, such as those involved in the electron transport chain. For pathogenic bacteria bio-available iron is a major growth limiting factor: most iron in the environment is present as insoluble ferric (Fe^{3+}) iron, and within the host iron may be scarce due to the competing nutritional requirements of the host, or as a defence mechanism (Ratledge and Dover, 2000).

Maintaining a suitable balance of iron compounds is important to all cells – too little iron affects energy production and metabolic capabilities, however free iron can catalyse the production of free radicals *via* Haber-Weiss/Fenton chemistry (Fones and Preston, 2013). Free ferric iron can react with hydrogen peroxide and superoxide to produce highly reactive hydroxyl ions, which can go on to cause DNA damage by attacking nucleotides (Kehrer, 2000). The sequestration of iron is both important to provide nutrients to the bacteria, and as a virulence factor by depriving host cells of iron. Many pseudomonads produce peptide-based siderophore compounds such as pyoverdine, which have higher affinity for ferric than ferrous (Fe^{2+}) iron, and are secreted to chelate environmental ferric iron (Cézard *et al.*, 2015). Iron is released from siderophores by a change in affinity by the reduction of ferric iron to ferrous iron. Ferrous iron may then be safely stored as ferric iron by incorporating it into protein-haem complexes such as bacterioferritin (Ratledge and Dover, 2000).

Bacterioferritin genes were up-regulated in rich media, indicating the stockpiling of abundant free iron. Adaption to iron restriction could be seen in starved and minimal media samples. Siderophore biosynthesis was up-regulated in minimal media (Tables 3.7, 3.8, 3.9), as were genes involved in iron uptake such as the FagA iron transporter and TonB-dependent outer membrane receptors, which in *Pseudomonas* typically function in siderophore transport and siderophore-induced signal transduction and regulation (Llamas *et al.*, 2014).

# DEGs	KEGG pathway number and description
278	pst01100 - Metabolic pathways
117	pst01110 - Biosynthesis of secondary metabolites
101	pst01120 - Microbial metabolism in diverse environments
82	pst01130 - Biosynthesis of antibiotics
72	pst02010 - ABC transporters
71	pst02020 - Two-component system
58	pst01230 - Biosynthesis of amino acids
42	pst01200 - Carbon metabolism
32	pst02024 - Quorum sensing
27	pst02030 - Bacterial chemotaxis
25	pst00230 - Purine metabolism
22	pst02025 - Biofilm formation
20	pst00270 - Cysteine and methionine metabolism
20	pst00620 - Pyruvate metabolism
20	pst00630 - Glyoxylate and dicarboxylate metabolism
19	pst00190 - Oxidative phosphorylation
18	pst00280 - Valine, leucine and isoleucine degradation
17	pst03070 - Bacterial secretion system
16	pst00260 - Glycine, serine and threonine metabolism
16	pst00650 - Butanoate metabolism
16	pst00860 - Porphyrin and chlorophyll metabolism
16	pst02040 - Flagellar assembly
15	pst00790 - Folate biosynthesis
15	pst00920 - Sulfur metabolism
14	pst00400 - Phenylalanine, tyrosine and tryptophan biosynthesis
13	pst00310 - Lysine degradation
13	pst01212 - Fatty acid metabolism
12	pst00250 - Alanine, aspartate and glutamate metabolism
12	pst00362 - Benzoate degradation
12	pst01210 - 2-Oxocarboxylic acid metabolism
12	pst03440 - Homologous recombination
11	pst00010 - Glycolysis / Gluconeogenesis
11	pst00071 - Fatty acid degradation
11	pst00220 - Arginine biosynthesis
10	pst00240 - Pyrimidine metabolism
10	pst00330 - Arginine and proline metabolism
10	pst00640 - Propanoate metabolism
10	pst00970 - Aminoacyl-tRNA biosynthesis

Table 3.6: KEGG pathways containing 10 or more significantly DEGs from comparisons of *in vitro* trasncriptomes. This is estimated from orthologous genes and pathways from *P. syringae* pv. *tomato* DC3000

The largest difference between the starved and minimal media samples was the shut-down of siderophore biosynthesis. Although iron is an essential nutritional requirement, siderophores are complex molecules which require large amounts of nitrogen, carbon and ATP to synthesise (Sexton and Schuster, 2017). Starved *Psa* down-regulated genes involved in siderophore biosynthesis and uptake.

Gene	Predicted function	l_2FC RvS	l_2FC RvM	l_2FC MvS	p_{adj} RvM
IYO_010820	Sigma factor PvdS	0.26	6.81	-6.55	1.81E-07
IYO_010830	PvdL Pyoverdine chromophore precursor synthetase	1.28	5.56	-4.29	8.67E-12
IYO_010840	Antibiotic synthesis protein MbtH	-0.49	4.21	-4.70	1.36E-19
IYO_010845	Metal ABC transporter substrate-binding protein	1.27	4.65	-3.39	4.49E-30
IYO_010850	ABC transporter in pyoverdine gene cluster	1.84	5.26	-3.42	4.09E-24
IYO_010855	ABC transporter in pyoverdine gene cluster	1.27	4.46	-3.2	7.84E-16
IYO_010880	Putative iron-regulated membrane protein	0.93	5.41	-4.48	1.69E-14
IYO_010885	SyrP	0.94	6.58	-5.64	4.52E-20
IYO_010890	Pyoverdine sidechain peptide synthetase	1.28	6.35	-5.07	2.16E-14
IYO_010895	PvdI Pyoverdine sidechain non-ribosomal peptide synthetase	1.21	5.15	-3.94	1.08E-12
IYO_010900	PvdJ Pyoverdine sidechain non-ribosomal peptide synthetase	1.83	6.28	-4.45	3.84E-27
IYO_010905	PvdD Pyoverdine sidechain non-ribosomal peptide synthetase	1.84	5.22	-3.38	1.87E-15
IYO_011245	TonB-dependent receptor	-0.29	5.89	-6.18	1.21E-29
IYO_011875	TonB-dependent receptor	0.88	5.33	-4.46	1.96E-35
IYO_012650	TonB-dependent outer membrane siderophore receptor precursor	0.47	5.21	-4.73	4.69E-29
IYO_012660	PvdE Pyoverdine ABC export system	1.74	6.71	-4.97	2.05E-18
IYO_012665	PvdO Pyoverdine responsive serine/threonine kinase	1.33	6.3	-4.97	1.9E-08
IYO_012670	PvdN Pyoverdine biosynthesis protein	0.72	6.29	-5.56	1.76E-15
IYO_012675	PvdM Putative pyoverdine biosynthesis dipeptidase	0.47	6.55	-6.08	1.9E-05
IYO_012680	PvdJ/PvdD/PvdP-like protein	-0.51	5.56	-6.07	1.86E-19
IYO_012685	Outer membrane pyoverdine efflux protein	0.95	5.17	-4.22	1.6E-48
IYO_012690	Pyoverdine efflux carrier and ATP binding protein	0.44	5.1	-4.66	1.25E-34

Table 3.7: DEGs up-regulated in *Psa* grown in minimal media annotated as being involved in pyoverdine biosynthesis, regulation, and iron transport. l $_2$FC is shown for media comparisons: RvS (Rich vs Starved), RvM (Rich vs Minimal) and MvS (Minimal vs Starved). P_{adj} is shown for Rich vs Minimal media comparisons.

Gene	Predicted function	l_2FC RvS	l_2FC RvM	l_2FC MvS	p_{adj} RvM
IYO_013860	Salicyl-AMP ligase	-0.89	2.03	-2.92	1.44E-08
IYO_013865	Yersiniabactin synthesis enzyme YbtT	0.62	3.12	-2.5	3.15E-13
IYO_013870	Iron aquisition yersiniabactin synthesis enzyme Irp3	-0.55	2.47	-3.02	1.69E-14
IYO_013450	TonB-dependent ferric achromobactin receptor protein	0.3	3.33	-3.03	6.59E-05
IYO_013875	Iron aquisition yersiniabactin synthesis enzyme Irp1	0.89	2.66	-1.77	2.37E-07
IYO_013885	Iron aquisition yersiniabactin synthesis enzyme Irp2	0.05	2.67	-2.62	8.24E-21
IYO_013900	Iron aquisition outermembrane yersiniabactin receptor	-0.58	5.24	-5.82	3.05E-78

Table 3.8: DEGs up-regulated in *Psa* grown in minimal media annotated as being involved in yersiniabactin biosynthesis, regulation, and iron transport. l $_2$FC is shown for media comparisons: RvS (Rich vs Starved), RvM (Rich vs Minimal) and MvS (Minimal vs Starved). P_{adj} is shown for Rich vs Minimal media comparisons.

Gene	Predicted function	l_2FC RvS	l_2FC RvM	l_2FC MvS	p_{adj} RvM
IYO_013460	Achromobactin biosynthesis protein AcsD	-0.04	2.3	-2.34	9.57E-06
IYO_013465	Achromobactin biosynthesis protein AcsE	-0.24	2.62	-2.85	4.88E-17
IYO_013475	Achromobactin biosynthesis protein AcsC	-0.37	2.44	-2.81	9.02E-20
IYO_013480	Achromobactin biosynthesis protein AcsB	-0.2	2.01	-2.21	1.39E-07

Table 3.9: DEGs up-regulated in *Psa* grown in minimal media annotated as being involved in achromobactin biosynthesis, regulation, and iron transport. l $_2$FC is shown for media comparisons: RvS (Rich vs Starved), RvM (Rich vs Minimal) and MvS (Minimal vs Starved). P_{adj} is shown for Rich vs Minimla media comparisons.

Nutritional stress *Psa* grown in minimal media and under starvation showed large adaptions to nutrient availability. Genes involved in oxidative phosphorylation were down-regulated and genes involved in the TCA cycle were highly up-regulated in both minimal and starved *Psa* compared to rich media, indicating the induction of carbon metabolism via the Entner-Doudoroff pathway by which pseudomonads utilise carbon sources (Lessie and Phibbs, 1984).

Glucose metabolism was up-regulated in minimal media compared to other samples (Table 3.10, where glucose was the main carbon source in the media. However *hexR*, which represses glucose metabolism and levansucrase secretion in *P. syringae* (Mehmood *et al.*, 2015), was also up-regulated, indicating active repression of phosphogluconate metabolism.

This may be required for redox balance, as HexR and glucose metabolism are in part controlled by the response to oxidative stress, as well as the intracellular abundance of 2-keto-3-deoxy-6-phosphogluconate (Daddaoua *et al.*, 2009).

Starved *Psa* up-regulated components of the beta-ketoadipate pathway, which converts cyclic aromatic amino acids to components of the citric acid cycle. One of the most highly expressed genes included phenylalanine hydroxylase (*pheA*), an uncommon step of the pathway present in pseudomonads which converts phenylalanine to tyrosine.

Many hydroxylases require iron as a cofactor; however, this particular reaction utilises an organic compound, pterin, as a cofactor. Pterin carbinolamine dehydratase, which recycles this pterin cofactor, was also highly up-regulated in starvation conditions. Previous transcriptomic studies of *P. syringae* pv. *syringae* B728a infection in *Arabidopsis* found phenylalanine degradation to be highly up-regulated in epiphytic populations, which may be a pre-emptive response for degrading phenylalanine-based plant defence molecules in the apoplast (Yu *et al.*, 2013).

Gene	Predicted function	l_2FC	l_2FC	l_2FC
		RvS	RvM	MvS
IYO_006365	Sugar ABC transporter substrate-binding protein	1.5	**4.37**	-2.88
IYO_006370	Glucose ABC transport system, inner membrane component 1	1.13	**4.4**	-3.27
IYO_006375	Glucose ABC transport system, inner membrane component 2	0.88	**4.13**	**-3.25**
IYO_006380	Glucose ABC transporter, ATP-binding subunit	0.71	**3.63**	**-2.92**
IYO_006390	Aldose epimerase	-0.65	**3.22**	**-3.87**
IYO_006395	Phosphogluconate repressor HexR	0.07	**1.33**	**-1.26**
IYO_006400	Glucose-6-phosphate dehydrogenase	-0.27	**1.35**	**-1.62**
IYO_006405	6-phosphogluconolactonase	0.4	**1.95**	**-1.55**
IYO_006410	Keto-deoxy-phosphogluconate aldolase	0.3	**2.31**	**-2.01**

Table 3.10: DEGs up-regulated in *Psa* grown in minimal media annotated as being involved in glucose metabolism, regulation, and transport. l_2FC is shown for media comparisons: RvS (Rich vs Starved), RvM (Rich vs Minimal) and MvS (Minimal vs Starved). Fold-changes highlighted in bold are highly significant ($p < 0.0001$).

PheA was down-regulated in minimal media, as were other components of amino acid degradation. This may be because these pathways are antagonistic to siderophore biosynthesis – PheA has been found have multiple roles in phenylalanine metabolism, including the degradation of chorismate, which is a key component of pyoverdine siderophores (Cézard *et al.*, 2015, Molina-Henares *et al.*, 2009).

Other enzymes involved in secondary metabolic pathways which generate feed-stocks for the TCA cycle, such as enzymes involved in the glyoxylate cycle, were also up-regulated under starvation conditions. This may be due to starved *Psa* utilising peptides from proteins or pyoverdine, or to utilise amino acids in the apoplastic fluid.

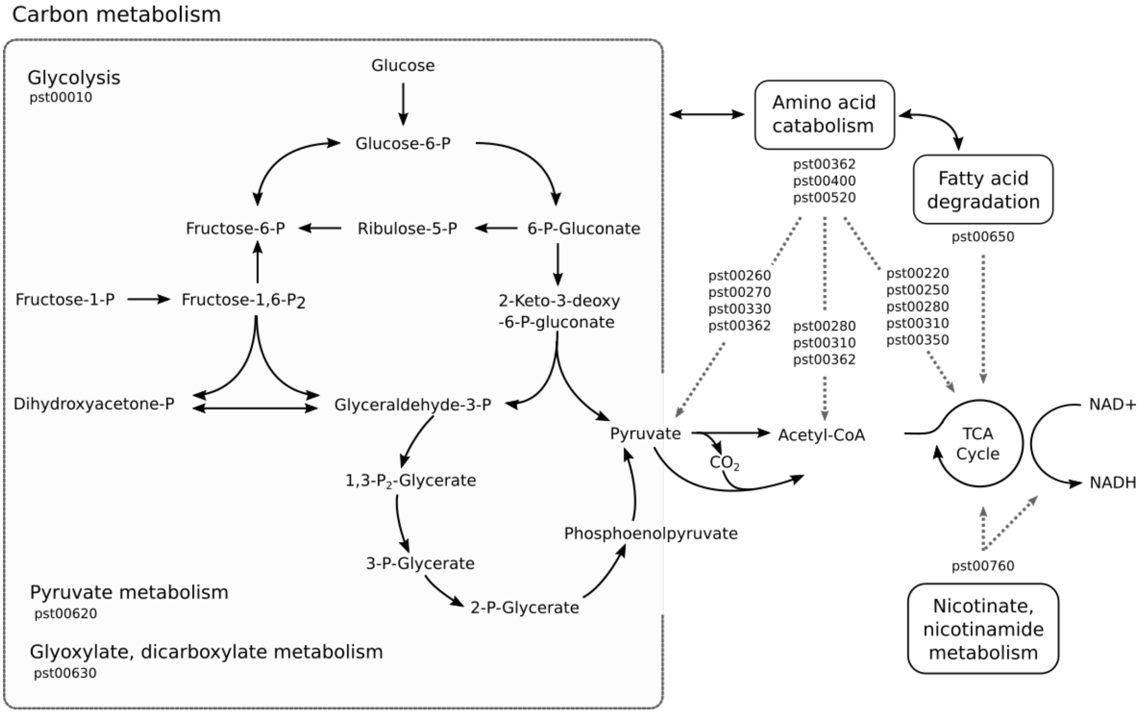


Figure 3.8: Diagram showing how differentially expressed KEGG metabolic pathways relate to energy metabolism via the TCA cycle in *Pseudomonas*. Major pathways feeding into the TCA cycle (carbon, amino acid and fatty acid metabolism) are shown contributing carbon feedstocks via pyruvate, acetyl-CoA or as TCA cycle intermediaries. Within carbon metabolism, the Pseudomonas Entner-Doudouroff pathway of hexose sugar metabolism is shown, based on Figure 1 by Morris *et al.* (1995). KEGG pathways containing genes that were significantly differentially expressed in the *in vitro Psa* RNA-seq are shown as ID numbers (full pathways can be found in Supplementary Materials).

Redox Genes involved in balancing reactive oxygen species were also significantly differentially expressed between sample groups. *P. syringae* and other plant pathogens experience redox stress *in planta* as ROS-generating molecules such as H_2O_2 are used as a plant defence mechanism (Jones *et al.*, 2007). Superoxides may also be generated during the extraction of iron from pyoverdine, or generated by reactions with imported free ferric iron. Many genes involved in the electron transport chain also contain iron-sulfur clusters or use iron as cofactors (Tuanyok *et al.*, 2005). The down-regulation of genes involved in oxidative phosphorylation, in addition to minimising the use of iron, may also protect the iron-sulphur clusters present in these proteins from oxidative damage (Kawakami *et al.*, 2010).

Several steps of Entner-Doudouroff metabolic pathways generate NADP$^+$, which can be reduced by ferrodoxins to NADPH, an important co-factor in the oxidative stress response. Many pseudomonads can redirect metabolic pathways towards NADPH production (Kivisaar, 2018), which has been found to promote resistance to oxidative stress in *Pseudomonas putida* (Chavarría *et al.*, 2013). This could be seen in starved *Psa*, which also up-regulated genes involved in NADP$^+$ generation, as well as a ferrodoxin NADP$^+$-reductase. This may be anticipatory of redox stress from plant defence, or to deal with endogenous stress from oxidative degradation of aromatic compounds (Kivisaar, 2018).

The largest fold change between *in vitro* media types were compensatory differences in the gene expression of superoxide dismutase during iron starvation (minimal and starved media), in that those using iron as a co-factor were down-regulated and those using manganese were up-regulated, which has previously been observed in *P. aeruginosa* (Vasil and Ochsner, 1999).

Virulence factors *Psa* grown *in vitro* did not express plant-induced *hrp* virulence factors such as the T3SS effectors, which are a major component of the *Psa* transcriptome *in planta* (McAtee *et al.*, 2018) (Table 3.12). However, in *Psa* grown in minimal media and in starvation conditions, some virulence factors were expressed that may be induced as part of stress responses.

Pseudomonas syringae is able to produce alginate, a polysaccharide composed of mannuronic and guluronic acid. Genes involved in alginate production were slightly up-regulated in minimal and starved compared to rich media, with the most expression in minimal media (Table 3.11).

While alginate biosynthesis was not significantly upregulated in starved *Psa* compared to minimal media, the alginate regulatory components *algU*, *algT* and *algZ* were more highly expressed in starvation conditions, as were several predicted alginate lyases, which are thought to play a role in restructuring alginate polymers (Preston *et al.*, 2000).

These factors were highly expressed independently of the alginate production operon, and have been found to play a role in stress tolerance and virulence and in *P. syringae* (Baynham *et al.*, 1999, Keith and Bender, 1999, Schenk *et al.*, 2008, Schurr *et al.*, 1996). The carbon storage regulator *csrA* was highly up-regulated in starved *Psa*, which in *P. syringae* pv. *tomato* DC3000 can control virulence expression, including the repression of alginate production (Ferreiro *et al.*, 2018).

Gene	Predicted function	l_2FC RvS	l_2FC RvM	l_2FC MvS
IYO_006020	Alginate o-acetyltransferase AlgF	-0.01	**1.05**	**-1.05**
IYO_006035	AlgL Alginate lyase precursor	**1.15**	**1.48**	-0.33
IYO_006040	Alginate O-acetyltransferase AlgX	-0.1	**1.2**	**-1.3**
IYO_006045	Poly(beta-D-mannuronate) C5 epimerase	0.7	**1.65**	**-0.94**
IYO_006050	alginate regulatory protein AlgE	0.67	**1.98**	**-1.31**
IYO_006055	Alginate biosynthesis protein AlgK precursor	**0.99**	**1.86**	-0.88
IYO_006060	Alginate biosynthesis protein Alg44	0.59	**1.73**	**-1.14**
IYO_006065	Glycosyl transferase Alginate biosynthesis protein Alg8	1.01	**2.96**	**-1.95**
IYO_006070	GDP-mannose 6-dehydrogenase	**1.41**	**3.2**	**-1.79**
IYO_021520	RNA polymerase sigma-H factor AlgU	**1.68**	0.32	**1.35**
IYO_021530	RNA polymerase sigma-H factor AlgT	**1.25**	0.03	1.22
IYO_000530	Alginate biosynthesis protein AlgZ/FimS	**2.84**	-0.7	**3.53**
IYO_001890	Alginate biosynthesis transcriptional regulator AlgB	**1.20**	-0.30	**1.50**

Table 3.11: DEGs up-regulated in *Psa* annotated as being involved in alginate biosynthesis. Log_2 fold-change is shown for each gene in for each pairwise media comparison. $1\ {}_2$FC is shown for media comparisons: RvS (Rich vs Starved), RvM (Rich vs Minimal) and MvS (Minimal vs Starved). Fold-changes highlighted in bold are significant ($p < 0.0001$).

Starved *Psa* up-regulated virulence factors involved in survival in ephiphytic environments, such as motility, chemotaxis and adhesion (Lindow *et al.*, 1993). Chemotaxis and flagella biosynthesis were up-regulated in starvation conditions, as were components of the type III secretion system, however the respective effectors were down-regulated, suggesting this may be involved with the transport of the flagellar components (Diepold and Armitage, 2015).

As well as the characteristic type III secretion system, *Psa* genomes contain components of type I, II, IV and VI secretion systems (Marcelletti *et al.*, 2011). Genes annotated as components of type VI, II and IV secretion systems were all down-regulated in starved *Psa*. The function of these systems in *Psa* are not well characterised, but are often found to have roles in effector and toxin secretion in pathogens (Marcelletti *et al.*, 2011, Records and Gross, 2010). One exception was a gene annotated as type V Aida, which is thought to be involved in adherence (Henderson *et al.*, 2004). Other genes which contribute to virulence and survival such as pili and fibriae were also up-regulated under starvation conditions.

Antibiotic resistance mechanisms were also up-regulated in starved *Psa*. An operon containing beta-lactamase (IYO_000615) and multidrug transporter genes was found to be highly up-regulated. Genes with beta-lactamase activities were down-regulated in minimal media, which may be to prevent degradation of secondary metabolites produced by *Psa*.

Psa genomes contains a multitude of transposons, group II introns and other mobile genetic elements, which are thought to contribute to their high genetic diversity. Many of these elements are highly multi-copy, making it difficult to quantify changes in gene expression for individual genes. Of the mobile genetic elements that were single-copy, several integrases and tyrosine recombinases were up-regulated in starved *Psa*

Comparison between *Psa* grown *in planta* and *in vitro*

Comparison of the *in vitro* and *in planta* experiments found that *in planta* experiments showed increased expression of the HrpL alternative σ-factor, which binds to the promoter region of *hrp*-genes (Fouts *et al.*, 2002), as were the type III secretion system effectors that are part of the HrpL regulon (Table 3.12). This reflects observations by previous studies, which show that *hrp*-regulated genes are not expressed in minimal media without additional stressors such as changes in pH and osmolyte content (McAtee *et al.*, 2018, Rahme *et al.*, 1992).

Principal component analysis (Figure 3.7) showed that these genes contributed the most to PC1, which clustered samples by experiment type, indicating that the induction of effectors and virulence factors *in planta* was the biggest difference between the *in vitro* and *in planta* transcriptomes. Effector genes were among the most abundant transcripts in *in planta* samples, and had higher read counts in *in planta* samples than *in vitro* samples prior to normalisation, despite the ~100-fold in sample read depth between the two experiment types.

The *in vitro* experiments in this study primarily capture transcriptional changes in response to nutrient availability. *In vitro* experiments showed similarities to several previous comparisons of *in vitro* and *in planta Psa* transcriptomes, and are not necessarily analogous to a particular stage of infection.

The PCA plot for *in vitro* and *in planta* samples (Figure 3.7) showed that early time-points (1.5–12hrs) clustered with samples grown in rich media along PC2. Overall, *hrp*-genes were among the most significantly differentially expressed genes in comparisons between *Psa* grown *in planta* and *in vitro* in rich media (Table 3.14). McAtee *et al.* (2018) found that fewer of these genes were expressed in early time-points, which may explain the closer association of rich media samples with these time-points.

Gene	PC1 contribution	l$_2$FC	Annotation
IYO_015535	1.00	11.77	hypothetical protein
IYO_005980	0.85	8.44	*
IYO_001355	0.85	8.44	*
IYO_027300	0.85	8.44	*
IYO_000990	0.85	8.44	*
IYO_027930	0.85	8.44	*
IYO_006790	0.85	8.38	Type III secretion HrpA pilin
IYO_006750	0.80	8.38	type III effector HrpW1 with pectin lyase domain
IYO_022020	0.79	8.24	hemolysin
IYO_015990	0.77	8.60	membrane protein
IYO_022025	0.77	8.72	glycerol acyltransferase
IYO_002645	0.71	10.02	*
IYO_003480	0.71	10.02	*
IYO_011415	0.71	10.02	*
IYO_028495	0.71	10.02	*
IYO_023550	0.71	10.02	*
IYO_000975	0.67	9.24	hypothetical protein
IYO_028770	0.66	6.82	LysR family transcriptional regulator
IYO_008385	0.64	7.14	hopAM1-2
IYO_023205	0.64	7.14	hopAM1-1
IYO_017375	0.61	7.98	Phosphonate ABC transporter
IYO_014970	0.57	6.60	hypothetical protein
IYO_005100	0.56	6.58	hypothetical protein
IYO_027905	0.54	6.15	Organosulfonate utilization protein SsuF
IYO_006735	0.54	6.11	hopN1
IYO_006880	0.54	6.76	type III secretion protein
IYO_026605	0.53	6.13	monooxygenase
IYO_006820	0.53	5.94	type III secretion protein HrpF
IYO_011995	0.52	6.03	phosphate ABC transporter substrate-binding protein
IYO_011310	0.51	6.28	NAD(P)H-dependent FMN reductase

Table 3.12: Top 30 genes contributing to PC1 variance in Figure 3.7. l $_2$FC *in planta* compared to *in vitro* and functional annotation are also shown (* denotes multi-copy transposons). The majority of these genes are hrp-controlled components of the type III secretion system and hop (Hrp-dependent outer protein) effectors.

Despite similarities between rich media and *in planta* samples, minimal media samples had the fewest significantly expressed genes of the *in vitro* samples (Table 3.13. Comparisons of rich and minimal media with *in planta* samples indicate that *Psa* nutrition *in planta* is well supported, however, similarities with minimal media samples indicate that stress responses are overall more highly expressed *in planta*. Siderophore synthesis was significantly up-regulated in minimal media, despite the weak expression of these genes *in planta* (McAtee *et al.*, 2018). Other differentially expressed genes relative to rich media included proteins involved in ion and sulfonate transport, which were more highly expressed *in planta*.

Media type	# DEGs compared to *in planta* samples
Starved	2194
Minimal	1237
Rich	2045

Table 3.13: Numbers of highly differentially expressed genes for pairwise comparisons of *in vitro* samples in different growth medias with *in planta* samples ($P_{adj} < 0.0001$)

M9 media has traditionally been used to simulate the nutritional profile of apoplastic fluid in studies of metabolic adaptions of *P. syringae* pathovars during early infection (O'Leary *et al.*, 2016, Rahme *et al.*, 1992). However, this does not entirely capture nutritional changes that may occur within the apoplast due to bacterial consumption of resources, as well as responses from the plant intended to restrict bacterial growth. For example, many genes involved in response to nutritional stress have been found to be down-regulated by *P. syringae* pv. *phaseolicola* grown in leaf and apoplastic fluid extracts compared to an M9 minimal media control Hernández-Morales *et al.* (2009). *P. syringae* pathovars are adapted to utilise sugars and carbohydrates in the apoplast (Rico and Preston, 2008), and preferred substrates such as amino acids and dicarboxylates are also found in high concentrations in apoplastic fluids (O'Leary *et al.*, 2016).

Psa grown in minimal media and under starvation conditions showed large-scale transcriptional changes to adapt to nutritional stress and iron restriction. With the exception of effectors, transcriptional differences found by Yu *et al.* (2013)'s comparison of *P. syringae* B728a in epiphytic vs apoplastic sites were similar to changes in *Psa* gene expression between starved and minimal media *in vitro* samples. *Psa* may also experience starvation within the plant as well as on the surface, as local resources are exhausted. Metabolic adaptions seen in starved *Psa*, where the Entner-Doudoroff pathway and TCA cycle were up-regulated, was also seen in *P. syringae* pv. *phaseolicola* in apoplastic fluid after all preferred carbon sources were consumed in apoplastic fluid (O'Leary *et al.*, 2016).

Comparison with *in planta*	Gene	Predicted function	l$_2$FC
Starved	IYO_015535	Hypothetical protein	10.86
	IYO_005475	Membrane protein	-5.19
	IYO_000485	Transposase	-2.84
	IYO_001195	Transposase	-2.84
	IYO_001615	Transposase	-2.84
	IYO_001730	Transposase	-2.84
	IYO_005000	Transposase	-2.84
	IYO_009075	Transposase	-2.84
	IYO_012240	Transposase	-2.84
	IYO_013520	Transposase	-2.84
Minimal	IYO_015535	Hypothetical protein	12.31
	IYO_006735	HopN1	6.35
	IYO_021635	Hypothetical protein	5.72
	IYO_004250	Hypothetical protein	5.05
	IYO_014970	Hypothetical protein	6
	IYO_006730	Type III chaperone protein ShcN	5.37
	IYO_005100	Hypothetical protein	6.86
	IYO_029645	LuxR family transcriptional regulator	3.98
	IYO_012700	Acyl-homoserine lactone acylase PvdQ	-5.69
Rich	IYO_015535	Hypothetical protein	12.6
	IYO_006735	HopN1	6.63
	IYO_021635	Hypothetical protein	5.76
	IYO_014970	Hypothetical protein	6.77
	IYO_006730	Type III chaperone protein ShcN	6.2
	IYO_004250	Hypothetical protein	4.8
	IYO_005100	Hypothetical protein	7.63
	IYO_006790	Type III secretion HrpA pilin	8.71
	IYO_010735	Hypothetical protein	4.69
	IYO_029700	Hypothetical protein	5.57

Table 3.14: Top 10 DEGs from comparisons of all *in planta* transcriptomes to *in vitro* samples different growth media (starved, minimal and rich). DEGs are chosen by the smallest P$_{adj}$ value, which are all below 1x10^{-50}. Log$_2$ fold change *in planta* relative to *in vitro* is shown.

Genes involved in alginate and extracellular polysaccharide biosynthesis and secretion expressed *in vitro* in minimal media and under starvation were also up-regulated expressed at late time-points *in planta* (McAtee *et al.*, 2018). However, the role of these compounds in virulence and ephiphytic survival remains unclear.

3.5 Conclusions and future directions

The ability to adapt to changeable and hostile environments is a major contributor to the success of pathogenic bacteria. *Psa* encounters several extreme environments and challenges during its life course, and must adapt its metabolism and homeostatic mechanisms appropriately. We present a high-depth RNA-seq dataset of the kiwifruit pathogen *Psa* grown *in vitro* in multiple growth media. Comparisons of these transcriptomes shows large-scale adaptions to cope with drastic changes in nutrient availability. These data also found that starvation conditions triggered several anticipatory responses to plant defence mechanisms, highlighting the regulatory connections between metabolism and virulence.

While the *in vitro* data generated for this chapter did not induce the expression of many gene cohorts found to be highly expressed *in planta* (McAtee *et al.*, 2018), these data show that *Psa* grown *in vitro* in nutrient-limiting conditions has similar transcriptional responses as other *P. syringae* pathovars in apoplastic and epiphytic environments. Methods for transposon insertion sequencing experiments in *Psa* are currently being developed (Mesarich *et al.*, 2017), which will provide a measure of gene essentiality for specific growth conditions. Current methods for dual RNA-seq studies of bacterial pathogens *in planta* are more challenging to perform than in animal cell lines, and have low bacterial transcript yields from plant tissue(Nobori and Tsuda, 2018). Identifying analogous *in vitro* growth conditions will allow us to identify and isolate environmental signals and stressors that trigger the expression of virulence genes, or other changes in gene expression important for survival.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, *215*(3), 403–410.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. (2008).

The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, *9*, 75.

Bae, C., Han, S. W., Song, Y.-R., Kim, B.-Y., Lee, H.-J., Lee, J.-M., Yeam, I., Heu, S., and Oh, C.-S. (2015). Infection processes of xylem-colonizing pathogenic bacteria: possible explanations for the scarcity of qualitative disease resistance genes against them in crops. *Theor. Appl. Genet.*, *128*(7), 1219–1229.

Baynham, P. J., Brown, A. L., Hall, L. L., and Wozniak, D. J. (1999). *Pseudomonas aeruginosa* AlgZ, a ribbon-helix-helix DNA-binding protein, is essential for alginate synthesis and *algD* transcriptional activation. *Mol. Microbiol.*, *33*(5), 1069–1080.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, *34*(5), 525–527.

Butler, M. I., Stockwell, P. A., Black, M. A., Day, R. C., Lamont, I. L., and Poulter, R. T. M. (2013). *Pseudomonas syringae* pv. *actinidiae* from recent outbreaks of kiwifruit bacterial canker belong to different clones that originated in China. *PloS One*, *8*(2), e57464.

Cézard, C., Farvacques, N., and Sonnet, P. (2015). Chemistry and biology of pyoverdines, *Pseudomonas* primary siderophores. *Curr. Med. Chem.*, *22*(2), 165–186.

Chapman, J. R., Taylor, R. K., Weir, B. S., Romberg, M. K., Vanneste, J. L., Luck, J., and Alexander, B. J. R. (2012). Phylogenetic relationships among global populations of *Pseudomonas syringae* pv. *actinidiae*. *Phytopathology*, *102*(11), 1034–1044.

Chavarría, M., Nikel, P. I., Pérez-Pantoja, D., and de Lorenzo, V. (2013). The Entner-Doudoroff pathway empowers *Pseudomonas putida* KT2440 with a high tolerance to oxidative stress. *Environ. Microbiol.*, *15*(6), 1772–1785.

Colombi, E., Straub, C., Künzel, S., Templeton, M. D., McCann, H. C., and Rainey, P. B. (2017). Evolution of copper resistance in the kiwifruit pathogen *Pseudomonas syringae* pv. *actinidiae* through acquisition of integrative conjugative elements and plasmids. *Environ. Microbiol.*, *19*(2), 819–832.

Daddaoua, A., Krell, T., and Ramos, J.-L. (2009). Regulation of glucose metabolism in *Pseudomonas*: the phosphorylative branch and Entner-Doudoroff enzymes are regulated by a repressor containing a sugar isomerase domain. *J. Biol. Chem.*, *284*(32), 21360–21368.

Diepold, A. and Armitage, J. P. (2015). Type III secretion systems: the bacterial flagellum and the injectisome. *Phil. Trans. R. Soc. B.*, *370*(1679).

Donati, I., Buriani, G., Cellini, A., Mauri, S., Costa, G., and Spinelli, F. (2014). New insights on the bacterial canker of kiwifruit (*Pseudomonas syringae* pv. *actinidiae*). *JBR-BTR*, *4*(2), 53–67.

Donati, I., Cellini, A., Buriani, G., Mauri, S., Kay, C., Tacconi, G., and Spinelli, F. (2018). Pathways of flower infection and pollen-mediated dispersion of *Pseudomonas syringae* pv. *actinidiae*, the causal agent of kiwifruit bacterial canker. *Hortic. Res.*, *5*, 56.

Ferreiro, M.-D., Nogales, J., Farias, G. A., Olmedilla, A., Sanjuán, J., and Gallegos, M. T. (2018). Multiple CsrA Proteins Control Key Virulence Traits in *Pseudomonas syringae* pv. *tomato* DC3000. *Mol. Plant. Microbe. Interact.*, *31*(5), 525–536.

Fones, H. and Preston, G. M. (2013). The impact of transition metals on bacterial plant disease. *FEMS Microbiol. Rev.*, *37*(4), 495–519.

Fouts, D. E., Abramovitch, R. B., Alfano, J. R., Baldo, A. M., Buell, C. R., Cartinhour, S., Chatterjee, A. K., D'Ascenzo, M., Gwinn, M. L., Lazarowitz, S. G., Lin, N.-C., Martin, G. B., Rehm, A. H., Schneider, D. J., van Dijk, K., Tang, X., and Collmer, A. (2002). Genomewide identification of *Pseudomonas syringae* pv. *tomato* DC3000 promoters controlled by the HrpL alternative sigma factor. *PNAS*, *99*(4), 2275–2280.

He, S. Y. (1996). Elicitation of Plant Hypersensitive Response by Bacteria. *Plant physiology*, *112*(3), 865–869.

He, S. Y., Huang, H. C., and Collmer, A. (1993). *Pseudomonas syringae* pv. *syringae* harpin$_{\mathrm{Pss}}$: a protein that is secreted via the *hrp* pathway and elicits the hypersensitive response in plants. *Cell*, *73*(7), 1255–1266.

Heath, M. C. (2000). Hypersensitive response-related death. *Plant Molecular Biology*, *44*(3), 321–334.

Henderson, I. R., Navarro-Garcia, F., Desvaux, M., Fernandez, R. C., and Ala'Aldeen, D. (2004). Type V protein secretion pathway: the autotransporter story. *Microbiol. Mol. Biol. Rev.*, *68*(4), 692–744.

Hernández-Morales, A., De la Torre-Zavala, S., Ibarra-Laclette, E., Hernández-Flores, J. L., Jofre-Garfias, A. E., Martínez-Antonio, A., and Alvarez-Morales, A. (2009). Transcriptional profile of *Pseudomonas syringae* pv. *phaseolicola* NPS3121 in response to tissue extracts from a susceptible *Phaseolus vulgaris* L. cultivar. *BMC Microbiol.*, *9*, 257.

Hirano, S. S. and Upper, C. D. (1990). Population Biology and Epidemiology of *Pseudomonas syringae*. *Annu. Rev. Phytopathol.*, *28*(1), 155–177.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., and Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.*, *34*(8), 2115–2122.

Jones, A. M., Lindow, S. E., and Wildermuth, M. C. (2007). Salicylic acid, yersiniabactin, and pyoverdin production by the model phytopathogen *Pseudomonas syringae* pv. *tomato* DC3000: synthesis, regulation, and impact on tomato and *Arabidopsis* host plants. *J. Bacteriol.*, *189*(19), 6773–6786.

Kawakami, T., Kuroki, M., Ishii, M., Igarashi, Y., and Arai, H. (2010). Differential expression of multiple terminal oxidases for aerobic respiration in *Pseudomonas aeruginosa*. *Environ. Microbiol.*, *12*(6), 1399–1412.

Keen, N. T. (1990). Gene-for-gene complementarity in plant-pathogen interactions. *Annu. Rev. Genet.*, *24*, 447–463.

Kehrer, J. P. (2000). The Haber–Weiss reaction and mechanisms of toxicity. *Toxicology*, *149*(1), 43–50.

Keith, L. M. and Bender, C. L. (1999). AlgT (ς^{22}) controls alginate production and tolerance to environmental stress in *Pseudomonas syringae*. *J. Bacteriol.*, *181*(23), 7176–7184.

Kivisaar, M. (2018). The Effect of Cellular Redox Status on the Evolvability of New Catabolic Pathways. *MBio*, *9*(5).

Lessie, T. G. and Phibbs, P. V. (1984). Alternative pathways of carbohydrate utilization in Pseudomonads. *Annu. Rev. Microbiol.*, *38*(1), 359–388. PMID: 6388497.

Lindgren, P. B. (1997). The role of *hrp* genes during plant-bacterial interactions. *Annu. Rev. Phytopathol.*, *35*(1), 129–152. PMID: 15012518.

Lindow, S. E., Andersen, G., and Beattie, G. A. (1993). Characteristics of Insertional Mutants of *Pseudomonas syringae* with Reduced Epiphytic Fitness. *Appl. Environ. Microbiol.*, *59*(5), 1593–1601.

Llamas, M. A., Imperi, F., Visca, P., and Lamont, I. L. (2014). Cell-surface signaling in *Pseudomonas*: stress responses, iron transport, and pathogenicity. *FEMS Microbiol. Rev.*, *38*(4), 569–597.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, *15*(12), 550.

Luo, W. and Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, *29*(14), 1830–1831.

Marcelletti, S., Ferrante, P., Petriccione, M., Firrao, G., and Scortichini, M. (2011). *Pseudomonas syringae* pv. *actinidiae* draft genomes comparison reveal strain-specific features involved in adaptation and virulence to *Actinidia* species. *PloS One*, *6*(11), e27297.

Mazzaglia, A., Studholme, D. J., Taratufolo, M. C., Cai, R., Almeida, N. F., Goodman, T., Guttman, D. S., Vinatzer, B. A., and Balestra, G. M. (2012). *Pseudomonas syringae* pv. *actinidiae* (PSA) isolates from recent bacterial canker of kiwifruit outbreaks belong to the same genetic lineage. *PLoS One*, *7*(5), e36518.

McAtee, P. A., Brian, L., Curran, B., van der Linden, O., Nieuwenhuizen, N. J., Chen, X., Henry-Kirk, R. A., Stroud, E. A., Nardozza, S., Jayaraman, J., Rikkerink, E. H. A., Print, C. G., Allan, A. C., and Templeton, M. D. (2018). Re-programming of *Pseudomonas syringae* pv. *actinidiae* gene expression during early stages of infection of kiwifruit. *BMC Genomics*, *19*(1), 822.

McCann, H. C., Li, L., Liu, Y., Li, D., Pan, H., Zhong, C., Rikkerink, E. H. A., Templeton, M. D., Straub, C., Colombi, E., Rainey, P. B., and Huang, H. (2017). Origin and Evolution of the Kiwifruit Canker Pandemic. *Genome Biol. Evol.*, *9*(4), 932–944.

McCann, H. C., Rikkerink, E. H. A., Bertels, F., Fiers, M., Lu, A., Rees-George, J., Andersen, M. T., Gleave, A. P., Haubold, B., Wohlers, M. W., Guttman, D. S., Wang, P. W., Straub, C., Vanneste, J. L., Rainey, P. B., and Templeton, M. D. (2013). Genomic analysis of the Kiwifruit pathogen *Pseudomonas syringae* pv. *actinidiae* provides insight into the origins of an emergent plant disease. *PLoS Pathog.*, *9*(7), e1003503.

Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E., and Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.*, *39*(Web Server issue), W339–46.

Mehmood, A., Abdallah, K., Khandekar, S., Zhurina, D., Srivastava, A., Al-Karablieh, N., Alfaro-Espinoza, G., Pletzer, D., and Ullrich, M. S. (2015). Expression of extra-cellular levansucrase in *Pseudomonas syringae* is controlled by the *in planta* fitness-promoting metabolic repressor HexR. *BMC Microbiol.*, *15*, 48.

Mesarich, C. H., Rees-George, J., Gardner, P. P., Ghomi, F. A., Gerth, M. L., Andersen, M. T., Rikkerink, E. H. A., Fineran, P. C., and Templeton, M. D. (2017). Transposon insertion libraries for the characterization of mutants from the kiwifruit pathogen *Pseudomonas syringae* pv. *actinidiae*. *PLoS One*, *12*(3), e0172790.

Molina-Henares, M. A., García-Salamanca, A., Molina-Henares, A. J., de la Torre, J., Herrera, M. C., Ramos, J. L., and Duque, E. (2009). Functional analysis of aromatic biosynthetic pathways in *Pseudomonas putida* KT2440. *Microb. Biotechnol.*, *2*(1), 91–100.

Monteil, C. L., Yahara, K., Studholme, D. J., Mageiros, L., Méric, G., Swingle, B., Morris, C. E., Vinatzer, B. A., and Sheppard, S. K. (2016). Population-genomic insights into emergence, crop adaptation and dissemination of *Pseudomonas syringae* pathogens. *Microb. Genom.*, *2*(10), e000089.

Morris, C. E., Monteil, C. L., and Berge, O. (2013). The life history of *Pseudomonas syringae*: linking agriculture to earth system processes. *Annu. Rev. Phytopathol.*, *51*, 85–104.

Morris, C. E., Sands, D. C., Vinatzer, B. A., Glaux, C., Guilbaud, C., Buffière, A., Yan, S., Dominguez, H., and Thompson, B. M. (2008). The life history of the plant pathogen *Pseudomonas syringae* is linked to the water cycle. *ISME J.*, *2*(3), 321–334.

Morris, V. L., Jackson, D. P., Grattan, M., Ainsworth, T., and Cuppels, D. A. (1995). Isolation and sequence analysis of the *Pseudomonas syringae* pv. *tomato* gene encoding a 2,3-diphosphoglycerate-independent phosphoglyceromutase. *J. Bacteriol.*, *177*(7), 1727–1733.

Nobori, T. and Tsuda, K. (2018). *In planta* Transcriptome Analysis of *Pseudomonas syringae*. *Bio-protocol*, *8*(17), e2987.

Nomura, K., Melotto, M., and He, S.-Y. (2005). Suppression of host defense in compatible plant-*Pseudomonas syringae* interactions. *Curr. Opin. Plant Biol.*, *8*(4), 361–368.

O'Leary, B. M., Neale, H. C., Geilfus, C.-M., Jackson, R. W., Arnold, D. L., and Preston, G. M. (2016). Early changes in apoplast composition associated with defence and disease in interactions between *Phaseolus vulgaris* and the halo blight pathogen *Pseudomonas syringae* pv. *phaseolicola*. *Plant, cell & environment*, *39*(10), 2172–2184.

Preston, L. A., Wong, T. Y., Bender, C. L., and Schiller, N. L. (2000). Characterization of alginate lyase from *Pseudomonas syringae* pv. *syringae*. *J. Bacteriol.*, *182*(21), 6268–6271.

Rahme, L. G., Mindrinos, M. N., and Panopoulos, N. J. (1992). Plant and environmental sensory signals control the expression of *hrp* genes in *Pseudomonas syringae* pv. *phaseolicola*. *J. Bacteriol.*, *174*(11), 3499–3507.

Ratledge, C. and Dover, L. G. (2000). Iron metabolism in pathogenic bacteria. *Annu. Rev. Microbiol.*, *54*, 881–941.

Records, A. R. and Gross, D. C. (2010). Sensor kinases RetS and LadS regulate *Pseudomonas syringae* type VI secretion and virulence factors. *J. Bacteriol.*, *192*(14), 3584–3596.

Rico, A. and Preston, G. M. (2008). *Pseudomonas syringae* pv. *tomato* DC3000 uses constitutive and apoplast-induced nutrient assimilation pathways to catabolize nutrients that are abundant in the tomato apoplast. *Mol. Plant. Microbe. Interact.*, *21*(2), 269–282.

Schenk, A., Weingart, H., and Ullrich, M. S. (2008). The alternative sigma factor AlgT, but not alginate synthesis, promotes *in planta* multiplication of *Pseudomonas syringae* pv. *glycinea*. *Microbiology*, *154*(Pt 2), 413–421.

Schurr, M. J., Yu, H., Martinez-Salazar, J. M., Boucher, J. C., and Deretic, V. (1996). Control of AlgU, a member of the σ^{E}-like family of stress sigma factors, by the negative regulators MucA and MucB and *Pseudomonas aeruginosa* conversion to mucoidy in cystic fibrosis. *J. Bacteriol.*, *178*(16), 4997–5004.

Scortichini, M., Marcelletti, S., Ferrante, P., Petriccione, M., and Firrao, G. (2012). *Pseudomonas syringae* pv. *actinidiae*: a re-emerging, multi-faceted, pandemic pathogen. *Mol. Plant Pathol.*, *13*(7), 631–640.

Serizawa, S., Ichikawa, T., Takikawa, Y., Tsuyumu, S., and Goto, M. (1989). Occurrence of Bacterial Canker of Kiwifruit in Japan. *Japanese Journal of Phytopathology*, *55*(4), 427–436.

Sexton, D. J. and Schuster, M. (2017). Nutrient limitation determines the fitness of cheaters in bacterial siderophore cooperation. *Nat. Commun.*, *8*(1), 230.

Spinelli, F., Donati, I., Vanneste, J. L., Costa, M., and Costa, G. (2011). Real time monitoring of the interactions between *Pseudomonas syringae* pv. *actinidiae* and *Actinidia* species. *Acta Hortic.*, *913*(913), 461–465.

Takikawa, Y., Serizawa, S., Ichikawa, T., Tsuyumu, S., and Goto, M. (1989). *Pseudomonas syringae* pv. *actinidiae* pv. nov. *Japanese Journal of Phytopathology*, *55*(4), 437–444.

Tanner, D. J. (2015). A biosecurity incursion: the impact of *Pseudomonas syringae* pv. *actinidiae* (*Psa*) on the New Zealand kiwifruit industry. *Acta Hortic.*, (1105), 379–384.

Templeton, M. D., Warren, B. A., Andersen, M. T., Rikkerink, E. H. A., and Fineran, P. C. (2015). Complete DNA Sequence of *Pseudomonas syringae* pv. *actinidiae*, the Causal Agent of Kiwifruit Canker Disease. *Genome Announc.*, *3*(5).

Tuanyok, A., Kim, H. S., Nierman, W. C., Yu, Y., Dunbar, J., Moore, R. A., Baker, P., Tom, M., Ling, J. M. L., and Woods, D. E. (2005). Genome-wide expression analysis of iron regulation in *Burkholderia pseudomallei* and *Burkholderia mallei* using DNA microarrays. *FEMS Microbiology Letters*, *252*(2), 327–335.

Vasil, M. L. and Ochsner, U. A. (1999). The response of *Pseudomonas aeruginosa* to iron: genetics, biochemistry and virulence. *Mol. Microbiol.*, *34*(3), 399–413.

Vinatzer, B. A. (2012). “Listening in” on how a bacterium takes over the plant vascular system. *MBio*, *3*(5), e00269–12.

Yu, X., Lund, S. P., Scott, R. A., Greenwald, J. W., Records, A. H., Nettleton, D., Lindow, S. E., Gross, D. C., and Beattie, G. A. (2013). Transcriptional responses of *Pseudomonas syringae* to growth in epiphytic versus apoplastic leaf sites. *Proceedings of the National Academy of Sciences*, *110*(5), E425–E434.

Chapter 4

Identifying candidate novel ncRNAs from *Psa* transcriptomes

4.1 Preface

Successful pathogenic bacteria must alter gene expression in response to changing and hostile environments. Non-coding RNAs (ncRNAs) contribute to adaptability by forming complex regulatory networks, and contribute to the survival and virulence of pathogens.

As of yet no studies have been undertaken to extensively characterise ncRNA genes in *Pseudomonas syringae* pv. *actinidiae*. We have generated transcriptomes of a New Zealand strain of *Psa* in multiple growth conditions *in vitro*, and used this data, as well as previously generated *in planta* transcriptomes, to identify candidate ncRNAs, with the aim of identifying those involved in infection and virulence. With this data we have identified transcripts that are highly expressed in specific growth conditions, and which have predicted secondary structure motifs characteristic to known functional ncRNAs. We have also characterised the expression pattern of annotated ncRNAs in *Psa*, and identified that an uncharacterised RNA motif, twoAYGGAY, is highly prevalent in the *Psa* genome.

4.1.1 Contributions

I performed all data analysis. RNA-seq generation is described in Chapter 3.

4.2 Introduction

Bacterial genomes contain a multitude of functional non-coding RNA (ncRNA) genes. These include small ncRNAs (sRNAs), short transcripts which often have some secondary structure, and predominantly function as regulators of gene or protein expression by base-pairing with mRNA transcripts. Many sRNAs are only expressed under specific growth conditions, and can be found as key regulators of pathogenic responses in bacteria, including those that infect plants.

Pseudomonas syringae pv. *actinidiae* (*Psa*) is a plant pathogen that causes kiwifruit canker disease, which has had a significant impact on the New Zealand kiwifruit industry. Mutations in the RNA chaperone Hfq, which is required to stabilise many sRNA-mRNA interactions, is linked to reduced virulence in the pandemic strain of *Psa*, suggesting sRNAs play a role in virulence in this organism (Zhao *et al.*, 2019).

Transcriptomic screens have identified sRNAs in other Pseudomonads (Gómez-Lozano *et al.*, 2012, 2015), notably in *P. syringae* pv. *tomato* DC3000, where a non-coding transcript was found to regulate the biosynthesis of the phytotoxin coronatine (Filiatrault *et al.*, 2010, Geng *et al.*, 2014). In other plant pathogens sRNAs have been found to be important regulators of pathogenicity. In *Erwinia carotovora*, RsmB regulates enzymes that degrade plant cell walls, and the horizontally-acquired sRNA-Xcc1 discovered in *Xanthomonas campestris* is a regulator of virulence factors (Chen *et al.*, 2011, Liu *et al.*, 1998).

Novel ncRNAs can be identified by measuring expression in intergenic regions (Lindgreen *et al.*, 2014, Rath *et al.*, 2017), with the incorporation of conservation and structural information providing useful signals to identify candidates that are likely to be functional (Stav *et al.*, 2019). Recent approaches for the comprehensive identification of ncRNAs have used exhaustive multi-condition transcriptomics, which has been effective in identifying ncRNAs relevant to pathogenicity in *E. coli* (Rau *et al.*, 2015) and *Salmonella* (Kröger *et al.*, 2013).

More targeted experiments have also been performed in *Pseudomonas aeruginosa* to identify strain and species-specific sRNAs (Ferrara *et al.*, 2012), as well as temperature-dependent sRNA expression (Wurtzel *et al.*, 2012). However, the phylogenetic distribution of many sRNAs are limited, and many sRNAs with specialised functions are genus, strain or niche-specific, due to both rapid evolution and sequence gain and loss of ncRNA genes through horizontal gene transfer (Jose *et al.*, 2019). This trend can also be seen in *P. aeruginosa* sRNAs, few of which have known homologues in *P. syringae* (Gómez-Lozano *et al.*, 2015).

In this study, transcriptomes from *Psa* generated in multiple growth conditions were used to predict candidate ncRNAs, aiming to identify ncRNAs relevant to pathogenicity. Sequences from manual ncRNA annotations were tested for differential gene expression between *Psa* transcriptomes grown *in vitro* in rich, starved and minimal media, and were also compared to transcriptomes of *Psa* grown *in planta* generated by McAtee *et al.* (2018).

4.3 Methods

4.3.1 ncRNA annotation

In vitro and *in planta* RNA-seq reads (see Chapter 3 for experimental details) were processed using `trimmomatic` v0.36 (Bolger *et al.*, 2014), and mapped to the *Pseudomonas syringae* pv. *actinidiae* ICMP 18884 chromosome and plasmid sequences (NCBI accession CP011972.2 and CP011973.1) (Templeton *et al.*, 2015) using Bowtie2 v2.3.4 (Langmead and Salzberg, 2012). `Samtools` v1.8 (Li *et al.*, 2009) was used to generate plot files representing the number of reads mapped to each nucleotide (Figure 4.1).

Candidate ncRNAs were manually annotated using the Artemis genome viewer (Carver *et al.*, 2012). Plot files were used to visualise read depth across the genome for each experiment. Expression peaks in intergenic regions, or opposite protein coding genes, that were over 50 nt in length and with a read depth of at least 50 in *in vitro* samples and over 10 in *in planta* samples that were similarly expressed between replicates were considered candidate ncRNAs.

While small protein-coding elements are a feature of some ncRNAs, this was considered a confounding factor for the conservation analysis. Expression peaks for likely protein-coding candidates and UTRs are included in supplemental materials. To ensure that predicted ncRNAs located in intergenic regions did not overlap with un-annotated genes, pseudogenes or small ORFs, the reference genome was re-annotated with `Prokka` (Seemann, 2014), `cmscan` with Rfam models (Nawrocki *et al.*, 2015) and `RNIE` (Gardner *et al.*, 2011). Protein domain annotations were generated from a 6-frame translation the genome, which was annotated using `HMMER` (Eddy, 2011) and Pfam HMMs (Punta *et al.*, 2012). For ncRNA annotations found only in the unstranded *in planta* RNA-seq data, terminators and protein coding genes were used to infer strand where applicable. Known ncRNAs were annotated using Rfam domains; as many of these annotations were partial, expression peaks were used to estimate transcription start sites.

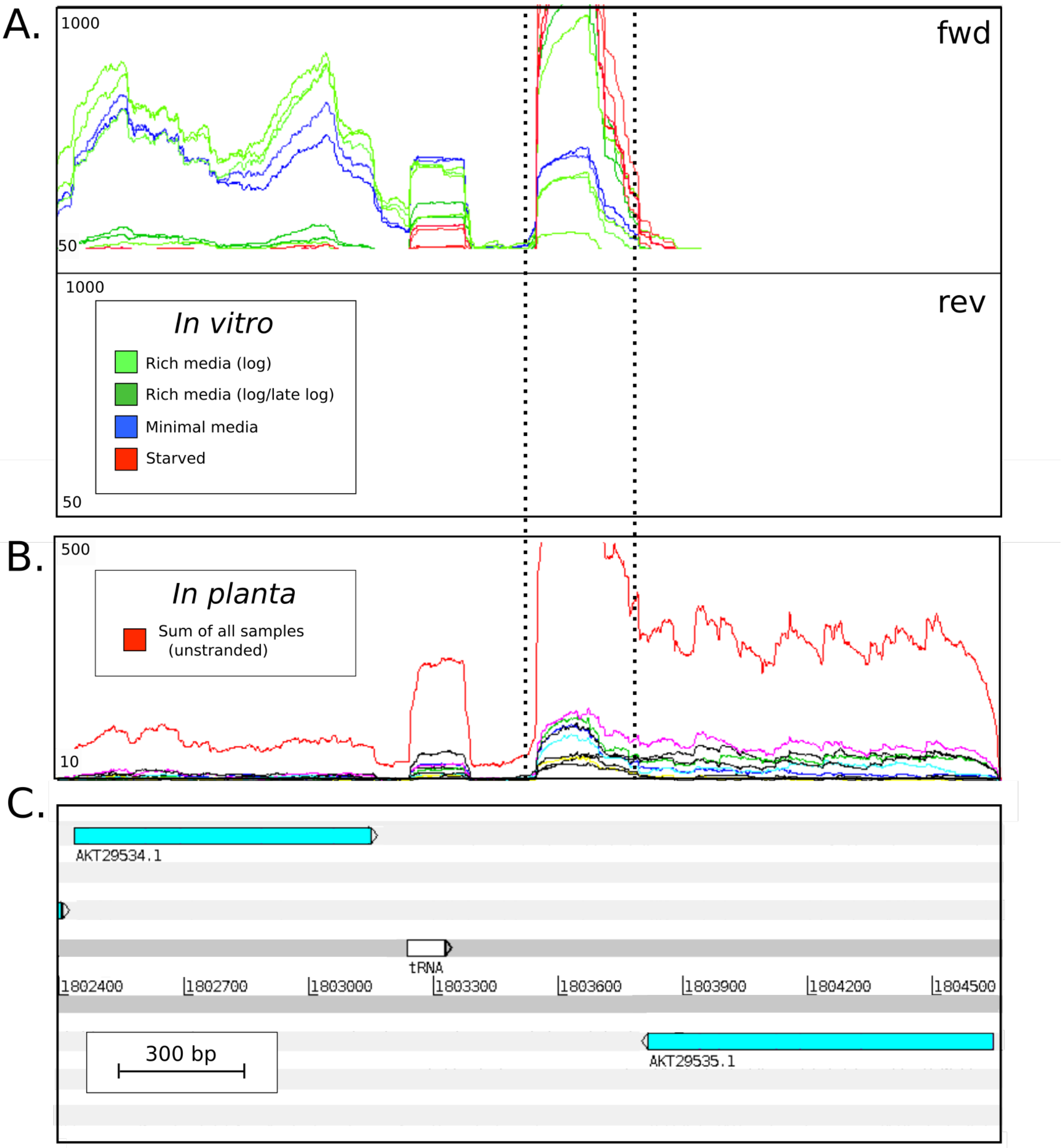


Figure 4.1: Example showing the visualisation of RNA-seq counts and annotations of candidate ncRNAs in the Artemis genome browser. Panels A and B show traces representing the number of RNA-seq reads (represented by the Y axis height) mapped to the position in the *Psa* genome (X axis), which is shown in Panel C. **(A)** *In vitro* traces each represent a single replicate and are coloured based on experiment conditions. The two windows show reads mapping to the forward and reverse strands. **(B)** Traces representing *in planta* samples. Due to the low read depth of the *in planta* data, each trace represents the sum across a set of replicates, with an additional trace added representing the sum across all *in planta* samples (shown in red). **(C)** Gene and genome feature annotations (shown in frame). Two protein-coding genes (blue) and an annotated tRNA are expressed. Dotted lines show an expression peak in an intergenic region, which would be classified as a candidate ncRNA.

4.3.2 Quality control

4.3.3 Sequence uniqueness

BLAST (Altschul *et al.*, 1990) was used to check whether candidate ncRNA sequences were homologous to each other, or contained sequences that were non-unique in the genome. A reciprocal `blastn` search was used between all candidate ncRNAs to look for multi-copy or common motifs. This showed that 18 sequences were homologous to at least one other candidate ncRNA. These clustered into five groups, which could be classified by shared sequence motifs and/or similarity of function of nearby genes. `Cytoscape` (Shannon *et al.*, 2003) was used to visualise groups of homologous sequences (Figure 4.2). The largest cluster of candidate ncRNAs were found upstream of group II introns. Several of these candidates also had homology to genomic regions containing an annotated Rfam motif, twoAYGGAY, which was also often upstream of group II introns.

A `blastn` search of these candidates against the genome identified that the ncRNA sequences in the UTRs of oxidoreductases and ABC transporters, and twoAYGGAY-like sequences were highly multi-copy throughout the genome. These were not included in the conservation analysis.

4.3.4 Conservation analysis

A homology search was performed to assess whether any candidate ncRNA sequences were conserved within *Pseudomonas* genomes, using an iterative `nhmmer`-based (Wheeler and Eddy, 2013) approach described previously (Methods, Chapter 2). A database of whole *Pseudomonas* genomes was created, consisting of 489 complete genome and plasmid sequences from NCBI. Candidate ncRNA sequences were used to generate single-sequence profile HMMs using `hmmbuild`, which were used in an `nhmmer` homology search of 63 *Pseudomonas syringae* genomes in the database. These sequences were then aligned using the ncRNA alignment tool `mafft-qinsi` (Katoh and Standley, 2013), built into new HMMs and searched over all *Pseudomonas* genomes. Multiple sequence alignments for each candidate were made from the final homology search results.

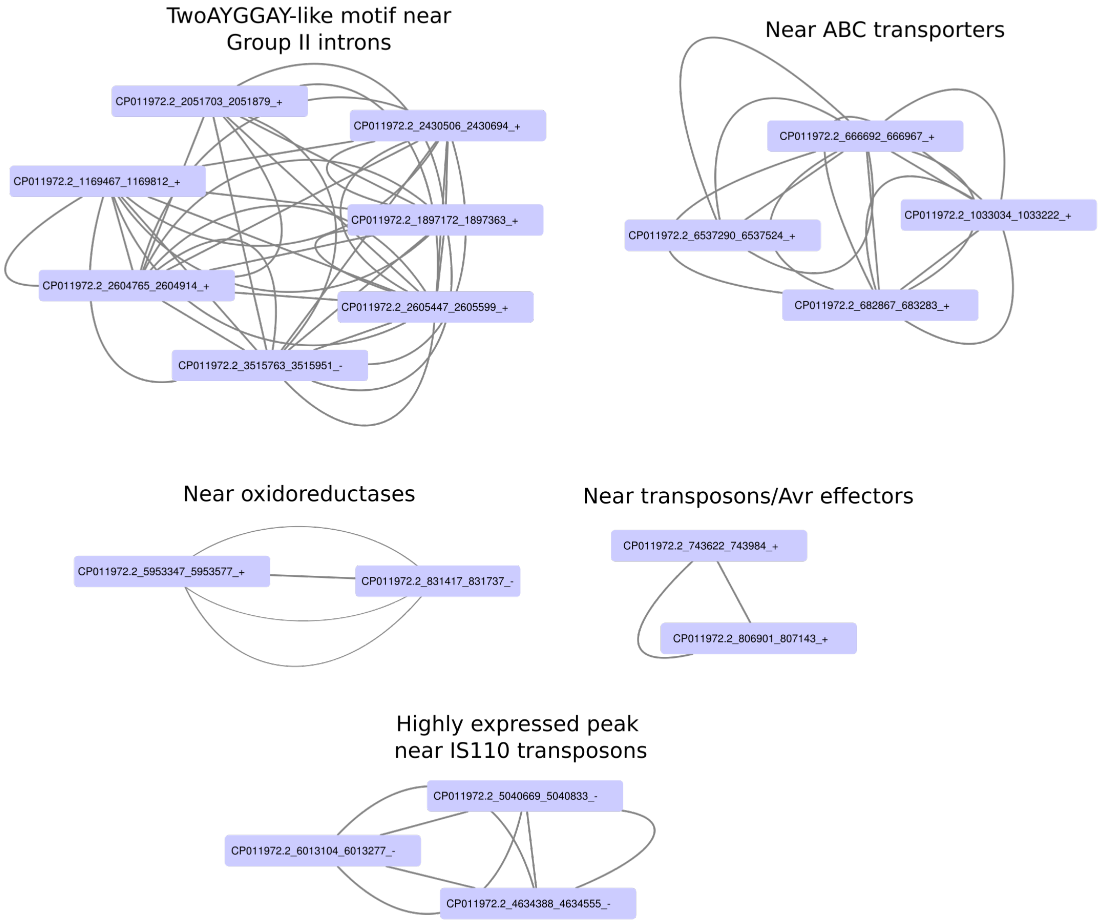

TwoAYGGAY-like motif near
Group II introns
CP011972.2_2051703_2051879_+
CP011972.2_2430506_2430694_+
CP011972.2_1169467_1169812_+
CP011972.2_1897172_1897363_+
CP011972.2_2604765_2604914_+
CP011972.2_2605447_2605599_+
CP011972.2_3515763_3515951_-
Near ABC transporters
CP011972.2_666692_666967_+
CP011972.2_1033034_1033222_+
CP011972.2_6537290_6537524_+
CP011972.2_682867_683283_+
Near oxidoreductases
CP011972.2_5953347_5953577_+
CP011972.2_831417_831737_-
Near transposons/Avr effectors
CP011972.2_743622_743984_+
CP011972.2_806901_807143_+
Highly expressed peak
near IS110 transposons
CP011972.2_5040669_5040833_-
CP011972.2_6013104_6013277_-
CP011972.2_4634388_4634555_-

Figure 4.2: Network diagram showing annotated ncRNAs with high sequence similarity to each other. Networks are generated from an all vs all blastn search of candidate ncRNAs. Nodes are individual ncRNAs, and edges represent instances of a significant blastn hit to another ncRNA. Clusters are labelled based on their context in the *Psa* genome and visualised in Cytoscape (Shannon *et al.*, 2003).

4.3.5 Alignment analysis and structure prediction

Several tools were used to identify alignments that were most likely to be non-coding RNAs (Figure 4.3). RNAcode was used to test if candidate ncRNAs had protein coding potential (Washietl *et al.*, 2011); 20 candidates with a p value < 0.01 were removed. Alifoldz was used to identify alignments that had significant signals of stable secondary structure (Washietl and Hofacker, 2004). Conserved candidate ncRNAs with negative z-scores were selected. For alignments with few sequences or little sequence variation, alifoldz z-score, RNAalifold secondary structures and alignments were manually inspected to select candidates.

`RNAalifold` (Bernhart *et al.*, 2008) was used to predict a consensus secondary structure for each alignment. Six conserved candidates that did not have a thermodynamically favourable secondary structure (minimum free energy of > -20) were removed. `R-scape` (Rivas *et al.*, 2017) was used to identify covarying base-pairs within predicted secondary structures. Six structures containing covarying base-pairs with an E-value < 0.01 were selected as final candidates. `R-scape` tests could not be applied to alignments with limited numbers of sequences. Thermodynamic stability and covariance scores, conservation and expression were used to select and rank the final ncRNA candidates. For candidates with limited phylogenetic distribution, MFE, overall expression and differential expression were used to select the final candidates. Secondary structures were visualised using R2R (Weinberg and Breaker, 2011).

4.3.6 Differential gene expression and expression in other data-sets

Candidate intergenic ncRNAs as well as Rfam annotations were tested for differential expression using `Kallisto` (settings: `-b 100 --rf-stranded`) (Bray *et al.*, 2016) and DESeq2 (Love *et al.*, 2014) using the same analysis pipeline as Chapter 2 (provided in Supplementary Materials).

Publicly available data on the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) was used to find paired-end, stranded *Pseudomonas* RNA-seq data-sets. Metadata from the NCBI FTP server was used to generate a table consisting of SRA accession, SRA experiment number, sequencing design (paired or single-end reads), read length, sequencing instrument, NCBI biosample, NCBI taxonomy ID, organism genome accession and species name. This table was then used to select seven *P. syringae* pv. *tabaci* ATCC 11528 (*Pta*) transcriptomes (Table 4.1, Table 4.2), generated *in vitro* in M9 media, which were used to test if candidate ncRNAs were expressed in other organisms.

Profile HMMs of candidate ncRNAs, which were generated from alignments generated during the conservation analysis, were used to annotate the *Pta* genome ATCC 11528 (NCBI Accession ACHU02.1). Sequencing reads were mapped to the genome and visualised using the same pipeline as the *Psa* transcriptomes (provided in Supplementary Materials). Homology search results were then used find sequences homologous to *Psa* candidate ncRNAs in the *Pta* genome that also showed similar levels of expression.

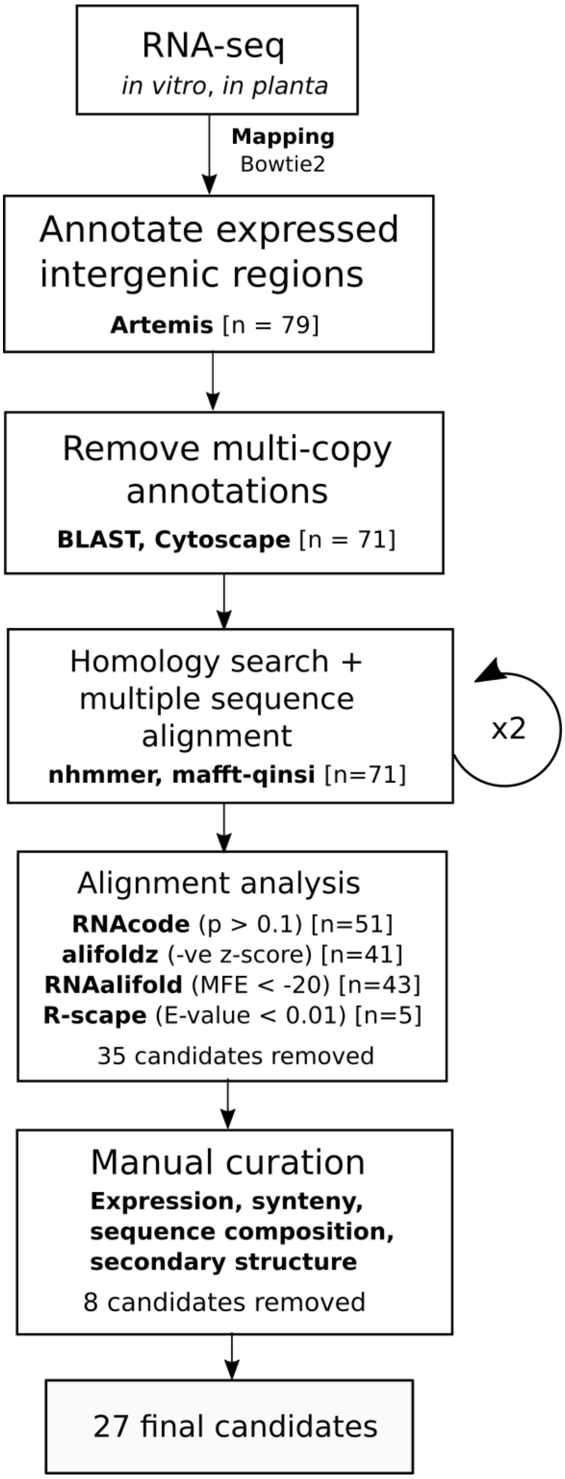

RNA-seq
in vitro, in planta
Mapping
Bowtie2
Annotate expressed intergenic regions
Artemis [n = 79]
Remove multi-copy annotations
BLAST, Cytoscape [n = 71]
Homology search + multiple sequence alignment
nhmmer, mafft-qinsi [n=71]
x2
Alignment analysis
RNAcode (p > 0.1) [n=51]
alifoldz (-ve z-score) [n=41]
RNAalifold (MFE < -20) [n=43]
R-scape (E-value < 0.01) [n=5]
35 candidates removed
Manual curation
Expression, synteny, sequence composition, secondary structure
8 candidates removed
27 final candidates

Figure 4.3: Flow-chart showing showing quality control steps used to select final candidate ncRNAs. Briefly, manual annotations that were single copy were selected for conservation analysis. Multiple sequence alignments from this homology search were then tested for signals that would identify the candidates as functional ncRNAs. `RNAcode` was used to test if sequences were likely to be protein coding (Washietl *et al.*, 2011), and `alifoldz` z-scores and `RNAalifold` minimum free energy (MFE) was used to identify candidates with thermodynamically stable predicted secondary structure (Bernhart *et al.*, 2008, Washietl and Hofacker, 2004). Additional tests were placed on conserved sequences (>12 sequences) using `R-scape` (Rivas *et al.*, 2017). Manual curation removed a further 8 candidates based on poor expression and conservation, maintenance of synteny, biased sequence composition and conservation of secondary structure.

SRA Accession	Platform	Read length	Paired?	Experiment conditions
SRR3747530, SRR3747531	HiSeq 2500	125	Yes	M9 media, 5mM mannitol
SRR3747532, SRR3747533	HiSeq 2500	125	Yes	M9 media, 5mM mannitol, C_6-HSL
SRR3747538, SRR3747539	HiSeq 2500	125	Yes	M9 media, 5mM mannitol, $3OC_6$-HSL

Table 4.1: Publicly available transcriptomes from *Pta* ATCC 11528 (SRP078028), used to validate expression of conserved candidate ncRNAs identified in *Psa*. Data generated by Cheng *et al.* (2018).

SRA Accession	Platform	Read length	Paired?	Experiment conditions
SRR3757176, SRR3757178, SRR3757179	HiSeq 2000	101	Yes	KB media, exponential/stationary phase
SRR3757150, SRR3757153, SRR3757154	HiSeq 2000	101	Yes	KB media, exponential phase
SRR3757099, SRR3757100, SRR3757106	HiSeq 2000	101	Yes	KB media, lag phase

Table 4.2: Publicly available transcriptomes from *Pta* ATCC 11528 (SRP078136), used to validate expression of conserved candidate ncRNAs identified in *Psa*. Data generated by Cheng *et al.* (2017).

4.4 Results and Discussion

Manual annotation identified 94 intergenic and 137 antisense expression peaks as candidate ncRNAs. Rfam-guided annotations of known ncRNAs identified gene boundaries for 54 small ncRNAs, as well as a class of 46 ncRNA sequences with twoAYGGAY-like motifs. Rfam domains were also used to identify 56 group II intron-derived sequences, although it is unclear what proportion of these remain intact or functional.

4.4.1 TwoAYGGAY-like motifs in *Psa* and *Pseudomonas*

Seven intergenic peaks annotated as novel ncRNAs were identified near group II introns. These sequences had high sequence similarity to each other, Rfam-annotated twoAYGGAY motifs, as well as intergenic regions which were not expressed in this experiment. Overall 43 motifs were annotated: Rfam identified 12 motifs, which increased to 36 after removing the –cut-ga thresholding during annotation (E-value threshold = 0.01). Others were identified by an nhmmer homology search, using a HMM generated from an alignment of manually annotated twoAYGGAY-like sequences. An alignment of all twoAYGGAY and twoAYGGAY-like sequences had a highly stable Y-shaped predicted secondary structure, which is similar to the Rfam twoAYGGAY motif (Figure 4.4). Motif-containing sequences within the *Psa* genome are generally highly similar, and have covarying sites maintaining secondary structure.

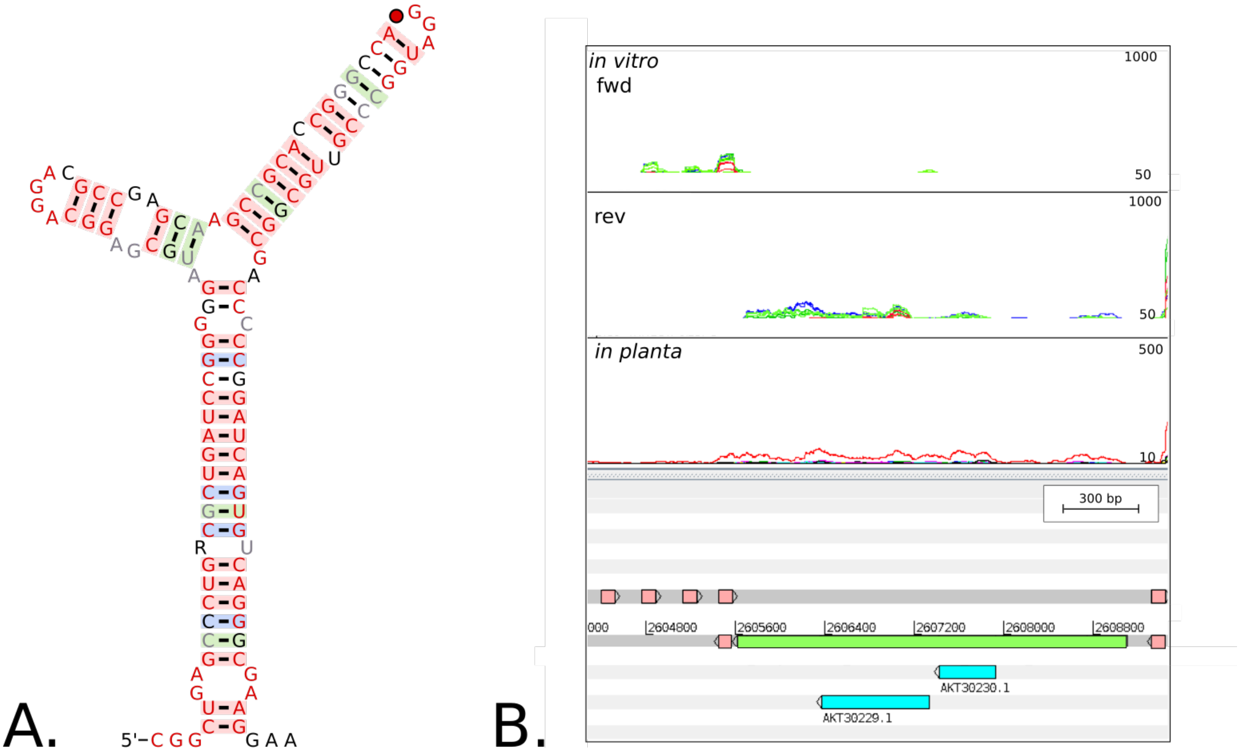


Figure 4.4: **(A)** Consensus predicted secondary structure of candidate ncRNAs homologous to annotated twoAYGGAY ncRNAs in the *Psa* genome. Candidates were aligned using `mafft-qinsi` and consensus secondary structure predicted with `RNAalifold`. **(B)** Example showing tandem repeats of the twoAYGGAY-like motif in *Psa*. Bottom panel shows genome annotations. **Pink:** Motifs, **Green:** Annotated group II intron, **Blue:** Protein annotations. Top panel shows RNA-seq counts for *in vitro* and *in planta* data-sets. A set of 4 tandem repeats can be seen on the forward strand at the 5′ end of the group II intron, and a single motif at the 3′ end. Motif annotations closest to the ends of the group II intron are palindromic, generating annotations on each strand. Annotations were visualised in the Artemis genome browser.

The *Psa* twoAYGGAY-like sequences do not have the 5′-AYGGAY-3′ sequence in the terminal loops, but do contain 5′-GGA-3′ motifs, which is consistent with the binding site of Csr/Rsm-binding RNAs (Gardner and Eldai, 2015). TwoAYGGAY-like motifs were commonly found throughout the *Psa* genome in intergenic regions, often near group II introns. Several intergenic regions have tandem repeats of this motif, and 26 sequences were sufficiently palindromic that they were annotated in the same location on both strands.

Conservation analysis of the most highly expressed sequence which contained the twoAYGGAY-like motif (CP011972.2, 1897172-1897363) found that this sequence is highly conserved throughout *Pseudomonas* genomes. Consensus secondary structure of sequences from the homology search is generally similar to the Rfam twoAYGGAY predicted secondary structure, but has a shorter and more variable P1 stem, a shorter P2 stem, small unpaired bulges, and a 5′-AUGGAU-3′ motif in the terminal loops (Figure 4.5).

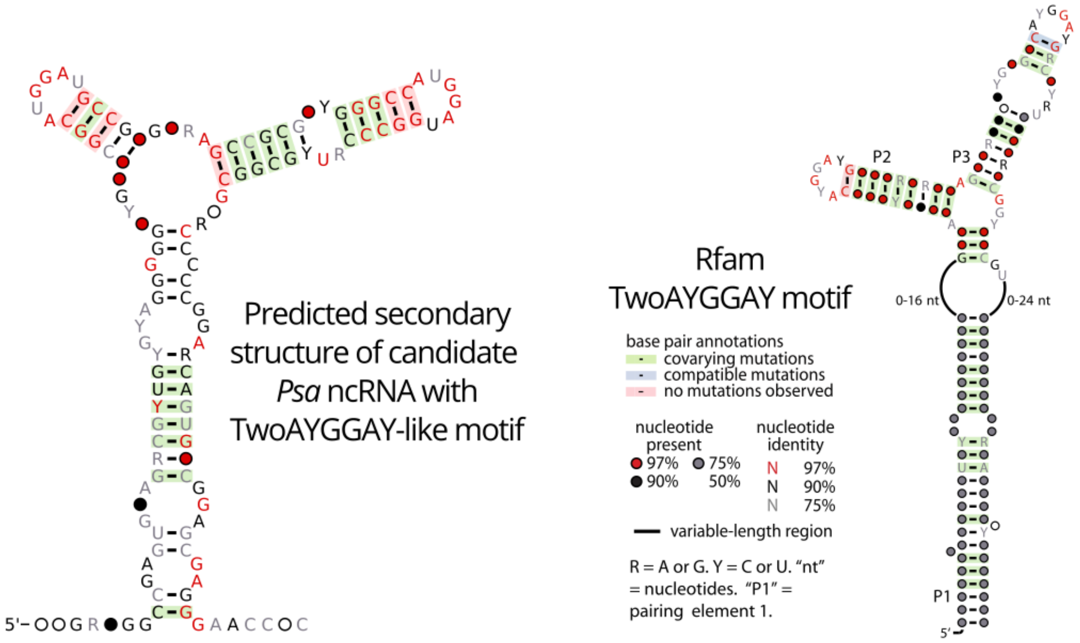


Figure 4.5: Left: Predicted consensus secondary structure of twoAYGGAY-like motifs in *Pseudomonas* genomes. An alignment was generated by an iterative nhmmer homology search from a single twoAYGGAY-like sequence over Pseudomonas genomes. **Right:** Rfam consensus secondary structure for the twoAYGGAY motif.

The function of these motifs are unclear, although the conservation of secondary structure and Csr/Rsm binding motif suggests they may function in the sequestration of Csr/Rsm RNA-binding proteins (Gardner and Eldai, 2015). These motifs were also associated with group II introns (p = 0.008, Fisher's exact test, Table 4.3).

Group II introns have been observed to insert near palindromic motifs in *P. putida* KT2440 (Yeo *et al.*, 2001), and similar palindromic elements, including ncRNAs, are often insertion sites for other mobile genetic elements (Dai and Zimmerly, 2002, Darmon and Leach, 2014). It may be that this sequence is associated with the re-homing of introns, however, the mechanisms of re-homing and transposition of bacterial group II introns are diverse and not well understood (Dai and Zimmerly, 2002), and their recognition sites are highly variable (Lambowitz and Zimmerly, 2011).

	TwoAYGGAY	Group II intron
Adjacent	28	21
Independent	15	35
Total	43	56

Table 4.3: Contingency table showing the association of twoAYGGAY-motifs with group II introns. These are classified into 'adjacent', where a twoAYGGAY annotation is in an intergenic region up- or down-stream of a group II intron, and 'independent' where flanking gene annotations are not part of group II introns. For regions containing multiple clustered twoAYGGAY repeats, all repeats in the same intergenic region are classed as adjacent or independent.

4.4.2 Intergenic candidate ncRNAs

Overall 71 annotated intergenic candidates were selected for conservation analysis. Quality control of the final alignments of homologous sequences removed 21 candidates that had predicted protein coding potential, 26 conserved ncRNAs with a positive `alifoldz` z-score, and 7 candidates that did not have thermodynamically stable predicted secondary structures. In total, 36 candidate ncRNAs passed alignment quality control. The conservation of these sequences and preservation of synteny was used to attempt to classify gene origins. During manual curation 4 candidates were classed as pseudogenes due to homology with proteins and tRNAs in other genomes and were removed, and another 4 were removed due to inconsistent conservation of secondary structure.

Most candidate ncRNAs were poorly conserved; 7 were specific to *Psa*, and another 7 associated with mobile genetic elements were primarily found in Psa with a small number of annotations to sequences near the same MGEs present in other species. Many poorly conserved candidates were located in regions created by recent re-arrangements in the *P. syringae* or *Psa* lineages. Several of these had highly stable predicted secondary structures, however these were part of highly repetitive and palindromic intergenic regions, which may be remnants of previous transposition events. While rearrangements can cause the creation of novel transcripts, it is difficult to identify if these are functional ncRNAs or transcriptional noise (Jose *et al.*, 2019). Other poorly conserved candidate ncRNAs were associated with mobile genetic elements such as transposons and integrases, reiterating results from the conservation analysis of *Salmonella* sRNAs in Chapter 2.

Name	Sequence	Start	End	Strand
Psa_sRNA_003	CP011973.1	21610	21774	-
Psa_sRNA_004*	CP011973.1	61084	61315	-
Psa_sRNA_005*	CP011973.1	61080	61307	+
Psa_sRNA_006	CP011972.2	170583	170708	-
Psa_sRNA_007*	CP011972.2	806901	807143	+
Psa_sRNA_009	CP011972.2	6082989	6083300	-
Psa_sRNA_010*	CP011972.2	6513384	6513635	-
Psa_sRNA_011	CP011972.2	743622	743984	+
Psa_sRNA_012	CP011972.2	3765999	3766298	-
Psa_sRNA_013*	CP011972.2	1752191	1752476	+
Psa_sRNA_014	CP011972.2	3260267	3260449	+
Psa_sRNA_015	CP011972.2	739233	739548	-
Psa_sRNA_016	CP011972.2	760028	760267	+
Psa_sRNA_017	CP011972.2	2294607	2294819	-
Psa_sRNA_018	CP011972.2	5148068	5148376	+
Psa_sRNA_019	CP011972.2	768152	768370	+
Psa_sRNA_020*	CP011972.2	2035568	2035765	+
Psa_sRNA_021*	CP011972.2	119433	119810	+
Psa_sRNA_022	CP011972.2	119724	119882	-
Psa_sRNA_023	CP011972.2	1494687	1495064	-
Psa_sRNA_024*	CP011972.2	4634388	4634555	-
Psa_sRNA_025*	CP011972.2	5040669	5040833	-
Psa_sRNA_026†	CP011972.2	1300899	1301216	-
Psa_sRNA_027*	CP011972.2	745275	745542	-
Psa_sRNA_028*†	CP011972.2	322030	322306	+
Psa_sRNA_029*	CP011972.2	5438181	5438417	-
Psa_sRNA_031†	CP011972.2	5459360	5459656	+
Psa_sRNA_032†	CP011972.2	5469984	5470208	+

Table 4.4: Final candidate ncRNAs and genome locations. * indicates that this transcript is either adjacent to or part of a mobile genetic element. † denotes candidates that were found to be expressed in *Pta*.

Final candidates (Table 4.4) were ranked by conservation, overall expression, expression in other data-sets, the presence of condition-specific expression and differential expression (Figure 4.6). Three were expressed only *in vitro*, four were only expressed *in planta*, two in minimal media only and one in rich log samples only. One candidate Psa_sRNA_012, was significantly differentially expressed across *in vitro* data-sets ($P_{adj} < 0.01$).

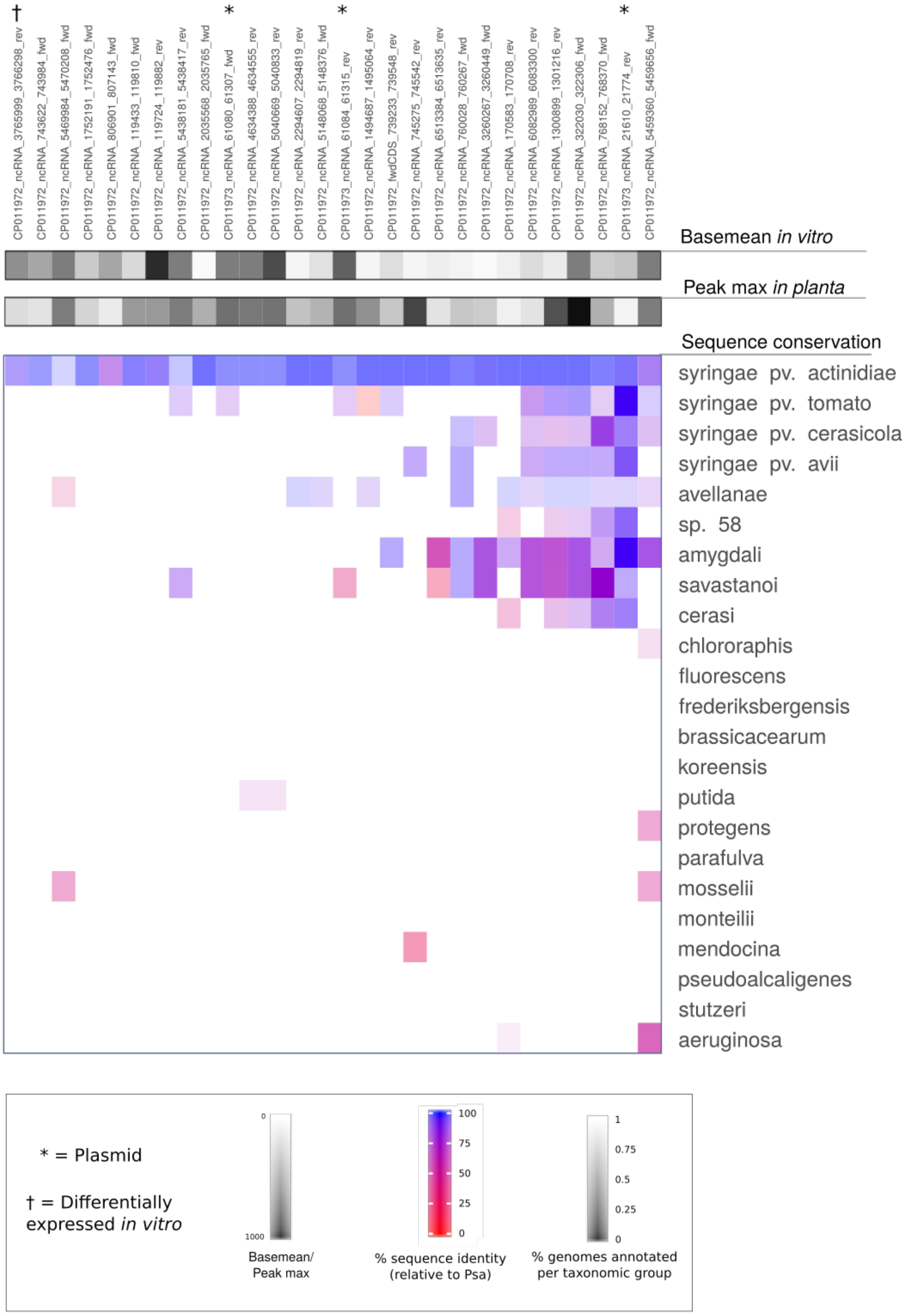


Figure 4.6: Heatmap showing the annotation range and sequence conservation of intergenic candidate ncRNAs annotated in *Psa* across *Pseudomonas* genomes. Sequence conservation is shown as a colour gradient from blue (100%) to red (40%), representing genus average percent sequence identity based on alignment to the sequence from *P. syringae* pv. *actinidiae* ICMP 18884. Gene conservation is shown as a change in opacity, represented by the percentage of genomes with an annotation within that taxonomic group. Genes are ordered based on overall conservation. Additional panels show expression data for these genes. **Top:** Basemean from *in vitro* RNA-seq calculated by DEseq2. **Middle:** Peak max of mapped reads from *in planta* plot files, which is the maximum combined read depth from all *in planta* plot files across the annotated region. Both panels are capped at 1000. * Denotes candidate ncRNAs located in the *Psa* ICMP 1884 plasmid, † denotes candidates that are differentially expressed *in vitro*

Candidates that were associated with mobile genetic elements were in general more highly expressed in starvation conditions. High expression of the protein components of these MGEs was also observed in the same conditions (Chapter 3), indicating these transcripts may be under the same regulatory mechanisms.

Psa_sRNA_022 was highly expressed under starvation conditions and *in planta*, and found next to a hypothetical protein. Additional annotation of this hypothetical protein found homology to the Abi/CAAX family of transmembrane metalloproteases (Kjos *et al.*, 2010), which are involved in self-immunity to bacteriocins, and have been found as part of a cassette of resistance genes carried by toxin-antitoxin systems in *Staphylococcus* (Bukowski *et al.*, 2017). A small expression peak adjacent to this candidate was also detected. Annotations in strains outside *Psa* found that Psa_sRNA_022 was located at the 5′ end of a CAAX protease. This locus appears to have been disrupted in *P. syringae* due to an integration or transposition event (Figure 4.7). This transcript had a conserved secondary structure with a GNRA tetraloop, which are structural motifs that can form and stabilise tertiary structure (Fiore and Nesbitt, 2013).

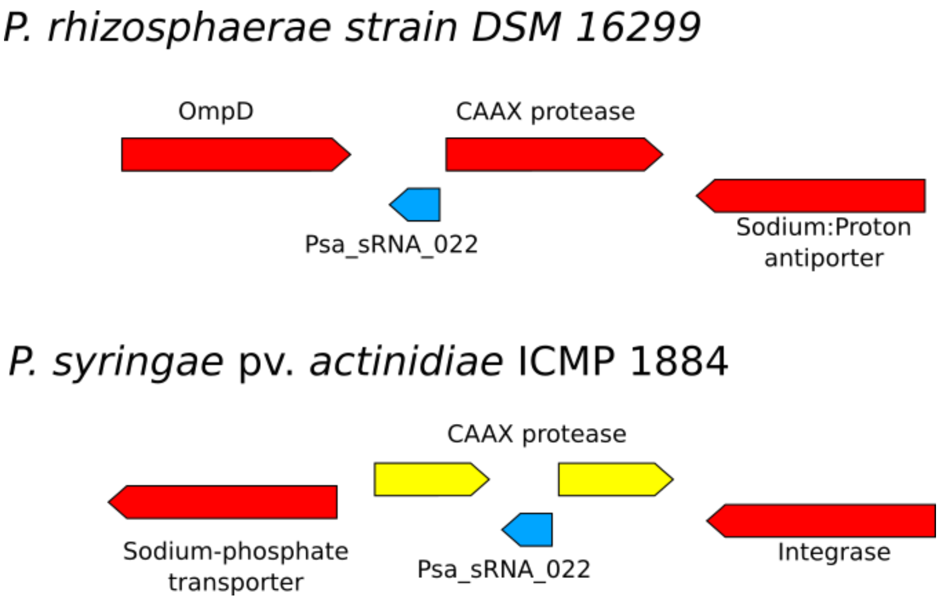
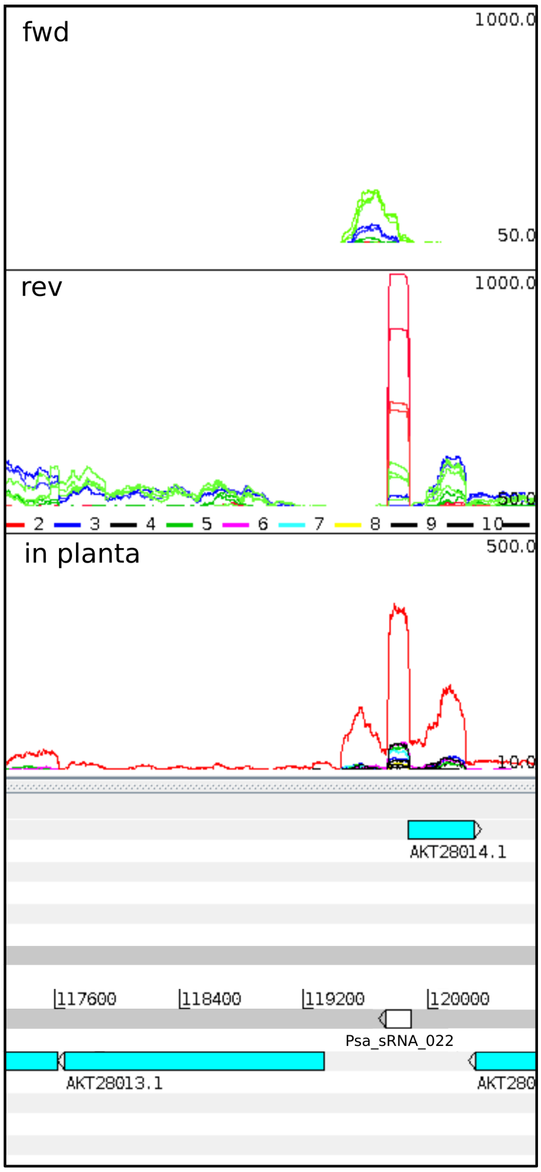

P. rhizosphaerae strain DSM 16299
OmpD
CAAX protease
Psa_sRNA_022
Sodium:Proton antiporter
P. syringae pv. actinidiae ICMP 1884
CAAX protease
Sodium-phosphate transporter
Psa_sRNA_022
Integrase

Figure 4.7: Comparison of the region containing Psa_sRNA_022 sequence in *Psa* and a homologous sequence in *P. rhizosphaerae* strain DSM 16299. **Top:** *P. rhizosphaerae* locus containing Psa_sRNA_022 (blue), next to a CAAX protease. **Bottom:** *Psa* locus containing Psa_sRNA_022, which is located between two hypothetical proteins (yellow) with homology to CAAX protease.



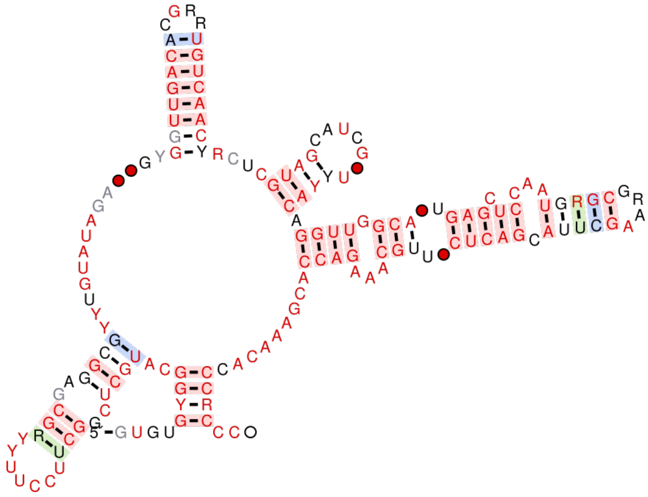
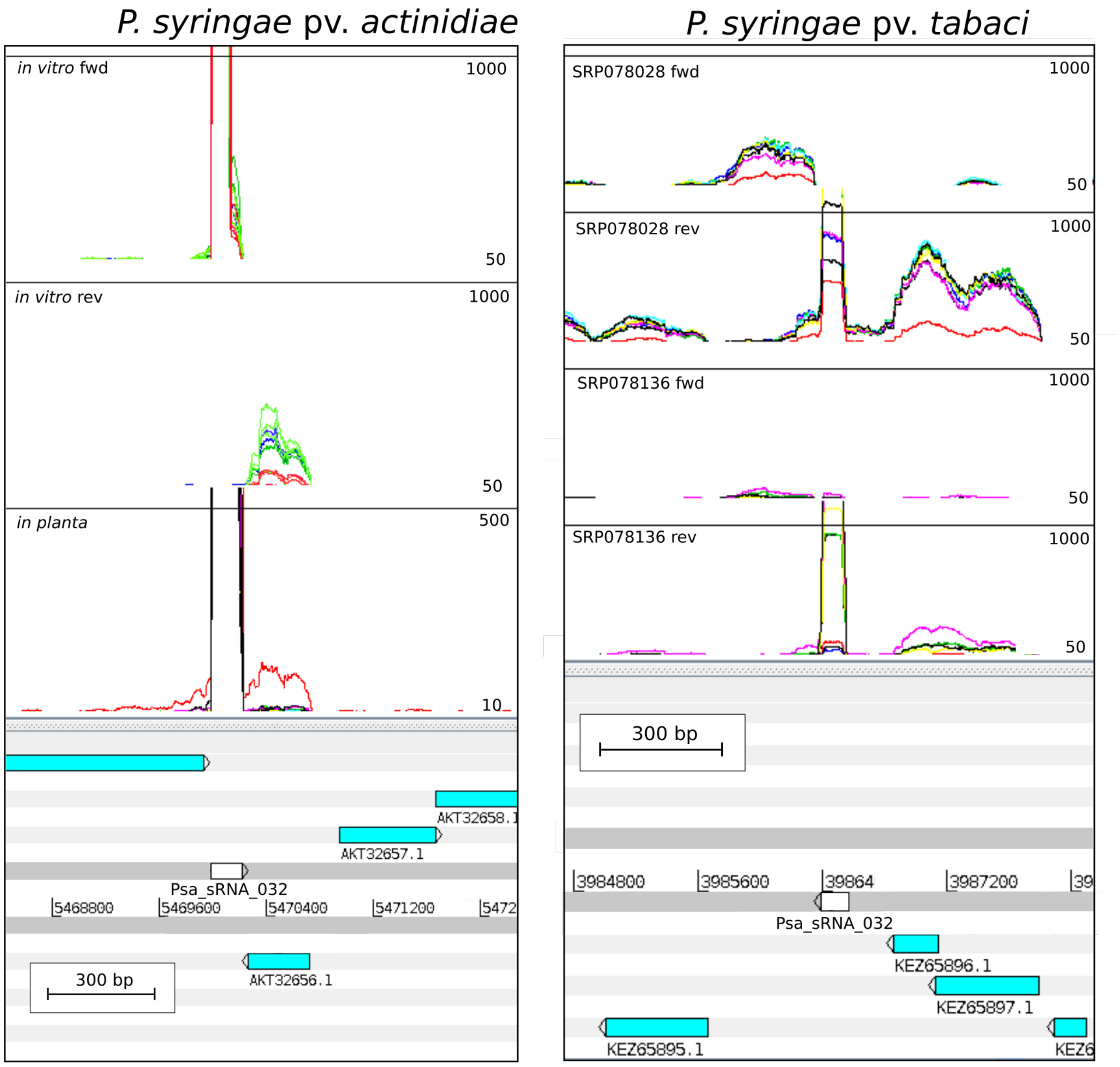


Figure 4.8: **Left:** Expression peak for Psa_sRNA_022. This region is highly expressed in starvation conditions *in vitro* and is expressed *in planta*. **Right:** Predicted consensus secondary structure for Psa_sRNA_022.

Psa_sRNA_032, located downstream of a plasmid partitioning protein ParA and conjugation machinery, was highly constitutively expressed in *Psa* and all *Pta* samples (Figure 4.9). This candidate was found primarily in *P. syringae* pv. *actinidiae*, and was homologous to intergenic regions near mobile genetic elements in 3 genomes outside of *P. syringae*. This transcript appears to have a stable secondary structure, containing conserved stem-loops (Figure 4.9). The context and high constitutive expression of this transcript suggests it may be a regulator or part of a Par plasmid-maintaining toxin-antitoxin system (Muthuramalingam *et al.*, 2019).



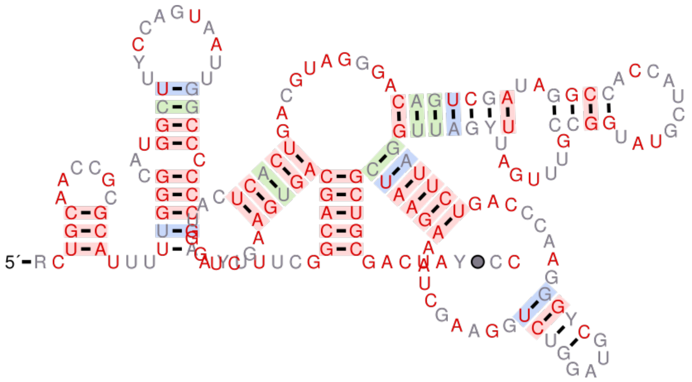


Figure 4.9: Top: Expression of Psa_sRNA_032 in *Psa* and expression of a homologous sequence in *Pta*. Psa_sRNA_032, which is in a locus that is conserved in both genomes is highly constitutively expressed in *Psa* grown *in vitro* and *in planta* (**left**), and also expressed in two data-sets of *Pta* grown *in vitro* (**right**). Peaks are smaller in *Pta* due to variations in read depth between data-sets. **Bottom:** Predicted consensus secondary structure for Psa_sRNA_032.

Psa_sRNA_006, which is weakly expressed *in vitro* and *in planta*, was found in multiple Pseudomonas species and had a highly conserved secondary structure (Figure 4.10), containing two GNRA tetraloops and a highly significant covarying base-pair (E-value = 0.0003). This candidate was located near siderophore export channels in *Psa*, and was homologous to a horizontally-acquired region in *P. aeruginosa* near the type IV pili regulator *hrpB*.

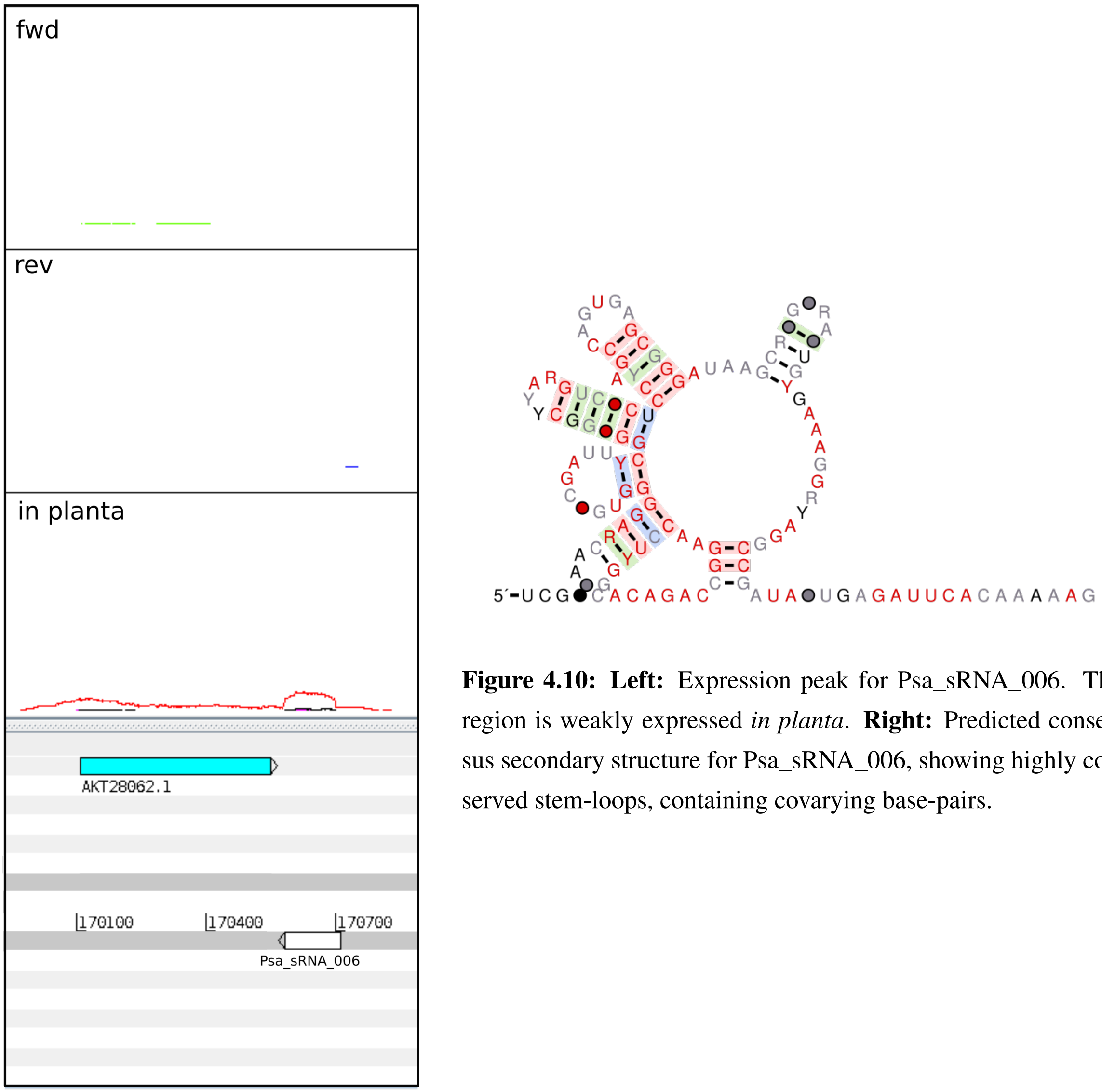


Figure 4.10: Left: Expression peak for Psa_sRNA_006. This region is weakly expressed *in planta*. **Right:** Predicted consensus secondary structure for Psa_sRNA_006, showing highly conserved stem-loops, containing covarying base-pairs.

Psa_sRNA_026 was conserved throughout *P. syringae* and found in several closely related Pseudomonas species. This candidate was only expressed in minimal media *in vitro* and was highly expressed *in planta*, and is located in a conserved intergenic region located upstream of alginate biosynthesis genes, and next to a metal-sensing peroxide stress protein YaaA (Liu *et al.*, 2011). The predicted consensus secondary structure for Psa_sRNA_026 was highly structured, however several peripheral stem-loops were not highly conserved (Figure 4.11).

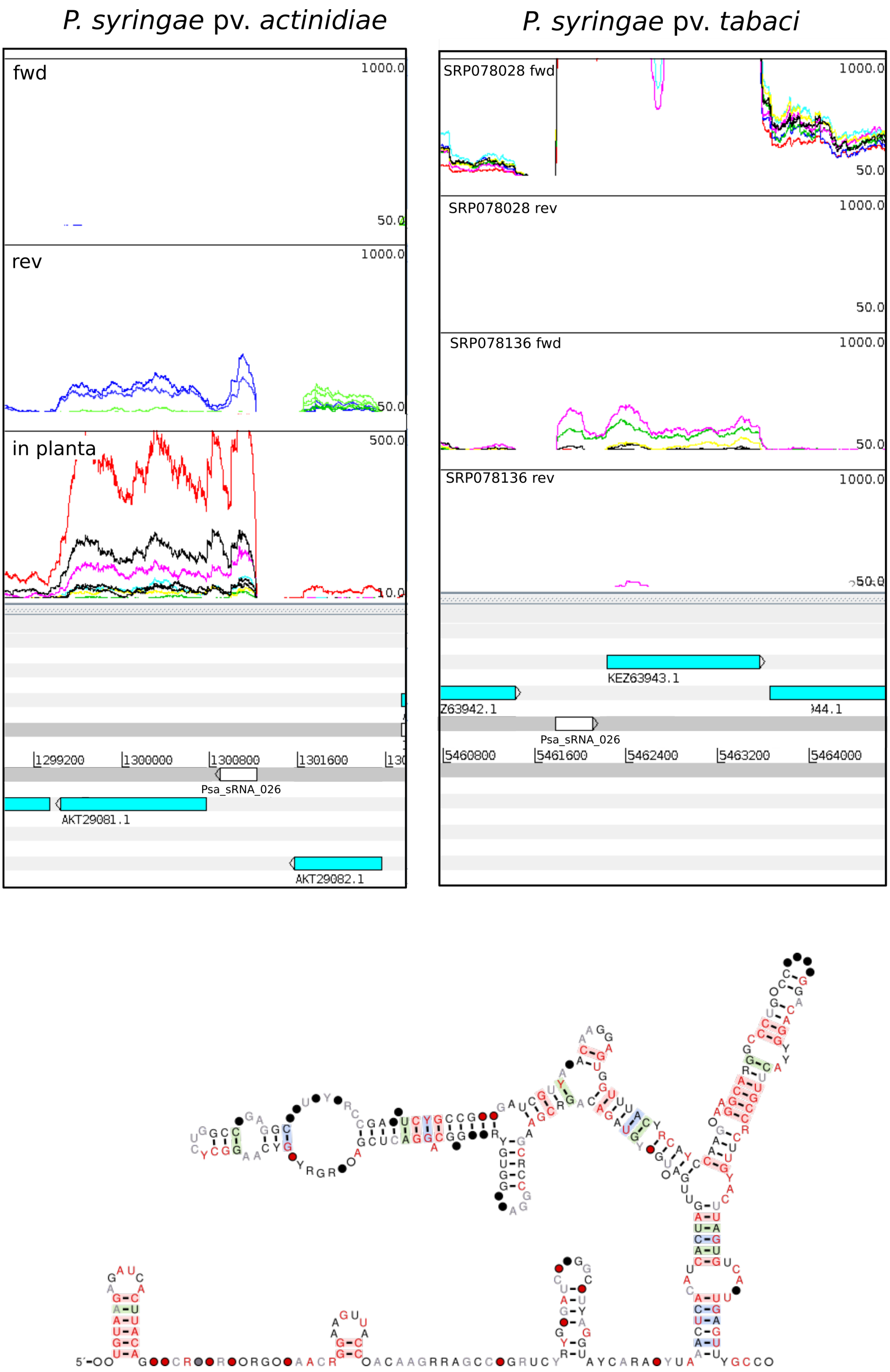


Figure 4.11: Top: Expression of Psa_sRNA_026 in *Psa*, where it is expressed in minimal media *in vitro*, and *in planta*. Expression in *Pta* is also shown. **Bottom**: Predicted consensus secondary structure for Psa_sRNA_026.

4.4.3 Psa_sRNA_031 is a *Pseudomonas syringae pesA* homologue

Homology search results showed that Psa_sRNA_031 was conserved across *Pseudomonas*, including several strains of *P. aeruginosa*, a well-studied human pathogen and model organism. The *Pseudomonas* transcriptome browser by Wurtzel *et al.* (2012) was used to identify that an intergenic region in *P. aeruginosa* str. PA14 that is homologous to Psa_sRNA_031 was highly expressed (Figure 4.12) and has been identified as a small RNA. This transcript, named PesA (previously known as SPA0021), is located in the PAPI-4 pathogenicity island found in many pathogenic strains of *P. aeruginosa*, and may act in *trans* to attenuate the expression of the bacteriocin, pyocin S3 (Ferrara *et al.*, 2017). PesA also improves survival and stress tolerance in response to UV, however the mechanism for this is unclear (Ferrara *et al.*, 2017).

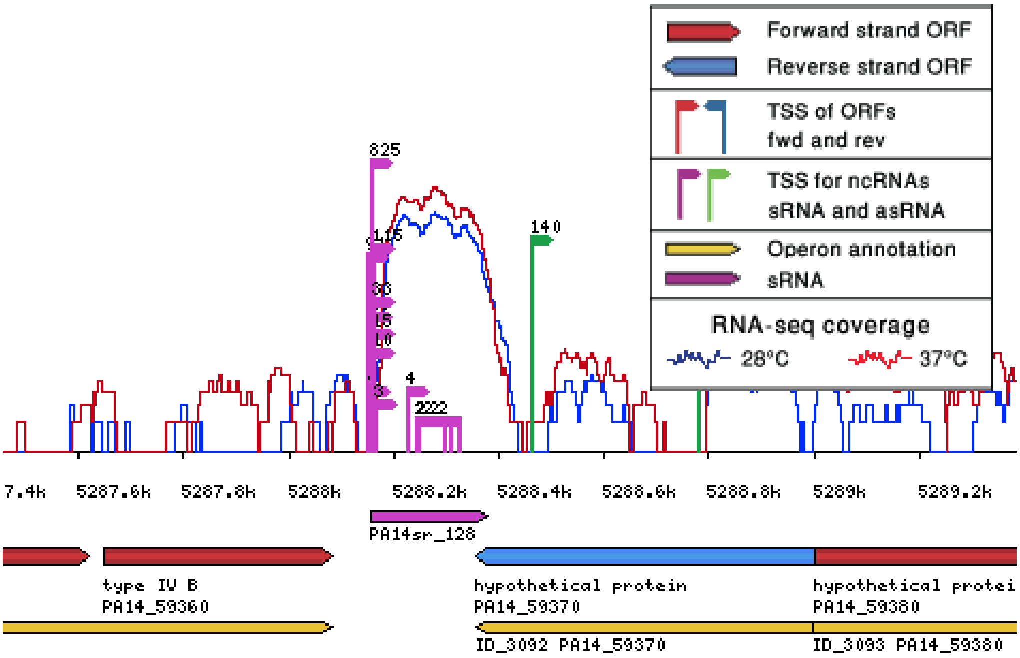


Figure 4.12: Data from the Pseudomonas RNA-seq browser (Wurtzel *et al.*, 2012)), showing expression of a region homologous to Psa_sRNA_031 the in *P. aeruginosa* str. PA14 (PA14_sr_128). This region overlaps with the annotated sRNA *pesA*, located at 5288100-5288500 in this genome (Pita *et al.*, 2018). The Pseudomonas RNA-seq browser is located at *http : //www.weizmann.ac.il/molgen/Sorek/pseudomonas_browser/*

The expression of *pesA* appears to be regulated differently in *P. syringae* strains than in *P. aeruginosa*. PesA expression is induced at body temperature in *P. aeruginosa*, however it is constitutively expressed across the *Psa* and *Pta* RNA-seq data-sets used in this study.

Conservation analysis and structural prediction showed that both the nucleotide sequence and secondary structure of PesA are conserved across *Pseudomonas* genomes (Figure 4.13), with an average sequence identity of 89% and lowest sequence identity of 71% between *Psa* and *P. protogens* sequences. A sequence region of *pesA* that has been predicted to be the interaction site with the pyocin S3 mRNA ribosomal binding site in *P. aeruginosa* (UUCUCCCUGUGUCUCCCUGUUCUUUGCUUGUC) was conserved across all sequences, with some variability in the 3′ end (Figures 4.13 and 4.14).

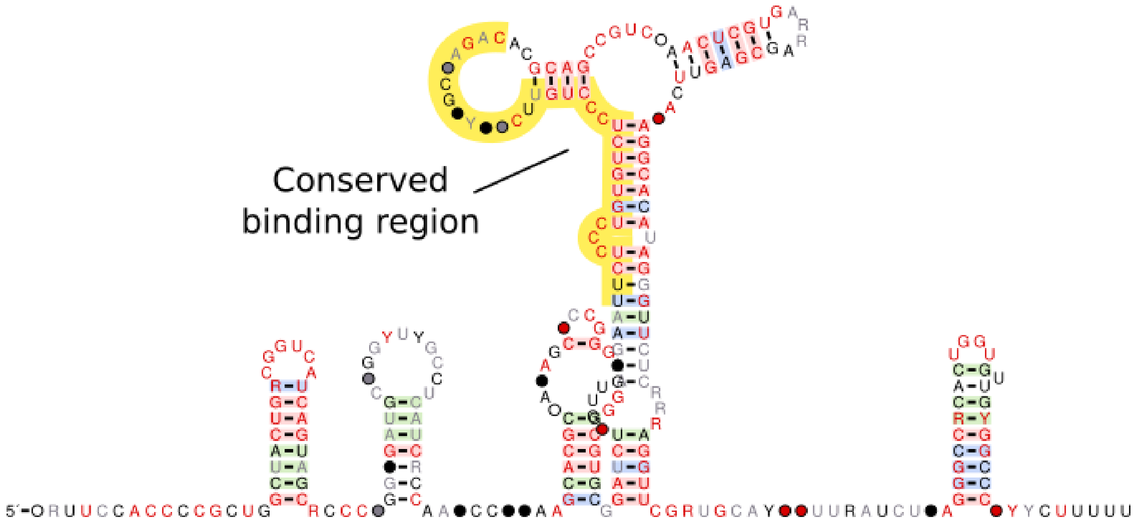


Figure 4.13: Predicted consensus structure of PesA with the experimentally confirmed mRNA binding region from *P. aeruginosa* highlighted (yellow).

In *P. aeruginosa* and *P. protegens*, *pesA* is found next to genes encoding Type IV pili, however in *P. syringae* and other *Pseudomonas* plant pathogens (*P. avenellae*, *P. amygdali* and *P. cerasicola*) *pesA* is located next to a DNA topoisomerase III in PacICE1, a integrative conjugative element circulating among *Pseudomonas* in the Pacific (Colombi, 2017, Colombi *et al.*, 2017).

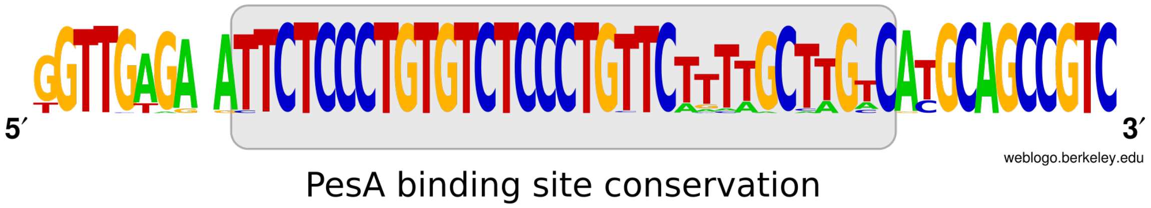


Figure 4.14: Sequence logo showing nucleotide conservation across the PesA binding site. Image generated by WebLogo (https://weblogo.berkeley.edu/logo.cgi)

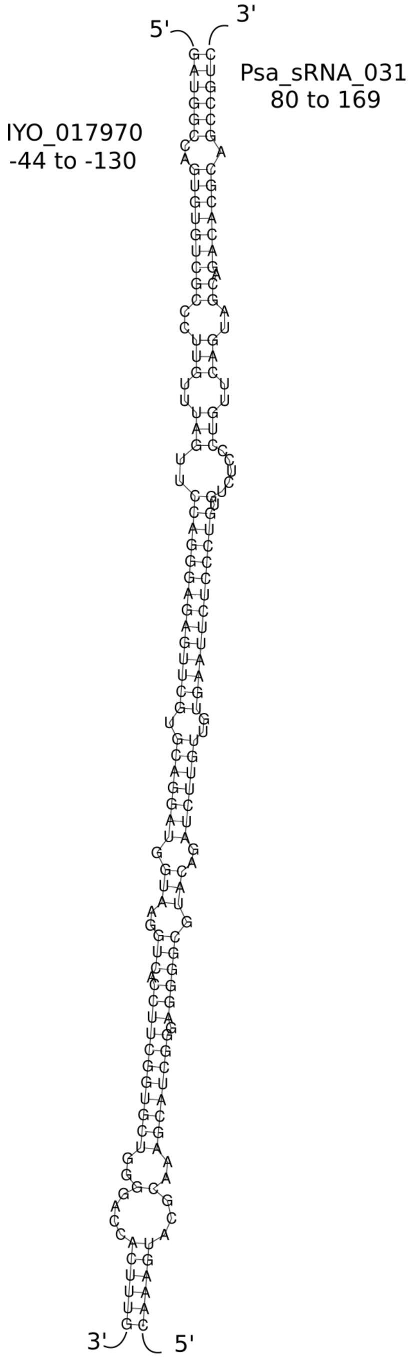


Figure 4.15: Visualisation showing the top predicted interaction between Psa_sRNA_031 (*pesA*) and *Psa* mRNAs generated by CopraRNA (Wright *et al.*, 2014, 2013). Interacting regions of the predicted conserved binding region of PesA and a region upstream of the start codon of IYO_017970 are shown. Figure generated with RNAplot (Lorenz *et al.*, 2011).

To investigate if *pesA* functions in pathogenicity in *P. syringae*, CopraRNA (Wright *et al.*, 2014, 2013) was used to predict conserved interactions between PesA and mRNA sequences in 4 *P. syringae* strains, *P. syringae* pv. *actinidiae* ICMP 18884 (NCBI accession: NZ_CP011972), *P. syringae* pv. *syringae* B728a (NCBI accession: NC_007005), *P. syringae* pv. *tomato* str. B13-200 (NCBI accession: NZ_CP019871) and *P. syringae* pv. *atrofaciens* str. LMG5095 (NCBI accession: NZ_CP028490). CopraRNA high scoring interactions (Table 4.5) include stress genes that are required for survival *in planta*, indicating that PesA may have an analogous role in promoting survival in plant hosts. The top-scoring interaction with an arylesterase, an enzyme which is involved in the degradation of carboxylic and phenolic esters (Wang *et al.*, 2010), which are present in many plant-derived compounds such as salicyclic acid and lignin. Phenolic compounds have been shown to induce the expression of *hrp* virulence and T3SS genes in *P. syringae* pv. *tomato* (Lee *et al.*, 2015). The ArnA interaction with PesA was predicted to interact with the previously identified PesA binding site (Figure 4.15). Other targes include UDP-glucoronic acid oxidase ArnA, which is involved in LPS synthesis (Breazeale *et al.*, 2002), and PhnA, which is involved in the uptake of alkylphosphonates, which are phosphate-containing compounds that include some fungicides and herbicides (Chen *et al.*, 1990). Other targets include the ribosomal protein RpsO, which binds to the 16S RNA and has been found to interact with sRNAs in *E. coli* (Fontaine *et al.*, 2016), and UspA, which is a conserved components of many stress responses (Kvint *et al.*, 2003).

CopraRNA p-value	Gene	Energy [kcal/mol]	Functional annotation of gene
9.43E-06	IYO_017970	-24.38	arylesterase
9.63E-05	IYO_016195	-26.59	UDP-glucuronic acid oxidase ArnA
0.000147	IYO_022295	-22.23	putative RNA-binding protein
0.000193	IYO_022725	-21.15	30S ribosomal protein S15 rpsO
0.000195	IYO_010790	-16.45	universal stress protein UspA
0.000239	IYO_018865	-19.45	hydroxyacylglutathione hydrolase
0.000248	IYO_021105	-20.62	methionine–tRNA ligase
0.00025	IYO_009170	-20.91	Free methionine-(R)-sulfoxide reductase
0.000275	IYO_014440	-18.96	DUF1993 domain-containing protein
0.000283	IYO_023490	-20.12	Alkylphosphonate utilization operon protein PhnA

Table 4.5: Top 10 PesA-mRNA interactions predicted by CopraRNA. CopraRNA p-value, gene and protein annotations, and predicted thermodynamic stability of the interaction are shown.

4.4.4 Antisense ncRNAs

Antisense transcription peaks were found for 137 genes (full list provided in Supplementary Materials). Most antisense peaks were short (~300 nt) and had low expression relative to their counterpart genes. Antisense transcripts appear to be abundant in bacterial transcriptomes (Georg and Hess, 2011), however, they are often poorly conserved and few have been functionally characterised (Cech and Steitz, 2014), leading to scepticism about what proportion of such transcripts act as functional ncRNAs (Lloréns-Rico *et al.*, 2016). Only 4 antisense transcripts were differentially expressed *in vitro* (Table 4.6), which were all more highly expressed in starvation conditions and down-regulated in minimal media.

Fifteen antisense transcripts were only expressed in specific growth conditions (Table 4.7). These were mostly expressed in rich media in log phase, and were antisense to several genes important for *Psa* pathogenicity and survival.

In rich media, antisense transcripts were mainly found opposite genes involved in iron metabolism, recombination and transposition. Antisense transcripts expressed in rich media in log phase were opposite genes involved in recombination and transposition, the biosynthesis of haem-containing molecules, as well as the haem-sensing TonB-dependent receptor, a membrane protein part of a lytic phage (IYO_005850), and a lipoprotein (IYO_003010) located next to anti-microbial peptide transporters. In rich media in log/late log phase, a transcript was found opposite a hypothetical protein, which was located in a locus containing tellurium resistance genes that may be involved in phage suppression (Anantharaman *et al.*, 2012). These may be involved in the regulation of iron homeostasis and the suppression of lytic phages and transposons.

Start	End	Strand	L_2FC (RvM)	L_2FC (MvS)	Opposite gene
2734548	2734968	-	**-1.8**	**2.0**	IYO_012460 enoyl-CoA hydratase
2385543	2385788	-	**-2.9**	-	IYO_011015 GNAT family acetyltransferase
4143876	4144167	-	-	**1.8**	IYO_018595 FAD-binding molybdopterin dehydrogenase
4614055	4614566	-	-	**1.2**	IYO_020700 peptide ABC transporter ATP-binding protein

Table 4.6: Antisense peaks in the *Psa* (CP011972.2) genome that are differentially expressed *in vitro*. Significant $\log_2$ fold changes ($p < 0.01$) are included for rich vs minimal and minimal vs starved media comparisons.

In minimal media and starved samples, antisense transcripts were found for genes involved in carbon metabolism, DNA repair and siderophore transport. In minimal media, a transcript was found opposite a glycine transporter (Figure 4.16). In starvation conditions, transcripts were found antisense to a transcriptional regulator of genes involved in the TCA cycle, as well as maltodextrin phosphorylase, which is involved in carbohydrate metabolism (Watson *et al.*, 1997). Starvation-specific antisense transcripts were also found for a transposon near the HopAM1-1 effector, and the DNA-repair protein mutL (Ban and Yang, 1998).

Differentially expressed antisense transcripts were opposite genes involved in acetyl-CoA synthesis, a xanthine dehydrogenase, which contains iron-sulphur clusters, and a peptide transporter (Table 4.6).

The changes in expression for the annotated antisense transcripts broadly reflect changes in gene expression discussed in Chapter 3, where rich media is characterised by growth and the suppression of siderophore production, and nutrient acquisition and the pivoting of carbon metabolism to the TCA cycle in minimal media and starvation samples.

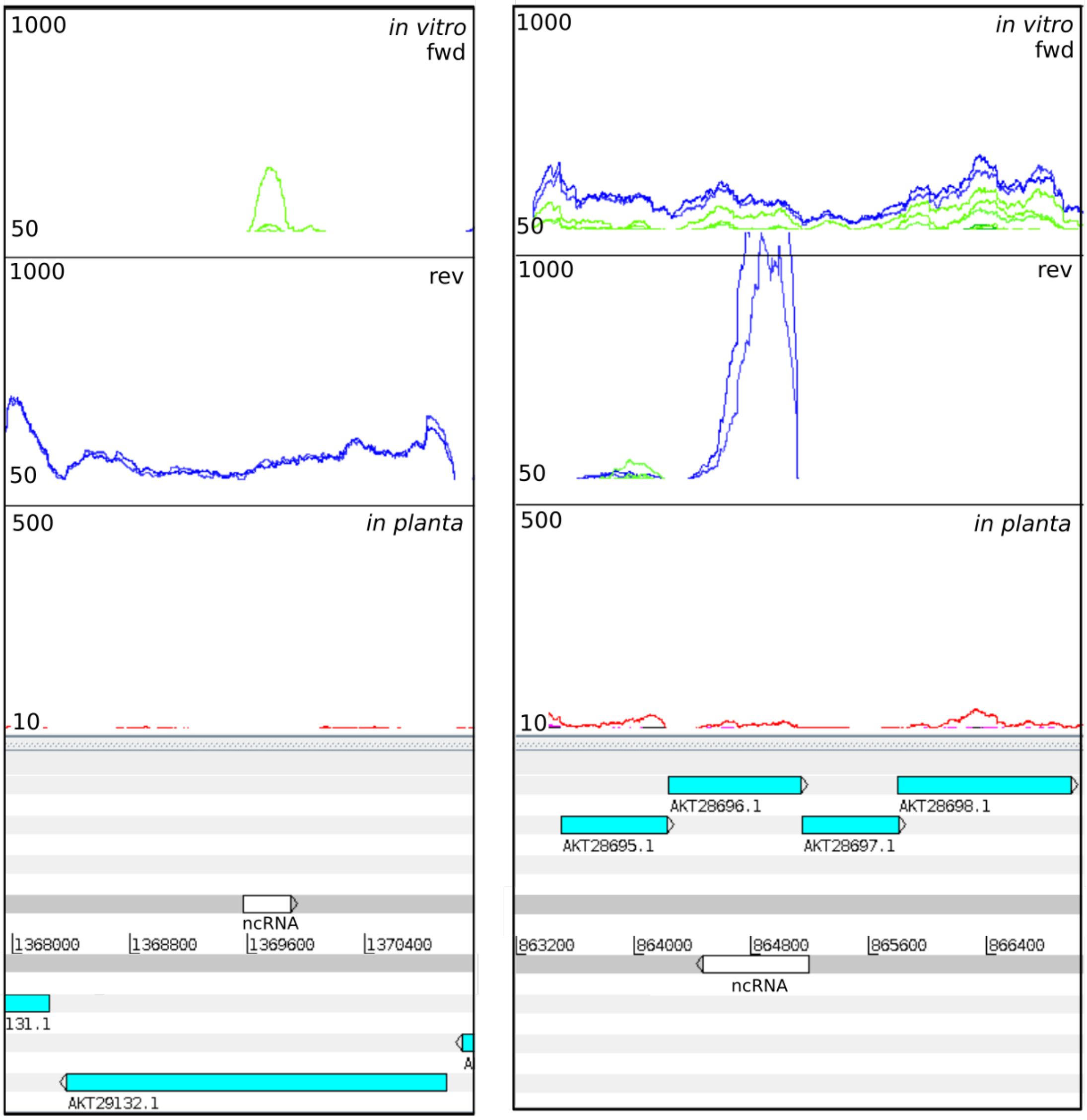


Figure 4.16: Examples of condition-specific antisense transcription in *Psa*. Read depth is visualised as plot files for each strand, which are coloured according to growth condition for the sample. **Left**: Antisense transcription opposite a TonB-dependent receptor (IYO_006325) occurring only in rich media samples in log phase. **Right**: Antisense transcription opposite a glycine ABC transporter (IYO_004105) occurring only in minimal media samples.

One antisense ncRNA spanning the 3$'$ of *ureA* gene and the 5$'$ of *pat* was highly expressed in *Psa*, particularly under starvation conditions, and was also expressed in *Pta*. A 5$'$ *ureB* transcript has previously been discovered in *Helicobacter pylori* (Wen *et al.*, 2011), where it acts to regulate *ureB* and *ureA* expression in neutral pH, when the requirement for urease is reduced. Interestingly, an Rfam search annotated a region antisense to *ureC* as homologous to the *H. pylori* 5$'$ ureB antisense RNA, however this region was not expressed under any RNA-seq conditions in this study.

Start	End	Strand	Opposite gene	Expressed in
5288805	5289233	-	IYO_023770 hypothetical protein	Rich late log
4886574	4887107	-	IYO_021900 membrane protein	Rich log
4791323	4791737	+	IYO_021485 DNA recombination protein RecO	Rich log
4478352	4479341	-	IYO_020055 spermidine/putrescine ABC transporter ATP-binding protein/IYO_020050 ABC transporter permease	Rich log
4296010	4296513	+	IYO_019195 RND transporter/IYO_019200 BNR/Asp-box repeat-containing protein	Rich log
4347701	4348589	+	IYO_019435 multifunctional fatty acid oxidation complex subunit alpha IYO_019440 hypothetical protein	Rich log
2522527	2522700	-	IYO_011620 sirohydrochlorin ferrochelatase	Rich log
1369572	1369895	+	IYO_006325 TonB-dependent receptor	Rich log
1245423	1245888	+	IYO_005850 membrane protein	Rich log
617339	617594	+	IYO_003010 lipoprotein	Rich log
864468	865185	-	IYO_004105 glycine/betaine ABC transporter substrate-binding protein	Minimal
2403507	2403680	-	IYO_011100 LacI family transcriptional regulator	Starved
5160963	5161205	+	IYO_023220 integrase	Starved
5601902	5602174	+	IYO_025370 DNA mismatch repair protein	Starved
5880993	5881310	-	IYO_026510 maltodextrin phosphorylase	Starved

Table 4.7: Antisense peaks in the *Psa* (CP011972.2) genome with condition-specific expression *in vitro*.

Urease is a highly conserved protein that functions in acid stress in aerobic bacteria, and is also a marker of virulence in fluorescent Pseudomonads (Bradbury *et al.*, 2014). The operon structure for the urease sub-units are significantly different between *Psa* and *H. pylori* (Figure 4.17). In *P. syringae* this locus has an insertion of two acetyltransferases; a tabtoxin-resistance protein, which acts to protect tabtoxin-producing *P. syringae* strains from their own phytotoxin (Wencewicz and Walsh, 2012), and a phosphinothricin N-acetyltransferase, which reduces toxicity of phosphinothicin, a phytotoxin and common herbicide (Davies *et al.*, 2007). It is unclear if this *Psa* antisense transcript has the same function as the *H. pylori ureB* antisense transcript, however the *pat* insertion between *ureA* and *ureB* is conserved across *Pseudomonas* genomes (Davies *et al.* (2007), $http://www.pseudomonas.com/orthologs/list?id = 112598$).

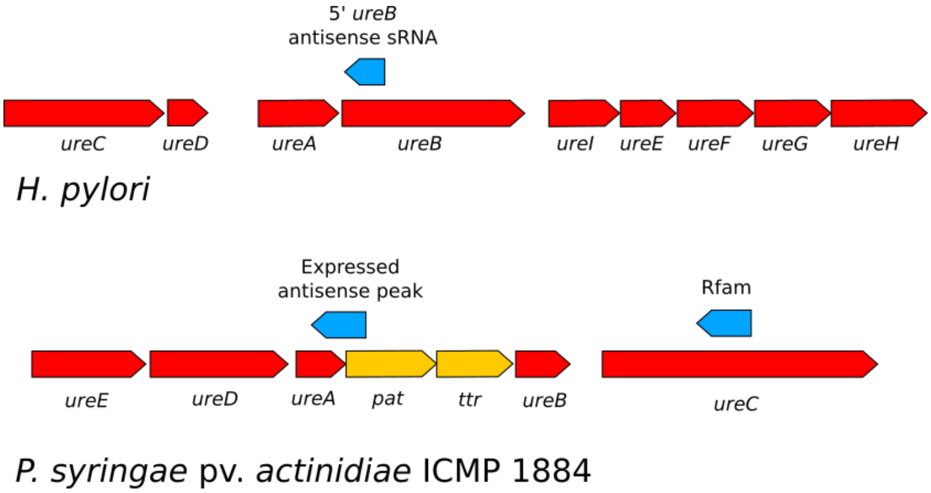


Figure 4.17: Structure of the urease operons in *Helicobacter pylori* (De Reuse *et al.*, 1997) and *Psa* (based on NCBI genome annotations), showing an insertion in between the *ureA* and *ureB* genes in *Psa*. **Blue**: Antisense sRNA annotations. **Red**: Components of the urease operon. **Yellow**: Toxin-resistance acetyltransferases *pat* and *ttr*.

Antisense expression in other *P. syringae* strains

Expression of antisense peaks in *Psa* were generally not conserved, as only 18/64 homologous sequences in *Pta* showed expression > 50 counts across in either of the RNA-seq data-sets; only 5 of these were consistent across both data-sets. Two antisense peaks showed high expression, which were opposite a peptidase (IYO_026530 in CP011972.2, and C1E_0209610 in ACHU02.1) and an AraC transcriptional regulator (IYO_019675 in CP011972.2, and C1E_0211300 in ACHU02.1).

Antisense sRNAs have been previously been characterised in *P. syringae* pv. *tomato* DC3000 in the *cmaU* and *aefR* genes (Filiatrault *et al.*, 2010), however these genes did not show antisense transcription in *Psa* in this study. The most well-characterised antisense sRNA, *fleQ*$_{\mathrm{as}}$, was predicted to have a conserved promoter region across Pseudomonas, including *Psa* (Markel *et al.*, 2018). In *P. syringae* pv. *tomato* DC3000 *fleQ*$_{\mathrm{as}}$ was found to be dependent on AlgU expression, however antisense transcription was not detected despite consitutive AlgU expression *in vitro* in *Psa*.

The poor conservation of antisense transcription in *P. syringae* is consistent with previous observations in *P. aeruginosa*, where antisense-expression was found to be strain and condition specific (Gómez-Lozano *et al.*, 2014), and with antisense transcription in general (Lloréns-Rico *et al.*, 2016).

Antisense transcripts for Hop effectors have previously been found in *P. syringae* pv. *tomato* DC3000 (Filiatrault *et al.*, 2010). Several expression peaks were also found opposite Hop effectors in *Psa* (Figure 4.18), however the *Psa* and *P. syringae* pv. *tomato* genomes contain different cohorts of effectors.

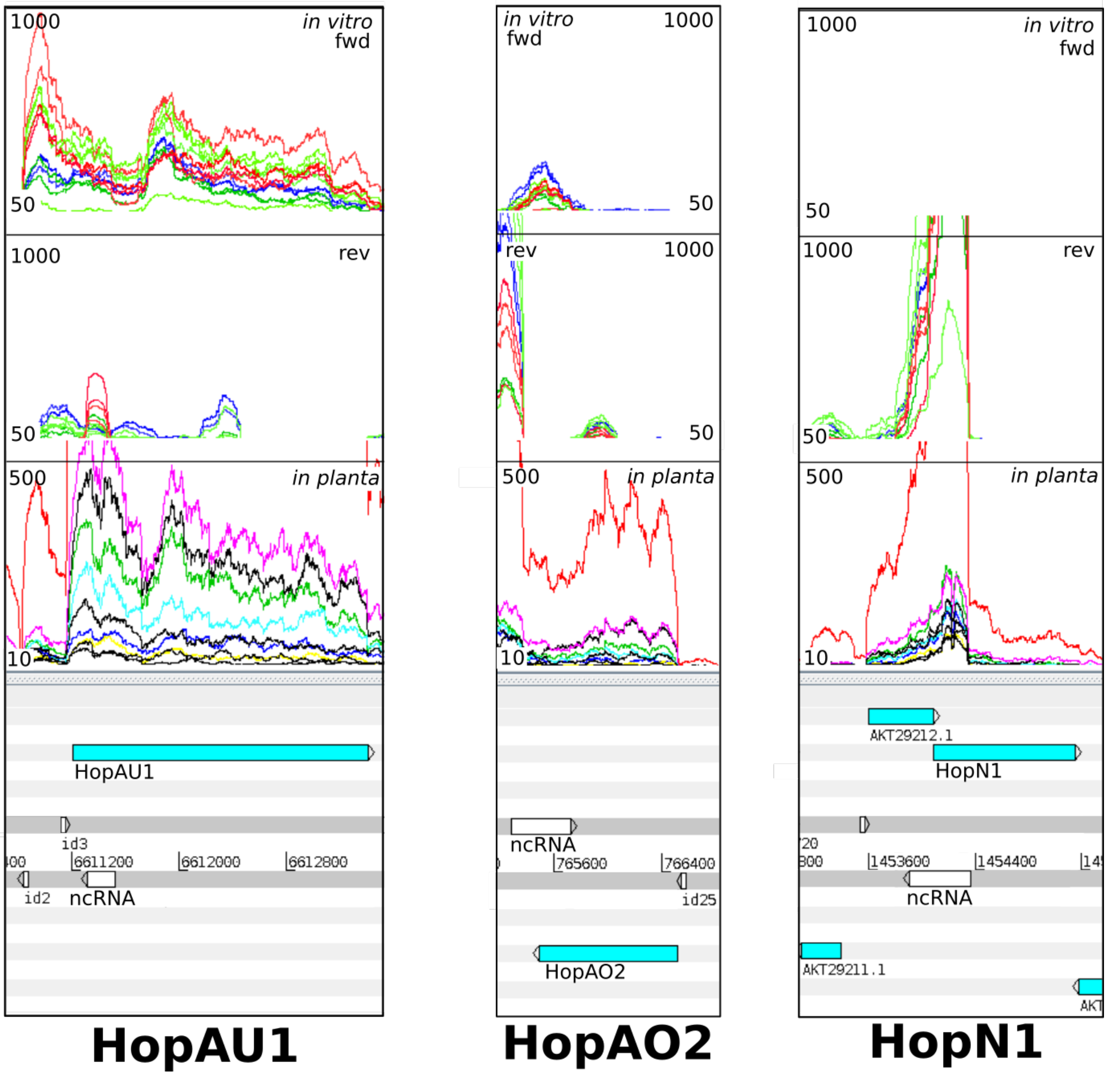


Figure 4.18: Antisense transcription to Hop effectors in *Psa*. The top two panels show read depth for *in vitro* experiments for each strand, coloured by experiment type. **Green:** Rich media, **Blue:** minimal media, **Red:** starvation. **Left**: A short peak expressed in starvation conditions opposite HopAU1. **Middle:** A peak overlapping with the 3′ end of HopAO2. **Right:** A highly expressed peak antisense to the 5′ end of HopN1.

4.4.5 Expression of Rfam annotated ncRNAs

In total, 50 Rfam annotated ncRNAs, excluding ribosomal RNAs, tRNAs and group II introns, were expressed *in vitro*. Two ncRNAs were only expressed *in planta*, and 29 ncRNAs were expressed in both *in vitro* and *in planta* data-sets.

The most highly expressed RNA in both *in vitro* and *in planta* samples was RsmY, which is part of the the Csr/Rsm family of ncRNAs. Csr/Rsm (carbon storage regulator, or repressor of secondary metabolism) is a widespread regulon of sRNAs and RNA-binding proteins that act to globally regulate metabolic pathways in response to environmental signals (Duss *et al.*, 2014). The Csr/Rsm family is comprised of a diverse set of ncRNAs, and contain GGA motifs in loops that allows them to sequester the RNA-binding proteins CsrA/RsmE. In *Pseudomonas syringae*, three Rsm ncRNAs, RsmX, RsmY and RsmZ (Janssen *et al.*, 2018, Moll *et al.*, 2010) are important regulators of metabolism, which are themselves under the control of the GacS/GacA two-component system (Brencic *et al.*, 2009).

Three ncRNAs annotated as CrcZ, a Csr/Rsm ncRNA that is a global suppressor of catabolism (Filiatrault *et al.*, 2013), were among the most highly expressed ncRNAs *in vitro* and *in planta*. These annotations likely represent the three *P. syringae* ncRNAs CrcX, CrcY and CrcZ (Filiatrault *et al.*, 2013), of which only CrcZ is represented in Rfam.

An annotation of the RMF RNA (location CP011972.2 4079040-407910), which is found in the 5′ UTR of ribosomal modulation factors in Pseudomonas (Weinberg *et al.*, 2010), was highly expressed at the 5′ end of a 500 bp expression peak overlapping an open reading frame. This region is likely to contain an un-annotated rmf gene or pseudogene. Several riboswitches and cis-regulatory elements, including FMN, Cobalamin, sucA-II and rne-II were also highly expressed *in vitro*.

RgsA is differentially expressed *in vitro*

P16, also called RgsA, was the only Rfam-annotated ncRNA that was significantly differentially expressed ($p_{adj} < 0.01$) *in vitro*, and was up-regulated in starvation conditions. RgsA is a short Hfq-binding ncRNA that is highly expressed in stationary phase, and forms a reciprocal regulatory circuit with the RpoS sigma factor (Lu *et al.*, 2018), which is a major transcriptional regulator during stationary phase. This sRNA has been found to play a role in protection from H_2O_2-induced oxidative stress in *P. aeruginosa*, *P. fluorescens* (González *et al.*, 2008), and in *P. syringae* pv. *tomato* (Park *et al.*, 2013), where it has been proposed to function in tolerance of plant defence response. RgsA also contains GGA loop motifs, suggesting it is part of the Csr/Rsm system.

Condition-specific expression

Three *Pseudomonas*-specific ncRNAs, P24, PhrS and GabT, were found to be expressed only under specific growth conditions *in vitro*. P24, an uncharacterised *Pseudomonas* ncRNA, was expressed in minimal media in *Psa*. Previously, P24 has been found to be expressed stationary phase in *P. aeruginosa* (Livny *et al.*, 2006), and in minimal media in *P. putida*, where it is down-regulated in response to osmotic stress (Bojanovič *et al.*, 2017).

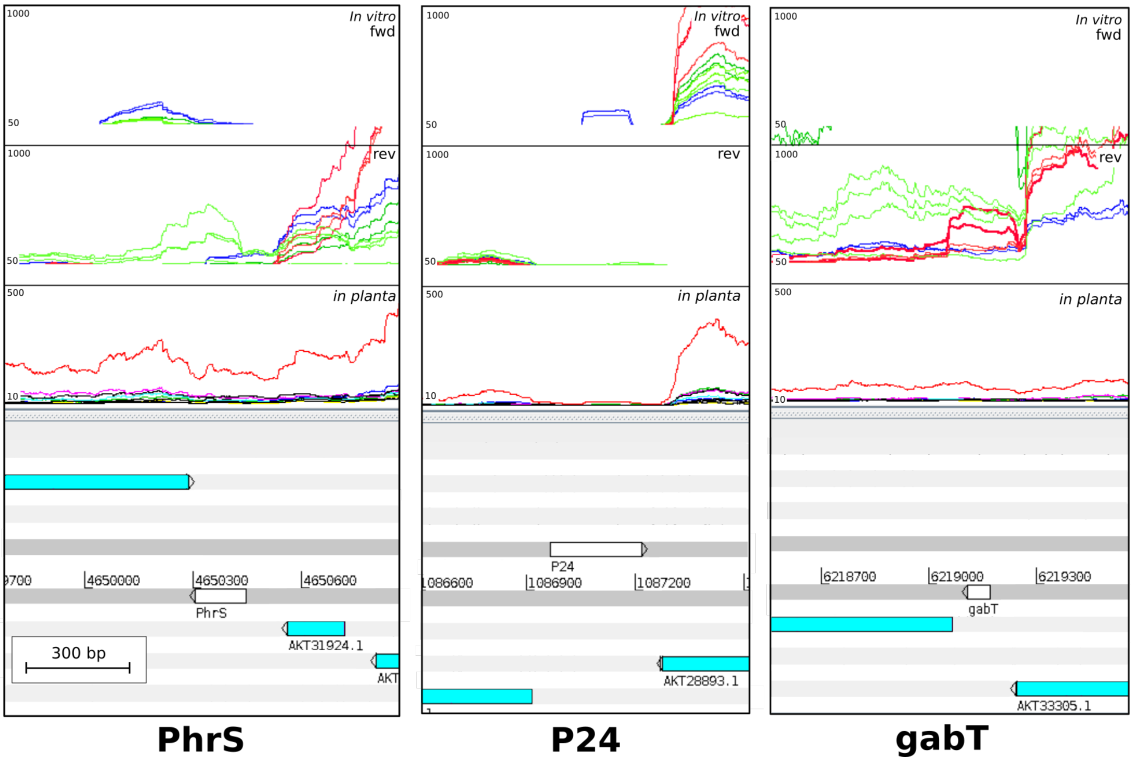


Figure 4.19: Condition-specific expression of Rfam annotated ncRNAs *in vitro*. **Left**: PhrS expressed in rich media in log phase (light green). **Middle**: P24 expressed in minimal media (blue). **Right**: gabT expressed in two samples under starvation conditions (red, bold line).

The PhrS RNA, located next to a pectate lyase in *Psa*, was only expressed in rich media during log phase. The PhrS RNA is involved in virulence in *P. aeruginosa*, where it is expressed in response to oxygen limitation and functions to regulate quorum sensing (Sonnleitner *et al.*, 2011) and CRISPR-cas systems (Lin *et al.*, 2019). However, its role in other Pseudomonads is unclear.

The gabT RNA, a *Pseudomonas*-specific putative *cis*-regulator of GabT (a transaminase) or GabD (Weinberg *et al.*, 2010), was only expressed in two replicates under starvation conditions (highlighted in Figure 4.19). GabT proteins are involved in the production of succinate from GABA, a common metabolite that also functions as a plant signalling molecule. In *P. syringae* pv. *tomato* DC3000, *gabT* knockouts have been found to have reduced virulence.

Three ncRNAs were highly expressed *in planta* during late infection (Figure 4.20). This late stage of infection is characterised the upregulation of genes involved in nutrient acquisition as well as genes involved in the production of alginate and biofilms (McAtee *et al.*, 2018).

The guanidine-I riboswitch ykkC-yxkD, was only expressed *in planta* at the 120hr time-point post-infection. This riboswitch binds selectively to guanidine, a product of arginine and guanine degradation, which is toxic to the cell (Nelson *et al.*, 2017). YkkC-yxkD is located upstream of urea carboyxlase transport genes in *Psa*, which have previously been found to selectively transport guanidine (Nelson *et al.*, 2017).

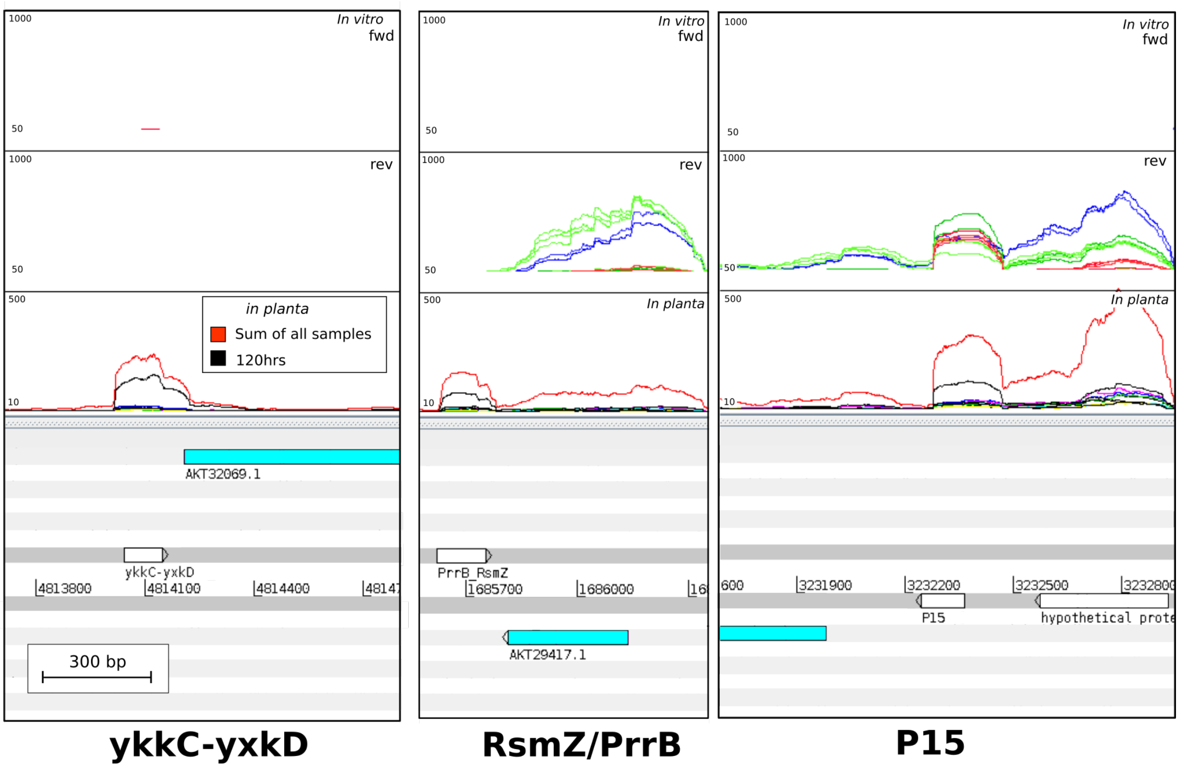


Figure 4.20: Condition-specific expression of Rfam-annotated ncRNAs *in planta*. Three ncRNAs, ykkC-yxkD, RsmZ/PrrB and P15 all show high expression 120 hours post-infection relative to other time-points. This can be seen as a black trace representing the sum across three replicates at this time-point. YkkC-yxkD and RsmZ/PrrB appear to be specifically expressed at this time-point *in planta*, whereas P15 is also expressed across all *in vitro* samples.

The P15 RNA, a *Pseudomonas*-specific sRNA with unknown function that has previously been identified as being expressed in stationary and exponential phase in *P. aeruginosa*, was highly expressed at 120hrs post infection relative to other *in planta* samples, and was also expressed *in vitro*.

The RsmZ sRNA (also called PrrB), a protein-sponge capable of sequestering up to multiple RsmE dimers per transcript (Duss *et al.*, 2014), was also only expressed *in planta* at 120hrs post-infection. RsmZ expression is induced by GacA, a response regulator with widespread and variable roles in virulence in *P. syringae* strains (Chatterjee *et al.*, 2003). Although RsmX, RsmY and RsmZ are all under the same regulatory control mechanism and appear to be functionally redundant (Moll *et al.*, 2010), RsmX and RsmY were expressed across all samples, whereas RsmZ was late infection-specific. This reflects previous observations of a fine-tuned regulatory cascade in *P. fluorescens* (Kay *et al.*, 2005). Kay *et al.* (2005) proposed that the presence of multiple redundant sRNA genes may be part of a gene dosage mechanism to allow the fine-tuning of the Gac/Rsm cascade. This regulon appears to be expanded in *P. syringae* pathovars, which have up to five copies of RsmX (Moll *et al.*, 2010). In *Psa*, the four copies of RsmX each showed different levels of expression relative to each other, which appears to support this hypothesis.

4.5 Conclusions and future directions

Bacterial ncRNAs are important components of gene regulation, however they are generally poorly conserved. We have performed comparative RNA-seq to identify expressed intergenic regions, and used computational measures of sequence and structural conservation to predict potential candidate ncRNAs.

Several of the candidates we identified were present on mobile genetic elements, including a plasmid and integrative conjugative element. *P. syringae* genomes are highly variable, and host to a diverse population of mobile genetic elements and plasmids (Bardaji *et al.*, 2011, Colombi *et al.*, 2017, Gutiérrez-Barranquero *et al.*, 2017). This is also true for *Psa* genomes, which has a large accessory genome (McCann *et al.*, 2017).

This exchange of genetic material plays an important role in pathogenicity via the transfer of virulence genes (Melnyk *et al.*, 2019), and by introducing structural changes due to their mobilisation (Baltrus *et al.*, 2017, Jackson *et al.*, 2011). The frequent conjugation and population of diverse mobile genetic elements circulating in environmental *Pseudomonads* are likely to introduce new ncRNAs at a high rate, and may require sampling of multiple strains within a species or population to identify ncRNAs in the accessory genome.

We have identified 27 candidate ncRNAs that are expressed in *Psa* in infection-relevant conditions, including 6 transcripts which are expressed in other pseudomonads. We have also explored the expression of known ncRNAs families in *Psa*. During candidate annotation we identified a twoAYGGAY-like motif, often associated with group II introns, that are abundant in *Pseudomonas* genomes. These motifs, some of which and are transcribed, appear to be Csr/Rsm-binding, however their function as transcripts or genomic elements is unclear. Work is planned with collaborators at the Plant and Food Research Centre, Auckland, to experimentally validate candidate intergenic ncRNAs identified in this study. This will include verification of the transcripts, knock-outs and fitness assays to identify if candidate ncRNAs are important for growth or pathogenicity.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, *215*(3), 403–410.

Anantharaman, V., Iyer, L. M., and Aravind, L. (2012). Ter-dependent stress response systems: novel pathways related to metal sensing, production of a nucleoside-like metabolite, and DNA-processing. *Mol. Biosyst.*, *8*(12), 3142–3165.

Baltrus, D. A., McCann, H. C., and Guttman, D. S. (2017). Evolution, genomics and epidemiology of Pseudomonas syringae: Challenges in Bacterial Molecular Plant Pathology. *Mol. Plant Pathol.*, *18*(1), 152–168.

Ban, C. and Yang, W. (1998). Crystal structure and ATPase activity of MutL: implications for DNA repair and mutagenesis. *Cell*, *95*(4), 541–552.

Bardaji, L., Añorga, M., Jackson, R. W., Martínez-Bilbao, A., Yanguas-Casás, N., and Murillo, J. (2011). Miniature transposable sequences are frequently mobilized in the bacterial plant pathogen *Pseudomonas syringae* pv. *phaseolicola*. *PLoS One*, *6*(10), e25773.

Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R., and Stadler, P. F. (2008). RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, *9*, 474.

Bojanovič, K., D'Arrigo, I., and Long, K. S. (2017). Global Transcriptional Responses to Osmotic, Oxidative, and Imipenem Stress Conditions in *Pseudomonas putida*. *Appl. Environ. Microbiol.*, *83*(7).

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Bradbury, R. S., Reid, D. W., and Champion, A. C. (2014). Urease production as a marker of virulence in *Pseudomonas aeruginosa*. *Br. J. Biomed. Sci.*, *71*(4), 175–177.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, *34*(5), 525–527.

Breazeale, S. D., Ribeiro, A. A., and Raetz, C. R. H. (2002). Oxidative decarboxylation of UDP-glucuronic acid in extracts of polymyxin-resistant *Escherichia coli*. Origin of lipid a species modified with 4-amino-4-deoxy-L-arabinose. *J. Biol. Chem.*, *277*(4), 2886–2896.

Brencic, A., McFarland, K. A., McManus, H. R., Castang, S., Mogno, I., Dove, S. L., and Lory, S. (2009). The GacS/GacA signal transduction system of *Pseudomonas aeruginosa* acts exclusively through its control over the transcription of the RsmY and RsmZ regulatory small RNAs. *Mol. Microbiol.*, *73*(3), 434–445.

Bukowski, M., Hyz, K., Janczak, M., Hydzik, M., Dubin, G., and Wladyka, B. (2017). Identification of novel *mazEF/pemIK* family toxin-antitoxin loci and their distribution in the *Staphylococcus* genus. *Sci. Rep.*, *7*(1), 13462.

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., and McQuillan, J. A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, *28*(4), 464–469.

Cech, T. R. and Steitz, J. A. (2014). The noncoding RNA revolution – trashing old rules to forge new ones. *Cell*, *157*(1), 77–94.

Chatterjee, A., Cui, Y., Yang, H., Collmer, A., Alfano, J. R., and Chatterjee, A. K. (2003). GacA, the response regulator of a two-component system, acts as a master regulator in *Pseudomonas syringae* pv. *tomato* DC3000 by controlling regulatory RNA, transcriptional activators, and alternate sigma factors. *Mol. Plant. Microbe. Interact.*, *16*(12), 1106–1117.

Chen, C. M., Ye, Q. Z., Zhu, Z. M., Wanner, B. L., and Walsh, C. T. (1990). Molecular biology of carbon-phosphorus bond cleavage. Cloning and sequencing of the *phn* (*psiD*) genes involved in alkylphosphonate uptake and C-P lyase activity in *Escherichia coli* B. *J. Biol. Chem.*, *265*(8), 4461–4471.

Chen, X.-L., Tang, D.-J., Jiang, R.-P., He, Y.-Q., Jiang, B.-L., Lu, G.-T., and Tang, J.-L. (2011). sRNA-Xcc1, an integron-encoded transposon- and plasmid-transferred trans-acting sRNA, is under the positive control of the key virulence regulators HrpG and HrpX of *Xanthomonas campestris* pathovar *campestris*. *RNA Biol.*, *8*(6), 947–953.

Cheng, F., Ma, A., Luo, J., Zhuang, X., and Zhuang, G. (2017). N-acylhomoserine lactone-regulation of genes mediating motility and pathogenicity in *Pseudomonas syringae* pathovar *tabaci* 11528. *MicrobiologyOpen*, *6*(3).

Cheng, F., Ma, A., Zhuang, G., and Fray, R. G. (2018). Exogenous N-acyl-homoserine lactones enhance the expression of flagella of *Pseudomonas syringae* and activate defence responses in plants. *Mol. Plant Pathol.*, *19*(1), 104–115.

Colombi, E. (2017). *The role of integrative conjugative elements in evolution of the kiwifruit pathogen Pseudomonas syringae pv. actinidiae*. Ph. D. thesis, Massey University.

Colombi, E., Straub, C., Künzel, S., Templeton, M. D., McCann, H. C., and Rainey, P. B. (2017). Evolution of copper resistance in the kiwifruit pathogen *Pseudomonas syringae* pv. *actinidiae* through acquisition of integrative conjugative elements and plasmids. *Environ. Microbiol.*, *19*(2), 819–832.

Dai, L. and Zimmerly, S. (2002). Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res.*, *30*(5), 1091–1102.

Darmon, E. and Leach, D. R. F. (2014). Bacterial genome instability. *Microbiol. Mol. Biol. Rev.*, *78*(1), 1–39.

Davies, A. M., Tata, R., Beavil, R. L., Sutton, B. J., and Brown, P. R. (2007). l-Methionine sulfoximine, but not phosphinothricin, is a substrate for an acetyltransferase (gene PA4866) from *Pseudomonas aeruginosa*: structural and functional studies. *Biochemistry*, *46*(7), 1829–1839.

De Reuse, H., Labigne, A., and Mengin-Lecreulx, D. (1997). The *Helicobacter pylori ureC* gene codes for a phosphoglucosamine mutase. *J. Bacteriol.*, *179*(11), 3488–3493.

Duss, O., Michel, E., Yulikov, M., Schubert, M., Jeschke, G., and Allain, F. H.-T. (2014). Structural basis of the non-coding RNA RsmZ acting as a protein sponge. *Nature*, *509*(7502), 588–592.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, *7*(10), e1002195.

Ferrara, S., Brugnoli, M., De Bonis, A., Righetti, F., Delvillani, F., Dehò, G., Horner, D., Briani, F., and Bertoni, G. (2012). Comparative profiling of *Pseudomonas aeruginosa* strains reveals differential expression of novel unique and conserved small RNAs. *PloS One*, *7*(5), e36553.

Ferrara, S., Falcone, M., Macchi, R., Bragonzi, A., Girelli, D., Cariani, L., Cigana, C., and Bertoni, G. (2017). The PAPI-1 pathogenicity island-encoded small RNA PesA influences *Pseudomonas aeruginosa* virulence and modulates pyocin S3 production. *PLoS One*, *12*(6), e0180386.

Filiatrault, M. J., Stodghill, P. V., Bronstein, P. A., Moll, S., Lindeberg, M., Grills, G., Schweitzer, P., Wang, W., Schroth, G. P., Luo, S., and et al. (2010). Transcriptome analysis of *Pseudomonas syringae* identifies new genes, noncoding RNAs, and antisense activity. *J. Bacteriol.*, *192*(9), 2359–2372.

Filiatrault, M. J., Stodghill, P. V., Wilson, J., Butcher, B. G., Chen, H., Myers, C. R., and Cartinhour, S. W. (2013). CrcZ and CrcX regulate carbon source utilization in *Pseudomonas syringae* pathovar *tomato* strain DC3000. *RNA Biol.*, *10*(2), 245–255.

Fiore, J. L. and Nesbitt, D. J. (2013). An RNA folding motif: GNRA tetraloop–receptor interactions. *Q. Rev. Biophys.*, *46*(3), 223–264.

Fontaine, F., Gasiorowski, E., Gracia, C., Ballouche, M., Caillet, J., Marchais, A., and Hajnsdorf, E. (2016). The small RNA SraG participates in PNPase homeostasis. *RNA*, *22*(10), 1560–1573.

Gardner, P. P., Barquist, L., Bateman, A., Nawrocki, E. P., and Weinberg, Z. (2011). RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic Acids Res.*, *39*(14), 5845–5852.

Gardner, P. P. and Eldai, H. (2015). Annotating RNA motifs in sequences and alignments. *Nucleic Acids Res.*, *43*(2), 691–698.

Geng, X., Jin, L., Shimada, M., Kim, M. G., and Mackey, D. (2014). The phytotoxin coronatine is a multifunctional component of the virulence armament of *Pseudomonas syringae*. *Planta*, *240*(6), 1149–1165.

Georg, J. and Hess, W. R. (2011). cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol. Mol. Biol. Rev.*, *75*(2), 286–300.

Gómez-Lozano, M., Marvig, R. L., Molin, S., and Long, K. S. (2012). Genome-wide identification of novel small RNAs in *Pseudomonas aeruginosa*. *Environ. Microbiol.*, *14*(8), 2006–2016.

Gómez-Lozano, M., Marvig, R. L., Molina-Santiago, C., Tribelli, P. M., Ramos, J.-L., and Molin, S. (2015). Diversity of small RNAs expressed in *Pseudomonas species*. *Environ. Microbiol. Rep.*, *7*(2), 227–236.

Gómez-Lozano, M., Marvig, R. L., Tulstrup, M. V. L., and Molin, S. (2014). Expression of antisense small RNAs in response to stress in *Pseudomonas aeruginosa*. *BMC Genomics*, *15*, 783.

González, N., Heeb, S., Valverde, C., Kay, E., Reimmann, C., Junier, T., and Haas, D. (2008). Genome-wide search reveals a novel GacA-regulated small RNA in *Pseudomonas* species. *BMC Genomics*, *9*, 167.

Gutiérrez-Barranquero, J. A., Cazorla, F. M., de Vicente, A., and Sundin, G. W. (2017). Complete sequence and comparative genomic analysis of eight native *Pseudomonas syringae* plasmids belonging to the pPT23A family. *BMC Genomics*, *18*(1), 365.

Jackson, R. W., Vinatzer, B., Arnold, D. L., Dorus, S., and Murillo, J. (2011). The influence of the accessory genome on bacterial pathogen evolution. *Mob. Genet. Elements*, *1*(1), 55–65.

Janssen, K. H., Diaz, M. R., Golden, M., Graham, J. W., Sanders, W., Wolfgang, M. C., and Yahr, T. L. (2018). Functional Analyses of the RsmY and RsmZ Small Noncoding Regulatory RNAs in *Pseudomonas aeruginosa*. *J. Bacteriol.*, *200*(11).

Jose, B. R., Gardner, P. P., and Barquist, L. (2019). Transcriptional noise and exaptation as sources for bacterial sRNAs. *Biochem. Soc. Trans.*, *47*(2), 527–539.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, *30*(4), 772–780.

Kay, E., Dubuis, C., and Haas, D. (2005). Three small RNAs jointly ensure secondary metabolism and biocontrol in *Pseudomonas fluorescens* CHA0. *PNAS*, *102*(47), 17136–17141.

Kjos, M., Snipen, L., Salehian, Z., Nes, I. F., and Diep, D. B. (2010). The Abi proteins and their involvement in bacteriocin self-immunity. *J. Bacteriol.*, *192*(8), 2068–2076.

Kröger, C., Colgan, A., Srikumar, S., Händler, K., Sivasankaran, S. K., Hammarlöf, D. L., Canals, R., Grissom, J. E., Conway, T., Hokamp, K., and Hinton, J. C. D. (2013). An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium. *Cell Host Microbe*, *14*(6), 683–695.

Kvint, K., Nachin, L., Diez, A., and Nyström, T. (2003). The bacterial universal stress protein: function and regulation. *Curr. Opin. Microbiol.*, *6*(2), 140–145.

Lambowitz, A. M. and Zimmerly, S. (2011). Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb. Perspect. Biol.*, *3*(8), a003616.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, *9*(4), 357–359.

Lee, J. S., Ryu, H. R., Cha, J. Y., and Baik, H. S. (2015). The *hrp* pathogenicity island of *Pseudomonas syringae* pv. *tomato* DC3000 is induced by plant phenolic acids. *J. Microbiol.*, *53*(10), 725–731.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Lin, P., Pu, Q., Wu, Q., Zhou, C., Wang, B., Schettler, J., Wang, Z., Qin, S., Gao, P., Li, R., Li, G., Cheng, Z., Lan, L., Jiang, J., and Wu, M. (2019). High-throughput screen reveals sRNAs regulating crRNA biogenesis by targeting CRISPR leader to repress Rho termination. *Nat. Commun.*, *10*(1), 3728.

Lindgreen, S., Umu, S. U., Lai, A. S.-W., Eldai, H., Liu, W., McGimpsey, S., Wheeler, N. E., Biggs, P. J., Thomson, N. R., Barquist, L., and et al. (2014). Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Comput. Biol.*, *10*(10), e1003907.

Liu, Y., Bauer, S. C., and Imlay, J. A. (2011). The YaaA protein of the *Escherichia coli* OxyR regulon lessens hydrogen peroxide toxicity by diminishing the amount of intracellular unincorporated iron. *J. Bacteriol.*, *193*(9), 2186–2196.

Liu, Y., Cui, Y., Mukherjee, A., and Chatterjee, A. K. (1998). Characterization of a novel RNA regulator of *Erwinia carotovora* ssp. *carotovora* that controls production of extracellular enzymes and secondary metabolites. *Mol. Microbiol.*, *29*(1), 219–234.

Livny, J., Brencic, A., Lory, S., and Waldor, M. K. (2006). Identification of 17 *Pseudomonas aeruginosa* sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. *Nucleic Acids Res.*, *34*(12), 3484–3493.

Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.-H., Bork, P., Glass, J. I., Serrano, L., and Lluch-Senar, M. (2016). Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci. Adv.*, *2*(3), e1501363.

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, *6*, 26.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, *15*(12), 550.

Lu, P., Wang, Y., Hu, Y., and Chen, S. (2018). RgsA, an RpoS-dependent sRNA, negatively regulates rpoS expression in *Pseudomonas aeruginosa*. *Microbiology*, *164*(4), 716–724.

Markel, E., Dalenberg, H., Monteil, C. L., Vinatzer, B. A., and Swingle, B. (2018). An AlgU-Regulated Antisense Transcript Encoded within the *Pseudomonas syringae fleQ* Gene Has a Positive Effect on Motility. *J. Bacteriol.*, *200*(7).

McAtee, P. A., Brian, L., Curran, B., van der Linden, O., Nieuwenhuizen, N. J., Chen, X., Henry-Kirk, R. A., Stroud, E. A., Nardozza, S., Jayaraman, J., Rikkerink, E. H. A., Print, C. G., Allan, A. C., and Templeton, M. D. (2018). Re-programming of *Pseudomonas syringae* pv. *actinidiae* gene expression during early stages of infection of kiwifruit. *BMC Genomics*, *19*(1), 822.

McCann, H. C., Li, L., Liu, Y., Li, D., Pan, H., Zhong, C., Rikkerink, E. H. A., Templeton, M. D., Straub, C., Colombi, E., Rainey, P. B., and Huang, H. (2017). Origin and evolution of the kiwifruit canker pandemic. *Genome Biol. Evol.*.

Melnyk, R. A., Hossain, S. S., and Haney, C. H. (2019). Convergent gain and loss of genomic islands drive lifestyle changes in plant-associated *Pseudomonas*. *ISME J.*, *13*(6), 1575–1588.

Moll, S., Schneider, D. J., Stodghill, P., Myers, C. R., Cartinhour, S. W., and Filiatrault, M. J. (2010). Construction of an *rsmX* co-variance model and identification of five *rsmX* non-coding RNAs in *Pseudomonas syringae* pv. *tomato* DC3000. *RNA Biol.*, *7*(5), 508–516.

Muthuramalingam, M., White, J. C., Murphy, T., Ames, J. R., and Bourne, C. R. (2019). The toxin from a ParDE toxin-antitoxin system found in *Pseudomonas aeruginosa* offers protection to cells challenged with anti-gyrase antibiotics. *Mol. Microbiol.*, *111*(2), 441–454.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and et al. (2015). Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, *43*(Database issue), D130–7.

Nelson, J. W., Atilho, R. M., Sherlock, M. E., Stockbridge, R. B., and Breaker, R. R. (2017). Metabolism of Free Guanidine in Bacteria Is Regulated by a Widespread Riboswitch Class. *Mol. Cell*, *65*(2), 220–230.

Park, S. H., Butcher, B. G., Anderson, Z., Pellegrini, N., Bao, Z., D'Amico, K., and Filiatrault, M. J. (2013). Analysis of the small RNA P16/RgsA in the plant pathogen *Pseudomonas syringae* pv. *tomato* strain DC3000. *Microbiology*, *159*(Pt 2), 296–306.

Pita, T., Feliciano, J. R., and Leitão, J. H. (2018). Small Noncoding Regulatory RNAs from *Pseudomonas aeruginosa* and *Burkholderia cepacia* Complex. *Int. J. Mol. Sci.*, *19*(12).

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., and et al. (2012). The Pfam protein families database. *Nucleic Acids Res.*, *40*(Database issue), D290–301.

Rath, E. C., Pitman, S., Cho, K. H., and Bai, Y. (2017). Identification of streptococcal small RNAs that are putative targets of RNase III through bioinformatics analysis of RNA sequencing data. *BMC Bioinformatics*, *18*(Suppl 14), 540.

Rau, M. H., Bojanovič, K., Nielsen, A. T., and Long, K. S. (2015). Differential expression of small RNAs under chemical stress and fed-batch fermentation in *E. coli*. *BMC Genomics*, *16*, 1051.

Rivas, E., Clements, J., and Eddy, S. R. (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature Methods*, *14*(1), 45–48.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, *13*(11), 2498–2504.

Sonnleitner, E., Gonzalez, N., Sorger-Domenigg, T., Heeb, S., Richter, A. S., Backofen, R., Williams, P., Hüttenhofer, A., Haas, D., and Bläsi, U. (2011). The small RNA PhrS stimulates synthesis of the *Pseudomonas aeruginosa* quinolone signal. *Mol. Microbiol.*, *80*(4), 868–885.

Stav, S., Atilho, R. M., Mirihana Arachchilage, G., Nguyen, G., Higgs, G., and Breaker, R. R. (2019). Genome-wide discovery of structured noncoding RNAs in bacteria. *BMC Microbiol.*, *19*(1), 66.

Templeton, M. D., Warren, B. A., Andersen, M. T., Rikkerink, E. H. A., and Fineran, P. C. (2015). Complete DNA Sequence of *Pseudomonas syringae* pv. *actinidiae*, the Causal Agent of Kiwifruit Canker Disease. *Genome Announc.*, *3*(5).

Wang, L., Mavisakalyan, V., Tillier, E. R. M., Clark, G. W., Savchenko, A. V., Yakunin, A. F., and Master, E. R. (2010). Mining bacterial genomes for novel arylesterase activity. *Microb. Biotechnol.*, *3*(6), 677–690.

Washietl, S., Findeiß, S., Müller, S. A., Kalkhof, S., von Bergen, M., Hofacker, I. L., Stadler, P. F., and Goldman, N. (2011). RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*, *17*(4), 578–594.

Washietl, S. and Hofacker, I. L. (2004). Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.*, *342*(1), 19–30.

Watson, K. A., Schinzel, R., Palm, D., and Johnson, L. N. (1997). The crystal structure of *Escherichia coli* maltodextrin phosphorylase provides an explanation for the activity without control in this basic archetype of a phosphorylase. *EMBO J.*, *16*(1), 1–14.

Weinberg, Z. and Breaker, R. R. (2011). R2R–software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics*, *12*, 3.

Weinberg, Z., Wang, J. X., Bogue, J., Yang, J., Corbino, K., Moy, R. H., and Breaker, R. R. (2010). Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.*, *11*(3), R31.

Wen, Y., Feng, J., Scott, D. R., Marcus, E. A., and Sachs, G. (2011). A cis-encoded antisense small RNA regulated by the HP0165-HP0166 two-component system controls expression of *ureB* in *Helicobacter pylori*. *J. Bacteriol.*, *193*(1), 40–51.

Wencewicz, T. A. and Walsh, C. T. (2012). *Pseudomonas syringae* self-protection from tabtoxinine-β-lactam by ligase TblF and acetylase Ttr. *Biochemistry*, *51*(39), 7712–7725.

Wheeler, T. J. and Eddy, S. R. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, *29*(19), 2487–2489.

Wright, P. R., Georg, J., Mann, M., Sorescu, D. A., Richter, A. S., Lott, S., Kleinkauf, R., Hess, W. R., and Backofen, R. (2014). CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains. *Nucleic Acids Res.*, *42*(Web Server issue), W119–23.

Wright, P. R., Richter, A. S., Papenfort, K., Mann, M., Vogel, J., Hess, W. R., Backofen, R., and Georg, J. (2013). Comparative genomics boosts target prediction for bacterial small RNAs. *PNAS*, *110*(37), E3487–96.

Wurtzel, O., Yoder-Himes, D. R., Han, K., Dandekar, A. A., Edelheit, S., Greenberg, E. P., Sorek, R., and Lory, S. (2012). The single-nucleotide resolution transcriptome of *Pseudomonas aeruginosa* grown in body temperature. *PLoS Pathog.*, *8*(9), e1002945.

Yeo, C. C., Yiin, S., Tan, B. H., and Poh, C. L. (2001). Isolation and characterization of group II introns from *Pseudomonas alcaligenes* and *Pseudomonas putida*. *Plasmid*, *45*(3), 233–239.

Zhao, Z., Chen, J., Gao, X., Zhang, D., Zhang, J., Wen, J., Qin, H., Guo, M., and Huang, L. (2019). Comparative genomics reveal pathogenicity-related loci in *Pseudomonas syringae* pv. *actinidiae* biovar 3. *Mol. Plant Pathol.*.

Chapter 5

Genome assembly of *Gemmata* isolates

5.1 Preface

This Chapter and Appendix A briefly describe my work on two genome sequencing projects during my time at the University of Canterbury. This work describes the genome assembly and preliminary analysis of *Gemmata* isolates, as part of a collaboration with Paul Gardner, Amy Osborne and Anthony Poole. This builds on previous work which explored the relationship between expression of ncRNAs and phylogenetic distance, which concluded that phylogeny-informed sampling is required for comparative methods to be effective. For this project, outgroups in the 'goldilocks zone', i.e phyla at a specific distance that will allow transcriptomic comparison, including *Gemmata*, have been selected for sequencing.

5.2 Contributions

I performed all assembly and data analysis. Strains were selected for sequencing by Sinan Umu and Stinus Lindgreen. Isolates were grown and prepared for sequencing by Amy Osborne.

5.3 Introduction

Gemmata are a genus of aerobic budding bacteria found in freshwater and in soil, which is of particular interest in evolutionary biology due to their phenotypic similarity to eukaryotes and archaea (Devos, 2013, Franzmann and Skerman, 1984). *Gemmata* are part of the Planctomycetes superphylum, which are a widespread group of environmental bacteria with unusual phenotypic and physiological characteristics reminiscent of eukaryotic cell. An especially noteable feature of planctomycetes is the presence of intracellular compartments (Lindsay *et al.*, 2001), including a large compartment enclosing the nucleoid that has been found in *Gemmata obscuriglobus* (Figure 5.1) (Fuerst, 2005).

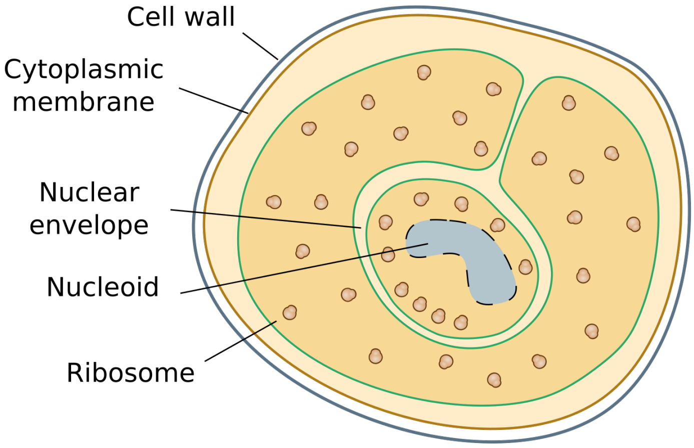


Figure 5.1: Diagram showing the intracellular compartments of a *Gemmata obscuriglobus* cell. Adapted from Fuerst and Sagulenko (2011).

The taxonomic placement of planctomycetes has been debated, as they share phenotypic features with eukaryotes, such as reproduction *via* yeast-like budding of the extracellular membrane, and also contain ammonia metabolic pathways normally found in archaea (Fuchsman and Rocap, 2006, Fuerst and Sagulenko, 2011). *Gemmata obscuriglobus* also has the unusual ability produce sterols (Rivas-Marin *et al.*, 2019), and utilises an endocytosis-like system for protein uptake (Lonhienne *et al.*, 2010).

However, there is some debate as to whether these features are structurally analagous rather than due to some shared homology (McInerney *et al.*, 2011). Planctomycetes have no substantial shared gene content with other kingdoms, aside from some small amount presumably obtained by HGT (Fuchsman and Rocap, 2006).

Currently, only a handful of representative strains and genome sequences are available for *Planctomycetes*, with *G. obscuriglobus* being the sole sequenced representative of the genus *Gemmata*. We have sequenced and annotated the genomes of a strain of *G. obscuriglobus*, and 4 other *Gemmata*-like isolates as part of an effort to increase the number of genomes available for comparison for this under-sampled bacterial phylum (Wu *et al.*, 2009). This Chapter describes the assembly, annotation and a preliminary comparative analysis of these genomes, which found large numbers of transposons in *G. obscuriglobus*, and high levels of rearrangements across the clade.

5.4 Methods

5.4.1 Sequencing

Gemmata isolates were collected from Queensland, Australia. *Gemmata*-like str. JW3-9s0 (aka Soil9) and *Gemmata obscuriglobus* were isolated from soil, *Gemmata*-like str. CJuq14 and *Gemmata*-like str. JW9-3f1 from a eutrophic lake and *Gemmata*-like str. JW11-2f5 from an ornamental fountain. Isolates were previously classified as *Gemmata*-like based on a phylogenetic analysis of 16S sequences (Wang *et al.*, 2002). Samples were isolated, cultured and prepared for sequencing as described in (Wang *et al.*, 2002). The genomes of all isolates were sequenced by Macrogen using Pacific Biosciences SMRT^{TM} sequencing.

5.4.2 Assembly and analysis

Raw bas.h5 sequencing files were converted to fastq reads with bash5tools.py from the pbh5tools package from PacificBiosciences (https://github.com/PacificBiosciences/pbh5tools/). These reads were trimmed and assembled using Canu v1.5 (default settings) (Koren *et al.*, 2017). Genomes were circularised with Circlator (Hunt *et al.*, 2015). All reads were mapped to the genome assemblies using BLASR (Chaisson and Tesler, 2012) to generate alignments for Qualimap, which was used to assess the coverage and quality of the assemblies (Okonechnikov *et al.*, 2016). The genome assemblies were annotated with Prokka (Seemann, 2014). To improve gene annotations, a custom database of planctomycetes protein annotations was used as an additional reference, consisting of all protein annotations of plantomycete genomes from NCBI. The gram-negative option for identifying signal peptides was used during annotation, as genes involved in outer membrane biogenesis have previously been identified in *Planctomycetes* (Speth *et al.*, 2012). Whole genome alignments were performed using mauve genome aligner (Darling *et al.*, 2010). Roary was used to identify the number of orthologous genes between the isolates (Page *et al.*, 2015).

5.5 Results and Discussion

The genome assembly tools completely resolved two genomes, *G. obscuriglobus* UQM2246 and *Gemmata*-like str. JW11, and generated large high-coverage contigs for other strains (Table 5.1). The genome sizes ranged between 7.86–10.14 Mb, with high % GC content (64–70%), which is consistent with the genomes of other *Gemmata* and *Planctomycete* genomes (Franke *et al.*, 2018). A small contig in the *Gemmata*-like str. CJuql4 assembly was classed as a plasmid by Circlator. Plasmid-encoded RepA proteins were annotated in larger contigs in *Gemmata*-like str. JW9 and *Gemmata*-like str. JW3-8s0, indicating that plasmid reads may have been mis-assembled into genome contigs. CRISPR-cas cassettes were also detected in all assemblies except *Gemmata*-like str. JW9, however *cas* endonucleases were found in all genome assemblies (Table 5.2).

Strain	Genome Contigs	Size (Mb)	% GC	Circular	Plasmid contigs	Mean coverage (X)
UQM2246	1	8.99	67.36	Yes	0	106
JW3-8s0	1	10.14	63.95	No	0	91
CJuql4	3	7.94	69.87	No	1	109
JW11	1	8.86	68.71	Yes	0	91
JW9	5	10.03	64.14	No	0	96

Table 5.1: Summary of *Gemmata* genome assemblies, percentage GC content, predicted plasmid contigs, and coverage.

Strain	Genes	CDS	rRNA	tRNA	tmRNA	misc RNA	Signal peptides	CRISPR regions	Cas proteins
UQM2246	7824	7702	12	101	1	8	912	4	1, 2, 3, 4, 6, 7
JW3-8s0	8994	8883	9	101	1	-	1075	4	1, 2, 3, 4
CJuql4	7115	6984	6	110	1	14	959	1	1, 2, 4
JW11	8003	7885	6	94	1	17	961	2	1, 2, 3, 4, 5, 6, 7, 8
JW9	9344	9213	6	114	1	10	1016	-	1, 2, 4

Table 5.2: Summary of Prokka genome annotations for *Gemmata* genome assemblies.

Whole-genome alignments of the genome assemblies showed the *Gemmata* genomes have highly diverse genome structures (Figure 5.4), despite high overall shared gene content. Previous estimates based on 16S sequences that *Gemmata*-like isolates JW9, Cjuql4 and JW3-8s0 are strains of a single *Gemmata* species (Wang *et al.*, 2002). A pan-genome `Roary` analysis found that 633 genes were shared between all 5 strains, and 21986 genes were unique to a single genome. These consisted of uncharacterised hypothetical proteins, and components of mobile genetic elements. The accessory gene tree (Figure 5.3), showing the similarity between cohorts of accessory genes between strains, also reflects the relationships shown in 16S tree produced here and by Wang *et al.* (2002).

Protein annotation found that our *Gemmata* genome assemblies contain high numbers of mobile genetic elements (Table 5.3), with *G. obscuriglobus* containing 217 predicted transposases. The true number of these elements may be larger, as many predicted proteins were not able to be functionally annotated.

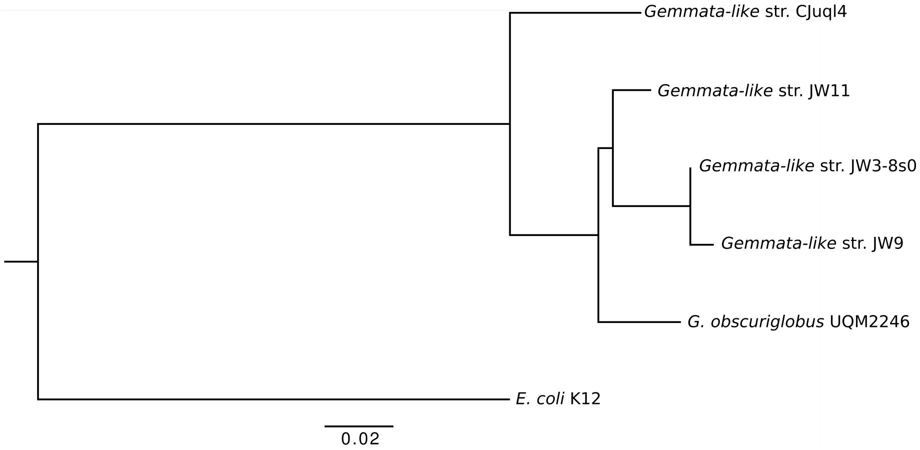

Gemmata-like str. CJuql4
Gemmata-like str. JW11
Gemmata-like str. JW3-8s0
Gemmata-like str. JW9
G. obscuriglobus UQM2246
E. coli K12
0.02

Figure 5.2: 16S tree of *Gemmata* genome assemblies, with E. coli K12 included as an outgroup. 16S sequences annotated by Prokka were aligned to a 16S covariance model sequences from Rfam. The *E. coli* K12 sequence was also included as an outgroup. Phylogenetic tree was generated by ClustalW2 (distance correction=on, type=neighbour-joining). Figure generated in Figtree ($https://github.com/rambaut/figtree$)

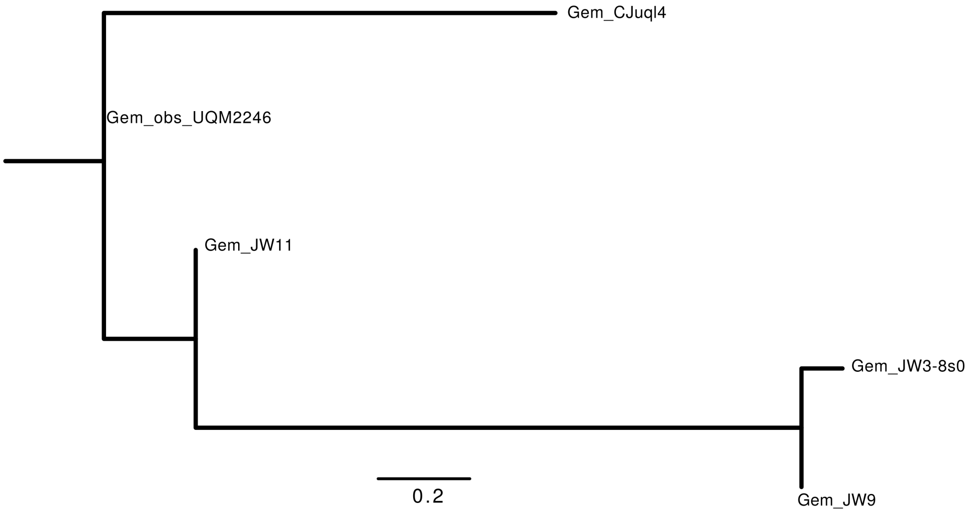


Figure 5.3: Phylogeny showing the relationships between the *Gemmata* genome assemblies based on gene content. Tree generated by `Roary` (Page *et al.*, 2015) and visualised in Figtree.

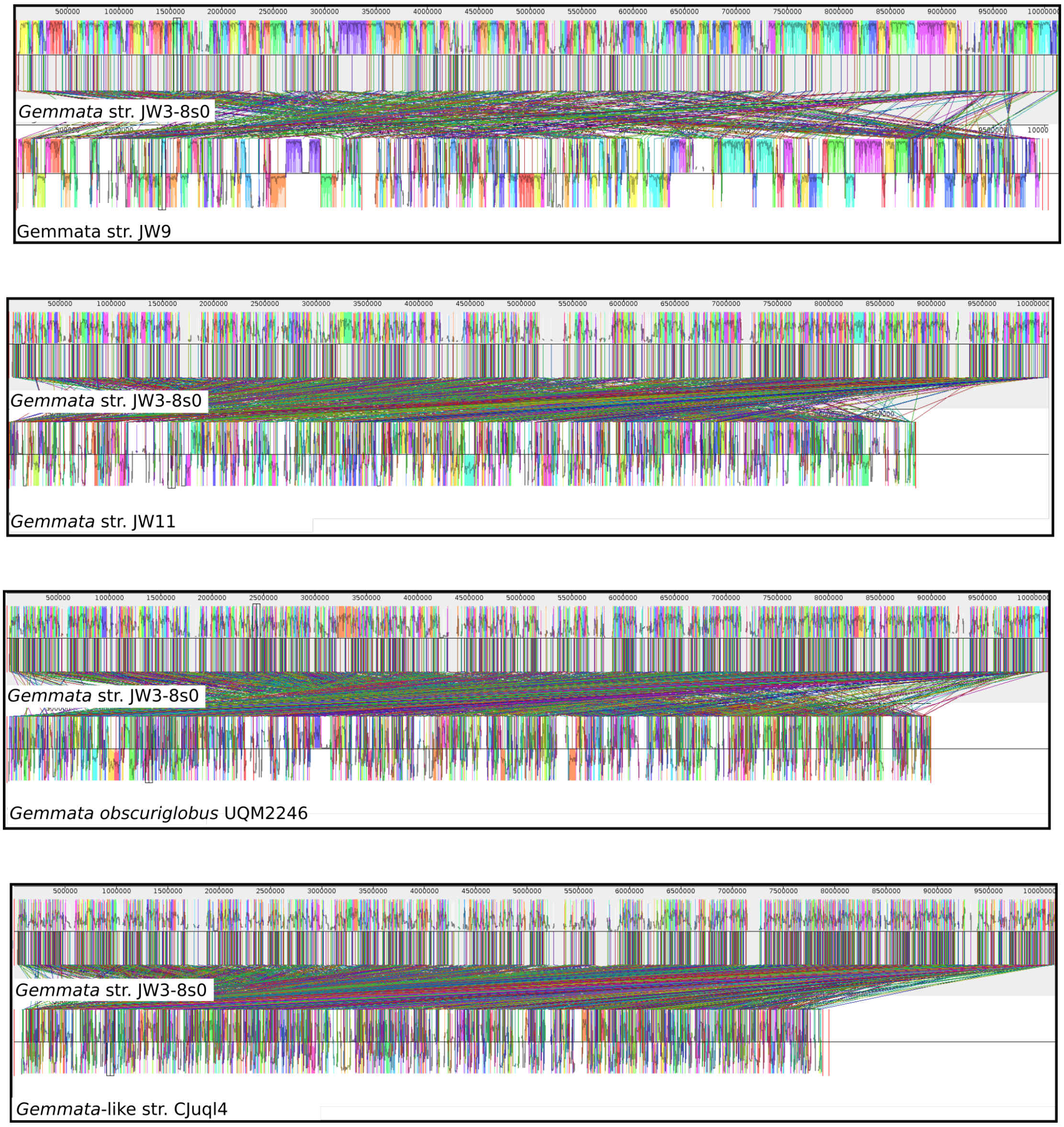


Figure 5.4: Mauve alignments showing pairwise alignments of the *Gemmata* genome assemblies from different phylogenetic distances (Figure 5.3). Local colinear blocks, consisting of shared sequence regions throughout the genomes, are shown as different colours and linked by connecting lines. This alignment shows extremely high levels of rearrangement between these genomes, despite high amount of shared gene content.

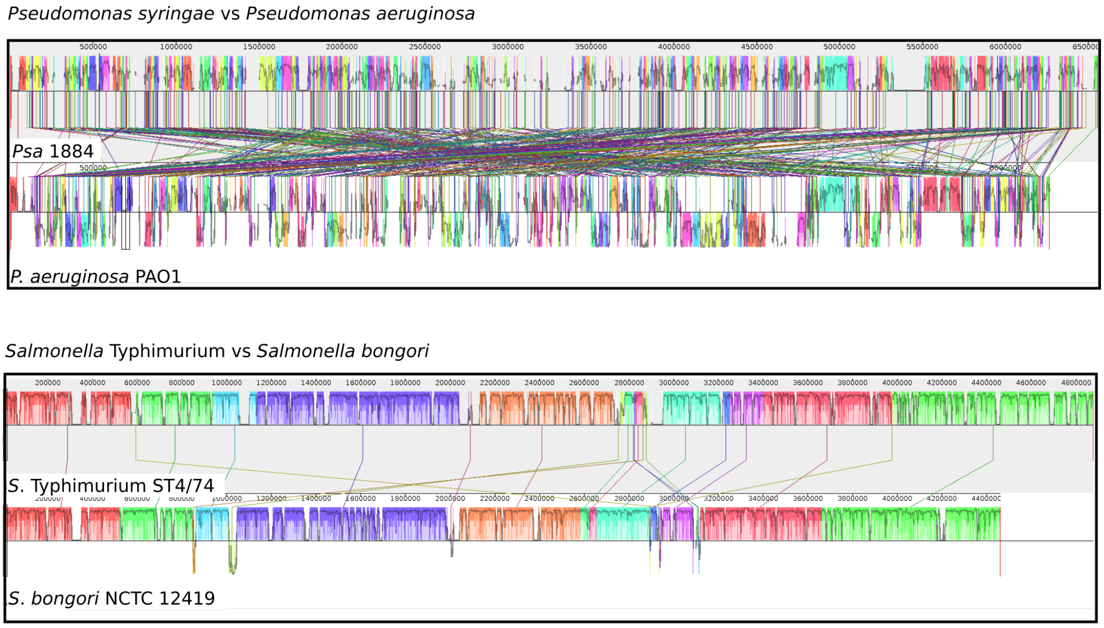


Figure 5.5: Mauve alignments of genomes from well-characterised organisms at various phylogenetic distances. To provide context for the alignment of the *Gemmata* genome assemblies, alignments are shown for within-genus comparisons of *Pseudomonas* and *Salmonella.* Top: Within-genus comparison of two genomes of the genus *Pseudomonas, P. aeruginosa* PAO1 (NCBI: NC_002516.2) and *P. syringae* pv. *actinidiae* ICMP 1884 (NCBI:CP011972.2). The genomes of the genus *Pseudomonas* are considered exceptionally dynamic, containing a large pan-genome and frequent rearrangements and horizontal gene transfer events (Silby *et al.*, 2011). Nonetheless, high amounts of genome content is shared between the two species, although they are substantially different in genome structure. Bottom: Within-genus comparison between *S.* Typhimurium ST4/74 (NCBI: CP002487.1) and *S. bongori* NCTC 12419 (NCBI: NC_015761.1). These genera retain large amounts of shared genome content and structure.

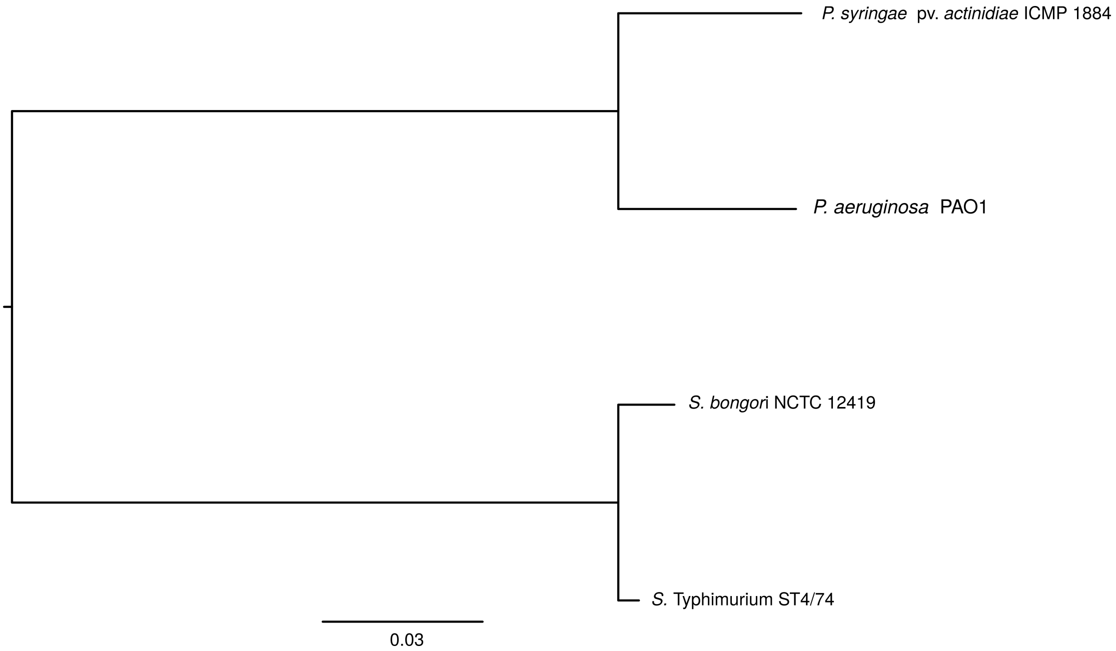


Figure 5.6: 16S tree of species used in Figure 5.5, generated as per Figure 5.2. This tree shows the relationships and approximate phylogenetic distance between the species used in Figure 5.5, which are more distantly related than the *Gemmata* assemblies.

Annotation	CJuql4	JW3-8s0	JW11	JW9	UQM2246
Transposase	36	88	67	81	217
Integrase	11	6	18	6	4
Recombinase	12	36	10	37	22
Prophage-like regions	4	3	4	3	7

Table 5.3: Summary of mobile genetic element annotations in *Gemmata* genome assemblies. Total number of proteins annotated as 'transposase', 'integrase' or 'recombinase' in the prokka annotations are shown for each genome. Prophage-like regions refers to the total number of predicted prophage containing regions generated by PHAST (Zhou *et al.*, 2011).

The remarkably high level of rearrangements in these genomes may be due to the high numbers of mobile genetic elements in these genomes (Table 5.3). This is likely in part due to the large size of *Gemmata* genomes, as genome size is thought to be the largest contributor for both the overall number and density of insertion sequence (IS) elements (Touchon and Rocha, 2007).

High levels of rearrangements associated with IS density have been found in various linages in Cyanobacteria (Bhaya *et al.*, 2007). Evidence that certain IS families are under positive selection in the ammonia-fixing cyanobacterium *Crocosphaera watsonii* (Mes and Doeleman, 2006), indicates that tranposition can provide a selective advantage in organisms living in similar oligotrophic environments and with similar genome sizes as *Gemmata* species. Kaneko *et al.* (2007) have suggested that lifestyle adaption may play a part in copy-number expansion in the Cyanobacteria *Microcystis aeruginosa*, which contains 452 ISs compared to < 100 in other related genera, as mutations and rearrangements associated with transposition were also enriched in the genome.

Rapid expansion of transposons has also been observed in the genomes of *Salinivibrio* symbionts in anglerfish, in which 28-31% of the coding sequences are transposase genes, compared to 2% in their non-symbiont counterparts (Hendry *et al.*, 2018). High transposon copy-number has been found in the genomes of Baltic Sea Cyanobacteria. The enrichment of transposase transcripts in ocean metagenomes suggests that these are also actively expressed. It has been suggested that these transposons can accelerate certain evolutionary processes (Pál and Papp, 2013), by providing genomic plasticity and hence an adaptive advantage in changeable coastal environments (Vigil-Stenman *et al.*, 2017)

Other examples of MGE enrichment can also be found in slow-growing free-living bacteria and symbionts (Schmitz-Esser *et al.*, 2011, Vigil-Stenman *et al.*, 2015). The unusual budding mode of reproduction and slow growth of *Gemmata* species may reduce their capacity to expunge such elements. An analagous example are the asexually reproducing eukaryotes such as bdelloid rotifers, which have developed additional defences to reduce transposon activity (Flot *et al.*, 2013) and prevent deleterious proliferation of transposons (Arkhipova and Meselson, 2005).

5.6 Future Directions

Many ncRNAs are poorly conserved at the sequence level, which makes it difficult to identify deeply-conserved ncRNAs that have changed significantly in nucleotide sequence. One strategy for discovering these ncRNAs is the comparison of transcriptomes across genomes which are phylogenetically distant, but retain enough shared conserved sequence that conserved transcripts can be detected. As part of a follow-on project to a previous study by Lindgreen *et al.* (2014), isolates from genera *Gemmata* and *Halococcus* were chosen as under-sampled taxa that provide a phylogenetically distant datapoint for comparison with Bacterial genomes.

This chapter describes the genome assembly, annotation and a preliminary comparative analysis of five *Gemmata* isolates. While these assemblies were generated as part of a larger project, the results from the initial annotation and comparison of these genomes highlighted significantly unusual genome architecture, which warranted further analysis in their own right. Most noteably, these genomes are highly enriched with mobile genetic elements, and appear to have highly dynamic genome structures. Future work aims to generate transcriptomes of these strains, with the primary aim of annotating novel ncRNAs by comparative transcriptomics as previously described (Lindgreen *et al.*, 2014). This project also includes the assembly of *Halococcus* genomes, which I have also assembled and annotated (unpublished data).

These genomes also provide a resource for further study of the enigmatic *Planctomycetes* bacteria, in particular the genomic basis for the eukaryotic and archaea-like phenotypes observed in *Gemmata* strains. Additional annotation of these *Gemmata* genomes is required to further characterise MGEs, and identify if these are active, and identify the mechanisms behind the apparent high rate of recombination in these genomes. These genomes also appear to replicate the previous observations of high transposon copy-number in bacteria living in complex aquatic environments, raising questions about their proliferation and possible beneficial roles within their host genomes.

References

Arkhipova, I. and Meselson, M. (2005). Deleterious transposable elements and the extinction of asexuals. *Bioessays*, *27*(1), 76–85.

Bhaya, D., Grossman, A. R., Steunou, A.-S., Khuri, N., Cohan, F. M., Hamamura, N., Melendrez, M. C., Bateson, M. M., Ward, D. M., and Heidelberg, J. F. (2007). Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME J.*, *1*(8), 703–713.

Chaisson, M. J. and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, *13*, 238.

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, *5*(6), e11147.

Devos, D. P. (2013). *Gemmata obscuriglobus*. *Curr. Biol.*, *23*(17), R705–7.

Flot, J.-F., Hespeels, B., Li, X., Noel, B., Arkhipova, I., Danchin, E. G. J., Hejnol, A., Henrissat, B., Koszul, R., Aury, J.-M., Barbe, V., Barthélémy, R.-M., Bast, J., Bazykin, G. A., Chabrol, O., Couloux, A., Da Rocha, M., Da Silva, C., Gladyshev, E., Gouret, P., Hallatschek, O., Hecox-Lea, B., Labadie, K., Lejeune, B., Piskurek, O., Poulain, J., Rodriguez, F., Ryan, J. F., Vakhrusheva, O. A., Wajnberg, E., Wirth, B., Yushenova, I., Kellis, M., Kondrashov, A. S., Mark Welch, D. B., Pontarotti, P., Weissenbach, J., Wincker, P., Jaillon, O., and Van Doninck, K. (2013). Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature*, *500*(7463), 453–457.

Franke, J. D., Blomberg, W. R., Todd, R. T., Thomas, R. W., and Selmecki, A. M. (2018). Assembly of a complete genome sequence for *Gemmata obscuriglobus* reveals a novel prokaryotic rRNA operon gene architecture. *Antonie Van Leeuwenhoek*, *111*(11), 2095–2105.

Franzmann, P. D. and Skerman, V. B. (1984). *Gemmata obscuriglobus*, a new genus and species of the budding bacteria. *Antonie Van Leeuwenhoek*, *50*(3), 261–268.

Fuchsman, C. A. and Rocap, G. (2006). Whole-genome reciprocal BLAST analysis reveals that *planctomycetes* do not share an unusually large number of genes with Eukarya and Archaea. *Appl. Environ. Microbiol.*, *72*(10), 6841–6844.

Fuerst, J. A. (2005). Intracellular compartmentation in *planctomycetes*. *Annu. Rev. Microbiol.*, *59*, 299–328.

Fuerst, J. A. and Sagulenko, E. (2011). Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function. *Nat. Rev. Microbiol.*, *9*(6), 403–413.

Hendry, T. A., Freed, L. L., Fader, D., Fenolio, D., Sutton, T. T., and Lopez, J. V. (2018). Ongoing Transposon-Mediated Genome Reduction in the Luminous Bacterial Symbionts of Deep-Sea Ceratioid Anglerfishes. *MBio*, *9*(3).

Hunt, M., Silva, N. D., Otto, T. D., Parkhill, J., Keane, J. A., and Harris, S. R. (2015). Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.*, *16*, 294.

Kaneko, T., Nakajima, N., Okamoto, S., Suzuki, I., Tanabe, Y., Tamaoki, M., Nakamura, Y., Kasai, F., Watanabe, A., Kawashima, K., Kishida, Y., Ono, A., Shimizu, Y., Takahashi, C., Minami, C., Fujishiro, T., Kohara, M., Katoh, M., Nakazaki, N., Nakayama, S., Yamada, M., Tabata, S., and Watanabe, M. M. (2007). Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Res.*, *14*(6), 247–256.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, *27*(5), 722–736.

Lindgreen, S., Umu, S. U., Lai, A. S.-W., Eldai, H., Liu, W., McGimpsey, S., Wheeler, N. E., Biggs, P. J., Thomson, N. R., Barquist, L., Poole, A. M., and Gardner, P. P. (2014). Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Comput. Biol.*, *10*(10), e1003907.

Lindsay, M. R., Webb, R. I., Strous, M., Jetten, M. S., Butler, M. K., Forde, R. J., and Fuerst, J. A. (2001). Cell compartmentalisation in *planctomycetes*: novel types of structural organisation for the bacterial cell. *Arch. Microbiol.*, *175*(6), 413–429.

Lonhienne, T. G. A., Sagulenko, E., Webb, R. I., Lee, K.-C., Franke, J., Devos, D. P., Nouwens, A., Carroll, B. J., and Fuerst, J. A. (2010). Endocytosis-like protein uptake in the bacterium *Gemmata obscuriglobus*. *PNAS*, *107*(29), 12883–12888.

McInerney, J. O., Martin, W. F., Koonin, E. V., Allen, J. F., Galperin, M. Y., Lane, N., Archibald, J. M., and Embley, T. M. (2011). Planctomycetes and eukaryotes: a case of analogy not homology. *Bioessays*, *33*(11), 810–817.

Mes, T. H. M. and Doeleman, M. (2006). Positive selection on transposase genes of insertion sequences in the Crocosphaera watsonii genome. *J. Bacteriol.*, *188*(20), 7176–7185.

Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*(2), 292–294.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693.

Pál, C. and Papp, B. (2013). From passengers to drivers: Impact of bacterial transposable elements on evolvability. *Mob. Genet. Elements*, *3*(1), e23617.

Rivas-Marin, E., Stettner, S., Gottshall, E. Y., Santana-Molina, C., Helling, M., Basile, F., Ward, N. L., and Devos, D. P. (2019). Essentiality of sterol synthesis genes in the *planctomycete* bacterium *Gemmata obscuriglobus*. *Nat. Commun.*, *10*(1), 2916.

Schmitz-Esser, S., Penz, T., Spang, A., and Horn, M. (2011). A bacterial genome in transition–an exceptional enrichment of IS elements but lack of evidence for recent transposition in the symbiont *Amoebophilus asiaticus*. *BMC Evol. Biol.*, *11*, 270.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069.

Silby, M. W., Winstanley, C., Godfrey, S. A. C., Levy, S. B., and Jackson, R. W. (2011). Pseudomonas genomes: diverse and adaptable. *FEMS Microbiol. Rev.*, *35*(4), 652–680.

Speth, D. R., van Teeseling, M. C. F., and Jetten, M. S. M. (2012). Genomic analysis indicates the presence of an asymmetric bilayer outer membrane in *Planctomycetes* and *Verrucomicrobia*. *Front. Microbiol.*, *3*, 304.

Touchon, M. and Rocha, E. P. C. (2007). Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.*, *24*(4), 969–981.

Vigil-Stenman, T., Ininbergs, K., Bergman, B., and Ekman, M. (2017). High abundance and expression of transposases in bacteria from the Baltic Sea. *ISME J.*, *11*(11), 2611–2623.

Vigil-Stenman, T., Larsson, J., Nylander, J. A. A., and Bergman, B. (2015). Local hopping mobile DNA implicated in pseudogene formation and reductive evolution in an obligate *cyanobacteria*-plant symbiosis. *BMC Genomics*, *16*, 193.

Wang, J., Jenkins, C., Webb, R. I., and Fuerst, J. A. (2002). Isolation of *Gemmata*-like and *Isosphaera*-like *planctomycete* bacteria from soil and freshwater. *Appl. Environ. Microbiol.*, *68*(1), 417–422.

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

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Res.*, *39*(Web Server issue), W347–52.

Chapter 6

Conclusion

6.1 Summary

This thesis describes work on several projects involved in the identification and comparison of bacterial ncRNA genes, using homology search and comparative transcriptomics (outlined in Figure 6.1). These genes are considerably more difficult to study than their proteinaceous bretheren, as they lack obvious genomic signals such as stop/start codons, codon bias. Features such as their short length and flexible requirements for function allow ncRNA nucleotide sequences to change rapidly over a small span of evolutionary time.

The potential reasons for observations of rapid sRNA sequence turnover are reviewed in the our previous work (Jose *et al.*, 2019), in which the relative probabilities of ncRNA gene acquisition due to *de novo* gene formation, exaptation and horizontal gene transfer are considered. In Chapter 2, we asked if this observation is due to the limitations of pairwise sequence alignment-based homology search methods. We developed a homology assignment pipeline based on an iterative profile HMM approach, which incorporates signals of synteny, to study the conservation of *Salmonella* Typhimurium sRNAs. Manual predictions of sRNA gene origins, using signals of sRNA conservation and the function of nearby genes identified instances of recent sRNA gene acquisitions in *Salmonella*. This work outlines a sensitive pipeline for annotating highly diverse sRNA sequences, and also identified sequence features of sRNAs which are likely to return high numbers of false positive homology search results.

Transcriptome data for an economically-relevant pathogen, *P. syringae* pv. *actinidiae* (*Psa*) grown *in vitro*, was generated for this project, primarily for the aim of identifying ncR-

NAs relevant to virulence. In Chapter 3, *in vitro* gene expression, under multiple growth conditions, was compared to data generated *in planta*, as well as to the literature. The *in vitro* data captured major responses to nutrient stress, namely the production of siderophores in nutrient-depleted conditions. Under starvation activation of metabolic adaptions to alternative carbon sources, as well as expression of virulence genes involved in adhesion, defence against plant defence molecules and physiological stress were observed.

Visualisations of *in vitro* and *in planta* transcriptomes were used to guide the annotation of novel ncRNAs expressed in *Psa*. The homology search pipeline developed in Chapter 2 was applied to identify conserved candidates and generate multiple sequence alignments, which were used to look for signals of secondary structure conservation that may indicate function.

Chapter 5 describes work on genome assemblies of five genomes of the genus *Gemmata*. Preliminary analysis of these genomes identified that these isolates have diverse genome structures and a high density of mobile genetic elements. This is part of a larger work aiming to generate genomes and transcriptomes for an ongoing project that aims to identify conserved ncRNAs using comparative transcriptomics.

6.1.1 Where do sRNA genes come from?

We first considered the current literature on sRNA evolution, which suggests that sRNAs are likely to be rapidly gained and and lost, and exhibit high sequence turnover. In our review, we considered the potential reasons for gene gain and loss, as the small size and minimal functional requirements for an sRNA make them attractive models for observing or modelling *de novo* gene evolution. We also explored the relative likelihood of different routes for the formation of *de novo* genes, either by a step-wise evolutionary process, or *via* the capture of stochastic transcription of intergenic regions, termed ’transcriptional noise’.

To explore if this observation is due to the limitations of pairwise sequence alignment-based homology search methods, we developed a pipeline based on iterative profile HMM searches, aiming to accurately identify highly divergent homologues from single sequences.

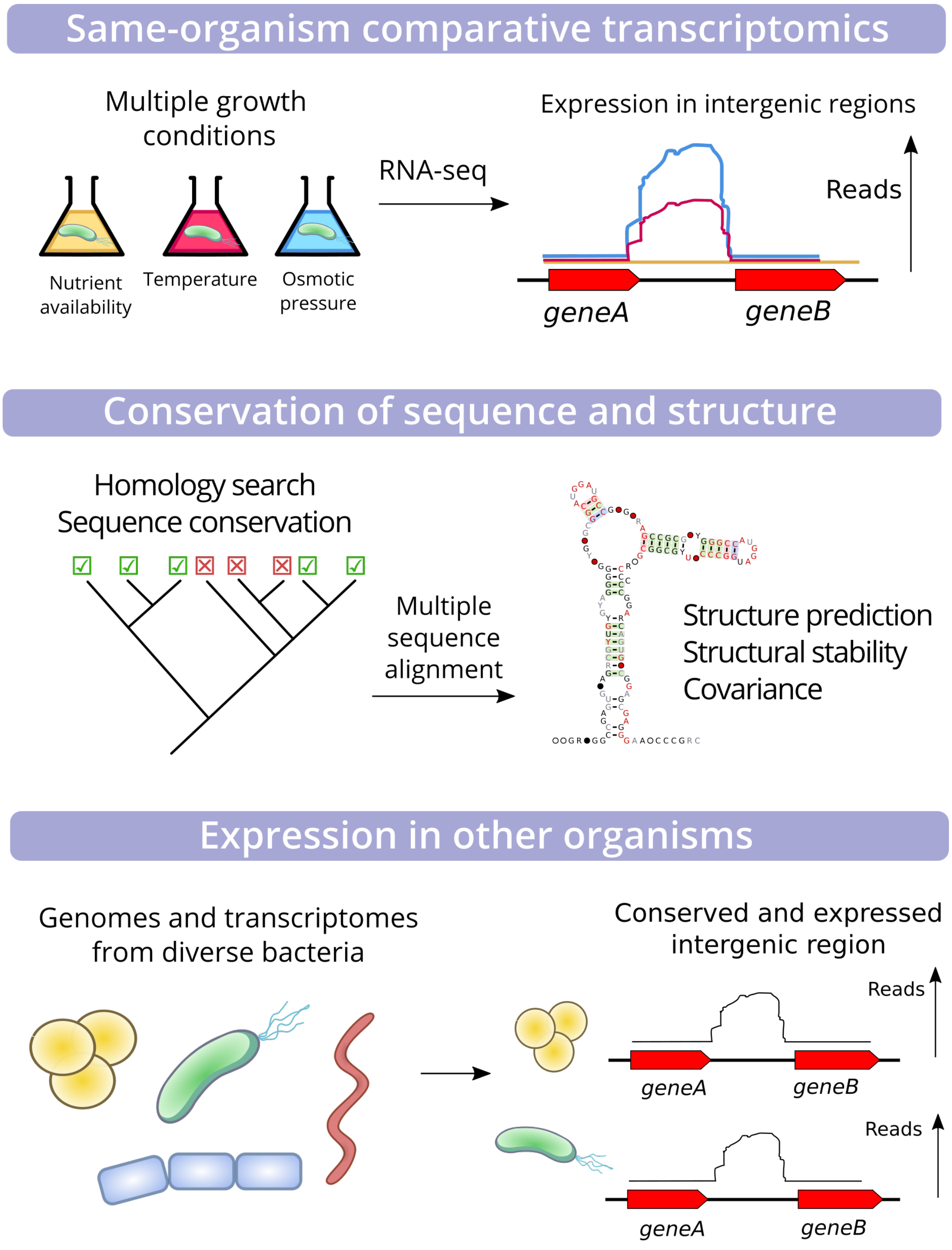

Same-organism comparative transcriptomics
Multiple growth conditions
Nutrient availability
Temperature
Osmotic pressure
RNA-seq
Expression in intergenic regions
Reads
geneA
geneB
Conservation of sequence and structure
Homology search
Sequence conservation
☑ ☑ ☑ ☒ ☒ ☒ ☑ ☑
Multiple sequence alignment
Structure prediction
Structural stability
Covariance
Expression in other organisms
Genomes and transcriptomes from diverse bacteria
Conserved and expressed intergenic region
Reads
geneA
geneB
Reads
geneA
geneB

Figure 6.1: Methods for the identification of ncRNAs used in this thesis. Top: Same-organism comparative transcriptomics, where a single strain is grown in multiple growth conditions to identify condition-specific ncRNA expression in intergenic regions (Chapter 4). **Middle:** Conservation of sequence and structure, using homology search to study the sequence conservation of ncRNAs (Chapter 2). This can then be used to identify structure conservation, which may be used to improve homology search or as an indicator of function for candidate ncRNAs (Chapter 4). **Bottom:** Expression in other organisms, where the conservation of expression homologous sequences can be used as a signal of function (Chapter 4), particularly over large phylogenetic distances (Chapter 5).

We focused on intergenic ncRNAs, which we expect to be more free to rapidly evolve than ncRNA genes opposite to, or that have substantial overlap with, protein-coding genes. Allowing for more sequence divergence presented several issues with models for paralogous ncRNAs with high similarity overlapping with each other, reverse strand matches from palindromes, and small sequences with low sequence complexity or similarity to sequence motifs generating highly-multi-copy annotations.

A 3-model approach was implemented, incorporating nearby sequences as synteny anchors. This was effective for resolving annotation conflicts and separating true and false positives. A flanking protein approach was also trialled, did not significantly improve the accuracy of results and was more computationally demanding.

The homology search pipeline has been automated, with a large scale analysis such as in Chapter 2 taking approximately 24hrs with a high-performance desktop computer. This method was also successfully applied to study the conservation of candidate ncRNAs in *Psa*, and identified a functionally characterised homologue in *P. aeruginosa*.

6.1.2 Methods to identify horizontally-acquired ncRNAs

Conservation of sRNA sequences and synteny anchors, as well as surrounding gene annotations were used to establish approximate ages and gene origins for *Salmonella* sRNAs. The analysis of gene origins was largely exploratory, aiming to identify if poor conservation was due to sequence divergence or the product of gene turnover. Initially, sRNA genes were classed manually using conservation and gene annotations. A large scale annotation of sRNA flanking proteins was subsequently added to identify the proportion of MGE-associated proteins near an sRNA, in an attempt to provide a metric for the gene origin classification. While this method was able to distinguish between poorly conserved sRNAs and HGT-associated sRNAs, vertically-inherited sRNAs were also associated with MGE insertions. The established method is significantly more computationally intensive than the homology search analysis, and required a curated list of protein descriptions associated with mobile genetic elements. In this case, a measurement of conservation and phylogenetic spread may be a simpler way to detect HGT-associated ncRNAs.

Chromosomal MGEs are likely to be a largely unexplored source of ncRNAs, however prokaryotic MGE annotation has lagged behind chromosomal genes (Frost *et al.*, 2005). MGE nomenclature and classification methods are varied (Piégu *et al.*, 2015), and annotation information is often poorly integrated across databases. Early large-scale MGE databases such as ACLAME (Leplae *et al.*, 2010) have not been maintained (last update 2013). Other databases are restricted to MGEs in a specific organism or clade (Partridge and Tsafnat, 2018), or focus on specific MGE types (Arndt *et al.*, 2016). Some poorly understood or diverse classes of MGEs, such as group II introns, also remain difficult to identify without manual curation steps (Candales *et al.*, 2012).

Automated generalised methods for MGE annotation are emerging, combining homology, repeat-identification and synteny information to find MGEs (Berthelier *et al.*, 2018). As annotation improves, we will be better able identify and study the spread and evolution of ncRNAs associated with MGEs through bacterial lineages.

6.1.3 Tracking ncRNAs through bacterial phylogenies

The study of many ancient ncRNAs and proteins is obfuscated by genetic drift (Hoeppner *et al.*, 2012). The rapid evolution of bacterial sRNAs may serve as a useful analogous system to test homology search tools designed to study deep time evolutionary processes, as the most rapidly evolving sRNAs have diverged on a timescale over which protein sequences and synteny remain stable. Incorporating synteny is a useful way to identify bacterial sRNAs, as nearby protein coding genes are likely to have a slower rate of sequence divergence. In Chapter 2 synteny information also provided an estimate of the conservation of the locus containing an sRNA, which allowed us to identify sRNAs that had diverged beyond sequence alignment. However, this was only found to have occurred in a small proportion of *Salmonella* sRNAs, and the apparent sequence divergence was often caused by insertions rather than significant amount of nucleotide substitutions.

Incorporating short nucleotide sequences flanking sRNAs in an independent homology search provided a way to track the locus containing an sRNA. In many cases this allowed us to increase sensitivity by allowing more divergent sequences, without losing the ability to discern false positives by decreasing specificity. Using the immediate nucleotide sequences allowed the detection of sRNAs sequences in small homologous MGEs that may not have been detected by using flanking proteins. Many sRNAs were located at sites of recombination or MGE insertion, and the short length of the flanking sequences allowed us to detection of the origin of an MGE after a locus had been disrupted or eroded due to purifying selection.

Homology search will have varying effectiveness depending on the selection pressures on the sRNA sequence. While structural requirements can be modelled, functional information is difficult to incorporate. We used expression and observed interactions with RNA-binding proteins as a proxy for regulon size, as these appear to correlate with the number of interactions that an sRNA participates in, which appears to restrict sequence conservation. However it is difficult to explore the relationship between sequence divergence and function without additional information in other species for comparison.

6.1.4 Integrative approaches to gene annotation

The lack of centralised resources for bacterial ncRNA annotations and function is a limiting factor for doing large-scale comparative studies. Currently, functional information for sRNAs depends on the availability of data-sets in the literature. Rfam (Nawrocki *et al.*, 2015) and RNAcentral (The RNAcentral Consortium, 2019) are well-maintained databases for ncRNA sequence and structural information, however these contain limited information about the functions and targets of individual ncRNAs besides literature references. Several attempts have been made to create resources for functionally characterised sRNAs, however at the time of writing these are not currently maintained (Huang *et al.*, 2009, Li *et al.*, 2013) or are no longer functional (Pischimarov *et al.*, 2012, Wang *et al.*, 2016).

Genome annotations for the location and targets of regulatory elements are provided for eukaryotic model organisms in NCBI (https://www.ncbi.nlm.nih.gov/refseq/functionalelements/), ENCODE (ENCODE Project Consortium, 2012) and ensembl (Zerbino *et al.*, 2016, 2015). Small databases of bacterial transcription factors binding sites such as Prodoric (Eckweiler *et al.*, 2018) and collectTF (Kiliç *et al.*, 2014), have yet to be included as part of prokaryotic annotation pipelines. As functional profiling methods for sRNAs and their regulons increase in popularity, a centralised resource for bacterial ncRNA annotations and functions will be vital for large-scale comparative studies.

6.1.5 Dense and varied sampling required to provide power to comparative approaches

The continued integration of new experimental technologies and computational methods are making the functional characterisation of ncRNAs easier, cheaper and more comprehensive (Georg *et al.*, 2019, Stav *et al.*, 2019). Work to generate data-sets in emergent strains of *S.* Typhimurium (Canals *et al.*, 2019) aims to provide insights into how rapidly changes in expression and regulation can occur.

We have used a complementary experimental and computational approach to identify and rank candidate ncRNAs in *Psa*. However, this analysis was limited in power as the design of this experiment was not complete, and had variable numbers of replicates. As this restricted the *in vitro* analysis to pairwise comparisons of growth media, growth phase is likely to be a confounding factor in this analysis. Additional replicates in minimal media in other growth phases would help to resolve this issue. The low read depth and unstranded nature of the *in planta* data was only able to confirm the expression of highly-expressed transcripts, and it is likely that weakly expressed ncRNA transcripts have been missed. This is also the case for the *P. syringae* pv. *tabaci* data. A replicated experiment in another *Psa* strain or *P. syringae* pathovar, or a same-family outgroup such as *P. fluorescens*, *P. putida* or *P. aeruginosa* would be useful to confirm the expression of homologous sequences, and to identify more novel ncRNAs.

Within-host RNA-seq of bacterial pathogens are useful for condition-specific expression of ncRNA genes and regulons important for pathogenicity (Westermann *et al.*, 2016). A number of experiments have now been performed *in vivo*, including work to explore the changes in both the bacteria and the host. However, these experiments are more difficult to achieve *in planta*, due to the larger size of plant cells, tough cell walls and difficulty in enriching bacterial reads (Nobori and Tsuda, 2018). Despite these limitations, we successfully identified a variety of expressed candidate ncRNAs in *Psa* grown *in planta*. These data were also useful for confirming the expression of candidate ncRNAs identified using *in vitro* data.

6.1.6 Future work

We plan to write a paper based on the work described in Chapter 2, and package the pipeline for wider use. We intend to submit this work to a specialist journal such as Nucleic Acids Research. I also aim to further explore sequence features and conservation patterns that are likely to cause issues with homology searches, and to identify genes within mobile genetic elements, to perform the gene origin analysis without large amounts of additional manual curation or increased computational load.

The analysis of the *Psa* transcriptomes in Chapter 3 and Chapter 4 will also provide the basis for a future publication. This will primarily focus on the *Psa* candidate ncRNAs identified in Chapter 4. Experiments are planned to experimentally verify and functionally characterise these candidates. Of particular interest is the *pesA* homologue present in *Psa*, which has a role in intracellular survival in *P. aeruginosa*. PesA shows different expression patterns in *Psa*, which has a different set of environmental stresses as a plant pathogen. Recent comparisons between Enterobacteriaceae pathogens that infect plant and animal hosts show that conserved sRNAs can acquire lifestyle-specific functions. ArcZ, which is involved in the anaerobic stress response in the gut pathogen *E. coli* K12 (Mandin and Gottesman, 2010), also has a function in surviving H_2O_2 plant defence in *Erwinia carotavora* (Schachterle *et al.*, 2019). The change of expression pattern of PesA suggests that a similar process of functional specialisation may have occurred in *Pseudomonas*.

The genome assembly and analysis of the planctomycetes genomes, as well as representatives of the genus *Halococcus* will continue as part of an ongoing collaboration to identify and compare ncRNA expression across short phylogenetic distances (Lindgreen *et al.*, 2014).

6.2 Concluding statement

Rapidly developing experimental and computational methods are increasing the power and scope of comparative methods. While the study of pathogens often focuses on the study of protein-coding genes, the presence and function of ncRNAs are increasingly becoming appreciated, as these can effect large-scale changes in gene regulation. Efforts to identify and characterise mobile genetic elements that are enriched for sRNA genes associated with virulence, in particular pathogenicity islands (Melnyk *et al.*, 2019), are likely to reveal many novel sRNAs and other ncRNAs.

For those sRNAs which are subject to rapid sequence change, the question remains as to how their regulons evolve and diversify. As conserved ncRNAs with divergent sequences can have conserved functions (Horler and Vanderpool, 2009), the comparison of rapidly-diversifying sRNAs may provide insight into the relationships between structure, sequence and function can co-evolve. Our increased ability to understand the function and essentiality of these genes will facilitate such studies.

Further work is needed to explore whether the rapid gene gain and loss of sRNAs occurs outside of fast-evolving pathogen species, or in less well-studied phyla. Using a starting point outside of the *Salmonella-Escherichia* lineage would be interesting to see if homology search annotations from different phylogenetic starting points show similar patterns of conservation. Comparisons of sRNA conservation within other families, such as *Pseudomonas*, *Streptococcus*, *Staphylococcus* or *Mycobacteria* are also needed to explore the evolutionary dynamics of sRNAs in organisms with different frequencies of horizontal gene transfer events, or with different selection pressures on genome architecture. The data-sets generated in *Psa* provide a comprehensive starting point for such a study, similar to the *S.* Typhimurium data-set by (Kröger *et al.*, 2013) used in Chapter 2.

References

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D. S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.*, *44*(W1), W16–21.

Berthelier, J., Casse, N., Daccord, N., Jamilloux, V., Saint-Jean, B., and Carrier, G. (2018). A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga Tisochrysis lutea. *BMC Genomics*, *19*(1), 378.

Canals, R., Hammarlöf, D. L., Kröger, C., Owen, S. V., Fong, W. Y., Lacharme-Lora, L., Zhu, X., Wenner, N., Carden, S. E., Honeycutt, J., Monack, D. M., Kingsley, R. A., Brownridge, P., Chaudhuri, R. R., Rowe, W. P. M., Predeus, A. V., Hokamp, K., Gordon, M. A., and Hinton, J. C. D. (2019). Adding function to the genome of African *Salmonella Typhimurium* ST313 strain D23580. *PLoS Biol.*, *17*(1), e3000059.

Candales, M. A., Duong, A., Hood, K. S., Li, T., Neufeld, R. A. E., Sun, R., McNeil, B. A., Wu, L., Jarding, A. M., and Zimmerly, S. (2012). Database for bacterial group II introns. *Nucleic Acids Res.*, *40*(Database issue), D187–90.

Eckweiler, D., Dudek, C.-A., Hartlich, J., Brötje, D., and Jahn, D. (2018). PRODORIC2: the bacterial gene regulation database in 2018. *Nucleic Acids Res.*, *46*(D1), D320–D326.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74.

Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.*, *3*(9), 722–732.

Georg, J., Lalaouna, D., Hou, S., Lott, S. C., Caldelari, I., Marzi, S., Hess, W. R., and Romby, P. (2019). The power of cooperation: Experimental and computational approaches in the functional characterization of bacterial sRNAs. *Mol. Microbiol.*.

Hoeppner, M. P., Gardner, P. P., and Poole, A. M. (2012). Comparative analysis of RNA families reveals distinct repertoires for each domain of life. *PLoS Comput. Biol.*, *8*(11), e1002752.

Horler, R. S. P. and Vanderpool, C. K. (2009). Homologs of the small RNA SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence. *Nucleic Acids Res.*, *37*(16), 5465–5476.

Huang, H.-Y., Chang, H.-Y., Chou, C.-H., Tseng, C.-P., Ho, S.-Y., Yang, C.-D., Ju, Y.-W., and Huang, H.-D. (2009). sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. *Nucleic Acids Res.*, *37*(Database issue), D150–4.

Jose, B. R., Gardner, P. P., and Barquist, L. (2019). Transcriptional noise and exaptation as sources for bacterial sRNAs. *Biochem. Soc. Trans.*.

Kiliç, S., White, E. R., Sagitova, D. M., Cornish, J. P., and Erill, I. (2014). CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. *Nucleic Acids Res.*, *42*(Database issue), D156–60.

Kröger, C., Colgan, A., Srikumar, S., Händler, K., Sivasankaran, S. K., Hammarlöf, D. L., Canals, R., Grissom, J. E., Conway, T., Hokamp, K., and Hinton, J. C. D. (2013). An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium. *Cell Host Microbe*, *14*(6), 683–695.

Leplae, R., Lima-Mendez, G., and Toussaint, A. (2010). ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res.*, *38*(Database issue), D57–61.

Li, L., Huang, D., Cheung, M. K., Nong, W., Huang, Q., and Kwan, H. S. (2013). BSRD: a repository for bacterial small regulatory RNA. *Nucleic Acids Res.*, *41*(Database issue), D233–8.

Lindgreen, S., Umu, S. U., Lai, A. S.-W., Eldai, H., Liu, W., McGimpsey, S., Wheeler, N. E., Biggs, P. J., Thomson, N. R., Barquist, L., Poole, A. M., and Gardner, P. P. (2014). Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Comput. Biol.*, *10*(10), e1003907.

Mandin, P. and Gottesman, S. (2010). Integrating anaerobic/aerobic sensing and the general stress response through the ArcZ small RNA. *EMBO J.*, *29*(18), 3094–3107.

Melnyk, R. A., Hossain, S. S., and Haney, C. H. (2019). Convergent gain and loss of genomic islands drive lifestyle changes in plant-associated Pseudomonas. *ISME J.*, *13*(6), 1575–1588.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D. (2015). Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, *43*(Database issue), D130–7.

Nobori, T. and Tsuda, K. (2018). *In planta* Transcriptome Analysis of *Pseudomonas syringae*. *Bio-protocol*, *8*(17), e2987.

Partridge, S. R. and Tsafnat, G. (2018). Automated annotation of mobile antibiotic resistance in Gram-negative bacteria: the Multiple Antibiotic Resistance Annotator (MARA) and database. *J. Antimicrob. Chemother.*, *73*(4), 883–890.

Piégu, B., Bire, S., Arensburger, P., and Bigot, Y. (2015). A survey of transposable element classification systems–a call for a fundamental update to meet the challenge of their diversity and complexity. *Mol. Phylogenet. Evol.*, *86*, 90–109.

Pischimarov, J., Kuenne, C., Billion, A., Hemberger, J., Cemič, F., Chakraborty, T., and Hain, T. (2012). sRNAdb: a small non-coding RNA database for gram-positive bacteria. *BMC Genomics*, *13*, 384.

Schachterle, J., Onsay, D., and Sundin, G. (2019). Small RNA ArcZ regulates oxidative stress response genes and regulons in Erwinia amylovora. *Front. Microbiol.*, *10*, 2775.

Stav, S., Atilho, R. M., Mirihana Arachchilage, G., Nguyen, G., Higgs, G., and Breaker, R. R. (2019). Genome-wide discovery of structured noncoding RNAs in bacteria. *BMC Microbiol.*, *19*(1), 66.

The RNAcentral Consortium (2019). RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res.*, *47*(D1), D221–D229.

Wang, J., Liu, T., Zhao, B., Lu, Q., Wang, Z., Cao, Y., and Li, W. (2016). sRNATarBase 3.0: an updated database for sRNA-target interactions in bacteria. *Nucleic Acids Res.*, *44*(D1), D248–53.

Westermann, A. J., Förstner, K. U., Amman, F., Barquist, L., Chao, Y., Schulte, L. N., Müller, L., Reinhardt, R., Stadler, P. F., and Vogel, J. (2016). Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. *Nature*, *529*(7587), 496–501.

Zerbino, D. R., Johnson, N., Juetteman, T., Sheppard, D., Wilder, S. P., Lavidas, I., Nuhn, M., Perry, E., Raffaillac-Desfosses, Q., Sobral, D., Keefe, D., Gräf, S., Ahmed, I., Kinsella, R., Pritchard, B., Brent, S., Amode, R., Parker, A., Trevanion, S., Birney, E., Dunham, I., and Flicek, P. (2016). Ensembl regulation resources. *Database*, *2016*.

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., and Flicek, P. R. (2015). The ensembl regulatory build. *Genome Biol.*, *16*, 56.

Appendix A

Papers associated with this thesis

A.1 Preface

This section consists of a literature review which discusses the evolutionary dynamics of bacterial small non-coding RNAs (sRNAs). In this review we discuss possible explanations for observed gain and loss of sRNAs over short evolutionary timescales, focusing on the mechanisms and rates by which these genes might originate and diversify. We consider the potential for bacterial genomes to generate and remove *de novo* genes by insertion and deletion events, recombination, or by the capture of a pool of stochastically-arising transcripts, termed "transcriptional noise". Examples of sRNA acquisition *via* exaptation, duplication and diversification, and horizontal gene transfer are also reviewed. We outline the selection pressures expected to act on each of these routes, the evidence supporting such events and their relative probabilities.

This paper is published as follows: Bethany R. Jose, Paul P. Gardner, Lars Barquist; Transcriptional noise and exaptation as sources for bacterial sRNAs. Biochem Soc Trans 30 April 2019; 47 (2): 527–539. doi: https://doi.org/10.1042/BST20180171

Data and scripts for Figure 2 from the paper associated with this work are located in : `lit_review/figure_2`

A.1.1 Contributions

Lars Barquist and I produced the final version of the manuscript. I wrote the initial drafts of the manuscript, and generated all figures. Lars Barquist and Paul Gardner developed the drake equation shown in Figure 4A.

A.2 Preface

The following paper is a summary of a *Legionella mcdadei* genome assembly project that I worked on with collaborators at the University of Canterbury and the University of Otago. *Legionella mcdadei* is a bacterial pathogen that can cause a severe form of pneumonia called Legionnaire's disease, which is especially prevalent in New Zealand. The genomes of two isolates of *L. mcdadei* from New Zealand patients presenting with Legionnaire's disease were sequenced, assembled and annotated.

This paper is published as follows: Osborne AJ, **Jose BR**, Perry J, Smeele Z, Aitken J, Gardner PP, Slow S. 2017. Complete genome sequences of two geographically distinct *Legionella micdadei* clinical isolates. *Genome Announc* 5:e00436-17. DOI: 10.1128/genomeA.00436-17

A.3 Contributions

I performed the genome assembly, annotation and comparison, submitted the data to NCBI and worked on the methods for the draft of the paper.

Appendix B

Chapter 2 Appendices

Dataset	E-value threshold	Number of results	Post-filtering
ST7/74 sRNAs	0.01	108543	32575
	0.001	86942	32511
sRNA flanking regions	0.01	97104	
	0.001	97104	

Table B.1: Effects of different E-value thresholding and filtering on homology search results.

sRNA	Other names
STnc1010	*STnc900*
DapZ	*STnc1020, STnc820*
STnc470	*STnc910*
SroA	*tpe79*
SgrS	*ryaA*
SraA	*psrA/t15*
ChiX	*MicM, SroB, RybC*
STnc480	*STnc970*
RybD	*STnc830*
RybB	*p25*
IsrB-1	*IS092*
SraB	*pke2*
IsrC	*IS102*
RyhB-2	*isrE, RfrB*
STnc540	*sRNA14*
RprA	*IS083*
RydB	*tpe7, IS082*
MgrR	*STnc560*
RyjB	*STnc1120*
RydC	*IS067*
MicC	*IS063, tke8*
FnrS	*STnc580*
SraC	*RyeA*
SdsR	*RyeB, tpke79*
STnc1690	*STnc1690*
STnc200	*STnc200*

RyeF	*STnc860*
RyeC	*tp11, SibA*
STnc1150	*STnc2060*
CyaR	*ryeE*
STnc2070	*STnc1370*
RyfA	*tp1, PAIR3*
GlmY	*tke1, sroF*
MicA	*sraD*
InvR	*STnc270*
GcvB	*IS145*
OmrA	*rygB*
OmrB	*t59, rygA, sraE*
SsrS	*6S*
RygC	*t27, SibC, QUAD1c*
STnc750	*sRNA8*
RygD	*tp8, SibD, C0730*
SraF	*tpk1, IS160, PRE-element*
ArcZ	*SraH, ryhA*
RyhB-1	*SraI, IS176, RfrA*
STnc770	*sRNA6*
STnc1430	*STM3624.1N*
GlmZ	*k19, ryiA, SraJ*
Spf	*Spot 42*
CsrC	*SraK, RyiB, tpk2*
STnc810	*STnc2120*
SraL	*ryjA*
STnc630	*STnc2140*

Table B.2: Alternative sRNA gene names

EggNOG MGE protein descriptions	
CcdB-like toxin protein	cell killing protein encoded within
cryptic prophage	Transposase
PipA protein	anti-termination protein
transposase	Integrase
Inherit from COG: Antirepressor	Inherit from COG: transposase
Integrase catalytic subunit	IS630 family transposase
Major tail protein	Phage transcriptional regulator, AlpA
Plasmid maintenance system antidote protein	plasmid maintenance system antidote protein, xre family
Protein of unknown function (DUF1019)	Transposase IS116 IS110 IS902
Transposase is3 is911	replication protein O
small terminase subunit	Tail assembly protein
tail length tape measure	Terminase, large subunit
bacteriophage protein	enhancing factor (Viral)
Antitermination protein	Minor Tail Protein
P2 GpU Family Protein	phage baseplate
phage minor tail protein L	phage protein
Prophage membrane protein	Qin prophage
CcdA protein	Excisionase
tail component of prophage	tail component of prophage CP-933K
Tail Fiber Assembly protein	tail fiber protein
Toxic component of a toxin-antitoxin (TA) module. A	NinB protein

late control	Late control D family protein
excisionase	Tail fiber protein
transcriptional activator, Ogr delta	lambda NinG
phage regulatory protein, rha family	Pfam:DUF1813
DNA-binding prophage protein	Pfam:Transposase
phage holin	is1 orf2
this blockage is overcome by subsequent expression of antitoxin HigA. Overexpression causes cleavage of a number of mRNAs in a translation-dependent fashion, suggesting this is an mRNA interferase	

Table B.3: Manually curated EggNOG MGE protein descriptions.

Appendix C

List of Supplementary Materials

Supplementary materials associated with this thesis can be found in the following github repository: https://github.com/Gardner-BinfLab/beth-thesis/tree/master/Supplementary_materials

C.0.1 Chapter 1

Data and scripts for Figure 1.1.

`introduction/figure_1_1`

C.0.2 Chapter 2

Figures 2.3, 2.4.

`sal_sRNA/HMM_specificity`

Figure 2.11.

`sal_sRNA/turnover_boxplots`

Figures 2.8, 2.7.

`sal_sRNA/heatmap`

Figures 2.12, 2.11.

`sal_sRNA/MGE_vs_inheritance`

Other files associated with homology search and gene origin classification.

`sal_sRNA/annotation_validation`

`sal_sRNA/bin`

`sal_sRNA/conservation`

`sal_sRNA/gene_origins`

`sal_sRNA/QC`

`sal_sRNA/structures.tar`

`sal_sRNA/summary.txt`

C.0.3 Chapter 3

Psa genome and plasmid annotations.

`psa_RNAseq/annotations`

Gene expression analysis Jupyter notebooks and data files.

`psa_RNAseq/DESeq_analysis`

`psa_RNAseq/In vitro and in planta RNAseq comparisons.ipynb`

`psa_RNAseq/In vitro RNAseq analysis.ipynb`

Transcriptome data.

`psa_RNAseq/mapping`

`psa_RNAseq/NZGL02428_sequencing-summary.html`

Pathview data and analysis

`psa_RNAseq/Pathview_analysis`

C.0.4 Chapter 4

Transcriptome plotfiles, manual ncRNA annotations, data for Figure 4.2.

`psa_ncRNA/manual_annotation`

`psa_ncRNA/sra2plot.sh`

`psa_ncRNA/curated_intergenic_de_novo_ncRNAs_final.gff`

Conservation analysis and selection of final candidate ncRNAs.

`psa_ncRNA/conservation`

`psa_ncRNA/alignments`

`psa_ncRNA/ncRNA_selection.ods`

`psa_ncRNA/pseudomonas_summary.txt`

Figure 4.6

`psa_ncRNA/heatmap`

Gene expression analysis of ncRNAs.

`psa_ncRNA/expression`

Data associated with PesA CopraRNA analysis, Figure 4.15

`psa_ncRNA/pesa`

Data associated with twoAYGGAY motif analysis, Figures 4.4 and 4.5.

`psa_ncRNA/twoAYGGAY`

C.0.5 Chapter 5

Summary statistics for genome assembly and annotation are provided.

`gemmata/`