

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

# **The evolution of selfish genetic elements within bacterial genomes**

A thesis submitted in partial fulfilment of the requirements for the degree  
of  
Ph.D.  
in  
Molecular Evolution

at Massey University, Auckland, New Zealand.

Frederic Bertels

2012



## Abstract

Genes that increase their copy number relative to that of the host genome are termed selfish. Selfish genes are found ubiquitously in bacterial genomes. Within genomes they can often be identified due to their repetitive nature. Short repetitive sequences such as repetitive extragenic palindromic (REP) sequences have been proposed to be selfish genetic elements. However, evidence for the selfishness of REPs is scarce due to the lack of knowledge about their origin, evolution and mechanisms of dispersal. Here, REPs are studied in the model bacterium *Pseudomonas fluorescens* SBW25. The evidence provided suggests that REPs are part of a greater mobile genetic element, which is termed REP doublet forming hairpins (REPINs).

Subsequently, I investigate the cause of REPIN dispersal: a putative transposase. The transposase, named REP-associated tyrosine transposase (RAYT) shares essential motifs with the IS200 family of insertion sequences. However, unlike insertion sequences, RAYTs are found only as single copy genes. This indicates that RAYTs may not be entirely selfish; instead they may have been co-opted by the host to perform a beneficial function.

Finally, two more repetitive sequence classes are studied in the SBW25 genome. Interestingly, both sequence classes consist of a protein coding sequence and a sequence that forms a stable secondary structure in single stranded DNA or RNA. This arrangement is reminiscent of bacterial toxin-antitoxin (TA) systems. Evidence from sequence analyses suggests that the repetitive nature of these elements in SBW25 may be the result of cooperation between REPINs or other replicative elements and the TA systems.

The presented analyses show that despite the streamlined nature of bacterial genomes selfish genetic elements frequently arise, replicate and probably increase their

persistence and spread through cooperation with addictive and duplicative elements respectively.

## **Acknowledgements**

Foremost, I would like to thank my supervisor Professor Paul Rainey for all the advice, guidance and inspiration he has given me over the past three years. Without him the research presented in this thesis would not have been possible, and I hope we keep in touch for years to come. Further, I would like to thank my co-supervisors Dr Justin O’Sullivan and Professor Allen Rodrigo for not only supporting me during the course of my PhD, but also for their help during my first year in New Zealand. Their fascination for research and science was one of the main reasons why I decided to stay in New Zealand for my PhD.

Another essential factor for the completion of a PhD is funding. Hence, I am very grateful for a doctoral scholarship from the Allan Wilson Centre. It was great being part of the Allan Wilson Centre. I especially enjoyed the scientific exchange and social activities at the annual meetings.

The scientific and social interactions with past and present members of the Rainey Lab also played an important role in my research. I had a great time and I hope to meet all of you again! I would like to especially thank Dr Jenna Gallie for endless advice and discussion. Jenna also proofread countless manuscripts and thesis attempts and taught me how to improve my English on many occasions. Dr Xue-Xian Zhang and Yunhao Liu were of great help during my attempts to conduct lab work.

Furthermore I would like to thank Ben Kerr and the rest of the Kerr lab for stimulating discussions and advice during my stay in Seattle, WA. The time in Seattle and the Kerr lab greatly supported my professional development by exposing me to new ideas and different types of thinking.

Finally, I would like to thank my parents Elke and Ralf Bertels, my sister Helen Bertels, my brothers Felix and Julian Bertels and again my fiancée Jenna Gallie for providing

the support I needed to finish this thesis. I also would like to thank Elaine Riley for helping me move around the world, and numerous other things that made my life much easier.

# Table of Contents

<b>ABSTRACT .....</b>	<b>I</b>
<b>ACKNOWLEDGEMENTS.....</b>	<b>III</b>
<b>TABLE OF CONTENTS.....</b>	<b>V</b>
<b>TABLE OF ABBREVIATIONS .....</b>	<b>X</b>
<b>CHAPTER 1: INTRODUCTION .....</b>	<b>1</b>
1.1 THE ROLE OF DNA SEQUENCE AMPLIFICATION IN LIFE.....	1
1.2 SELFISH GENETIC ELEMENTS .....	2
1.2.1 <i>Defining selfish genetic elements</i> .....	2
1.2.2 <i>Duplicative selfish genetic elements</i> .....	3
1.2.3 <i>Addictive selfish genetic elements</i> .....	3
1.3 DUPLICATIVE SELFISH GENETIC ELEMENTS .....	4
1.3.1 <i>Autonomous and non-autonomous transposons</i> .....	4
1.3.2 <i>Retrotransposons</i> .....	5
1.3.3 <i>DNA transposons</i> .....	7
1.3.4 <i>Short repetitive sequences in bacteria</i> .....	8
1.3.5 <i>Plasmids</i> .....	9
1.4 ADDICTIVE SELFISH GENETIC ELEMENTS .....	9
1.4.1 <i>Toxin-antitoxin (TA) systems</i> .....	9
1.4.2 <i>Bacteriocins</i> .....	12
1.4.3 <i>Restriction-modification systems (RMS)</i> .....	13
1.5 OTHER SELFISH GENETIC ELEMENTS .....	14
1.6 CHARACTERISTICS OF THE MODEL ORGANISM PSEUDOMONAS FLUORESCENS SBW25.....	15
1.7 SUMMARY AND OBJECTIVES OF THIS STUDY .....	16
<b>CHAPTER 2: METHODS.....</b>	<b>18</b>
2.1 GENERAL METHODS.....	18
2.1.1 <i>Bioinformatics</i> .....	18
2.1.2 <i>Specific genomes used for analyses</i> .....	18



2.2 METHODS CHAPTER 3 .....	19
2.2.1 <i>Bioinformatics and phylogenies</i> .....	19
2.2.2 <i>Generation of randomized genomes</i> .....	19
2.2.3 <i>Frequency determination of most abundant oligonucleotides</i> .....	19
2.2.4 <i>Grouping of highly abundant oligonucleotides in SBW25</i> .....	20
2.2.5 <i>Extending REP sequence groups and identifying the frequency of false positives</i> .....	21
2.2.6 <i>Distribution simulation</i> .....	22
2.2.7 <i>Singlet decay</i> .....	23
2.2.8 <i>Population sequencing</i> .....	23
2.2.9 <i>Testing for excision of REP singlets</i> .....	24
2.3 METHODS CHAPTER 4 .....	24
2.3.1 <i>Bioinformatics and phylogenies</i> .....	24
2.3.2 <i>REP sequence selection in other genomes</i> .....	25
2.4 METHODS CHAPTER 5 .....	25
2.4.1 <i>Genomes</i> .....	25
2.4.2 <i>BLAST search</i> .....	25
2.4.3 <i>Identifying duplications</i> .....	26
2.4.4 <i>Taxonomy information</i> .....	26
2.4.5 <i>Frequency determination of flanking 16-mers</i> .....	26
2.4.6 <i>Calculating the pairwise identity for amino acid sequences and its significance</i> .....	26
2.4.7 <i>Calculating phylogenetic clusters</i> .....	27
2.5 METHODS CHAPTER 6 .....	27
2.5.1 <i>Bioinformatics</i> .....	27
2.5.2 <i>Pairwise identities for R200 sequences</i> .....	27
<b>CHAPTER 3: WITHIN-GENOME EVOLUTION OF REPINS: A NEW CLASS OF BACTERIAL MOBILE DNA .....</b>	<b>29</b>
3.1 INTRODUCTION .....	29
3.1.1 <i>Interspersed repetitive sequences</i> .....	29
3.1.2 <i>Non-autonomous DNA transposons (MITEs)</i> .....	30
3.1.3 <i>Evolution and origin of repetitive sequences in bacteria</i> .....	30

3.1.4 Overview.....	31
3.1.5 Aims .....	31
3.2 RESULTS.....	33
3.2.1 Short sequence frequencies in <i>P. fluorescens</i> SBW25 and <i>P. fluorescens</i> Pf0-1 .....	33
3.2.2 The distribution of REP sequences in the genome of SBW25 .....	36
3.2.3 The replicative unit.....	38
3.2.4 Higher order arrangements of REP sequences.....	48
3.3 DISCUSSION.....	51
3.3.1 Short repetitive sequences .....	51
3.3.2 The replicative unit.....	51
3.3.3 Higher order arrangements of REPs and REPINs .....	53
3.3.4 Concluding comment .....	54
<b>CHAPTER 4: THE CAUSE OF REPIN DISSEMINATION.....</b>	<b>55</b>
4.1 INTRODUCTION.....	55
4.1.3 Aims .....	57
4.2 RESULTS.....	58
4.2.1 Detection of RAYTs, a class of genes linked to REPINs in SBW25 .....	58
4.2.2 Similarities between IS200 transposases and RAYTs .....	60
4.2.3 Association between RAYTs and REPINs in other genomes.....	62
4.3 DISCUSSION.....	66
4.3.1 Overview of the discovery of REPIN-RAYT systems in SBW25 .....	66
4.3.2 Summary of the similarities between the REPIN-RAYT system and IS200/IS605 insertion sequences.....	66
4.3.3 Analysis of higher order arrangements of REPs in different bacterial genomes.....	66
4.3.4 Concluding comments.....	67
<b>CHAPTER 5: EVOLUTIONARY CHARACTERIZATION OF RAYTs, A NOVEL CLASS OF REP AND REPIN-ASSOCIATED GENES .....</b>	<b>69</b>
5.1 INTRODUCTION.....	69
5.1.1 Molecular characteristics of RAYTs and IS200 transposases .....	69
5.1.2 Genomic distribution of housekeeping genes versus insertion sequences .....	71
5.1.3 Phylogenetic methodology.....	77

5.1.4 Aims.....	78
5.2 RESULTS .....	79
5.2.1 Comparison of the genomic distribution of four gene families: RAYTs, IS200, IS110 and def79	
5.2.2 Phylogenetic comparisons between IS200 and RAYT proteins .....	86
5.2.3 The four phylogenetic RAYT clusters and their characteristics .....	90
5.3 DISCUSSION .....	100
5.3.1 Overview of the results.....	100
5.3.2 The genomic distribution of the RAYT gene family.....	100
5.3.3 The relationship between the RAYT and the IS200 family .....	101
5.3.4 RAYT subfamilies and their genomic distribution.....	102
5.3.5 Conclusion.....	103
<b>CHAPTER 6: EVOLUTIONARY CHARACTERIZATION OF TWO REPETITIVE SEQUENCE CLASSES IN THE GENOME OF SBW25 .....</b>	<b>105</b>
6.1 INTRODUCTION .....	105
6.1.1 Regulatory antisense RNA in bacteria .....	105
6.1.2 Computational approaches for identifying non-coding RNAs within bacterial genomes .....	107
6.1.3 Repetitive sequence analysis in the SBW25 genome .....	109
6.1.4 Aims.....	109
6.2 RESULTS .....	110
6.2.1 Characterization of R178 repeat sequences.....	110
6.2.2 R200 repeat sequences.....	121
6.2.3 Association between R200 repeats and REPs/REPINs .....	126
6.3 DISCUSSION .....	130
6.3.1 Overview of the results.....	130
6.3.2 Cooperation of selfish genetic elements.....	130
6.3.3 R178 repeats.....	131
6.3.4 R200 repeats.....	133
6.3.5 Association between R200 repeats and REP/REPIN structures.....	134
6.3.6 Concluding comments .....	135
<b>CHAPTER 7: DISCUSSION .....</b>	<b>136</b>

---

7.1 OVERVIEW OF THE RESULTS .....	136
7.1.1 <i>Summary of Chapter 3: Within-genome evolution of REPINs</i> .....	136
7.1.2 <i>Summary of Chapter 4: Cause of within-genome REPIN dispersal</i> .....	137
7.1.3 <i>Summary of Chapter 5: Characterization of the RAYT family</i> .....	138
7.1.4 <i>Summary of Chapter 6: Novel repetitive elements in the genome of SBW25</i> .....	140
7.2 EVALUATION OF THE IMPLICATIONS .....	142
7.2.1 <i>Technological advances that made this work possible</i> .....	142
7.2.2 <i>Relevance of the developed approaches to the field</i> .....	143
7.2.3 <i>Relevance of the described results to the field</i> .....	145
7.3 FUTURE DIRECTIONS.....	147
7.3.1 <i>REPINs and their associated RAYTs</i> .....	147
7.3.2 <i>Research opportunities arising from studying cluster (c) and (d) RAYTs</i> .....	149
7.3.3 <i>R178 and R200 repeats</i> .....	150
7.4 FINAL COMMENT .....	150
<b>REFERENCES</b> .....	<b>152</b>
<b>APPENDICES</b> .....	<b>174</b>

## Table of Abbreviations

Abbreviation	Meaning
BIMEs	Bacterial interspersed mosaic elements
BLAST	Basic local alignment search tool
BLASTP	BLAST Protein (protein query against protein database)
BLASTN	BLASTN Nucleotide (nucleotide query against nucleotide database)
TBLASTN	Translated BLASTN (protein query against nucleotide database)
bp	base pairs
CAS genes	CRISPR associated genes
CRISPRs	Clustered regularly interspaces short palindromic repeats
ERICs	Enterobacterial repetitive intergenic consensus sequences
IR	Inverted repeat
IS	Insertion sequence
LARDs	Large retrotransposon derivatives
LINEs	Long interspersed elements
LTR	Long terminal repeat
MITE	Miniature inverted repeat transposable elements
NEMISs	<i>Neisseria</i> miniature insertion sequences
NGS	Next-Generation Sequencing
PSK	Post segregational killing
PU	Palindromic units
ORF	Open reading frame
RAYTs	REP-associated tyrosine transposase
REPs	Repetitive extragenic palindromic sequences
REPINs	REP doublets forming hairpins
RMS	Restriction modification system
RNAi	RNA interference
RUP	Repeat unit of pneumococcus
SDR	Small dispersed repeats
SINEs	Short interspersed element
ssDNA	Single stranded DNA
TA	Toxin-antitoxin system
TPRT	Target primed reverse transcription
TRIMs	Terminal-repeat retrotransposons in miniature