COMPUTATIONALLY IDENTIFYING NOVEL ESSENTIAL GENES, INCLUDING NON-CODING RNAs, IN INTERGENIC SEQUENCES IN BACTERIAL GENOMES

by

Chi Kin Ho
B.Sc. University of Warwick 2005

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the
Department of Molecular Biology and Biochemistry

© Chi Kin Ho 2008

SIMON FRASER UNIVERSITY

Summer 2008

All rights reserved. This work may not be reproduced in whole or in part, by photocopy or other means, without permission of the author.
Approval

Name: Chi Kin Ho
Degree: Master of Science
Title of Thesis: Computationally identifying novel essential genes, including non-coding RNAs, in intergenic sequences in bacterial genomes

Examining Committee:

Chair: Dr. Bruce P. Brandhorst
Professor, Department of Molecular Biology and Biochemistry

Dr. Fiona S. L. Brinkman
Senior Supervisor
Associate Professor, Department of Molecular Biology and Biochemistry

Dr. David L. Baillie
Supervisor
Professor, Department of Molecular Biology and Biochemistry

Dr. S. Cenk Sahinalp
Supervisor
Professor, School of Computing Science

Dr. Peter J. Unrau
Internal Examiner
Associate Professor, Department of Molecular Biology and Biochemistry

Date Defended/Approved:  June 12, 2006
Declaration of
Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the right to lend this thesis, project or extended essay to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users.

The author has further granted permission to Simon Fraser University to keep or make a digital copy for use in its circulating collection (currently available to the public at the "Institutional Repository" link of the SFU Library website <www.lib.sfu.ca> at: <http://ir.lib.sfu.ca/handle/1892/112>) and, without changing the content, to translate the thesis/project or extended essays, if technically possible, to any medium or format for the purpose of preservation of the digital work.

The author has further agreed that permission for multiple copying of this work for scholarly purposes may be granted by either the author or the Dean of Graduate Studies.

It is understood that copying or publication of this work for financial gain shall not be allowed without the author’s written permission.

Permission for public performance, or limited permission for private scholarly use, of any multimedia materials forming part of this work, may have been granted by the author. This information may be found on the separately catalogued multimedia material and in the signed Partial Copyright Licence.

While licensing SFU to permit the above uses, the author retains copyright in the thesis, project or extended essays, including the right to change the work for subsequent purposes, including editing and publishing the work in whole or in part, and licensing other parties, as the author may desire.

The original Partial Copyright Licence attesting to these terms, and signed by this author, may be found in the original bound copy of this work, retained in the Simon Fraser University Archive.
ABSTRACT

The increasing availability of bacterial genome sequences and genome-wide laboratory analyses has opened the door for high-throughput bioinformatic approaches to accelerate the discovery of antimicrobial drug targets. To expand the list of such possible drug targets, I have developed an approach to identify novel essential genes encoded in unusually large intergenic regions between previously annotated genes. Applying this approach to the analysis of the intrinsically antibiotic resistant pathogen *P. aeruginosa* PAO1, I computationally predicted at least 5 novel protein-coding genes that were also confirmed to be transcribed by RT-PCR. In addition, at least 5 novel non-coding RNAs were predicted. I used the same computational pipeline to predict such novel putative genes in *Mycobacterium tuberculosis* H37Rv, reflecting the general applicability of the method. Finally, I analyzed the phylogenetic distribution of the essential genes, providing insight into their evolutionary origins.

Keywords:
Bioinformatics; genomics; bacteria; pathogens; antibacterial targets; essential genes; intergenic regions; gene prediction; non-coding RNAs

Subject Terms:
Pathogenic microorganisms; Bioinformatics; Genomics; Drugs – Research – Data processing
DEDICATION

To my dearest Father and Mother

whose Sacrifice, Love and Support have given me all that I have.
ACKNOWLEDGEMENTS

I would like to acknowledge and express my sincere thanks to everyone who has helped and guided me on my thesis work. First and foremost, my deepest gratitude goes to Dr. Fiona Brinkman, for her endless support, guidance, kindness and patience over these past few years. I would also like to thank my supervisory committee, Drs. David Baillie and Cenk Sahinalp, and formerly, Drs. Peter Unrau and Eldon Emberly, for overlooking my progress and their insightful advice. I want to thank everyone in the Brinkman Lab – past and present – who have been involved in the BCID project: David Mulder, Diane Wu, Matthew Laird and Dr. Shannan Ho Sui, and also to all others in the lab who have made my time so enjoyable as a graduate student.

I would like to thank all the funding agencies who have financially supported me over these years. These include the MSFHR/CIHR Bioinformatics Training Program, the Simon Fraser University Community Trust Endowment Fund, Simon Fraser University, and the donor of the Weyerhaeuser Molecular Biology Graduate Scholarship.

I love you, Mum and Dad, my church fellowship, and all my friends who have given me so many words of encouragement, and in particular Karen and Felix, for your emotional and spiritual support during my times of weakness and stress. And my Lord Jesus Christ, whose unending Grace has again and again been the source of my refuge and strength, and an ever-present help in trouble.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approval</td>
<td>ii</td>
</tr>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>v</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xv</td>
</tr>
<tr>
<td>Glossary</td>
<td>xvi</td>
</tr>
</tbody>
</table>

## Chapter 1  Introduction

1. The burden of infectious diseases and antimicrobial resistance ......... 1
   1.1 A brief history ........................................................................... 1
   1.1.1 How antibiotics work, and how bacteria fight back .................... 6
1.2 The need for identifying novel therapeutic targets – genomics
   provides new hope ........................................................................ 9
1.3 Approaches to identifying essential bacterial genes ........................ 10
   1.3.1 Summary ................................................................................ 10
   1.3.2 Definition of essential genes and their interest ....................... 10
   1.3.3 Traditional methods ................................................................ 12
   1.3.4 High throughput transposon mutagenesis .................................. 14
   1.3.5 Other genome-wide studies of essential genes ........................... 19
   1.3.6 Computational methods ........................................................... 20
1.4 Identifying non-coding RNAs in bacteria ........................................ 21
   1.4.1 Computational approaches in identifying ncRNAs ....................... 24
   1.4.2 Experimental approaches in identifying ncRNAs .......................... 26
1.5 Insights gained regarding essential genes and ncRNAs ...................... 28
1.6 An introduction to Pseudomonas aeruginosa: an intrinsically
   antibiotic-resistant pathogen .......................................................... 31
   1.6.1 Transposon mutagenesis studies in Pseudomonas
   aeruginosa ...................................................................................... 33
1.7 Goal of the present research ........................................................... 34

## Chapter 2  Identifying intergenic regions that may contain novel essential genes

2.1 Summary ......................................................................................... 36
2.2 Rationale ....................................................................................... 36
Chapter 3 Identifying novel putative essential protein coding genes

3.1 Summary ................................................................................................................. 58
3.2 Materials and Methods......................................................................................... 58
  3.2.1 BLASTX ........................................................................................................... 58
  3.2.2 Extracting the coding sequence within the IGR ........................................... 60
  3.2.3 Automatic characterization of putatively novel protein-coding genes ........... 61
  3.2.4 Manual inspection .......................................................................................... 63
  3.2.5 Prioritization of predictions for transcription verification ............................. 63
  3.2.6 Cross-referencing with the P. aeruginosa PA14 transposon mutant library ...... 64
  3.2.7 Test data set .................................................................................................... 65
  3.2.8 Lab verification methods .................................................................................. 66
3.3 Results and Discussion ......................................................................................... 67
  3.3.1 Test performance ............................................................................................. 68
  3.3.2 Prediction of 34 novel putative essential protein-coding genes in P. aeruginosa PAO1 .................................................................................................................. 69
  3.3.3 Functional characterization of novel gene predictions ..................................... 70
  3.3.4 Probability of essentiality of novel gene predictions ..................................... 71
  3.3.5 Descriptive summary of 14 transcribed novel protein-coding genes .......... 72
  3.3.6 Previously identified novel non-essential genes ............................................. 88
  3.3.7 High proportion of GC-rich codons correlates with false positives and amplification failure of protein-coding gene predictions ................................. 88
3.4 Concluding remarks ............................................................................................ 91

Chapter 4 Identifying novel putative essential non-coding RNAs .......................... 94
4.1 Summary ............................................................................................................... 94
4.2 Materials and Methods ....................................................................................... 94
LIST OF FIGURES

Figure 1-1: A timeline of the history in the research and development of antibiotics.................................................................3

Figure 1-2: Current antibiotics target three main bacterial mechanisms: cell-wall biosynthesis, protein synthesis and DNA replication and repair.................................................................6

Figure 1-3: Identification of essential and non-essential genes using transposon mutagenesis (A) in a specified region of interest; (B) global transposon mutagenesis (descriptions in text). Red arrows denote essential genes, black arrows non-essential genes...............16

Figure 2-1: Given a transposon gap size cutoff L, this would be an example of a putatively essential intergenic sequence (PEIS) which would be extracted for further analysis. Green horizontal arrows denote annotated genes, and vertical black lines represent transposon insertion locations........................................43

Figure 2-2: (A) The distribution of sizes of intergenic regions (IGRs) in P. aeruginosa PAO1, and (B) a zoomed version of the same. The dotted line denotes the 75% quantile (181bp), the dashed line the 90% quantile (323bp), and the solid line the 95% quantile (426bp).................................................................44

Figure 2-3: Transposon gap sizes along the P. aeruginosa PAO1 chromosome, with notable regions of low insertion density..............47

Figure 2-4: Histogram of transposon gap sizes in the P. aeruginosa PAO1 genome........................................................................48

Figure 2-5: Quantile-quantile plot of transposon gap sizes within ORFs vs. IGRs ........................................................................50

Figure 2-6: A comparison of the distribution of transposon gap sizes resulting from a simulation of 30,100 random insertions (solid bars), with the exponential distribution of the same rate parameter (dashed line).................................................................53

Figure 3-1: Schematic diagram of the genomic context of PA2173.1, a newly identified putative essential protein-coding gene, denoted by the red arrow. Blue arrows denote previously annotated protein-coding genes, and a black triangle denotes the presence
of a transposon mutant in that location. The direction of the arrow denotes the direction of transcription.

Figure 3-2: Schematic diagram of the genomic context of PA0836.2, a newly identified putative essential protein-coding gene, denoted by the red arrow. Green arrow denotes an sRNA.

Figure 3-3: Schematic diagram of the genomic context of PA2566.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.

Figure 3-4: Schematic diagram of the genomic context of PA2747.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.

Figure 3-5: Schematic diagram of the genomic context of PA2926.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.

Figure 3-6: Schematic diagram of the genomic context of PA0457.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.

Figure 3-7: Schematic diagram of the genomic context of PA1366.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.

Figure 3-8: Schematic diagram of the genomic context of PA0613.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.

Figure 3-9: Schematic diagram of the genomic context of PA1152.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.

Figure 3-10: Schematic diagram of the genomic context of PA1838.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.

Figure 3-11: Schematic diagram of the genomic context of PA2118.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.

Figure 3-12: Schematic diagram of the genomic context of PA3762.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.

Figure 3-13: Schematic diagram of the genomic context of PA3414.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.

Figure 3-14: Schematic diagram of the genomic context of PA0050.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.
Figure 3-15: Histogram of GC scores (proportion of GC-rich codons in a given gene) for all annotated genes in the *P. aeruginosa* PA01 genome.

Figure 4-1: Schematic diagram of the genomic context of PA1270.1 (putative cobalamin riboswitch), a previously unannotated and putatively essential ncRNA, denoted by the green arrow. Blue arrows denote previously annotated protein-coding genes, and a black triangle denotes the presence of a transposon mutant in that location. The direction of the arrow denotes the direction of transcription.

Figure 4-1: Schematic diagram of the genomic context of PA1270.1 (putative cobalamin riboswitch), a previously unannotated and putatively essential ncRNA, denoted by the green arrow.

Figure 4-2: Schematic diagram of the genomic context of PA2910.1 (putative yybP-ykoY leader), a previously unannotated and putatively essential ncRNA, denoted by the green arrow.

Figure 4-3: Schematic diagram of the genomic context of PA4055.1 (putative FMN riboswitch), a previously unannotated and putatively essential ncRNA, denoted by the green arrow.

Figure 4-4: Schematic diagram of the genomic context of PA4421.1 (putative RNase P), a previously unannotated and putatively essential ncRNA, denoted by the green dashed arrow. The green solid arrow denotes an annotated ncRNA in the *P. aeruginosa* PA01 genome.

Figure 4-5: Schematic diagram of the genomic context of PA4973.1 (putative TPP riboswitch), a previously unannotated and putatively essential ncRNA, denoted by the green arrow.

Figure 4-6: Distribution of number of members in orthologous sequence sets.

Figure 4-7: Distribution of windows analyzed by RNAz in native vs. shuffled multiple sequence alignments in terms of of (A) Consensus minimum free energy (MFE); (B) SVM classification score; (C) Structure Conservation Index (SCI); and (D) z-score.

Figure 4-8: Distribution of scores of native vs. shuffled windows, in terms of (A) raw RNA score; (B) log-odds RNA score; and (C) sigmoidal RNA score. (D) is the zoomed version of C.

Figure 4-9: Multiple sequence alignment and predicted secondary structure of PA0844.1, a novel putative ncRNA. Alignment columns and secondary structure base-pairs are color-coded according to RNAz conventions (Washietl et al., 2005). Yellow indicates that the base pair is supported by 2 types of pairs. Red indicates that only a single pair-type supports the interaction. The intensity of the color is proportional to the number of sequences having that base pair.
Figure 4-10: Multiple sequence alignment and predicted secondary structure of 3 terminator structures as annotated by (Benen et al., 1989), just upstream of PA1587.1, a novel putative ncRNA. Yellow indicates that the base pair is supported by 2 types of pairs. Red indicates that only a single pair-type supports the interaction. The intensity of the color is proportional to the number of sequences having that base pair.

Figure 4-11: Multiple sequence alignment and predicted secondary structure of PA1587.1, a novel putative ncRNA. Alignment columns and secondary structure base-pairs are color-coded according to RNAz conventions (Washietl et al., 2005). Yellow indicates that the base pair is supported by 2 types of pairs. Red indicates that only a single pair-type supports the interaction. The intensity of the color is proportional to the number of sequences having that base pair.

Figure 4-12: Multiple sequence alignment and predicted secondary structure of PA1655.1, a novel putative ncRNA. Alignment columns and secondary structure base-pairs are color-coded according to RNAz conventions (Washietl et al., 2005). Yellow indicates that the base pair is supported by 2 types of pairs. Red indicates that only a single pair-type supports the interaction. The intensity of the color is proportional to the number of sequences having that base pair.

Figure 4-13: Multiple sequence alignment and predicted secondary structure of PA2291.1, a novel putative ncRNA. Alignment columns and secondary structure base-pairs are color-coded according to RNAz conventions (Washietl et al., 2005). Yellow indicates that the base pair is supported by 2 types of pairs. Red indicates that only a single pair-type supports the interaction. The intensity of the color is proportional to the number of sequences having that base pair.

Figure 4-14: Multiple sequence alignment and predicted secondary structure of PA3919.1, a novel putative ncRNA. Alignment columns and secondary structure base-pairs are color-coded according to RNAz conventions (Washietl et al., 2005). Yellow indicates that the base pair is supported by 2 types of pairs. Red indicates that only a single pair-type supports the interaction. The intensity of the color is proportional to the number of sequences having that base pair.

Figure 5-1: Venn diagram of 155 putatively essential genes in *P. aeruginosa* PAO1, and their essentiality status of orthologs in *P. aeruginosa* PA14 and *E. coli* K-12.

Figure 5-2: Histogram of the phylogenetic distribution of essential genes vs. non-essential genes at the Superkingdom level.
Figure 5-3: Histogram of the phylogenetic distribution of essential genes vs. non-essential genes at the Phylum level ........................................ 148

Figure 5-4: Histogram of the phylogenetic distribution of essential genes vs. non-essential genes at the Class level .................................................. 148

Figure 5-5: Histogram of the phylogenetic distribution of essential genes vs. non-essential genes at the Order level ............................................ 149

Figure 5-6: Histogram of the phylogenetic distribution of essential genes vs. non-essential genes at the Family level ........................................... 149

Figure 5-7: Histogram of the phylogenetic distribution of essential genes vs. non-essential genes at the Genera level ............................................ 150

Figure 5-8: Phylogenetic distribution of essential genes along the chromosome of P. aeruginosa PA01 ................................................................. 151

Figure 5-9: Essential genes that are not phylogenetically distributed. Fourteen genes within the yellow triangle are essential in both P. aeruginosa PA01 and PA14, but are distributed in less than 50 different genera ................................................................. 153

Figure 5-10: Histogram of phylogenetic distribution of essential ncRNAs vs. non-essential ncRNAs at the Superkingdom level .......................... 155

Figure 5-11: Histogram of phylogenetic distribution of essential ncRNAs vs. non-essential ncRNAs at the Genus level ........................................ 156

Figure 6-1: Automated computational pipeline for the identification of putatively essential protein-coding genes and ncRNAs .............. 162

Figure 6-2: Distribution of (A) RNAz z-scores and (B) QRNA sigmoidal RNA scores in native windows vs. shuffled windows .................. 168
LIST OF TABLES

Table 2-1: Fourteen largest intergenic regions in *P. aeruginosa* PA01 (>1kb). The average %G+C in entire genome is 66.6%................. 46

Table 2-2: Comparison of transposon gap size statistics within ORFs vs. IGRs................................................................. 51

Table 2-3: Transposon gap size cutoff for given p-values. Two insertion rates are compared.................................................. 55

Table 3-1: Thirty protein-coding gene predictions using the original IGR extraction algorithm, in addition to a positive control OprF, sorted by proportion of GC-rich codons, and their RT-PCR outcome. Gene prediction IDs in bold are the fourteen transcribed novel putative essential genes described in section 3.3 (Results and Discussion). Those with a proportion of GC rich codons in the deduced coding region highly deviated from the mean are more likely not true genes. .......................................................................................... 90

Table 5-1: Mean, median and maximum number of taxa represented by orthologs of essential genes vs. non-essential genes............... 147

Table 5-2: Mean, median and maximum number of taxa represented by homologs of essential ncRNAs vs. non-essential ncRNAs......... 154

Table 6-1: Six predicted ncRNA candidates in *Mycobacterium tuberculosis* H37Rv with RNAz z-score < -2.0, and QRNA sigmoidal RNA score > 40. ............................................................................................................. 170
GLOSSARY

BLAST  basic local alignment search tool
bp     base-pair(s)
CDD    Conserved Domain Database
COG    Cluster of Orthologous Groups
CRISPR regularly interspaced short palindromic repeat
FMN    flavin mononucleotide
HMM    Hidden Markov Model
IGR    intergenic region
ITS    internal transcribed spacer
kb     kilobase (1,000 nucleotides)
LUCA   last universal common ancestor
Mb     megabase (1,000,000 nucleotides)
MDR    multi-drug resistant
MFE    minimum free energy
MRSA   methicillin-resistant Staphylococcus aureus
NCBI   National Center for Biotechnology Information
ncRNA  non-coding RNA
ORF    open reading frame
PCR    polymerase chain reaction
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEC</td>
<td>Profiling of the <em>E. coli</em> Chromosome</td>
</tr>
<tr>
<td>PEIS</td>
<td>putatively essential intergenic sequence</td>
</tr>
<tr>
<td>PSSM</td>
<td>Position specific scoring matrices</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RPS-BLAST</td>
<td>reverse position-specific BLAST</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>SCFG</td>
<td>stochastic context-free grammar</td>
</tr>
<tr>
<td>SCI</td>
<td>structure conservation index</td>
</tr>
<tr>
<td>SELEX</td>
<td><em>in vitro</em> selection</td>
</tr>
<tr>
<td>sRNA</td>
<td>small RNA</td>
</tr>
<tr>
<td>TB</td>
<td>tuberculosis</td>
</tr>
<tr>
<td>tmRNA</td>
<td>transfer-messenger RNA</td>
</tr>
<tr>
<td>TPP</td>
<td>thiamine pyrophosphate</td>
</tr>
<tr>
<td>TraSH</td>
<td>transposon site hybridization</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>ts</td>
<td>temperature-sensitive</td>
</tr>
<tr>
<td>VRE</td>
<td>vancomycin-resistant <em>enterococcus</em></td>
</tr>
<tr>
<td>XDR</td>
<td>extensively drug resistant</td>
</tr>
</tbody>
</table>
CHAPTER 1  INTRODUCTION

1.1 The burden of infectious diseases and antimicrobial resistance

1.1.1 A brief history

Infectious diseases have plagued humanity since the dawn of civilization. Pandemics such as cholera, plague, influenza, typhoid and tuberculosis were so widespread in ancient times that people rarely lived past middle age. Entire families and villages being wiped out by disease were common, as was evident by the Black Death in the Middle Ages that will always be in history books as one of the most horrific pandemics that wiped out a third of Europe's population (Gottfried, 1983).

Against this backdrop of helplessness towards infectious diseases well into the 19th century, the discovery of antibiotics in the 20th century was almost a miracle to behold. With the first discovery of the antibiotic effects of penicillin by British scientist Alexander Fleming in 1929 (Fleming, 1929), tireless but costly endeavours were begun, which have resulted in a plethora of antibiotic compounds that revolutionized modern medicine.

When the production of penicillin was first industrialized in the 1940s, successful efforts were underway in isolating and developing the early major classes of antibiotics (Fernandes, 2006). In these early days, the discovery process involved screening soil samples for natural products that kill bacteria,
first *in vitro* and then *in vivo*. These include the penicillin and cephalosporins isolated from fungi, as well as streptomycin, tetracycline, erythromycin and vancomycin screened from different strains of the filamentous bacterium *Streptomyces*. In the 1960s, with the failure to isolate more antibiotics from natural sources, scientists began to develop semi-synthetic analogs to existing antibiotics, such as second- and third-generation β-lactams of penicillin and cephalosporin, which added to the arsenal of antimicrobials further with broader spectrum and increased potency. In the 1980s, with the exhaustion of natural product analogs, modifications to the antibacterial agent nalidixic acid introduced the fluoroquinolone class of antibiotics into the antibiotic scene, leading to several new compounds, such as ciprofloxacin, gaining regulatory approval around the world (Fernandes and Chu, 1987). In the late 1980s, with the limited success from microbial fermentation screening, efforts were redirected in the screening of small molecule libraries, leading to the discovery of a new class of synthetic antimicrobial agents, oxazolidinone (Slee et al., 1987), used today as a last resort to treating antibiotic-resistant bacterial infections. However, by the end of the 20th century it seemed that the discovery of new antibacterial agents was grinding to a halt.
Since the advent of antibiotics, millions of lives around the world have been saved, as once deadly infectious diseases seemed to be on the verge of being eradicated or controlled (Hopkins et al., 2005; Heymann, 2006). Despite the success of antibiotics, it is therefore perhaps surprising that infectious diseases are in the present day still such a leading cause of death worldwide, killing more than 14.6 million people each year (WHO, 2004). On the one hand, extreme poverty, the lack of basic sanitation and inadequate funding, coupled with the inability to deliver adequate health care to the most vulnerable in developing countries, are still causing millions of needless deaths from completely treatable diseases. On the other hand, bacteria themselves are fast developing resistance to existing drugs and emerging in both developed and
developing countries, eroding away many hard-won efforts gained in the past and causing major concerns in health care today.

The possibility that bacteria could acquire drug resistance had already been famously foretold by Fleming in 1946: "There is probably no chemotherapeutic drug to which in suitable circumstances the bacteria cannot react by in some way acquiring ‘fastness’ [resistance]" (Fleming, 1946). Scientists soon began noticing the emergence of penicillin-resistant _Staphylococcus pyogenes_ in 1947, just a few years after penicillin was put into market (Barber, 1947). In the 1950s and 1960s, enteric bacteria such as _Escherichia coli, Shigella_ and _Salmonella_ were first found to be multidrug resistant (MDR) (Watanabe, 1963). Later, ampicillin-resistant strains of _Haemophilus influenzae_ and _Neisseria gonorrhoeae_ were discovered (De Graaff et al., 1976; Elwell et al., 1977), and in the case of _Haemophilus_, chloramphenicol- and tetracycline- resistant strains as well (van Klingeren et al., 1977). The re-emergence of multidrug resistant tuberculosis in the 1980s required difficult treatments using six to seven different drugs (Bloom and Murray, 1992; Iseman, 1993). Today, the vast majority of _Staphylococcus aureus_ found in hospitals is penicillin-resistant, and up to 50% are resistant to stronger drugs such as methicillin (Palumbi, 2001; Weinstein, 2001). More recently, methicillin-resistant _S. aureus_ (MRSA) were found in hospitals to be also resistant to vancomycin, the drug of choice for treating MRSA, causing treatment complications (Hiramatsu, 1998; Fridkin, 2001). Among the Gram-negative bacteria, strains of _Pseudomonas aeruginosa_ and _Acinetobacter baumannii_ were
found to be resistant to all antibiotics, seriously impairing the treatment of immunodeficient patients (Levin et al., 1999). At the time of writing, the World Health Organization had just recorded the highest rates ever of multidrug resistant tuberculosis (MDR-TB), and reported the emergence of a virtually unremitting extensively drug-resistant tuberculosis (XDR-TB) in 45 different countries (WHO, 2008).

Canada is no exception in this worldwide phenomenon. The incidence of MRSA, in proportion to the total number of S. aureus isolates, has increased from 1% in 1995, to 8% in 2000, and vancomycin-resistant enterococcus (VRE) has been reported in all provinces since the first report in 1995 (Conly, 2002). It is estimated that the economic burden in Canadian hospitals for isolating MRSA and treating infected patients yearly is CAD$ 42-59 million (Kim et al., 2001).

The consequences of multidrug resistant bacteria are dire. Hospitalized patients who succumb to multidrug resistant bacterial infections face fewer effective treatment options, have more prolonged periods of infection, and have higher rates of death than patients infected by drug-sensitive strains (Cosgrove et al., 2003). It is estimated that drug-resistant infections double the hospital stay, double the mortality, and probably also double the morbidity (Holmberg et al., 1987). Another study has shown that the economic costs associated with treating drug-resistant infections is almost a threefold increase from drug-sensitive infections, and an estimated US$ 1 billion to combat MDR-TB and XDR-TB (Rubin et al., 1999).
Clearly, there is a need for the identification of genuinely new antimicrobial drug targets.

1.1.2 How antibiotics work, and how bacteria fight back

Antibiotics work in either one of the two ways: they kill bacteria (bacteriocidal), or they hamper their growth (bacteriostatic). They can also be classified in terms of their targets. Antibiotics mainly target one of these three bacterial mechanisms (Figure 1-2): 1) bacterial cell-wall biosynthesis; 2) bacterial protein synthesis; and 3) bacterial DNA replication and repair.

Figure 1-2: Current antibiotics target three main bacterial mechanisms: cell-wall biosynthesis, protein synthesis and DNA replication and repair.

β-lactams-containing antibiotics, such as penicillins and cephalosporins, and glycopeptides, such as vancomycin, inhibit the synthesis of the bacterial cell wall (Spratt and Cromie, 1988; Williams, 1996). The cell wall, whether in Gram-positive or Gram-negative bacteria (though substantially thicker in Gram-positive bacteria), contains a peptidoglycan layer, a mesh-like structure whose covalent
crosslinks are formed with the help of the enzyme transpeptidase (also known as penicillin-binding proteins, or PBPs). By occupying the active site of transpeptidases (as in the case of penicillin and cephalosporins) or by binding to the peptide substrate (as in the case of vancomycin), peptidoglycan crosslinks are prevented to be formed, leading to a weaker bacterial cell wall that becomes vulnerable to changes in osmotic pressure. However, since Gram-negative bacteria possess an additional outer membrane, penicillins and vancomycin are mainly effective against Gram-positive bacteria.

Antibiotics that target protein synthesis include the macrolides (Brisson-Noël et al., 1988), tetracyclines (Chopra and Roberts, 2001), aminoglycosides (such as streptomycin) (Fourmy et al., 1996) and oxazolidinones (Xiong et al., 2000), for instance. Specifically, they all act by inhibiting the peptidyl transferase activity of the ribosome, preventing peptide bonds from forming between adjacent amino acids.

Finally, the fluoroquinolones target a type II topoisomerase called DNA gyrase (Shen et al., 1989), an enzyme responsible for uncoiling supercoils introduced inevitably in the process of DNA replication. By inhibiting this enzyme, the accumulation of superhelical tension caused by supercoils results in a build-up of strand breaks and eventually, bacterial cell death.

However, bacteria have also found ways to evade the attack of antibiotics. Development of clinically significant resistance in bacteria can appear in years, even months (Davies, 1996), thanks to the process of natural selection. With a mutation rate in the magnitudes of 1 in $10^8$ per base pair per generation in
bacteria (Drake et al., 1998), a large bacterial population coupled with a short generation time could lead to mutants able to resist the applied antibiotic under certain conditions. In particular, if an antibiotic treatment is at subtherapeutic levels, resistant bacteria can quickly outgrow antibiotic-sensitive bacteria and become the prevalent population.

There are three main mechanisms bacteria have developed to resist antibiotics. Firstly, both Gram-positive and Gram-negative bacteria have developed efflux pumps to pump out antibiotics that have penetrated into the cytoplasm, resisting antibiotics such as tetracyclines that target protein synthesis, a process which only occurs in the cytoplasm (Levy, 1992). By pumping out drugs faster than it can diffuse in, antibiotic concentrations within the cell are kept too low to be effective.

A second antibiotic resistance mechanism involves developing means to chemically disabling the action of the antibiotic, as in the case of the production of β-lactamases by resistant bacteria (Bradford, 2001). β-lactamases hydrolyze antibiotics such as penicillins and cephalosporins, and disable their ability to compromise the formation of crosslinks in the bacterial cell wall.

Finally, bacteria have also learned to modify antibiotic targets, such as that carried out by the Erm methyl transferase enzyme, which methylates a particular adenine residue in the 23S rRNA (Bussiere et al., 1998). As a result, the binding affinity of all members of erythromycin is greatly reduced. A similar example is in the modification of the peptidoglycan terminus from D,D-dipeptide
to D,D-depsipeptide, which lowers the binding affinity of vancomycin by 1,000-fold (Bugg et al., 1991).

1.2 The need for identifying novel therapeutic targets – genomics provides new hope

In the light of the growing ineffectiveness and depletion of existing antibacterial agents against the emergence of increasingly resistant pathogenic bacteria, there are increasing calls to accelerate the discovery of novel antibacterial agents. At a time when pharmaceutical companies are pulling out from the unattractive antibacterial market (Projan, 2003), this need is ever more urgent.

The age of bacterial genomics, which started by the sequencing of the first bacterium, *Haemophilus influenzae*, in 1995 (Fleischmann et al., 1995), opened the door to new opportunities in antibacterial research. As of March 23rd 2008, the NCBI Microbial Genome Resources contains a total of 649 completely sequenced microbial genomes, with an additional 973 microbial genome projects in progress. This represents an unprecedented wealth of genomic data that naturally lends to the development of a novel genomics-based approach to antibacterial drug discovery.

One particular emphasis has been in the search of novel antibacterial targets (Hughes, 2003), as drug-discovery programmes are shifting from a traditional top-down approach of screening compound libraries in search for the ‘magic’ compound, to a bottom-up approach of first identifying targets. With the availability of complete genomic sequences and new advances in high-
throughput mutagenesis techniques, the genome-wide identification of genes essential for bacterial viability can now be performed. Bioinformatics tools can be used to characterize and prioritize potential target candidates, and the increasing availability of genomic sequence of related bacterial strains also leads to the use of comparative genomics techniques to identify the evolutionary distribution and conservation of potential targets (Becker et al., 2006). The latter is of particular importance for developing antibiotics that are either broad-spectrum or narrow-spectrum.

1.3 Approaches to identifying essential bacterial genes

1.3.1 Summary

Genes that are involved in processes essential for a bacterium’s viability are both of academic and practical interest. Researchers have devised different approaches, both experimentally and computationally, for identifying essential genes. In this section, I will review the definition of what is meant by essential genes, as well as the different methods in identifying them, with particular emphasis on transposon mutagenesis, since this approach was an important component of my research.

1.3.2 Definition of essential genes and their interest

Different bacterial genomes have varying sizes, and differ considerably in their gene content and structure. Rearrangements shuffle the genomic order of genes (Sankoff and Nadeau, 2003), and often cause gene deletions (Mira et al., 2001; Rocha, 2003). Genomes increase in size by the process of gene
duplication, or can acquire genes from external sources by the process of horizontal gene transfer (Lawrence and Hendrickson, 2003). However, there is also a subset of genes that are rarely gained or lost because they code for functions required for growth or viability of the organism under almost any condition.

These "essential genes" are of interest in various respects. From a pharmaceutical point of view, genes involved in essential metabolic and biosynthetic pathways in pathogenic organisms are good targets for broad-spectrum antibacterial drugs (Chalker and Lunsford, 2002). Academically, they have sparked excitement in attempts to compile the hypothetical "minimal gene set", the repertoire of genes that is "necessary and sufficient to support cellular life" (Mushegian and Koonin, 1996), and in the case of synthetic biologists, the set of genes using which the simplest free-living organism can be assembled (Ferber, 2004). From the evolutionary biologist's perspective, the set of essential genes is also of considerable interest regarding the reconstruction of the last universal common ancestor (LUCA) of all life forms (Woese, 1998; Lazcano and Forterre, 1999).

Even though essential genes can be semantically understood as those required for viability and growth of the organism, in practice assaying for gene essentiality is highly dependent on the conditions under which the assay is being performed. For example, whereas assaying for colony formation of bacteria grown in "rich media" would tell us about the minimum requirements of an organism to live in a relatively nutrient abundant environment, assaying for
growth on a minimal medium would identify genes essential in basic metabolism under more nutrient limiting conditions. One can also assay for essentiality in \textit{in vivo} conditions, or identify genes essential in particular cellular processes, such as cell division or motility.

It is therefore important to define gene essentiality operationally in the context of its assaying conditions. In the following chapters, in light of the way essential genes are inferred in the methods described below, I hereby adopt a definition of essential genes as those for which knockout is lethal under the chosen condition (Koonin, 2003). This is most often under standard LB media growth conditions for the bacterium, but can vary due to the media requirements for the bacteria in question. Similarly, non-essential genes are defined as those for which a knockout yields viable bacteria that are able to grow up and form a colony under the defined conditions.

1.3.3 Traditional methods

The most basic and direct method for probing for gene essentiality is in the “loss of function” approach: a deliberate mutagenesis attempt to disrupt the function of the gene resulting in the loss of viability of the organism under a chosen condition.

One of the earliest examples was the isolation of auxotrophic strains of \textit{E. coli} that were observed to be unable to grow under minimal media, suggesting that they require various amino acids, nucleic acid components and vitamins for viability (Davis, 1949; Lederberg and Lederberg, 1952). Another approach is the
generation of “conditional lethal” mutants that can grow in one condition but not another, such as temperature-sensitive (ts) mutants (Horowitz and Leupold, 1951). Ts mutants are first generated by chemical mutagenesis resulting in random point mutations, and are subsequently allowed to grow at a low (permissive) temperature, followed by a higher (non-permissive) temperature. Essential genes are identified by growth at the permissive temperature but not at the non-permissive temperature, and the site of mutation can be determined by a mapping strategy, or by complementation with a library of wild-type genes. For instance, temperature-sensitive mutants were used at a classic large-scale study to obtain ts mutants in *E. coli* in enzymes responsible for DNA replication (Sevastopoulos et al., 1977). It was later also applied to *Salmonella typhimurium* (Schmid et al., 1989). A problem with this approach is the uncertainty of the extent to which mutated essential genes are affected by changes in temperature, which makes it difficult to determine the coverage of this technique.

Recombinant DNA technology also enabled the development of site-directed mutagenesis techniques characterized by its ability to target the mutation to a specific gene leading to its knockout. The most commonly used method is oligonucleotide-directed mutagenesis (Sevastopoulos et al., 1977). A plasmid vector is first used to clone a gene of interest. The plasmid is denatured to a single strand, and an oligonucleotide containing the desired mutation is annealed to one strand of the gene of interest, and serves as a primer for DNA synthesis. The newly synthesized strand therefore contains the mutagenic oligonucleotide. The plasmid is then introduced into the bacterial host cell in the
hope that it will be integrated into the chromosome via transformation (Méjean et al., 1981; Lee et al., 1998).

In order to identify mutants which have taken up the plasmid DNA and therefore the mutated gene of interest, various techniques exist in selecting for mutants. For instance, a selection marker can be ligated to the mutagenic oligonucleotide to repair an antibiotic resistance gene (Lewis and Thompson, 1990; Bohnsack, 1996). Introducing the antibiotic will select for the mutant strain. Another commonly used technique is PCR-based. Among the many different variations, the common theme is in the use of one or more mutagenic oligonucleotide(s) as primer(s) for PCR (Higuchi et al., 1988). After multiple amplification cycles, the mutated gene would exponentially outnumber the wild-type gene, creating a homogeneous solution of the former.

1.3.4 High throughput transposon mutagenesis

Transposons, first discovered by Barbara McClintock (McClintock, 1950) in maize, are mobile DNA fragments that can jump to new locations within the same genome or between different genomes, in a process called transposition. There are two major classes of transposons according to their mechanism of transposition. On the one hand, Class I transposons, also known as retrotransposons, work by first copying themselves using an RNA intermediate, which is then copied back to DNA with the help of reverse transcriptase and inserted back to the genome in a different location. Examples of Class I transposons include the Alu elements present in humans and other primates, as well as intra-cisternal A particles (IAPs) in rodents, Ty elements in yeast and
gypsy and copia-like elements in the fruit fly. Class II elements, on the other hand, are DNA transposons which in contrast do not use an RNA intermediate, but transpose directly from DNA to DNA. With the help of the transposase enzyme, the transposon excises itself from a donor site, and ligates itself to another location. They are characterized by flanking inverted repeats which are recognition sites for transposase. Examples of Class II transposons include P-elements in Drosophila, the Ac/Ds system in maize described by McClintock, and Tn elements in bacteria.

Transposons are also mutagens, causing disruptions in the region of the genome where it is inserted. They have therefore been widely used for more than 20 years as a mutagenesis method, both in bacteria and eukaryotes, for the identification of essential genes. The basic concept is the following: if the transposon is inserted within a gene, the inability of the bacterium to survive due to the disruption would suggest that the gene is essential for the bacterium's viability. Typically, essential genes are identified by a “negative” screen: mutants which are recovered are defined as being in non-essential regions; conversely, in any statistically significant region that does not contain a mutant (under saturation mutagenesis conditions), the region is assumed to be essential.
Figure 1-3: Identification of essential and non-essential genes using transposon mutagenesis (A) in a specified region of interest; (B) global transposon mutagenesis (descriptions in text). Red arrows denote essential genes, black arrows non-essential genes.

Figure 1-3 illustrates a typical protocol for transposon mutagenesis. In Figure 1-3 A, the essentiality of one specific region of the chromosome is probed. The region is first amplified by PCR, then subjected to saturated \textit{in vitro} transposon mutagenesis using a transposon containing a selectable marker, for instance an antibiotic resistance cassette, and re-introduced into the bacterial cell via transformation. After subjecting to antibiotic selection, viable mutants are recovered, then subjected to PCR analysis to identify the location of the insertion (by using one primer in the transposon and the other in the chromosome). Insertion locations that show up in the agarose gel are defined as non-essential.
This process can also be extended to the whole genome, as shown in Figure 1-3 B. Instead of amplifying a specific region, in high-throughput transposon mutagenesis the whole chromosome is subjected to a large number of transposon insertions. Viable mutants are recovered after selection and insertion locations are identified by PCR. Recovered insertion locations are defined as non-essential, while the rest are putatively essential. However, for any statistical significance of gene essentiality to be drawn, achieving near-saturation levels is required, which can be very costly when applied in a genomic scale. Even when near-saturation levels are achieved, essential regions can still only be estimated and assigned a probability.

Variations exist concerning the type of transposon used and mutant selection methods, however the theme is the same. Transposon mutagenesis has been applied in identifying essential genes in many different bacterial organisms. For instance, Hutchison et al. (Hutchison et al., 1999) have performed a global transposon mutagenesis in two Mycoplasma genomes, *M. genitalium* and *M. pneumoniae*, which identified about 265-350 genes that are essential in *M. genitalium*. The authors claim this as an approximation to the "minimal gene set", given that *M. genitalium* is currently the bacterium with the smallest known genome. Transposon mutant libraries have also been constructed in *Mycobacterium tuberculosis* H37Rv (McAdam et al., 2002; Lamichhane et al., 2003), *H. influenzae* strains BC200 (Reich et al., 1999) and Rd (Akerley et al., 2002), *E. coli* MG1655 (Gerdes et al., 2003; Kang et al., 2004), *P. aeruginosa* strains PAO1 (Jacobs et al., 2003) and PA14 (Liberati et
al., 2006), *Helicobacter pylori* (Salama et al., 2004), and *Francisella tularensis novicida* (Gallagher et al., 2007).

Microarrays have also been used in variations of transposon mutagenesis. For instance, the transposon site hybridization (TraSH) technique was developed, which used gene probes in microarrays to hybridize with the mutant pool as a negative screen of essential genes – this was applied to *M. tuberculosis* H37Rv (Sassetti et al., 2003). Recently, Weiss et al. (Weiss et al., 2007) have also employed a similar microarray-based negative selection screen to identify *F. tularensis* genes essential for growth and survival in mice, while Smith et al. have devised a microarray-based technique called “Monitoring of Gene Knockouts” (MGK) to identify genes essential in aromatic metabolism as well as antibiotic resistance (Smith et al., 2007).

There are many advantages to transposon mutagenesis as a method in identifying essential genes. Mutants with disrupted genes can be generated very quickly which lends itself to high-throughput application to whole genomes. Due to the near-randomness of a transposon insertion event, there is no need to design probes to target a specific location. It also does not require that the organism be competent. For probing essentiality in small regions, saturation can also be achieved more easily than targeted mutagenesis methods, though this is more difficult for genome-wide studies.

There are some drawbacks, however. One of the biggest concerns is that until saturation levels are reached, it is very difficult to say to any degree of certainty that a region absent of transposon insertions is essential. Even if near-
saturation levels are reached, one can only assign a probability, and there is still
the possibility that the region is non-essential, until a biological experiment
demonstrating essentiality is performed. Furthermore, insertions in particular
gene locations, such as at 3’-ends, do not necessarily disrupt their function.
Finally, polar effects due to operon structure will likely introduce some bias in
where transposons are inserted, leading to certain “hot spots” or sequence
preferences. When there is a gap (regarding where transposon mutants are
mapped to that genome) that involves more than one gene that may be in an
operon, it is hard to determine whether the lack of mutants in the upstream gene
is due to polar effects on the downstream gene(s) or whether that gene is indeed
truly essential. Incorporating a strong, outward-facing promoter and translation
signals can help reduce polar effects due to transposon insertions on the
expression of downstream genes (Badarinarayana et al., 2001).

1.3.5 Other genome-wide studies of essential genes

Other than transposon mutagenesis, researchers have also used other
genome-wide methods to identify essential genes. For instance, Baba et al. have
systematically made precise, single-gene deletions in 4288 genes in E. coli K-12,
identifying 303 candidates essential genes (Baba et al., 2006).

Antisense RNA, which can hybridize and inactivate a target gene, have
also been employed in the search for essential genes. Two studies have applied
this technique to S. aureus, identifying 658 essential gene candidates, among
which only 168 are conserved in M. genitalium (Ji et al., 2001; Forsyth et al.,
2002). There is however an inherent limitation with the antisense RNA method, in
that the use of antisense RNA is limited to those genes for which an adequate expression of the inhibitory RNA can be obtained.

1.3.6 Computational methods

The availability of many completely sequenced bacterial genomes have led to various computational methods such as using comparative genomics to identify essential genes (Bruccoleri et al., 1998). Both Fraser et al. and Mushegian and Koonin have compared the genomes of *H. influenzae* (1703 genes) and *M. genitalium* (468 genes), hypothesizing that a "minimal gene set for cellular life" might be approximated by finding the conserved genes between these two organisms belonging to two ancient bacterial lineages: Gram-negative and Gram-positive, respectively (Fraser et al., 1995; Mushegian and Koonin, 1996). Arigoni et al. have subsequently compared the genomes of *M. genitalium* and *E. coli* in search for essential genes (Arigoni et al., 1998). Gil et al. have extended the comparative genomics approach to eight bacterial genomes (Gil et al., 2004). The sequencing of multiple strains of certain organisms have enabled the analysis of multiple essentiality datasets, such as in *P. aeruginosa* (Sakharkar et al., 2004), *Burkholderia pseudomallei* (Chong et al., 2006) and *H. pylori* (Dutta et al., 2006).

Using a machine learning approach, Gustafson et al. have attempted to extract genomic features such as "phyletic retention" (number of orthologs present in other organisms), protein features such as codon bias, and experimental features such as involvement in protein interactions, in order to explore their predictive power of gene essentiality. By combining different
features, essential genes in *E. coli* and *Saccharomyces cerevisiae* were predicted (Gustafson et al., 2006).

Databases of bacterial essential genes have also been developed. For instance, the Profiling of *E. coli* Chromosomes (PEC) database compiles evidence from literature and other sources concerning the essentiality of genes in *E. coli* K-12 (Hashimoto et al., 2005). Another example is the Database of Essential Genes containing information on essential genes in various organisms (Zhang et al., 2004).

1.4 Identifying non-coding RNAs in bacteria

In terms of composition, RNA is a nucleic acid like DNA, however RNA differs from DNA in two main aspects: RNA nucleotides contain ribose instead of deoxyribose, and RNA contains the nucleotide uracil (denoted by the letter ‘U’) in place of thymine. Canonical base-pairs include cytosine (C) with guanine (G), and adenine (A) with uracil (U). In addition, guanine can form an hydrogen bond with uracil, forming a non-canonical G-U base pair. Many non-canonical base pairs exist for RNA, however are more rare (Nagaswamy et al., 2002).

Non-coding RNAs (ncRNAs) are classes of genes which, in contrast to protein-coding genes, produce transcripts that function directly as structural, catalytic or regulatory RNAs, rather than expressing mRNAs encoding proteins. In1952, the Central Dogma put forward by James Watson held the then prevailing “one gene, one ribosome, one protein” hypothesis: that mRNAs were only acting as intermediates for protein translation (Crick, 1958). Later,
ribosomes were found to be composed not only ribosomal proteins, but also contained ribosomal RNAs (Brenner et al., 1961). Crick's "adaptor" hypothesis, that a molecule existed that can map the triplet genetic code to the corresponding amino acid, led to the discovery of transfer RNAs by Mahlon Hoagland and co-workers (Hoagland et al., 1958). Later, the Central Dogma broke down even further when more types of RNAs were being discovered, such as the Uridine-rich RNAs (U RNAs) associated with ribonucleoprotein complexes (Zieve, 1981), or the discovery of the Ribonuclease P enzyme containing an essential RNA component called SRP-RNA (Stark et al., 1978). In the nucleolus of eukaryote cells, a plethora of small nucleolar RNAs (snoRNAs) such as the "C/D box" snoRNAs and the "H/ACA" snoRNAs were being isolated (Balakin et al., 1996) and shown to have essential catalytic functions (Eliceiri, 1999). Short RNAs were being discovered to inhibit translation of target mRNAs in eukaryotes, known as microRNAs (miRNAs), and the discovery of small-interfering RNAs (siRNAs) (Fire et al., 1998) involved in RNA interference pathway even paved the way to a Nobel Prize in 2006.

It is not surprising that ncRNAs also have an important roles in bacteria. It's been known for many years that cis-encoded antisense RNAs have roles in regulating plasmid replication and maintenance. One well-known example is the copA antisense RNA in the plasmid R1, which is encoded in the opposite strand of its target mRNA copT, and binds to it via a long stretch of full complementarity, inhibiting its translation (Blomberg et al., 1990). Antisense RNAs have also been found to be trans-encoded, targeting mRNAs in a separate locus by partial
complementarity. The first such RNA to be characterized was the MicF RNA in *E. coli* (Mizuno et al., 1984). Upon induction by environmental stress conditions, the MicF RNA is expressed and blocks translation of its target mRNA ompF via base complementarity. Another well-known example is the OxyS RNA, also in *E. coli*, which is strongly induced by oxidative stress, and basepairs with its target mRNA fhlA, a transcriptional activator, repressing its translation (Altuvia et al., 1997). Later, the DsrA RNA was also found to target the rpoS mRNA at low temperatures, which encodes the stationary phase sigma factor $\sigma^S$, leading to increased translation by preventing the formation of a structure occluding the ribosomal binding site (Sledjeski et al., 1996). These regulatory RNAs can be as small as 50 nucleotides (typically 50 to 300 nucleotides), thus they have also been termed “small RNAs” (sRNAs).

Other non-coding RNAs exist in bacteria, however. One RNA found in all bacteria is the transfer-messenger RNA (tmRNA), which acts both as a tRNA and mRNA to mediate the release of stalled ribosomes (Williams and Bartel, 1996). Recently, dozens of classes of phylogenetically prevalent “riboswitches” have been found in 5’-untranslated regions of certain types of mRNAs encoding important metabolic functions. Riboswitches are encoded within the same transcript of the mRNA, and regulates it by forming different structures that inhibit or induce the transcription or translation process, depending on the presence or absence of binding of small metabolites (Winkler and Breaker, 2005).

NcRNAs have been found to be much more prevalent and important in bacteria than previously thought, and their knowledge is little, anecdotal, and
limited to biochemically abundant species; until 2001 only ten such genes were known (Wassarman et al., 1999). The prevalence, diversity and the functional relevance of ncRNAs in bacteria remain largely unknown, and their detection has been difficult. The difficulty in detecting ncRNAs has been due to several factors, including a lack clear of signals (start and stop signals) and their short size, which makes them hard to distinguish from random sequence.

However, with the increasing availability of complete bacterial genome sequence information, methods have been developed to perform systematic searches of ncRNAs, which have recently led to the discovery of hundreds of sRNAs. These are described below.

1.4.1 Computational approaches in identifying ncRNAs

The first systematic genome-wide screens for ncRNAs have been performed in *E. coli*, characterized by three pioneering studies in 2001. In the first study, Argaman and colleagues computationally analyzed intergenic regions in *E. coli* to identify phylogenetically conserved loci that have a predicted promoter and rho-independent terminator 50-400 nucleotides apart, assuming that sRNA transcripts can be found in between (Argaman et al., 2001). Twenty four candidate sRNAs were predicted, of which 14 were indeed shown to produce small transcripts by northern blot analysis.

In a second separate study, Wassarman et al. analyzed intergenic regions that showed conservation to other genomes, ranking them using criteria such as the distance from a nearby ORF, the presence of promoter or rho-independent
termination signals, and RNA expression in high-density microarrays
(Wassarman et al., 2001). 59 candidate sRNAs were predicted, of which 17 were
shown to be transcribed in stationary phase. As with Argaman’s study, however,
the major drawback is the assumption that sRNAs are transcribed between the
promoter and rho-independent termination signals. Accurate promoter prediction
is still an open problem, and sRNAs having termination signals other than rho-
independent terminators will be missed.

In the third study, Rivas and colleagues have developed a statistically-
based computational algorithm (QRNA) for finding ncRNAs, which uses
comparative genomics to detect the conservation of RNA secondary structure
signals within a pair of sequences, as opposed to the conservation of amino-acid
coding sequence or a null hypothesis (Rivas and Eddy, 2001). Screening the E.
coli genome, 275 candidate loci were predicted and 49 were analyzed by
northern blot analysis in one condition (exponential growth in rich media). Of
these, 11 were found to produce small RNAs (Rivas et al., 2001). These three
pioneering studies have identified a total of 34 sRNAs in E. coli.

Other algorithms to predict novel sRNAs have later also been developed.
For instance, Washietl et al. have developed RNAz, which combines comparative
sequence analysis and RNA secondary structure prediction to identify potential
non-coding RNAs (Washietl et al., 2005). Livny et al. have developed a program,
sRNAPredict, which integrates the output of different transcription signal
prediction programs to predict putative sRNAs in intergenic regions (Livny et al.,
2005). This has been applied to genome-wide screens in 11 different organisms,
predicting 2759 candidate sRNA loci. Among 34 *P. aeruginosa* predictions, 31 were analyzed by Northern blot and 17 were confirmed to encode sRNA transcripts (Livny et al., 2006). More recently, Yao et al. developed a computational pipeline analyzing sequences upstream of conserved orthologs searching for cis-regulatory RNAs (Yao et al., 2007), and have used it to identify 22 novel candidate RNA motifs, including six riboswitches (Weinberg et al., 2007).

A common caveat to all these computational approaches is the difficulty in assessing their predictive power given the lack of both positive and negative controls. Computational predictions need to be verified experimentally regarding evidence of their transcription, however this also comes with certain caveats. Specifically, a negative outcome does not necessarily prove the absence of an sRNA, since it might expressed only in specific growth conditions. For instance, oxyS is only known to be expressed in oxidative stress (Altuvia et al., 1998), and csrB only expressed in stationary-phase cells (Romeo, 1998). Therefore, verifying the presence of transcription is not straightforward.

### 1.4.2 Experimental approaches in identifying ncRNAs

Apart from computational approaches, many experimental approaches have also been developed in identifying sRNAs. The first sRNAs that were determined experimentally were identified by gel fractionation of total cellular RNA that was first labelled, either metabolically using $^{32}$P-orthophosphate, or by chemical staining, and the sequence identified by nuclease digestion. For instance, the sRNAs BS190 and BS203 in *B. subtilis* were identified this way.
(Silvaggi et al., 2006), as well as three *S. aureus* sRNAs (Pichon and Felden, 2005). An obvious drawback to this technique is its low-throughput nature, and the fact that it can only detect sRNAs which are visible and can be separated as a single species on the gels.

Detection of sRNAs using DNA microarrays containing probes specific to intergenic regions have been proposed (Selinger et al., 2000; Tjaden et al., 2002). Alternatively, RNAs can first be isolated by co-immunoprecipitation with Hfq, a protein known to bind numerous sRNAs (Zhang et al., 2003), or using an antibody-based approach (Hu et al., 2006). DNA microarrays enable a high-throughput method in identifying putative sRNAs, however it suffers from challenges in preparation and labelling of RNA samples. The small size and structure of sRNAs would also make these transcripts poor substrates for amplification and labelling.

Thirdly, shotgun cloning has been employed in screening sRNAs in *E. coli*. In one study, total RNA from three different growth phrases that meet a specific size (50-500nt) were first reverse-transcribed to cDNA. Abundant RNAs such as rRNAs and tRNAs were filtered out, and remaining low-signal clones were chosen for sequencing, leading to the discovery of 20 previously identified sRNAs and 7 novel sRNAs (Vogel et al., 2003).

A genomic *in vitro* selection (SELEX) approach has also been developed (Lorenz et al., 2006). A library of random sequences within the *E. coli* genome were first transcribed *in vitro*, then incubated with Hfq. The sequences were filtered for Hfq binding, converted to cDNA, and finally its interaction with Hfq was
determined *in vivo* using the yeast three-hybrid system. Since sRNAs were generated *in vitro*, growth conditions need not to be taken into account with this approach, and is capable in detecting sRNAs expressed in low levels. However, for the same reason the functional relevance of sRNAs within the bacteria needs to be further examined. The protocol also takes several months to complete.

1.5 Insights gained regarding essential genes and ncRNAs

Given a large number of studies attempting to identify essential genes described above, many insights have been gained that have led to a better understanding of their properties and nature. It is perhaps important to emphasize at this point that all studies that have attempted to systematically characterize essential genes have focused on protein-coding genes. I am not aware of any systematic study on the essentiality of ncRNAs.

The Profiling of the *E. coli* Genome (PEC) database has been used by various studies, classifying *E. coli* genes as being either “essential”, “non-essential”, or “undetermined”. The method by which curators use to assign the classification provides some insight about the functions of essential genes and non-essential genes. For instance, genes falling into known categories of ribosomal proteins, tRNA synthesis and aminoacyl tRNA synthases are considered to be essential. Genes with null-type mutants available (deletion or transposon insertion mutations) are considered to be non-essential, as well as genes involved in flagellation, motility, and chemotaxis.
Of theoretical importance, one would like to ask: what is the minimum number of genes essential for cellular life? The number of essential genes identified in studies described in previous chapters sometimes vary considerably, however some insights on the range can perhaps be gained. In PEC, for instance, out of 4,761 protein coding genes in *E. coli* K-12 characterized, 302 (6.3%) were classified as essential, while 4,455 (93.7%) were non-essential (last updated February 15th, 2008, ver. 4.2.0). In *B. subtilis*, Kobayashi et al. demonstrated that out of a total of 4,101 genes, 271 (6.6%) were indispensable for growth when inactivated singly, while 3,830 (94.4%) were not (Kobayashi et al., 2003). In a transposon mutagenesis study in *H. influenzae*, 484 (28.5%) genes out of a total 1,700 were found to be critical for growth or viability and did not have any transposon insertions (Akerley et al., 2002). Similarly, transposon mutagenesis in *P. aeruginosa* PA01 has identified around 678 (12%) genes (out of 5,568) without insertions (Jacobs et al., 2003). In *H. pylori*, transposon mutagenesis yielded 344 putatively essential genes (Salama et al., 2004). Furthermore, by performing global transposon mutagenesis in two of the smallest known genomes, *M. genitalium* (480 genes) and *M. pneumoniiae* (677), about 265-350 genes were essential in *M. genitalium* (Hutchison et al., 1999). Also working with *M. genitalium*, a Gram-positive bacterium, Mushegian and Koonin compared it to *H. influenzae*, a Gram-negative bacterium, and identified 256 putatively essential genes common to both lineages (Mushegian and Koonin, 1996). From these studies, it seems reasonable to estimate that the minimum number of genes essential for cellular life is approximately in the range of 200-
500. These studies also seem to suggest that the number of essential genes in any given bacterium is more or less constant, regardless of the size of the genome, which seems reasonable because the number of absolutely essential housekeeping functions such as protein synthesis will unlikely to differ much in small or large genomes alike. However, number discrepancies in different studies could not only be due to differences of methodology, but also could reflect true differences in the organisms’ survival needs due to different living environments.

It is also of interest to ask how essential genes are phylogenetically distributed among the Tree of Life. In *B. subtilis*, for instance, over 80% of essential genes had homologs in all bacteria with genomes above 3 Mb, and 57% were found in bacteria with the smallest genomes (e.g. *Mycoplasma*) (Kobayashi et al., 2003). Almost 70% of genes were present in at least one kingdom other than Bacteria (Kobayashi et al., 2003). Another study sought to determine densities of essential genes in *E. coli*, and their phylogenetic distribution using evolutionary retention indexes (ERIs) (Gerdes et al., 2003). ERIs were calculated as the fraction of orthologs present in 33 phylogenetically diverse organisms. It was found that the density of essential genes was significantly positively correlated with the ERI, suggesting that cluster regions with essential genes are more likely to be evolutionarily retained. Finally, Jordan et al. counted the number of taxonomic groups for each *E. coli* gene that had an ortholog in the Cluster of Orthologous Groups (COG) database. This analysis showed that, from a total of 26 distinct taxonomic groups in the COG database at
that time, essential genes had a higher average phylogenetic distribution parameter value (20-21) than non-essential genes (13-17) (Jordan et al., 2002).

A related question is whether essential genes are more conserved in terms of sequence, than non-essential genes. Jordan et al. used a data set that contained only essential and non-essential genes in *E. coli* for which experimental evidence existed, and compared the synonymous (Ks), non-synonymous (Ka) rates and Ka/Ks ratio (Jordan et al., 2002). The authors show that the average Ks and Ka were significantly lower for essential genes than for non-essential genes, suggesting that essential genes undergo a significant purifying selection in bacteria. Orthologs of *E. coli* genes used in the study were identified in *H. pylori* and *N. meningitidis*, and for both species, the rates of Ks and Ka of essential genes were significantly lower. However, both species had significantly greater values of Ks, Ka and Ka/Ks than those in *E. coli* (with essential genes in *H. pylori* appearing to evolve even faster than *E. coli* non-essential genes), suggesting that even though essential genes evolve significantly slower relative to non-essential genes in the same genome, the rates of evolution differ between species, which might reflect the distinct bacterial lifestyles.

1.6 An introduction to *Pseudomonas aeruginosa*: an intrinsically antibiotic-resistant pathogen

*P. aeruginosa* is a Gram-negative, rod-shaped aerobic bacterium in the Gamma proteobacteria Class of the Bacterial Domain of life. It's a ubiquitous free-living bacterium that can be found in such diverse environments such as soil
and water, as well as plant and animal tissues. It is also very versatile, being capable of growing on very simple nutritional requirements (Stover et al., 2000).

*P. aeruginosa* is perhaps most well known, not only for its ability to infect a wide range of hosts including insects and plants, but also as a notorious opportunistic human pathogen, infecting in particular individuals with compromised immune systems, such as cystic fibrosis and cancer patients, burn victims and patients in hospital intensive care units. In humans, this pathogen is capable of causing infections in any mucosal surface, including urinary tract, burns and wounds, as well as blood. Infections that occur in the lower respiratory tract are particularly difficult, if not impossible, to treat, leading to decades of chronic lung infections in cystic fibrosis patients, usually resulting in death (Lyczak et al., 2002).

Its clinical importance is also due to its notorious intrinsic antibiotic resistance. The possession of no less than four multidrug efflux systems (Nikaido, 1998), antibiotic resistance genes, as well as the permeability barrier of its outer membrane makes this pathogen resistant to virtually all existing antibiotics (Hancock, 1998). Its ability to form biofilm also adds to increased resistance, even if antibiotics are in therapeutic levels.

Due to its clinical importance, *P. aeruginosa* strain PAO1 has also been one of the first bacterial organisms to be sequenced (Stover et al., 2000). It was the largest bacterial genome sequenced to date at the time, and remains one of the largest, at 6.3Mbp in size. It is also one of the most G+C rich bacterial genomes (66.6% G+C content). The large number of predicted ORFs (5,570) is
comparable to that in the higher eukaryote yeast. Two other strains were subsequently completely sequenced: *P. aeruginosa* strain UCBPP-PA14 (Lee et al., 2006) (more commonly referred to as strain PA14), and *P. aeruginosa* strain PA7 (NCBI genomes database; unpublished). Several additional clinical isolates are also in the progress of being completely sequenced.

The Pseudomonas Genome Database (Winsor et al., 2005), maintained and updated by the Brinkman Lab, includes genes annotations and analysis tools for more than a dozen *Pseudomonas* genomes, providing valuable data and tools for researchers to conduct comparative analysis.

### 1.6.1 Transposon mutagenesis studies in *Pseudomonas aeruginosa*

Due to the clinical importance of *P. aeruginosa*, two comprehensive global transposon mutagenesis studies have already been performed, first for *P. aeruginosa* PAO1 (Jacobs et al., 2003), then for *P. aeruginosa* UCBPP-PA14 (Liberati et al., 2006). Due to their near-saturation nature of the transposon mutagenesis protocol, both studies provide very valuable information in terms of the essentiality of any given region within the two genomes. Analysis in *P. aeruginosa* PAO1 put an initial estimate of the number of essential genes to be between 300 and 400. Combining information from both libraries, Liberati *et al.* narrowed the estimate of putative essential genes in *P. aeruginosa* to 335.

Both studies focused on probing the essentiality of annotated genes, however they also provide valuable information regarding the identification of intergenic regions that are likely to be essential.
The clinical importance of *P. aeruginosa*, its intrinsic antimicrobial resistance, the presence of a transposon mutant library for the species, the fact that the genome was annotated so early, as well as our access to the Pseudomonas Genome Database and the Brinkman lab's familiarity with *P. aeruginosa*, made this species' genome an attractive model for testing an approach to identify new essential genes within previously annotated intergenic regions which may represent novel antimicrobial drug targets.

### 1.7 Goal of the present research

Motivated by the need to accelerate the search for novel antibacterial targets, the goal of the present research involves developing a computational approach, which should be applicable generally to any bacterial organism, to identify novel essential genes within intergenic sequences in bacterial genomes. An additional goal was to gain insight into the nature of bacterial essential genes from an evolutionary perspective.

Herein, I describe the various steps involved in the extraction of essential intergenic sequences (Chapter 2), and the subsequent prediction of novel essential protein coding genes (Chapter 3), as well as novel essential ncRNAs (Chapter 4) in *P. aeruginosa*. Moreover, I describe additional analyses from an evolutionary perspective of essential genes in bacteria (Chapter 5). Finally, I describe the integration of these steps into a computational package, and its application to *M. tuberculosis* H37Rv (Chapter 6).
Ultimately, this work aims to aid discovery of additional antimicrobial targets that warrant further study.
CHAPTER 2  IDENTIFYING INTERGENIC REGIONS THAT MAY CONTAIN NOVEL ESSENTIAL GENES

2.1 Summary

The first component of my research involved the extraction of intergenic regions within a bacterial genome of interest that appear to be essential. This has been implemented in silico, and tested on the P. aeruginosa PA01 genome, using statistical models. The intergenic regions identified as being putatively essential were predicted, erring on the side of high sensitivity to avoid missing notable new genes. The resulting list of intergenic regions provide a starting point for finding novel essential protein coding genes and ncRNAs, described in later chapters.

2.2 Rationale

Intergenic regions are stretches of DNA sequences within a genome between two known or predicted genes, based on the current genome annotation. Many of these may contain regulatory modules such as transcriptional promoters, terminators or other regulator binding sites.

They may also contain novel genes that have not yet been identified. Reasons for this are many and could be attributed to fallacies introduced in the process of genome annotation. In a typical bacterial annotation pipeline, ORFs are first predicted using gene prediction programs such as GeneMark (Lukashin and Borodovsky, 1998) or Glimmer (Salzberg et al., 1998; Delcher et al., 1999),
which can correctly detect 85-98.1% of ORFs in benchmark tests (i.e. predicted ORF matches the 3’-end of a reference ORF) (Besemer et al., 2001). However, the correct prediction of gene start sites is not as precise (only 62.4-83.2% of predicted ORFs had the precise start sites predicted compared to a reference genome), resulting in at least 16.8% of the genome with incorrectly predicted start sites. Furthermore, short genes (less than 300nt) are notoriously difficult to detect for gene finders, as they cannot be easily distinguished from random ORFs (Besemer et al., 2001).

These problems can be overcome to some certain extent by the additional use of a comparative genomics approach. Typically, as a subsequent step to ORF finding, a BLASTx search is performed to query each predicted ORF against a protein database. ORFs with statistically significant matches to characterized genes in other bacterial organisms are more likely to be true ORFs. Correct gene boundaries can also be determined by comparing to boundaries of existing proteins.

However, protein databases are far from comprehensive, and still growing at an exponential rate. From July 2005 to July 2006, NCBI has reported an increase of 62% in its non-redundant protein collection, with the majority consisting of microbial proteins (1,990,849 out of the total 2,762,164 records) (Pruitt et al., 2007). This signifies that, for older bacterial genomes that were first sequenced and annotated, a huge number of proteins were not available at that time in the database for querying, leading to many possible missed gene
predictions, wrongly labelling a region as 'intergenic' where it should in fact contain a gene.

As a result of the constant growth of protein databases, I hypothesized that novel genes could now be found in annotated genomes that hadn't been detected before. This in particular applies to the test organism, *P. aeruginosa* PAO1, since its genome sequence was reported in 2000, with the actual annotation of the sequence occurring in the late 1990s when only a few bacterial genomes had been sequenced (Stover et al., 2000). Functional assignments were also performed based by relatively strict criteria (Winsor et al., 2005), leading to many possible missed genes.

In addition to novel protein-coding genes, intergenic regions may also harbour previously undiscovered ncRNAs. As mentioned in the introductory chapter, only recently have scientists realized the involvement of many sRNAs and *cis*-regulatory RNAs in several important biological functions in bacteria. In a typical bacterial annotation pipeline, no component is devoted to the systematic search of ncRNAs, except for scanning for tRNAs, rRNAs and tmRNAs. This is true particularly for older bacterial genomes. Therefore, I also hypothesize that regions labelled as 'intergenic' in older annotated bacterial genome sequences may harbour previously undetected ncRNAs.

### 2.3 Materials and Methods

The process of extraction of essential intergenic sequences was implemented as a software module written in the Perl programming language.
2.3.1 Extraction of intergenic sequences

In order to extract intergenic sequences, two pieces of information were needed: 1) a file containing the complete genomic DNA sequence of the genome of interest in FASTA format, and 2) a file of gene annotations in GFF format, which contains the coordinates of gene boundaries used to delineate gene regions and intergenic regions. For *P. aeruginosa PAO1*, its genomic DNA sequence and gene annotations were downloaded from the NCBI Microbial Genome Resources website (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi; accessed 2008-01-10).

The coordinates of gene boundaries defined in the genome annotation file were used to delineate annotated gene regions from intergenic regions. Namely, an intergenic sequence starts at one bp to the right of the right boundary of a gene, and ends one bp to the left of the left boundary of the subsequent gene, assuming we are counting genes from left to right.

2.3.2 Defining essential regions using transposon mutant library

In order to determine whether a given region in the genome was essential, a transposon mutant library was used. For *P. aeruginosa PAO1*, the transposon mutant data constructed by Jacobs et al. (Jacobs et al., 2003), which contains chromosome coordinates of transposon insertion sites of recovered mutants, was downloaded from the University of Washington Genome Sequencing Center website (http://www.genome.washington.edu/UWGC/pseudomonas/index.cfm; accessed 2008-01-10).
As mentioned in the introductory chapter, essential regions in transposon mutagenesis studies are characterized by the notable absence of insertions, since they become lethal following the disruption of the region and cannot be recovered. The coordinates of the transposon mutant insertions can thus be used to calculate a statistically significant cutoff for the minimum size of a “transposon gap” (the distance between two adjacent transposon insertions) above what can be expected by chance alone. In other words, gaps that are larger than the cutoff are more likely to be truly essential, and can be tentatively assigned as putative essential genes.

2.3.3 Statistical modelling

Random transposon insertions in a genome can be considered as random events occurring within a finite spatial region, and thus can be modelled as a Poisson process. The probability that exactly \( k \) insertion events occur within a region of length \( L \), given an average rate of insertion \( r \), is therefore given by:

\[
Pr(k, r, L) = \frac{(rL)^k e^{-rL}}{k!}
\]  

(E1)

(\( rL \) also denotes the expected number of insertions within a region of length \( L \), conventionally known as \( \lambda \)).

Since we are interested in identifying essential regions free of insertions, we are interested in the case where \( k = 0 \), i.e., the probability that no insertions occur within a region of length \( L \), given an average rate of insertion \( r \). Equation E1 thus becomes:
\[ \Pr (0, r, L) = e^{-rL} \] (E2)

The above equation follows an exponential distribution with rate parameter \( r \) (this is expected, since the distances between two successive events guided by a Poisson process theoretically follows an exponential distribution). Thus, the sizes of “transposon gaps”, i.e. the distance between two adjacent transposon insertions, will theoretically follow an exponential distribution. In order to determine whether this is the case empirically, a computational simulation of random insertions was performed, as described later in the Results section.

Equation E2 can also be used as a p-value cutoff below which we reject the null hypothesis, the null hypothesis being that the region is non-essential:

\[ \Pr(0, r, L) = p < e^{-rL} \] (E3)

Rearranging for \( L \), we get:

\[ L > \frac{-\ln (p)}{r} \] (E4)

In other words, a transposon gap needs to be larger than \( L \) to be observed with a probability below \( p \). Thus the essentiality of a region is determined with respect to a given p-value cutoff, which the user determines beforehand: regions containing transposon gaps larger than \( L \) are labelled as essential.

Based on equation E4, only two parameters need to be calculated in order to determine the transposon gap size cutoff above which we define a region to be essential: \( p \) and \( r \). \( p \) is defined by the user. The average rate of insertion \( r \) (insertions per kb) can be estimated directly from the transposon mutant data by...
simply calculating $N/L_g \times 1000$, where $N$ is the number of transposon insertions, and $L_g$ is size (in bp) of the genome.

2.3.4 Extracting essential intergenic sequences

Based on a value of $p$ given by the user, the transposon gap size cutoff $L$ is calculated using equation E4. Subsequently, all nucleotide sequences within intergenic regions that correspond to the locations of transposon gaps larger than $L$, are extracted, as demonstrated in Figure 2-1. These will be called “putatively essential intergenic sequences (PEIS)”.

Note that a PEIS does not necessarily span the entire intergenic region (from one gene boundary to another), as shown in Figure 2-1, but spans only the region corresponding to a transposon gap larger than $L$, since we are interested in only identifying essential genes. As such, one entire intergenic region may contain multiple PEISs, which will be analyzed separately. If the entire intergenic region is free of transposon insertions and is larger than $L$, then the entire region is considered a PEIS and extracted for analysis.

All PEISs extracted were then saved into a single FASTA file and subject to subsequent analysis.
Figure 2-1: Given a transposon gap size cutoff L, this would be an example of a putatively essential intergenic sequence (PEIS) which would be extracted for further analysis. Green horizontal arrows denote annotated genes, and vertical black lines represent transposon insertion locations.

2.4 Results and Discussion

2.4.1 Intergenic sequence statistics in *P. aeruginosa PAO1*

There are a total of 4,662 intergenic regions in the *P. aeruginosa PAO1* genome. They comprise 10.27% (643,541bp) of the total genome size of 6,264,404bp. The sizes of intergenic regions were plotted in a histogram, as shown in Figure 2-2. Unsurprisingly, the vast majority are only a few nucleotides long, as it’s typical in bacterial genomes to have compact operon structures (Mira et al., 2001). The mean intergenic size is 138.13bp, with a standard deviation of 160.29, showing that intergenic regions range markedly in size, as is evident by the presence of numerous very small regions coupled with some very large intergenic regions forming the long tail of the distribution, which can skew the mean value. The median intergenic size is 92bp.
Figure 2-2: (A) The distribution of sizes of intergenic regions (IGRs) in *P. aeruginosa* PA01, and (B) a zoomed version of the same. The dotted line denotes the 75% quantile (181bp), the dashed line the 90% quantile (323bp), and the solid line the 95% quantile (426bp).
Notably, fourteen intergenic regions, listed in Table 2-1 are larger than 1kb. The largest intergenic region, which is 3024bp long, lies between genes PA0041 and PA0042, which encode a 3536-amino-acid-long secreted hemagglutinin, and a small 132 amino-acid long hypothetical protein, respectively.
Table 2-1: Fourteen largest intergenic regions in *P. aeruginosa* PAO1 (>1kb). The average %G+C in entire genome is 66.6%.

<table>
<thead>
<tr>
<th>Left flanking gene (strand)</th>
<th>Right flanking gene (strand)</th>
<th>Size (bp)</th>
<th>%G+C</th>
<th>Size of largest transposon gap (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA0041 - probable hemagglutinin (+)</td>
<td>PA0042 - hypothetical protein (+)</td>
<td>3024</td>
<td>50.3%</td>
<td>694</td>
</tr>
<tr>
<td>PA3835 - hypothetical protein (-)</td>
<td>PA3836 - hypothetical protein (+)</td>
<td>2644</td>
<td>47.5%</td>
<td>879</td>
</tr>
<tr>
<td>PA5412 - hypothetical protein (-)</td>
<td>PA5413 - low specificity l-threonine aldolase (-)</td>
<td>1339</td>
<td>55.9%</td>
<td>903</td>
</tr>
<tr>
<td>PA3865 - probable amino acid binding protein (-)</td>
<td>PA3866 - pyocin protein (-)</td>
<td>1301</td>
<td>54.6%</td>
<td>549</td>
</tr>
<tr>
<td>PA0263 - secreted protein Hcp (-)</td>
<td>PA0263.1 - tRNA-Arg (-)</td>
<td>1254</td>
<td>53.7%</td>
<td>460</td>
</tr>
<tr>
<td>PA5321 - deoxyuridine 5'-triphosphate nucleotidohydrolase (+)</td>
<td>PA5322 - phosphomannomutase AlgC (+)</td>
<td>1251</td>
<td>70.5%</td>
<td>255</td>
</tr>
<tr>
<td>PA1655 - probable glutathione S-transferase (+)</td>
<td>PA1656 - hypothetical protein (+)</td>
<td>1171</td>
<td>50.9%</td>
<td>490</td>
</tr>
<tr>
<td>PA2729 - hypothetical protein (+)</td>
<td>PA2730 - hypothetical protein (-)</td>
<td>1168</td>
<td>51.0%</td>
<td>557</td>
</tr>
<tr>
<td>PA1152 - hypothetical protein (+)</td>
<td>PA1153 - hypothetical protein (-)</td>
<td>1143</td>
<td>42.3%</td>
<td>606</td>
</tr>
<tr>
<td>PA2046 - hypothetical protein (-)</td>
<td>PA2047 - probable transcriptional regulator (-)</td>
<td>1034</td>
<td>71.5%</td>
<td>421</td>
</tr>
<tr>
<td>PA2754 - hypothetical protein (+)</td>
<td>PA2755 - ecotin precursor (+)</td>
<td>1020</td>
<td>62.2%</td>
<td>452</td>
</tr>
<tr>
<td>PA1244 - hypothetical protein (-)</td>
<td>PA1245 - hypothetical protein (+)</td>
<td>1014</td>
<td>55.1%</td>
<td>628</td>
</tr>
<tr>
<td>PA2763 - hypothetical protein (-)</td>
<td>PA2764 - hypothetical protein (+)</td>
<td>1012</td>
<td>60.5%</td>
<td>345</td>
</tr>
<tr>
<td>PA0981 - hypothetical protein (+)</td>
<td>PA0982 - hypothetical protein (-)</td>
<td>1010</td>
<td>51.9%</td>
<td>329</td>
</tr>
</tbody>
</table>
2.4.2 Global transposon insertion statistics in *P. aeruginosa* PAO1

The transposon mutant library used (Jacobs et al., 2003) contained 30,100 unique transposon insertion locations where mutants could be recovered. 90.6% (27,263) of transposons were inserted inside ORFs, while the rest 9.4% (2,837) fell within intergenic regions.

![Scatterplot of all transposon insertion gaps](image)

Figure 2-3: Transposon gap sizes along the *P. aeruginosa* PAO1 chromosome, with notable regions of low insertion density.
In order to perform an initial exploratory data analysis to get a general overview of the distribution and density of transposon gap sizes throughout the genome, transposon gap sizes were calculated and plotted along the entire chromosome (Figure 2-3), and their distribution shown in Figure 2-4. The mean transposon gap size is 209bp, and the median is 118bp. The sizes vary markedly, ranging from 1bp to 13,093bp.

Notable features that can be observed immediately are several regions of low transposon insertion density as pointed by the arrows, and in particular, a fairly large region approximately between coordinates 2M and 3M. In fact, this
particular region was documented in the transposon mutagenesis study, however
the cause for the low insertion density was unknown (Jacobs et al., 2003).
Regions with low transposon insertion density may explain the presence of small
clusters of essential genes, as is evident by the sporadic presence of unusually
large gaps. One extreme example is the largest transposon gap in the whole
dataset, spanning 13,093bp between coordinates 4772974 and 4786066, which
includes 9 genes encoding 50S ribosomal proteins and DNA-directed RNA
polymerases, as well as 4 tRNAs. This region is very likely to be truly essential
since it is well known that such ribosomal proteins cannot be easily disrupted
(Hare et al., 2001; Brodersen and Nissen, 2005).

The low insertion density region between 2M and 3M, nonetheless, is not
characterized by sporadically large gaps. It is possible that essential genes in this
region are more spread out leading to a smoother distribution of transposon gap
sizes, however further analysis showed that the low insertion density is
consistent within both intergenic and ORFs, thus it seemed likely that the
explanation might be due to other effects. Polar effects could be playing a role,
as could be biases inherent in the methodology used. Finally, accessibility of that
region to transposons within the bacterial cell could also have an effect.

2.4.3 Transposon gap statistics in ORFs vs. IGRs

The 30,100 transposon gaps were classified into three classes: 1) 
transposon gaps that were exclusively within ORFs (23,205); 2) transposon gaps
that were exclusively within IGRs (1,222); and 3) transposon gaps that crossed
between ORF regions and IGRs (5,673), which were discarded. Statistics for the first two classes were calculated.

Comparing the statistics of transposon gap sizes within ORFs with those within IGRs in the quantile-quantile plot (Figure 2-5), it can be observed that transposon gaps are disproportionately larger within ORFs than within intergenic regions. The mean transposon gap size in ORFs is just less than double that in intergenic regions, while the median size is more than double. The largest
transposon gap within an ORF is 4,479bp (4942853..4947331), spanning four genes within a fourteen-gene operon, encoding cell division proteins and several transferases and ligases. This is more than 5-fold larger than the largest transposon gap (880bp) within an IGR that is located between two genes encoding hypothetical proteins. Transposon gaps being disproportionately larger within ORFs is expected, as insertions within ORFs are more likely to disrupt the production of functionally important proteins and the size of intergenic sequences in general is smaller. Also, the presence of ubiquitous transcription and translation machinery in highly expressed ORFs might have a negative effect on the success of a transposon insertion.

<table>
<thead>
<tr>
<th>Transposon gap size</th>
<th>ORFs</th>
<th>IGRs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean size (bp)</td>
<td>151.64</td>
<td>80.53</td>
</tr>
<tr>
<td>Median size (bp)</td>
<td>94</td>
<td>44.5</td>
</tr>
<tr>
<td>Max. Size (bp)</td>
<td>4479</td>
<td>880</td>
</tr>
<tr>
<td>75% quantile</td>
<td>203 bp</td>
<td>110 bp</td>
</tr>
<tr>
<td>90% quantile</td>
<td>361 bp</td>
<td>200 bp</td>
</tr>
<tr>
<td>95% quantile</td>
<td>487 bp</td>
<td>261 bp</td>
</tr>
</tbody>
</table>

2.4.4 Transposon gap sizes follow an exponential distribution

Assuming an exponential distribution (See Materials and Methods in Chapter 2.3), the statistical significance of observing a transposon gap of a certain size can now be assessed using a p-value. A p-value cutoff can then be
chosen to filter transposon gaps large enough to be statistically significant (see Materials and Methods in Chapter 2.3).

The transposon insertion rate, \( r \), was estimated by dividing the total number of transposon insertions by the length of the genome:

\[
r = \frac{30,100}{6,264,404} \times 1000 = 4.80 \text{ insertions per kb}
\]

However, this would be an underestimate of the true insertion rate, since the 30,100 insertions only include the ones that could be recovered, i.e. all insertions that fell within essential regions would have been "lost". In order to measure the extent of under-estimation, and to approximate the "true" background insertion rate, genes that are very likely to be essential (and thus the most likely to "lose" insertions) were removed: all genes larger than 1kb that did not contain any transposons were thus removed, and the insertion rate recalculated.

A total of 158 genes, comprising 225,685bp, were removed, so an adjusted transposon insertion rate was calculated which would be a closer estimate of the true background insertion rate:

\[
r = \frac{30,100}{6,264,404 - 225,685} \times 1000 = 4.98 \text{ insertions per kb}
\]

In order to see whether transposon gaps sizes really follow an exponential distribution empirically, the sizes of transposon gaps calculated using the \( P. aeruginosa \) PAO1 transposon mutant library were plotted in a histogram and
compared to a theoretical exponential distribution using the estimated background insertion rate of 4.98 (Figure 2-6).

Figure 2-6: (A) Comparing the distribution of all transposon gap sizes in the *P. aeruginosa* PA01 transposon mutant library, with a theoretical exponential distribution of insertion rate 4.98 per kb; values for transposon gap sizes >2000bp were omitted due to image size constraints; (B) residues resulting from subtracting the theoretical distribution from the empirical distribution. The residue value for the 0-20 bin is 1,946, and has been omitted due to image size constraints.
As can be seen in Figure 2-6A, transposon gap sizes in *P. aeruginosa* PAO1 closely follows the exponential distribution. The significant exception is for transposon gaps of sizes 0 to 20bp, which are overrepresented in the empirical data. A possible explanation is the preference of transposons to repeatedly insert into certain ‘hotspot’ regions within the genome, resulting in insertions that are very close to each other. Analyzing the residues of the two distributions (Figure 2-6 B), it can be observed that transposon gap sizes larger than 600bp are overrepresented in the empirical distribution. This suggests that these larger gaps are likely to not have arisen by chance alone, but introduced due to the presence of essential genes. Transposon gaps larger than 2000 bp only appear in the empirical distribution (these were not shown in Figure 2-6 due to image size constraints).

Using r, equation E4 (see Materials and Methods, chapter 2.3) was then applied to calculate transposon gap size cutoffs for chosen p-values:
Table 2-3: Transposon gap size cutoff for given p-values. Two insertion rates are compared.

<table>
<thead>
<tr>
<th>P-value</th>
<th>Transposon gap size with ( r=4.80 ) insertions per kb</th>
<th># “essential” intergenic sequences</th>
<th>Transposon gap size with adjusted ( r=4.98 ) insertions per kb</th>
<th># “essential” intergenic sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.5</td>
<td>144 bp</td>
<td>1430</td>
<td>139 bp</td>
<td>1496</td>
</tr>
<tr>
<td>&lt; 0.4</td>
<td>191 bp</td>
<td>863</td>
<td>184 bp</td>
<td>928</td>
</tr>
<tr>
<td>&lt; 0.3</td>
<td>251 bp</td>
<td>474</td>
<td>242 bp</td>
<td>511</td>
</tr>
<tr>
<td>&lt; 0.25</td>
<td>288 bp</td>
<td>336</td>
<td>278 bp</td>
<td>368</td>
</tr>
<tr>
<td>&lt; 0.2</td>
<td>335 bp</td>
<td>225</td>
<td>323 bp</td>
<td>252</td>
</tr>
<tr>
<td>&lt; 0.15</td>
<td>395 bp</td>
<td>143</td>
<td>381 bp</td>
<td>152</td>
</tr>
<tr>
<td>&lt; 0.1</td>
<td>480 bp</td>
<td>71</td>
<td>462 bp</td>
<td>83</td>
</tr>
<tr>
<td>&lt; 0.05</td>
<td>624 bp</td>
<td>17</td>
<td>601 bp</td>
<td>21</td>
</tr>
</tbody>
</table>

As can be seen from Table 2-3, the transposon gap size cutoff calculated using the adjusted insertion rate does not differ significantly from that using the original insertion rate (a change of 3.4-3.6% for all p-values). This suggests that the bias introduced by essential genes does not significantly affect the overall background insertion rate, which is likely due to the merits of using a near-saturation approach. However, given the relatively large genome of *P. aeruginosa*, the proportion of essential genes relative to the whole genome is likely to be small. Thus it is possible that for small genomes, the bias effect of essential genes on the estimation of the background transposon insertion rate would be greater, however further investigation would be needed.
Given that the insertion rate calculated using the original data is generally a good approximation to the true background insertion rate, and in order to simplify the process of determining the transposon gap size cutoff given its automated nature, no bias adjustments were made. The insertion rate of \( r = 4.80 \) insertions/kb was used in all subsequent steps.

### 2.4.5 Essential intergenic sequences in *Pseudomonas aeruginosa* PAO1

In order to choose a suitable p-value cutoff, several considerations were made. First of all, as will be described in subsequent chapters, positive and negative data sets of protein-coding genes and ncRNAs were compiled in order to assess the accuracy of predictions. In these test data sets, most ncRNAs were located within relatively small transposon gaps (\(~300\text{bp}\) ), thus in order to be able to perform the test phase, a p-value cutoff corresponding to a transposon gap size no larger than 300bp must be chosen. Moreover, choosing a less stringent cutoff would increase the sensitivity of gene prediction.

Based on these considerations, a p-value cutoff of \( p < 0.25 \) was therefore chosen. This corresponds to a transposon gap size cutoff of 288bp. All intergenic sequences corresponding to transposon gaps larger than 288bp were extracted for further analysis. A total of 336 such sequences were extracted, of which 75 had a p-value less than 0.1.

### 2.4.6 Improvements over an original faulty IGR extraction algorithm

At the outset of the project, the original IGR extraction algorithm identified IGRs harboring unusually large transposon gaps, however instead of returning
only the subsequence absent of insertions, it returned the entire IGR for further analysis. As a result, this led to the prediction of many non-essential genes in subsequent analysis steps, which are not within the scope of the goal of identifying essential genes in this project. The algorithm was thus later modified into the current form described in Materials and Methods section in chapter 2.3.

Furthermore, the original methodology used a hard-coded transposon gap size cutoff of 300bp to define the essentiality of a region. Considering the applicability to other genomes, using a hard-coded size cutoff was inflexible, as different transposon mutagenesis studies will inevitably have different transposon insertion rates. As a result, statistical modelling was incorporated into the new algorithm in order to generate a more accurate cutoff.

2.5 Concluding remarks

In this chapter I have illustrated the importance of defining essential regions operationally: using statistical analysis of data from a transposon mutagenesis study, essential regions were defined as those containing significantly large transposon gaps according to a set $p$ value. I have developed a statistical model to estimate the statistical significance of a transposon gap, and used a cutoff of $p < 0.25$ to extract intergenic sequences which are likely to be essential, to err on the side of increased sensitivity. I hypothesized that within such regions, some novel essential protein coding genes and/or ncRNAs could be identified. As such, 336 putative essential intergenic sequences were extracted for further analysis.
CHAPTER 3 IDENTIFYING NOVEL PUTATIVE ESSENTIAL PROTEIN CODING GENES

3.1 Summary

The 336 putatively essential intergenic sequences (PEISs) extracted from the procedure described in Chapter 2 were subjected to a search for novel protein-coding genes. Using test datasets, the accuracy of the method used was estimated and improvements made in gene detection. Using my method, novel protein-coding genes that were confirmed experimentally to be transcribed were identified. However, the results illustrated that the PAO1 strain genome annotation is of good quality with regard to its annotation of probable essential genes, since very few new genes were detected. Still some new putatively essential genes were identified that warrant further study as potential antimicrobial drug targets.

3.2 Materials and Methods

3.2.1 BLASTX

A BLASTX search was performed for every sequence against the NCBI 'nr' protein database. A choice was made at the implementation stage to use remote BLASTX instead of a local BLASTX search. Remote BLASTX would involve both submitting the sequences and retrieving the results over the Internet, which requires an Internet connection and slower speeds due to the need of queuing for jobs and passing data back and forth over the Internet. Local
BLASTX, on the other hand, would consist of first downloading the complete 'nr' database from the NCBI FTP site, setting up the BLAST database locally, and running a standalone BLASTX search. It would be faster, however it requires significant computer space and network bandwidth requirements (the compressed tar.gz version of the protein 'nr' database was 1.87GB as of April 7th, 2008, and will continue to grow), and would constantly need to be updated to obtain the latest version for every run. As the nature of our novel gene finding requires the most up-to-date protein database, and to save the labour of frequent downloads and updates, the remote BLASTX approach was thus chosen. This approach will be amenable for incorporation into a pipeline for the identification of such regions – a pipeline that would be useable by a researcher with a basic internet connection, with no requirements for significant disk storage for local installation of the nr database.

A Perl script implementing the BioPerl RemoteBLAST package (Bio::Tools::Run::RemoteBlast) was used for this purpose. A single FASTA file, containing all essential intergenic sequences extracted, was submitted as a batch BLASTX job to the NCBI 'nr' protein database, using an e-value cutoff of 1E-10 (which can be adjusted by the user). For each sequence, results were retrieved as they became ready. In the case of at least one hit above the e-value cutoff, the BLASTX report is first saved in HTML format for the purpose of further manual inspections.
3.2.2 Extracting the coding sequence within the IGR

The BLASTX report was then computationally parsed to identify the coding sequence of the putative protein-coding gene. For practical purposes, only the top BLAST hit was used in this automated procedure, however as noted above the whole BLAST report was saved for further manual perusal.

The coding sequence of the putative protein-coding gene is computed using the alignment(s) of the Highest Scoring Pair(s) (HSP) of the top protein hit. In the simplest case that the top hit only contains a single HSP, and the HSP spans the entire length of the protein (including the start and stop codons), then the portion of the query IGR that participates in the alignment is the coding sequence for the novel gene candidate.

When the top BLAST hit has one single HSP, but the HSP does not span the whole length of the protein (either the start or the end of the protein is missing, or both), then an attempt is made to extend the alignment towards the missing end(s) to obtain the start or end of the coding sequence, until the boundaries of the intergenic sequence are reached. This scenario occurs when either ends of the sequences are not similar enough to be included in the alignment. If the boundaries of the intergenic sequence are reached but the putative coding sequence lacks a start codon or a stop codon, or both, then this is labelled as a “truncated hit”.

In the case that the top BLAST hit has more than one HSP, these HSPs may 1) be in the same frame but do not overlap, or 2) are in different frames. If they are in the same frame but do not overlap, then if they are near enough to
each other (default is 50bp but can be adjusted by the user), they are bridged to obtain a single coding sequence. This case happens when the middle portion of the protein alignment is not similar enough and thus is broken up into two separate HSP alignments. If the HSPs are in different frames, however, then this is labelled as a "frameshifted hit" (likely a pseudogene). If a premature stop codon is observed in the putative coding sequence with respect to the protein hit, it's labelled as "disrupted" (likely a pseudogene).

If any of the boundaries of the predicted gene overlaps with that of an upstream or downstream gene, which can sometimes happen, an allowance of 3bp is made to accommodate these cases.

When the putative coding sequence is identified, its relative coordinates within the intergenic region are then mapped into the absolute coordinates of the genome. It is also translated into its deduced protein sequence, which is then subjected to further characterization.

3.2.3 Automatic characterization of putatively novel protein-coding genes

The list of putative novel protein-coding genes were then characterized in terms of their possession of conserved protein domains or motifs, or their membership in protein families or orthologous groups, and subcellular localization. Information that can be obtained for any of these can be regarded as evidence of function.

Putative protein-coding genes were first searched with a remote version of Reverse Position-Specific BLAST (RPS-BLAST) against the NCBI Conserved
Domain Database (CDD) version 2.13 (last accessed 15th April 2008). The CDD (Marchler-Bauer et al., 2005) contains Position Specific Scoring Matrices (PSSMs) for protein domain collections from 1) Pfam, a database of protein families (Finn et al., 2006); 2) SMART, a database of protein domains (Letunic et al., 2006), 3) the Cluster of Orthologous Groups (COG) database (Tatusov et al., 2003), and 4) the NCBI Protein Clusters (PRK) database (Wheeler et al., 2008). CDD version 2.13 contains 9,318 PSSMs from Pfam, 663 from SMART, 4,873 from COG, and 6,043 from PRK. All hits were returned with an e-value cutoff of 1E-4 for consistency with annotation guidelines in the *Pseudomonas* Genome Database. Putative protein-coding genes were also searched against the TIGRFAM database (Selengut et al., 2007), a collection of protein families defined based on Hidden Markov Models (HMMs), using hmmpfam (http://tigrblast.tigr.org/web-hmm/) with an e-value cutoff of 1E-4. Hmmpfam also searches the PFAM database simultaneously, though here it uses HMMs instead of RPS-BLAST.

Protein subcellular localization was also predicted for each putative novel protein-coding gene using PSORTb version 2.04 (Gardy et al., 2005).

A spreadsheet was then automatically generated with a script summarizing the results of the analysis for all putative novel protein coding genes. The fields of the spreadsheet are according to the following list: 1) PEIS ID; 2) Size of transposon gap sequence; 3) Transposon gap essentiality p-value; 4) Left gene locus; 5) Left gene product name; 6) Left gene strand; 7) Right gene locus; 8) Right gene product name; 9) Right gene strand; 10) Number of BLASTX
3.2.4 Manual inspection

BLASTX reports for each of the putative protein-coding genes were then manually inspected for possible errors or hits that might have been missed. For instance, genes that overlap more than 3bp with an adjacent flanking gene are not predicted automatically, but can be recovered by the manual inspection of BLAST alignments. New locus IDs were then assigned to each of the putative protein-coding genes.

3.2.5 Prioritization of predictions for transcription verification

An approach developed by David Mulder in the Brinkman lab was used to sort and prioritize the list of putative protein-coding genes for transcription verification. A score \( S \) was computed for each gene based on the following formula, which calculates the percentage of GC-rich codons (consisting of three G's or C's) comprising the protein sequence:
where \( N_{GC} \) is the number of GC-rich codons, and \( N \) is the total number of amino acid residues in the protein. It was observed during the transcription verification process that such a score was good at discriminating between genes that had a positive transcriptional signal from those that didn’t, under rtPCR or Northern blot analysis conditions. This was subsequently used to prioritize genes (discussed further in chapter 3.3.7).

3.2.6 Cross-referencing with the \textit{P.aeruginosa} PA14 transposon mutant library

Approximately 96.3\% of the DNA sequence found in \textit{P. aeruginosa} PA01 is also found in \textit{P. aeruginosa} PA14, and the presence of comprehensive transposon mutant libraries for both strains facilitates the cross-referencing between the two sets of data to gain additional insight into the likelihood of true essentiality of novel predicted genes. In essence, I hypothesize that predicted genes that have no transposon mutant in strain PA01, would also likely have no transposon mutant in the orthologous gene in strain PA14, if an orthologous gene in strain PA14 is present. Such genes would be hypothesized to be more likely true essential, based on my study of the degree of conservation of essential genes (see Chapter Chapter 5). However, since these are different strains, I opted not to pool the transposon mutant data from the two strains for my analysis.
Nucleotide sequences of novel gene predictions in PAO1 were queried using BLASTN to obtain the start and end coordinates of the corresponding orthologous sequence in PA14, if any. The PA14 transposon mutant library data was then examined for any insertion located between the start and the end of the specified PA14 sequence.

3.2.7 Test data set

In order to test the performance of the protein-coding gene prediction procedure, positive and negative data sets were compiled.

The positive data set was composed of 14 experimentally verified protein-coding genes (classified as Class 1 in the *Pseudomonas* database) having no transposon insertions, and whose sizes are between 289bp and 500bp. The upper bound of 500bp was chosen to limit the data set to shorter genes, as these are more likely to be missed in original gene prediction procedures. The lower bound of 289bp was chosen so that after their removal, the resulting “intergenic region” would be above the size cutoff to be labelled as “essential” and thus subjected to analysis. Following is the list of 14 genes comprising the positive data set: PA0408 (πIG, 408bp), PA0610 (prtN, 315bp), PA0970 (tolR, 441bp), PA1456 (cheY, 375bp), PA1705 (pcrG, 297bp), PA1715 (pscB, 423bp), PA1721 (pscH, 432bp), PA1722 (pscl, 339bp), PA2960 (πIZ, 357bp), PA3807 (ndk, 432bp), PA4230 (pchB, 306bp), PA4232 (ssb, 498bp), PA4386 (groES, 294bp) and PA4847 (accB, 471bp).
The negative data set was composed of five annotated non-coding RNAs, three of which were experimentally verified sRNAs (PrrF1, PrrF2 and PA5181.1), in addition to one 16S rRNA (PA5369.5) as well as one 23S rRNA (PA0668.4). These were chosen due to their sizes, such that their removal from the annotations would result in a PEIS.

All 14 genes in the positive data set and the 6 in the negative data set were removed from the annotations before the IGR extraction step and BLASTX analysis. To prevent the BLASTX search from returning the exact same protein entry, a line of code was added to the script to ignore such cases.

3.2.8 Lab verification methods

Putative novel essential protein-coding genes were subjected to transcription verification performed by David Mulder, of the Brinkman Laboratory. This was done using RT-PCR after a total RNA extraction step as described previously (Brinkman et al., 1999), but outlines briefly below.

Total RNA was extracted from a liquid LB broth culture of *P. aeruginosa* PAO1 (37C with shaking) at OD600 = 1.35 (or at earlier OD’s to characterize cells under different stages of log phase growth) using RNeasy spin columns (Qiagen). To enhance small RNA recovery the RW1 wash was omitted. The RNA was treated with DNase I (Qiagen) and stored for later use at -80C. RNA quality and integrity were assessed by denaturing agarose-formaldehyde gel electrophoresis (examining the intensity of the ribosomal RNA bands) in addition to Nanodrop spectrophotometer measurement.
RT-PCR was then performed using the following steps. First strand complementary cDNAs were generated from total RNA using reverse transcriptase (Fermentas) and complementary DNA oligo primers. First strand cDNA was amplified by PCR (Fermentas) using two internal primers to generate internal amplicons that were assessed by gel electrophoresis. Amplicon identity was further assessed by specific restriction endonuclease (Fermentas) digestion. Short and long target DNA contamination controls were performed concurrently.

3.3 Results and Discussion

Using a p-value cutoff of \( p < 0.25 \), corresponding to a transposon gap size cutoff of 289bp, a total of 336 PEISs were extracted and subjected to BLASTX analysis. Of the 336 PEISs, 205 did not return any BLAST hits, while 131 returned at least one hit.

Of the 131 sequences which returned at least one hit, 73 were labelled as “truncated hits”, meaning that only a portion of the top protein hit was predicted within the intergenic region and not the whole. Most of these are likely negatives (pseudogenes), but it is possible that some of these are *bona fide* proteins but extend beyond the boundaries of the “essential intergenic sequence”; thus, even if they were *bona fide*, they would not be essential and so were ignored in this study. Many of these “truncated hits” contained hits to the upper portion of a protein, whose lower portion coincides with the downstream annotated gene flanking the intergenic region. These represent cases where an alternative start site of the downstream gene is possible, and highlight the common difficulty for annotators in identifying the correct start codon for a gene (Nielsen and Krogh,
This problem particularly affects GC-rich genomes such as *P. aeruginosa*, since there is a higher frequency of possible start codons (in particular the GTG triplet occurs randomly much more often) and the AT rich stop codons (TAA, TGA, TAG) do not occur as frequently as in AT rich genomes, making it hard to distinguish whether the lack of a stop codon in a region is due to chance, or due to the presence of a functional gene.

Eleven out of the 131 sequences were labelled “frameshifted hits”, where coding sequences were predicted for different portions of a protein hit, but in different frames. Furthermore, nine out of the 131 sequences were labelled “disrupted hits”, where a premature stop codon was predicted relative to the stop codon of the protein hit. It is possible some of these are true pseudogenes, regardless both sets were ignored in this study.

Excluding all sequences that were labelled either as “truncated”, “frameshifted” or “disrupted”, after a manual inspection of the BLAST reports a total of 46 protein-coding gene predictions were made, comprising 12 in the positive dataset and 34 novel predictions.

### 3.3.1 Test performance

Of the fourteen genes in the positive dataset, twelve were recovered, yielding an 86% true positive rate. Of the two false negatives, PA1456 could not be recovered because the start codon of the top BLAST hit was incorrectly predicted. Its first 6 amino acids overlap with the last 6 amino acids of its upstream gene, PA1455, which exceeds the 3bp allowance for such gene
overlap cases. This is a feature that could be easily modified. The other gene, PA1722, could not be recovered because it was directly adjacent to PA1721, which was also removed from the annotations. The simultaneous removal of these two adjacent genes resulted in one single PEIS, which by default makes only one gene prediction since it only bases its prediction on the top hit. However, manual inspection of the BLAST reports recovered these two genes that could not be recovered by the automated process. This highlights the importance of saving the BLAST reports and their manual review and illustrates the importance of performing an evaluation of your method using both positive and negative controls. These scenarios, resulting in two positive control genes that weren't detected, could be easily modified to ensure such cases were detected in the automated analysis.

Of the five genes in the negative data set, all of them were found to be correctly identified as true negatives. The three sRNAs returned no protein hits, however both the 16S and the 23S rRNAs returned many "protein hits" labelled as "disrupted hits" due to the presence of many premature codons, and thus no protein-coding gene prediction was made.

3.3.2 Prediction of 34 novel putative essential protein-coding genes in Pseudomonas aeruginosa PAO1

A total of 34 novel putative essential protein-coding genes were predicted, with locations spread throughout the genome. In general, all predicted genes were conserved in a limited number of organisms, and in particular within the P. aeruginosa species itself. Twenty nine returned less than 10 BLAST hits, and
none of them returned more than 25 hits. In comparison, half of the genes in the positive dataset returned the maximum limit of 100 BLAST hits, with only two genes having less than 10 hits. The predicted genes are also relatively short in size, ranging from 129 to 459bp (corresponding to 43 and 153 amino acids respectively), with an average of 232bp and a standard deviation of +/- 75.

Being the first genome to be sequenced in the *P. aeruginosa* species, the limited phylogenetic distribution of these genes and their small sizes may have contributed to their missed annotation. However, the increasing availability of genomes from different strains and species have allowed these now to be detected, highlighting the importance of sequencing related genomes for comparative genomics. Still the lack of homologs in diverse species leads to some scepticism that most of these genes are truly putatively essential, since, as illustrated in Chapter 5, many genes that are essential are highly conserved across diverse phylogenetic lineages. Further functional analysis is required.

### 3.3.3 Functional characterization of novel gene predictions

The attempt to predict function for these novel genes returned limited results, perhaps not surprisingly, given that these may be genes that were missed through initial annotation efforts. Whereas more than 75% of the genes in the positive data set could be characterized by at least one of COG, PFAM, SMART or PRK, only 4 of the 34 (12%) of the novel predicted genes could be characterized in this way. Two returned hits from the COG database, while an additional two returned hits from the PFAM collection. None of them returned hits from SMART or PRK. Six of the genes in the positive dataset (50%) returned hits
to TIGRFAM families, but none for the novel predicted genes. The limited ability
to characterize novel gene predictions by functional motifs could be a direct
result of the limited phylogenetic distribution shown, since all five protein
functional motif collections were built by clustering profile similarities using
comparative genomics data.

One observation made was that the two genes which returned PFAM hits
were searched using HMMs, whereas no results for these genes were returned
using RPS-BLAST. This is suggestive of a generally higher sensitivity of HMM­
based searches, though this would warrant further study with a larger dataset.

PSORTb returned slightly better results than motif finding, being able to
make subcellular localization predictions for 10 genes (29%), and prediction of
signal peptides for 2 additional genes (6%). PSORTb is notable for its high
specificity (>95%) (Gardy et al., 2005), and so it is likely that many of these
predictions are bona fide.

3.3.4 Probability of essentiality of novel gene predictions

Concerning the probability of essentiality of the 34 novel predicted genes,
the average size of the transposon gap within which novel genes were predicted
was 450.82, corresponding to a p-value of 0.145. Five (14.7%) were located
within transposon gaps with p-values less than 0.05; nine (26.5%) with p-values
between 0.05 and 0.1; seven (20.6%) with p-values between 0.1 and 0.15; Five
(14.7%) with p-values between 0.15 and 0.2; and finally eight (23.5%) with p-
values between 0.2 and 0.25.
The size of a transposon gap appears to be correlated with whether or not BLAST hits were returned, suggesting that those with higher p values are likely not true genes. The average size of transposon gaps for which no BLAST hit was returned is 391.41bp (corresponding to an essentiality p-value of 0.167), whereas the average for the ones with at least one BLAST hit is 459.70bp (essentiality p-value of 0.146).

Cross-referencing with data from the *P. aeruginosa* PA14 transposon mutant library, of the 34 novel putative gene predictions in *P. aeruginosa* PA01, 15 (44.1%) also did not have any transposon insertion in the orthologous sequence in *P. aeruginosa* PA14, suggesting these are likely to be truly essential. Another 15 had at least one transposon insertion in the PA14 ortholog, while the remaining 4 did not have a detectable ortholog in PA14 and thus represent putative unique genes in *P. aeruginosa* PA01.

Of the 14 genes that were found to be transcribed, seven (50%) did not have any transposon insertion in the PA14 ortholog: PA0457.1, PA0613.1, PA0836.2, PA2173.1, PA2566.1, PA2747.1, PA2926.1; five (35.7%) had at least one insertion: PA0050.1, PA1838.1, PA2118.1, PA3414.1, PA3762.1; finally, two (14.3%) were unique in PA01: PA1152.1 and PA1366.1.

### 3.3.5 Descriptive summary of 14 transcribed novel protein-coding genes

At the time of writing, 14 out of the 34 predicted protein-coding genes were experimentally tested for transcription, performed by David Mulder, and all of them were found to be transcribed. The remaining 20 genes were still in the
progress of being verified. Below is a descriptive summary of each of the genes, classified into three categories, with regard to the likelihood that the predicted gene is likely essential, sorted in order of decreasing likelihood: 1) Genes with evidence of essentiality in both strains of *P. aeruginosa*; 2) Genes that are relatively unique to strain PAO1; and 3) Genes for which the ortholog in strain PA14 are not essential.
1) *Multi-strain evidence of essentiality*

**PA2173.1** – This novel gene is located within a cluster of genes of unknown function, and also coincides with the region of low transposon insertion density. It is predicted to be localized in the cytoplasm, and is conserved within the *P. aeruginosa* species only. Despite the low phylogenetic distribution, no transposon mutant was observed in *P. aeruginosa* PA14, adding confidence to its likelihood to be essential.

![Schematic diagram of the genomic context of PA2173.1](image)

*Figure 3-1: Schematic diagram of the genomic context of PA2173.1, a newly identified putative essential protein-coding gene, denoted by the red arrow. Blue arrows denote previously annotated protein-coding genes, and a black triangle denotes the presence of a transposon mutant in that location. The direction of the arrow denotes the direction of transcription.*
PA0836.2 – This novel 201bp gene is transcribed directly upstream of ackA, a gene encoding acetate kinase, within a 306bp IGR absent of transposon insertions. A recently discovered sRNA (Livny et al., 2005) is transcribed directly in the opposite strand overlapping this predicted gene, suggesting the possibility of regulation by antisense RNA, though of course this warrants further study before making any conclusions. This novel gene is conserved extensively amongst the Pseudomonas genus, as well as Ralstonia and Shewanella, and is likely essential in P. aeruginosa PA14.

Figure 3-2: Schematic diagram of the genomic context of PA0836.2, a newly identified putative essential protein-coding gene, denoted by the red arrow. Green arrow denotes an sRNA.
PA2566.1 – This small 210bp gene is transcribed within a large transposon gap of 658bp whose orthologous region in *P. aeruginosa* PA14 is also not disrupted by any transposon insertion, suggesting this region is truly essential. However, it is conserved only within *P. aeruginosa*.

Figure 3-3: Schematic diagram of the genomic context of PA2566.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.
**PA2747.1** – This novel 234bp gene is transcribed in the same direction as both flanking genes, but is located closer to the upstream gene, which encodes a probable methionine aminopeptidase involved in protein modification. No transposon mutant was identified in the orthologous region in *P. aeruginosa* PA14, suggesting it is likely to be essential, however it is conserved only within the *P. aeruginosa* species. The function of this novel gene is unknown.

![Image showing genomic context of PA2747.1](image-url)

**Figure 3-4: Schematic diagram of the genomic context of PA2747.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.**
**PA2926.1** – This small 183bp gene is transcribed in the antisense direction as both flanking genes. Predicted to be localized in the cytoplasm, it is conserved within the *Pseudomonas* genus, with hits to putative acetyltransferases. The orthologous region in *P. aeruginosa* PA14 was not disrupted by any transposon insertion.

![Diagram of genomic context of PA2926.1](image)

*Figure 3-5: Schematic diagram of the genomic context of PA2926.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.*
2) Relatively unique to strain PA01

**PA0457.1** – This novel gene is transcribed in the minus strand, directly downstream of a probable major facilitator superfamily transporter, likely within the same operon. It spans the whole IGR of 369bp absent of any transposon insertion, and overlaps with the last 7bp of PA0457. Localized in the cytoplasmic membrane, this gene lacks an annotated ortholog in *P. aeruginosa* PA14, however is conserved within 5 *Pseudomonas* species as well as numerous species in the Burkholderiales order.

---

**Figure 3-6:** Schematic diagram of the genomic context of PA0457.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.
PA1366.1 – Transcribed in the same direction as both flanking genes and directly upstream of PA1366 encoding an hypothetical protein likely involved in cell division, this novel 279bp gene lies within a 423bp IGR absent of any transposon insertions, and likely within the same operon. It has a very limited phylogenetic distribution, with significant hits only within *P. aeruginosa*, however no ortholog is present in *P. aeruginosa* PA14.

Figure 3-7: Schematic diagram of the genomic context of PA1366.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.
PA0613.1 – A 225bp gene located within a 363bp transposon gap, transcribed in the same forward direction as the two flanking genes. Predicted to be localized in the cytoplasm, this gene does not show any significant conservation in other species.

Figure 3-8: Schematic diagram of the genomic context of PA0613.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.
PA1152.1 – Predicted within one of the two large transposon gaps within the 1143bp IGR, this novel 273bp gene contains a motif similar to that of Colicin D proteins, which typically have RNase activity. Conserved only in *P. aeruginosa* PA7 within the *Pseudomonas* genus, it is poorly but widely conserved within *Yersinia* spp.

Figure 3-9: Schematic diagram of the genomic context of PA1152.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.
3) Orthologs in strain PA14 are not essential

**PA1838.1** – This small 249bp gene is transcribed in the same direction as both flanking genes, directly downstream of PA1839, and lies within a 460bp transposon gap. It is conserved within the *Pseudomonas* genus, however it is disrupted by a transposon insertion in *P. aeruginosa* PA14, suggesting it could also be non-essential in *P. aeruginosa* PA01.

Figure 3-10: Schematic diagram of the genomic context of PA1838.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.
**PA2118.1** – Located within a relatively long 534bp IGR absent of any transposon insertions, this novel gene is transcribed downstream of a gene encoding alcohol dehydrogenase. Its subcellular localization is uncertain, however it’s predicted to be non-cytoplasmic, and a signal peptide is predicted for residues 1 to 22. A rho-independent transcription terminator is predicted just downstream. Its phylogenetic distribution is limited to within the *P. aeruginosa* species, however the orthologous region in *P. aeruginosa* PA14 has been disrupted by one transposon insertion, providing some scepticism that this is a truly essential gene.

![Figure 3-11](image)

*Figure 3-11: Schematic diagram of the genomic context of PA2118.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.*
PA3762.1 – This 279bp gene is transcribed in the same direction as both flanking genes, downstream of a Phosphoribosylformylglycinamidine synthase gene involved in the metabolism of nucleotides. It is conserved only within the *Pseudomonas* genus, and its *P. aeruginosa* PA14 ortholog is disrupted by one transposon insertion.

Figure 3-12: Schematic diagram of the genomic context of PA3762.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.
PA3414.1 – This novel 459bp gene is transcribed downstream of PA3414 encoding a hypothetical protein, and spans almost the whole of the 492bp IGR with no transposon insertions. It is conserved only within the *Pseudomonas* genus. However, the *P. aeruginosa* PA14 ortholog contains three transposon mutants, suggesting that it might not be an essential gene.

Figure 3-13: Schematic diagram of the genomic context of PA3414.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.
PA0050.1 – Located within a large 674bp transposon gap within an IGR of 823bp, between a hypothetical protein and a potential phenazine-modifying enzyme, this short gene of 138bp is only conserved in *P. aeruginosa* PA14, *P. fluorescens* Pf-5 and *P. entomophila* L48. Its product is unknown, however one signal peptide was predicted. This gene is disrupted in the orthologous region in *P. aeruginosa* PA14 by one transposon insertion.

Figure 3-14: Schematic diagram of the genomic context of PA0050.1, a newly identified putative essential protein-coding gene, denoted by the red arrow.
3.3.6 Previously identified novel non-essential genes

At the outset of the project, using the original intergenic region extraction algorithm as described in Chapter 2, using a transposon gap size cut-off of 300bp, a total of 31 novel putative protein-coding genes were predicted by this old method. Of these, 30 were tested for transcription; 22 were found to be transcribed (true positives), 6 were not found to be transcribed (false positives), and the remaining 2 failed to generate an amplification signal.

Of the 22 genes that were found to be transcribed, 14 were essential and predicted by the newer method, and have just been described above. The remaining 8 represent novel protein-coding genes that were found to be transcribed but were non-essential.

The newer IGR extraction algorithm generated improvements in terms of essential gene prediction accuracy over the original algorithm. Firstly, all non-essential genes were no longer predicted using the newer IGR extraction method. Furthermore, the 6 false positives were also no longer predicted using the newer IGR extraction method.

3.3.7 High proportion of GC-rich codons correlates with false positives and amplification failure of protein-coding gene predictions

As the 30 original gene predictions were tested for transcription, six failed to generate a transcription signal, and two could not be generate an amplification signal. Comparing with the 22 true positives, it was anecdotally observed that (David Mulder, personal communication) the eight negatives possessed a
relatively high proportion of GC-rich codons (containing all G's and C's),
compared to the positive controls.

A scoring method was thus developed by Mulder (see section 3.2.5 in
Materials and Methods), which calculates the proportion of GC-rich codons of a
given gene. Table 3-1: ranks the thirty gene predictions (plus a positive control,
OprF) by their proportion of GC-rich codons, and it can be seen that the eight
negatives are amongst the 12 predicted genes with the highest scores,
possessing 30-39% GC-rich codons.

Figure 3-15 is a histogram of GC scores (proportion of GC-rich codons for
any given gene) computed for all annotated genes in the *P. aeruginosa* PA01
genome, showing that the majority of genes have GC scores within a range of
0.24 to 0.36, with a skewness towards smaller values. Genes with GC scores
above 0.38 is quite unusual.

This scoring function was incorporated to filter for potential false positives
and to prioritize gene predictions for experimental verification.
Table 3-1: Thirty protein-coding gene predictions using the original IGR extraction algorithm, in addition to a positive control OprF, sorted by proportion of GC-rich codons, and their RT-PCR outcome. Gene prediction IDs in bold are the fourteen transcribed novel putative essential genes described in section 3.3 (Results and Discussion). Those with a proportion of GC rich codons in the deduced coding region highly deviated from the mean are more likely not true genes.

<table>
<thead>
<tr>
<th>Gene prediction ID</th>
<th>Proportion of GC-rich codons (%)</th>
<th>RT-PCR outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 PA3969.1</td>
<td>38.28</td>
<td>Failure to detect transcript</td>
</tr>
<tr>
<td>2 PA1483.1</td>
<td>36.49</td>
<td>Failure to amplify</td>
</tr>
<tr>
<td>3 PA4821.1</td>
<td>35.77</td>
<td>Failure to detect transcript</td>
</tr>
<tr>
<td>4 PA0701.1</td>
<td>34.68</td>
<td>Failure to amplify</td>
</tr>
<tr>
<td>5 PA2747.1</td>
<td>34.62</td>
<td>Transcribed</td>
</tr>
<tr>
<td>6 PA3762.1</td>
<td>34.41</td>
<td>Transcribed</td>
</tr>
<tr>
<td>7 PA3217.1</td>
<td>34.38</td>
<td>Failure to detect transcript</td>
</tr>
<tr>
<td>8 PA2192.1</td>
<td>33.71</td>
<td>Failure to detect transcript</td>
</tr>
<tr>
<td>9 PA3414.1</td>
<td>33.08</td>
<td>Transcribed</td>
</tr>
<tr>
<td>10 PA3781.1</td>
<td>31.89</td>
<td>Transcribed</td>
</tr>
<tr>
<td>11 PA2178.1</td>
<td>31.17</td>
<td>Failure to detect transcript</td>
</tr>
<tr>
<td>12 PA4100.1</td>
<td>31.03</td>
<td>Failure to detect transcript</td>
</tr>
<tr>
<td>13 PA1427.1</td>
<td>29.06</td>
<td>Transcribed</td>
</tr>
<tr>
<td>14 PA2926.1</td>
<td>27.87</td>
<td>Transcribed</td>
</tr>
<tr>
<td>15 PA2754.1</td>
<td>27.27</td>
<td>Transcribed</td>
</tr>
<tr>
<td>16 PA3261.1</td>
<td>25.69</td>
<td>Transcribed</td>
</tr>
<tr>
<td>17 PA1938.1</td>
<td>24.48</td>
<td>Transcribed</td>
</tr>
<tr>
<td>18 PA0457.1</td>
<td>24.0</td>
<td>Transcribed</td>
</tr>
<tr>
<td>19 PA0050.1</td>
<td>23.91</td>
<td>Transcribed</td>
</tr>
<tr>
<td>20 PA3733.1</td>
<td>23.16</td>
<td>Transcribed</td>
</tr>
<tr>
<td>21 PA4635.1</td>
<td>22.95</td>
<td>Transcribed</td>
</tr>
<tr>
<td>22 PA2566.1</td>
<td>22.86</td>
<td>Transcribed</td>
</tr>
<tr>
<td>23 PA2118.1</td>
<td>22.22</td>
<td>Transcribed</td>
</tr>
<tr>
<td>24 PA0613.1</td>
<td>21.33</td>
<td>Transcribed</td>
</tr>
<tr>
<td>25 PA0836.2</td>
<td>20.90</td>
<td>Transcribed</td>
</tr>
<tr>
<td>26 PA0041.1</td>
<td>20.51</td>
<td>Transcribed</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>27</td>
<td>PA1838.1</td>
<td>20.48</td>
</tr>
<tr>
<td>28</td>
<td>PA2173.1</td>
<td>18.92</td>
</tr>
<tr>
<td>29</td>
<td>PA1366.1</td>
<td>17.54</td>
</tr>
<tr>
<td>30</td>
<td>OprF (control)</td>
<td>17.09</td>
</tr>
<tr>
<td>31</td>
<td>PA1152.1</td>
<td>5.49</td>
</tr>
</tbody>
</table>

Figure 3-15: Histogram of GC scores (proportion of GC-rich codons in a given gene) for all annotated genes in the *P. aeruginosa* PAO1 genome.

### 3.4 Concluding remarks

In this chapter we have described an approach to predict novel protein-coding genes, that may be essential, within previously annotated intergenic sequences. This method errs on the side of high sensitivity, with the assumption that further downstream testing of candidates will be performed. Using positive and negative datasets to test the method, I was able to detect some conditions
leading to the formation of false negatives (i.e. not allowing for a greater than 3 bp overlap for genes), I was able to modify the method accordingly.

Applying this method to *P. aeruginosa* PAO1, we have predicted putatively essential protein-coding genes, of which to date 14 have been found to be transcribed, according to RT-PCR experiments. Examining the orthologs of these predicted genes in strain PA14, and incorporating data regarding a GC score for codon usage, provided additional evidence that at least five of the novel transcribed genes (PA2173.1, PA0836.2, PA2566.1, PA2747.1 and PA2926.1) are likely to be truly essential in *P. aeruginosa*. These represent an additional list of possible antibacterial targets that could warrant further study.

An additional four are potentially essential (PA0457.1, PA1366.1, PA0613.1, and PA1152.1), but relatively unique and so supporting data of their essentiality from PA14 orthology data is not possible. They also warrant further study, however the lack of an ortholog in a closely related strain leads to some scepticism that they would be useful drug targets.

An additional five (PA1838.1, PA2118.1, PA3762.1, PA3414.1 and PA0050.1) have orthologs in PA14 that are not essential, but this is not surprising given the high specificity of the p value cutoff in identifying putative essential genes. Some predicted genes have GC-rich codon usage deviated from the norm, suggestive that they may not be true genes, however transcription was detected in these regions. This may be because the region is transcribed between genes, but there is no translated product from the transcript in this area of the sequence. Further evidence of a translated product is required.
Based on these results, it would appear that the method is identifying new genes that may be essential (with supporting evidence from the PA14 strain) however, incorporation of methods such as the GC score, to identify unusual codon usage in some open reading frames could be useful. Due to the small size of the genes being detected, traditional gene predictors would have trouble predicting them, however another option (instead of using a simplistic GC scoring system) would be to train a gene predictor using a smaller, more representative gene set that may increase sensitivity. Of course, any such scoring system needs to reflect the genome being analyzed.

Regardless, any genes predicted require more investigation in the laboratory to confirm whether they are indeed true genes. However, what is notable is that any genes were identified at all. It was certainly not clear if, based on the conservation of essential genes, that any novel essential genes would be predicted. We have indeed identified some promising candidates that appear to be essential (based on evidence involving multiple strains), are transcribed, and have a codon usage consistent with a true gene. The low number of candidates suggests, however, that the annotation of the *P. aeruginosa* PA01 genome is excellent.

Even if only a few genes are identified that are likely of real interest as potential novel essential genes, this is significant because this method could be potentially applied to a wide range of bacterial genomes. Any such genes identified represent new antimicrobial drug targets that will not likely have been examined to date.
CHAPTER 4 IDENTIFYING NOVEL PUTATIVE ESSENTIAL NON-CODING RNAS

4.1 Summary

In addition to protein-coding genes, as described above, I also hypothesize that some of the 336 putatively essential intergenic sequences (PEISs) in *P. aeruginosa* PAO1 extracted using the methodology described in Chapter 2 may harbour non-coding RNAs (ncRNAs). In this chapter I describe the methodology that was developed to identify such ncRNAs, and report my findings. Novel ncRNAs that were confirmed to be experimentally expressed were identified, however the analysis illustrated the difficulties that we still face when trying to predict novel ncRNA genes.

4.2 Materials and Methods

4.2.1 Searching for ncRNAs in known families in RFAM

As of October 2007, RFAM version 8.1 (Griffiths-Jones et al., 2005) contains 607 RNA families. Since the RFAM database does not support the remote submission of jobs, the database was downloaded to the local machine.

PEISs were queried against the RFAM database version 8.1 using the INFERNAL software version 0.72 (Nawrocki and Eddy, 2007). INFERNAL uses covariance models (CMs), similar to a sequence profile, which takes into account both sequence and RNA secondary structure. Compared to just sequence
search, using INFERNAL increases the accuracy of finding an RNA (Nawrocki and Eddy, 2007). Hits returned represent unannotated putative ncRNAs that can be classified into an ncRNA family.

4.2.2 Obtaining orthologous sequences using blastn

PEISs extracted using the methods described in Chapter 2 were submitted remotely via web services using the BioPerl RemoteBlast package, to perform a BLASTN search against the NCBI nucleotide ‘nr’ database. Since nucleotides in orthologous RNA sequences tend to have compensatory mutations, using a smaller word size during the BLAST search would increase sensitivity. Thus a word size of 7 was used instead of the default 11. Also, an e-value cutoff of 1E-3 was used, as well as a percentage identity cutoff of 70%. The percentage identity cutoff was used because benchmark studies have shown that in general, using sequences below a 50-60% identity resulted in a drastic performance drop in RNA sequence alignments (Gardner et al., 2005).

During the time of the project, a new P. aeruginosa Liverpool Epidemic strain (LES) was in the process of being sequenced by the Sanger Institute Pathogen Genomics group, and the sequence of the unfinished genome were being deposited on the Sanger website (http://www.sanger.ac.uk/Projects/P_aeruginosa/). With around 99% similarity to P. aeruginosa PAO1, this was an additional related genome which could provide useful comparative data. However, since sequences haven’t been deposited into the NCBI ‘nr’ database yet, the unfinished genome FASTA sequence was downloaded from the website, and converted into a local BLAST database using formatdb. Subsequent to the
remote BLASTN search against NCBI, a local BLASTN was performed against the *P. aeruginosa* LES genome to obtain additional orthologous sequences.

### 4.2.3 Filtering orthologous sequences for multiple sequence alignment

For each PEIS that acted as the query, a filtering step was then performed on returned BLAST hits in order to retain the best sequences for the multiple sequence alignment step. HSPs that were shorter than 60nt were discarded, as I opted to focus on an analysis involving larger sequences initially. In addition, HSPs that were 100% identical were also discarded since they had no useful comparative information. If a given BLAST hit contained multiple HSPs, then two approaches are used: if any given pair of HSPs is less than 150nt apart from each other then the middle missing sequence is retrieved, and the three are concatenated into one single final sequence; however if two HSPs are more than 150nt apart, then they are regarded as two separate final sequences. If there are more than 20 final sequences after the filtering step, only the first 20 sequences (ranked by ascending BLAST e-value) were taken as input to the multiple sequence alignment step, in order to impose an upper limit to the number of input sequences.

### 4.2.4 Multiple sequence alignment of orthologous sequences using MAFFT

The MAFFT software package version 6.240 (Katoh and Toh, 2008) was used to perform multiple sequence alignments on each putative essential intergenic sequence and its orthologous sequences returned by the BLASTN search.
The choice of MAFFT, as opposed to other multiple sequence alignment programs, was based on benchmark studies comparing the performance of different programs on RNA alignments (Wilm et al., 2006). The choice of a good multiple sequence alignment is important, since RNA prediction performance is highly dependent on the correct alignment of orthologous bases (or columns in a multiple sequence alignment).

Performance benchmark studies have been done for aligning sets of protein sequences (Thompson et al., 1999), however only recently have they been tested on RNA alignments (Gardner et al., 2005; Wilm et al., 2006). General findings include the observation that the "twilight zone" for RNA alignment is between 50% and 60% sequence identity (Gardner et al., 2005), meaning multiple sequence alignment programs perform badly as an RNA alignment tool for sets of sequences below this identity threshold.

Performance rankings of various multiple sequence programs were also compiled. In general, iterative approaches such as MUSCLE performed better than progressive approaches such as CLUSTALW; the iterative option of MAFFT also clearly outperforms its progressive option. MAFFT version 5.0 and up ranked consistently at the top throughout all test sets in the benchmark study (Wilm et al., 2006).

Based on these benchmark studies, I therefore chose MAFFT version 6.240 as the multiple sequence alignment program. The "L-INS-i" option was used, denoting the better-performing "iterative" approach, and a "local" multiple sequence alignment instead of a global one. The default 100 iterations was used.
The local alignment option was chosen because sequences returned from the BLAST search usually do not span the entire length of the query sequence, but align to only a portion of it. Therefore a local alignment would be more appropriate.

The ease and flexibility of MAFFT in specifying different alignment options, for instance whether to use an iterative/progressive method, or to perform a local/global alignment, was also a factor in choosing MAFFT as the multiple sequence alignment program.

This step produces a set of multiple sequence alignments, comprising the original putative essential intergenic sequence and its putative orthologs. Due to the requirement of QRNA to use pairwise alignments as inputs, each multiple sequence alignment was then broken down into their projected pairwise alignments, using the original query sequence as reference.

4.2.5 Non-coding RNA prediction software: RNAz and QRNA

Two RNA prediction software programs using complementary approaches were used to predict putative ncRNAs within the multiple sequence alignments: RNAz version 1.0 (Washietl et al., 2005), and QRNA version 2.0.3c (Rivas and Eddy, 2001).

RNAz uses thermodynamic stability and structure conservation to predict ncRNAs within a multiple sequence alignment. It first calculates a consensus secondary structure and a consensus minimum free energy (MFE) from the set of sequences. A z-score is then obtained by comparing the thermodynamic
stability of the native fold with a large number of random sequences; the smaller
the z-score, the more stable it is than expected by chance. A structure
conservation index (SCI) is also calculated by comparing the consensus MFE to
the average MFE of individual sequences; the higher the SCI, the more closely
the individual folds resemble the consensus structure.

QRNA on the other hand uses a complementary approach to RNAz. It
uses Stochastic Context-Free Grammars (SCFGs) to test patterns of
substitutions in a pairwise alignment, distinguishing between synonymous
substitutions typical in coding sequences, from compensatory mutations typical in
a conserved RNA structure (e.g. a G-C base pair to a A-U base pair).

Both RNAz and QRNA have previously been used in genome-wide
screens. RNAz has been used predict thousands of functional ncRNAs within the
human genome (Washietl et al., 2005), and also within the ENCODE regions
(Washietl et al., 2007). QRNA was used to successfully identify 11 expressed
ncRNAs in E. coli during a genomic screen (Rivas et al., 2001).

Multiple sequence alignments were sliced into windows of size 120bp,
with a sliding offset of 20bp, and input into both programs.

4.2.6 Estimating false discovery rate

To control for false positives in the predicted ncRNAs, a multiple sequence
alignment shuffling approach was employed. The concept is in attempting to
destroy the signal arising from RNA structure, but keeping other sequence
properties intact, for instance GC content, proportion of gaps, and the proportion of conserved columns.

A mono-nucleotide shuffling procedure was used to generate artificial negative alignments. A perl script implementing this procedure (shuffle-aln.pl), part of supplementary data from AlifoldZ, was downloaded from the following website (http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/Alifoldz/).

For each of RNAz and QRNA, two different runs were performed: a native run, and a random run. In the native run, the original set of multiple sequence alignments were used as input to the programs. In the shuffled run, shuffled multiple sequence alignments were first obtained using shuffle-aln.pl, and were then used as input to both programs for an additional run.

The distribution of scores of all windows in the shuffled run were first plotted, representing the background score distribution. Score distributions for the native run was then overlayed in the same graph. A score cutoff was then determined by visual inspection with the desired false discovery rate.

Graphs were automatically generated using R statistical package called within the Perl script.

Coloured RNA alignments and RNA secondary structure drawings were generated using RNAz.

4.2.7 Test data set

To test the performance of ncRNA prediction using the approach above, a test data set was generated, using the same test data set as in Chapter 3, but
now with protein-coding genes as negative dataset, and the ncRNAs as the positive dataset.

The 14 protein-coding genes comprising the negative data set for the ncRNA prediction step were as follows: PA0408 (pilG, 408bp), PA0610 (prtN, 315bp), PA0970 (tolR, 441bp), PA1456 (cheY, 375bp), PA1705 (pcrG, 297bp), PA1715 (pscB, 423bp), PA1721 (pscH, 432bp), PA1722 (pscl, 339bp), PA2960 (pilZ, 357bp), PA3807 (ndk, 432bp), PA4230 (pchB, 306bp), PA4232 (ssb, 498bp), PA4386 (groES, 294bp) and PA4847 (accB, 471bp).

The positive data set comprised of five annotated non-coding RNAs, three of which were experimentally verified sRNAs (PrrF1, PrrF2 and PA5181.1), in addition to one 16S rRNA (PA5369.5) as well as one 23S rRNA (PA0668.4). These were chosen due to their sizes, such that their removal from the annotations would result in a PEIS, as described earlier.

All 14 genes in the positive data set and 6 in the negative data set were removed from annotations at the outset.

4.2.8 Laboratory verification methods for the presence of ncRNAs

Lab verification of candidate ncRNAs was performed by David Mulder. RNA extraction was done using the same methodology as described in section 3.2.8. Northern blot analysis was performed similar to as was described previously (Brinkman et al., 1999) but using an electroblotter and oligo end labeling. Briefly, total RNA was separated on a denaturing agarose-formaldehyde gel in formamide loading buffer and transferred to positively charged membrane
(Roche) by electroblot transfer (Bio-Rad). RNA was UV cross-linked using auto settings (Stratagene). The membrane was hybridized with complementary 30bp DNA oligo probes 3’-end labeled with DIG using terminal transferase (Roche) overnight at 40°C in DIG Hyb Buffer (Roche). The signal was detected using an ELISA based luminescent reaction (Roche).

4.3 Results and Discussion

4.3.1 Novel ncRNAs with known RNA families

Searching all 336 putative essential intergenic sequences against RFAM, a total of eight significant hits were returned, of which 3 were in the positive dataset: PrrF1, PrrF2 and the 23S rRNA. Notably, the remaining two sRNAs in the positive dataset had no hits and apparently haven’t yet been classified into any RNA family in RFAM. This illustrates the need for additional complementary methods for ncRNA detection, as of course was suspected (and desired, given the interest in detecting genuinely new ncRNAs).

In addition to the detection of the above mentioned positive controls, five RFAM hits were detected in this 336 intergenic sequences that had not yet been annotated in the P. aeruginosa PA01 genome, and thus represent novel putative ncRNAs. These include three riboswitches (cobalamin, FMN and TPP), one yybP-ykoY leader, and one bacterial RNase P Class A. Below is a discussion of each.
**PA1270.1 (cobalamin riboswitch)** – Cobalamin riboswitches are cis-regulatory RNA elements which have a metabolite binding domain, in this case for cobalamin. They are typically located in the 5' untranslated region of genes related to vitamin B$_{12}$ biosynthesis. In this case, PA1270.1 presumably regulates the tonB-dependent vitamin B$_{12}$ receptor directly downstream. Upon ligand binding, the riboswitch undergoes an allosteric structural rearrangement, which prevents ribosome binding and thus translation of the regulated gene. This novel cis-regulatory ncRNA was found to be expressed, according to Northern analysis, together with PA1271 as a single transcript.

![Diagram](image)

**Figure 4-1:** Schematic diagram of the genomic context of PA1270.1 (putative cobalamin riboswitch), a previously unannotated and putatively essential ncRNA, denoted by the green arrow. Blue arrows denote previously annotated protein-coding genes, and a black triangle denotes the presence of a transposon mutant in that location. The direction of the arrow denotes the direction of transcription.
PA2910.1 (yybP-ykoY leader) – This small putative 117nt cis-regulatory ncRNA is typically found upstream of related families of protein-coding genes in many bacteria, though the function of these genes is unknown. It has been suggested that this ncRNA element acts as a riboswitch (Barrick et al., 2004).

Figure 4-2: Schematic diagram of the genomic context of PA2910.1 (putative yybP-ykoY leader), a previously unannotated and putatively essential ncRNA, denoted by the green arrow.
PA4055.1 (FMN riboswitch) – This predicted riboswitch lies upstream of ribC, a gene within a 4-gene cluster ribEBCD involved in riboflavin biosynthesis. Typically, the FMN riboswitch is a cis-regulatory ncRNA found in 5' untranslated regions of prokaryotic genes involved in flavin mononucleotide (FMN) biosynthesis. It binds the FMN metabolite and controls the gene expression of the downstream genes by an allosteric structural rearrangement causing premature transcription termination (Winkler et al., 2002). This unannotated putative riboswitch was found to be expressed, according to Northern analysis.

![Schematic diagram of the genomic context of PA4055.1 (putative FMN riboswitch), a previously unannotated and putatively essential ncRNA, denoted by the green arrow.](image-url)
PA4421.2 (Bacterial RNase P Class A) – This unannotated ncRNA partially overlaps an annotated ncRNA (PA4421.1) in the opposite strand, which is also described as the RNA component of RNase P. However, comparing with the sequence and secondary structure published in the Ribonuclease P database (Brown et al., 1994), it seems that the newly predicted PA4421.2 had the more accurate coordinates. RNase P is a ubiquitous endoribonuclease, distributed phylogenetically in archaea, bacteria and eukaryotes, and acts as a ribozyme, an RNA molecule with catalytic function (for review, see (Bartel and Unrau, 1999)). It functions by cleaving the 5'-leaders of precursor-tRNAs, generating mature 5'-ends of tRNAs (Frank and Pace, 1998). In bacteria, RNase P consists of an RNA component (M1 RNA), as well as a polypeptide chain (C5 protein) whose function is to increase the binding affinity of the M1 RNA to the substrate. Located within a transposon gap of 395bp, it is predicted to be essential.
Figure 4-4: Schematic diagram of the genomic context of PA4421.1 (putative RNase P), a previously unannotated and putatively essential ncRNA, denoted by the green dashed arrow. The green solid arrow denotes an annotated ncRNA in the *P. aeruginosa* PA01 genome.

**PA4973.1 (TPP riboswitch)** – This predicted TPP riboswitch is located upstream of thiC, a thiamine biosynthesis protein. TPP riboswitches bind directly to thiamine pyrophosphate (TPP), and regulates the expression of the downstream gene (Winkler et al., 2002). It is phylogenetically widely distributed, being found in archaea, bacteria and even eukaryotes. This novel unannotated riboswitch was found to be expressed, according to Northern analysis, in *P. aeruginosa* PA01, together with thiC in one single transcript.
4.3.2 Blastn and multiple sequence alignment

Of the 336 PEISs that were submitted to BLASTN, only four did not return any hits, while the remaining sequences returned at least one hit. After the filtering and HSP-concatenating step (described in Materials and Methods in section 4.2.3), a total of 328 orthologous sequence sets (comprising the PEIS and at least one other orthologous sequence) were retained.

Of the 328 orthologous sequence sets, 244 sets (74.4%) had less than 10 member sequences, 34 sets (10.4%) had between 10 and 20 members, 35 sets
(10.7%) between 20 and 30 members, and 15 sets (4.6%) had more than 30 member orthologous sequences (Figure 4-6).

The fifty sets that had more than 20 members were slimmed down to 20 members (retaining the first 20). After this filtering step, multiple sequence alignments were computed for all 328 orthologous sequence sets, yielding 328 sets of “native” multiple sequence alignments. Each multiple sequence alignment is also shuffled, yielding an additional 328 sets of “shuffled” multiple sequence alignments.

![Figure 4-6: Distribution of number of members in orthologous sequence sets](image)

Figure 4-6: Distribution of number of members in orthologous sequence sets
4.3.3 RNAz ncRNA predictions

RNAz was run on each of the 328 native and 328 shuffled multiple sequence alignments. If the alignment is more than 120 columns long, it is sliced into sliding windows of 120 columns each, and sliding by 20 columns each time. Each window was then fed to RNAz individually. As such, a total of 9520 windows in the set of native multiple alignments were analyzed by RNAz, while a total of 9755 windows in the set of shuffled multiple alignments were analyzed. The reason for the difference is due to the filtering of low-quality alignments by RNAz itself. For instance, sequences with more than 25% gaps are discarded, and the whole window will be discarded if only the reference sequence (the PEIS) is left.

Figure 4-7 illustrates the distribution of the four different RNAz scores obtained for all windows in native and shuffled multiple sequence alignments. In general, the consensus minimum free energy (MFE) does not have a large discrepancy between native and shuffled windows, though it can be observed that for MFES below -50, there are consistently more native windows than shuffled windows in any given bin, suggesting more stable secondary structures in native windows. Moreover, structures with MFE below -65 are almost exclusively dominated by native windows.

The structure conservation index (SCI) measures the average deviation of the structure predicted for each individual sequence in the window with respect to the consensus structure. No observed discrepancy can be observed between native and shuffled windows, and a large number of windows have a SCI of 1.0.
This likely reflects the fact that many sequences have a very high sequence identity, as evident by the presence of sequences in extremely related *P. aeruginosa* strains such as PA14, PA7 and the newly sequenced LES strain.
RNAz consensus MFE (native vs shuffled)

![Graph A](image)

RNAz SVM classification score (native vs shuffled)

![Graph B](image)
Figure 4-7: Distribution of windows analyzed by RNAz in native vs. shuffled multiple sequence alignments in terms of of (A) Consensus minimum free energy (MFE); (B) SVM classification score; (C) Structure Conservation Index (SCI); and (D) z-score.
The SVM classification score seems to be a more discriminatory feature between native and shuffled windows. For SVM classification scores above 0, which corresponds to the RNAz P-value > 0.5, the false discovery rate drops to less than 50%. For SVM classification scores greater than 2, corresponding to RNAz P-value > 0.9, the false discovery rate drops to less than 10%.

Similarly, the distribution of z-scores for native windows is noticeably shifted towards more negative values than shuffled windows. Z-score measures how stable a structure is compared to what is expected by chance, and thus is related to consensus MFE scores. For z-scores < -2.0, the false discovery rate falls to less than 50%, and for z-scores < -3.0, the false discovery rate drops to less than 10%.

Windows that were scored with a P-value < 0.9 were discarded, then any overlapping windows were then clustered into “loci”. MFE scores, SCI scores, and z-scores were averaged out to produce a single average MFE, SCI and z-score for a given loci. This resulted in a total of 223 predicted loci with an average P-value greater than 9.

The 223 predicted loci have a mean consensus MFE of -42.06 (standard deviation of 11.53), and a mean z-score of -2.60 (standard deviation of 0.64), suggesting that most predicted loci do generally have a more stable secondary structure than expected by chance. The mean P-value for all predicted loci is 0.966 (standard deviation of 0.02).
4.3.4 QRNA ncRNA predictions

328 native and 328 shuffled multiple sequence alignments were first broken into pairwise alignments, then sliced into sliding windows of 120 columns, with an offset of 20. This produced 55,394 native windows and 55,394 shuffled windows. QRNA was then run for each window.

Figure 4-8 illustrates the distribution of scores of windows in native multiple sequence alignments compared to shuffled multiple sequence alignments. “Raw” RNA scores are the scores generated under the RNA model. The “log odds” RNA score calculates the log-odds posterior probability of the alignment being generated under the RNA model as opposed to the OTH (i.e. no specific substitution pattern) model. The “sigmoidal” RNA score is the score calculated by taking both OTH and COD (coding model) as null models. An alignment will be labelled as “RNA” if the sigmoidal RNA score is greater than both the OTH and COD sigmoidal scores.

Looking at the general distribution of sigmoidal RNA scores for all native windows compared to shuffled windows in Figure 4-8 C, there seems to be little discrepancy between the two. Nonetheless, particularly large sigmoidal RNA scores (scores > 60) are typically dominated by native windows, suggesting that they could harbour regions likely to contain a detectable RNA secondary structure.
Figure 4-8: Distribution of scores of native vs. shuffled windows, in terms of (A) raw RNA score; (B) log-odds RNA score; and (C) sigmoidal RNA score. (D) is the zoomed version of C.
All native windows labelled as either "OTH" or "COD" were discarded. Out of the 55,394 windows, 39,585 were discarded, and the remaining 15,809 labelled as "RNA" were retained. Since all windows were pairwise alignments, equivalent windows were combined, and the sigmoidal RNA scores averaged across the equivalent windows.

Windows which overlapped were then clustered into "loci", and sigmoidal RNA scores averaged to produce a final mean sigmoidal score for a given loci. The maximum sigmoidal score over all windows in the loci was also retained. As such, 364 RNA loci was predicted by QRNA.

Over all 364 predicted loci, the mean sigmoidal RNA score was 5.74 (standard deviation 8.06), and the mean value of the maximum sigmoidal RNA score was 9.78 (standard deviation 14.3).

4.3.5 RNAz/QRNA overlapping predictions reduces false positive rate

RNAz and QRNA represent different but complementary approaches to ncRNA prediction; RNAz primarily bases predictions on how thermodynamically stable a structure is than expected by chance, while QRNA makes predictions by finding patterns of substitution most likely caused by RNA secondary structure as opposed to a coding sequence or a position-independent sequence. I propose that analyzing the overlap of their RNA predictions will thus identify predictions that possess both qualities of RNA secondary structure, which could be used to filter out false positives.
Out of the total 223 loci predicted by RNAz, and 364 loci predicted by QRNA, a total of 169 loci overlap (here defined as overlapping by more than 60nt). QRNA predicted 195 unique loci not predicted by RNAz (54% of total QRNA predictions), while RNAz predicted 54 unique loci not predicted by QRNA (24.2% of total RNAz predictions). This indicates that a larger proportion of total RNAz predictions are also predicted by QRNA than vice-versa, suggesting that QRNA predictions are likely to harbour a large proportion of false positives, as can also be observed in the lack of discrepancy between sigmoidal RNA score distributions in native vs. shuffled windows except for extremely large values. This could probably be explained by the very high sequence identity observed in some alignments (typically 90-100% for sequences between P. aeruginosa PAO1, PA14, PA7 and LES strains), providing very few patterns of substitutions that can be confidently distinguished between the various QRNA sequence models. For these cases of extreme sequence identity, the measure of thermodynamic stability would be a more useful distinguishing feature, as evident by the shift of the distribution of RNAz z-scores of native windows towards more negative values relative to shuffled windows.

To analyze whether overlapping predictions are more likely to harbour RNA structures, RNAz and QRNA scores in overlapping predictions were compared to scores in non-overlapping predictions. In terms of thermodynamic stability, the mean consensus MFE over the 169 overlapping loci is -42.91, while the mean consensus MFE over the 54 unique loci predicted by RNAz is -39.37, a statistically significant difference (P-value < 0.0494 using an unpaired t-test). The
mean z-score over the 169 overlapping loci is -2.625, while the mean z-score over the 54 unique loci predicted by RNAz is slightly greater, with a value of -2.532 (P-value of 0.3551). Out of the 169 overlapping loci, 43 had a z-score less than -3. The mean consensus MFE over these 43 predictions is -46.54 (P=0.003), while the mean z-score is -3.53 (P<0.0001). These results suggest that overlapping RNAz/QRNA predictions are more likely to harbour more thermodynamically stable RNA structures.

For loci predicted by QRNA, the mean sigmoidal RNA score over the 195 unique QRNA predicted loci is 8.134, while the mean sigmoidal RNA score over the 169 overlapping QRNA/RNAz loci is 11.675, a statistically significant increase (P=0.0183), suggesting that overlapping loci contain a statistically higher proportion of sequences containing patterns of substitutions typical of RNAs. If only predictions with a sigmoidal RNA score > 10 in the overlapping set is retained, 57 loci satisfy this criteria, with an average sigmoidal RNA score of 27.81, which is highly statistically significant (P<0.0001).

4.3.6 Test results

To measure the accuracy of RNA predictions by QRNA and RNAz, I cross-referenced my results with the 5 ncRNAs in the positive dataset, and the 14 protein-coding genes in the negative dataset.

Out of the five positives, RNAz was able to recover all of them, even though specific start and end coordinates were not entirely the same, resulting in discrepancies in the number of ncRNAs that would be predicted. PrrF1 and
PrrF2, which are adjacent to each other, were predicted as one single locus by RNAz instead of two separate loci. The 23S rRNA was predicted as five loci, the 16S rRNA as three loci, and the PA5181.1 sRNA was predicted as two loci. Out of the 14 negatives, RNAz falsely made predictions for 6 of them, indicating that this method had a notable 43% false positive rate.

QRNA also recovered 100% of all positives. PrrF1 and PrrF2 again were predicted as one single loci, the 16S rRNA as five loci, and the PA5181.1 sRNA as two loci. Nonetheless, only approximately 150nt of the 23S rRNA was predicted. Out of the negatives, QRNA falsely made predictions for 4 of them, giving a slightly lower false positive rate of 28.6%.

The false positive rate was lowered when the "overlapping" RNAz/QRNA predicted loci were considered. While all true positives rate remained at 100%, only two negatives were falsely predicted by both QRNA and RNAz, yielding a false negative rate of 14.3%. This supports my hypothesis that combining predictions from RNAz and QRNA lowers the false positive rate while keeping the true positive rate constant.

Of the 5 ncRNA predictions that returned significant hits in RFAM, RNAz predicted 4 of them, while QRNA predicted all of them; 4 out of 5 were thus predicted by both RNAz and QRNA. Only the cobalamin riboswitch was uniquely predicted by QRNA but not by RNAz. If we considered the three transcriptionally verified riboswitches as additional positives in addition to the original five ncRNAs in the positive dataset, this gives an overall true positive rate of 87.5% (5 + 2 / 8).
This again suggests that combining RNAz and QRNA predictions can effectively lower the false positive rate while retaining a high true positive rate.

After the removal of positive and negative controls, loci coinciding with protein-coding gene predictions, and loci coinciding with predictions via RFAM, a total of 385 loci were predicted as RNA. Of these, 217 were uniquely predicted by one of RNAz or QRNA (46 unique RNAz loci and 171 unique QRNA loci), while the remaining 148 were predicted by both QRNA and RNAz (the "overlapping" set).

4.3.7 Filtering non-ncRNA sequences

We expect that many loci predicted either by RNAz or QRNA will contain sequences with stable or conserved RNA secondary structures, but which is not of interest for the current study. These include rho-independent terminators, rRNA spacers and certain repetitive elements. I performed a filtering step to filter out predicted loci that coincided with any of these elements. Since we expect the loci which were uniquely predicted by either RNAz or QRNA to contain a large proportion of false positives, this filtering step was performed only on the "overlapping" set.

The list of rho-independent terminators for *P. aeruginosa* PAO1 were obtained from the Comprehensive Microbial Resource of the J. Craig Venter Institute (http://cmr.jcvi.org). Of the 148 loci in the "overlapping" set, 19 coincided with a rho-independent terminator. However, three of these deserved special attention. For the terminator located between genes PA0611 and PA0612, both
flanking genes are transcribed in the direction away from the terminator. Thus if the terminator is functional, then this region might encode a novel transcript. The terminator located between genes PA0806 and PA0807, and the terminator between PA5429 and PA5430, are very far away (239bp and 220bp respectively) from the upstream gene of the same direction (PA0806 and PA5429 respectively), suggesting there is a possibility that these regions are expressed. These three special cases were retained, while the remaining 16 were removed, leaving 132 candidate loci.

We also filtered out possible RNA structures arising from rRNA spacers. Five loci fell into this class. One of these is a 16S-23S rRNA spacer, commonly known as an internal transcribed spacer (ITS), located within the rRNA cluster of PA0668.1-PA0668.5. Three loci were directly upstream of an rRNA cluster, and one was directly downstream of an rRNA cluster, presumably control regions or transcribed spacers containing RNA secondary structure. These five loci were also removed, leaving 127 candidate loci.

Candidate loci were also filtered for repeated regions such as clustered regularly interspaced short palindromic repeats (CRISPRs) (Grissa et al., 2007), and tandem repeats from the Tandem Repeats Database (Denœud and Vergnaud, 2004). No CRISPR was found, but seven loci coincided with tandem repeat regions. These 7 loci were removed, leaving a final count of 120 candidate loci.
4.3.8 Characterization of selected ncRNA candidates

It is likely that some of these 120 candidate ncRNA loci will contain *bona fide* ncRNAs. Indeed, all five ncRNAs in the positive dataset, as well as 4 out of the 5 loci coinciding with predictions via RFAM, would be within this set before their removal. Moreover, as noted above, 7 out of the 28 negatives would be within this set.

In order to obtain the most significant ncRNA candidates from the 120 candidate loci, I have sorted the 120 candidate loci by ascending mean RNAz z-score and descending mean QRNA sigmoidal RNA score, and have chosen the top 5 candidates for characterization. All of these had a mean RNAz z-score < -3, and a mean QRNA sigmoidal RNA score > 10.
PA0844.1 – Located within an intergenic region of 341bp (p=0.19) that does not contain any transposon insertion, this candidate ncRNA has a predicted length of 199nt. RNAz scored this loci over 5 windows averaging a z-score of -4.331, which is statistically significant. The RNA signal is stronger in the reverse direction. QRNA also predicted this loci as RNA with an average sigmoidal RNA score of 13.45 in the native alignment, while the shuffled alignment was labelled as OTH. It is located in the upstream region of a phospholipase C (heat-labile hemolysin) precursor involved in phosphate regulation. The transcription start site of this gene was determined previously to be just ~25 bases upstream of the gene start codon (Pritchard and Vasil, 1986). This candidate ncRNA is therefore unlikely to be transcribed in the same transcript as the hemolysin gene. The orthologous region of this candidate ncRNA in P. aeruginosa PA14 also does not contain any transposon insertion, suggesting it is likely to be essential. Phylogenetically, this nucleotide sequence seems only to be conserved in P. aeruginosa PA14 and PA7.
Figure 4-9: Multiple sequence alignment and predicted secondary structure of PA0844.1, a novel putative ncRNA. Alignment columns and secondary structure base-pairs are color-coded according to RNAz conventions (Washietl et al., 2005). Yellow indicates that the base pair is supported by 2 types of pairs. Red indicates that only a single pair-type supports the interaction. The intensity of the color is proportional to the number of sequences having that base pair.
PA1587.1 – For the entire 328bp intergenic sequence between genes lpdG (lipoamide dehydrogenase-glc) and sucC (succinyl-CoA synthetase beta chain), RNAz predicted the presence of RNA secondary structure throughout the entire IGR, while QRNA made prediction between 180 and 324 bases from the end of lpdG. A closer look at the BLASTN report revealed orthology of the first 150 bases of the IGR to the 3'-untranslated region (UTR) of the lpd gene in P. fluorescens (Benen et al., 1989), which is annotated with three terminator sequences (Genbank accession M28356). These three terminator sequences correspond to positions 8-34, 62-84, and 113-149 of the P. aeruginosa IGR, which can also be seen in Figure 4-10.

The PA1587.1 candidate ncRNA (Figure 4-11) with predicted size of 145nt is located between positions 180 to the end of the IGR, with a mean RNAz z-score of -3.27 and mean sigmoidal RNA score 15.78. It has a very stable predicted structure of -73.21, which is rare for shuffled alignments (see Figure 4-8 A), and contains four locations of covarying base pairs. RNA signal is strongest in the forward direction, consistent with the transcribed direction of both flanking genes. The orthologous region in P. aeruginosa PA14 is also flanked by the corresponding orthologs of lpdG and sucC, and does not contain any transposon insertion, suggesting this region is likely to be essential for P. aeruginosa.

Phylogenetically, this ncRNA candidate is conserved in all strains of P. aeruginosa, however only the region between positions 254-321 are conserved in other Pseudomonas species such as P. mendocina, P. entomophila and
various strains of *P. putida*. Given the proximity of this candidate ncRNA to the 5’end of the sucC gene, it is possible that it is a *cis*-regulatory.
Figure 4-10: Multiple sequence alignment and predicted secondary structure of 3 terminator structures as annotated by (Benen et al., 1989), just upstream of PA1587.1, a novel putative ncRNA. Yellow indicates that the base pair is supported by 2 types of pairs. Red indicates that only a single pair-type supports the interaction. The intensity of the color is proportional to the number of sequences having that base pair.
Figure 4-11: Multiple sequence alignment and predicted secondary structure of PA1587.1, a novel putative ncRNA. Alignment columns and secondary structure base-pairs are color-coded according to RNAz conventions (Washietl et al., 2005). Yellow indicates that the base pair is supported by 2 types of pairs. Red indicates that only a single pair-type supports the interaction. The intensity of the color is proportional to the number of sequences having that base pair.
**PA1655.1** - This predicted ncRNA candidate of approximately 120nt is located within a transposon gap of 391bp in an IGR of size 1171bp, the seventh largest IGR in the *P. aeruginosa* PA01 genome. It is located far away from any flanking gene: 564bp from the end of PA1655 encoding a probable glutathione S-transferase, and 491bp from the start of PA1656, a conserved hypothetical protein. RNAz scored this candidate locus with a mean z-score of -3.26, and QRNA with a sigmoidal RNA score of 10.75. The orthologous sequence in *P. aeruginosa* PA14 lacks any transposon insertion, and in fact is located within a large transposon gap of 650bp. Phylogenetically, it is conserved in *P. aeruginosa* PA14 and LES strains, however only the long hairpin loop is conserved in *P. fluorescens* PfO-1 and *P. mendocina*. 
Figure 4-12: Multiple sequence alignment and predicted secondary structure of PA1655.1, a novel putative ncRNA. Alignment columns and secondary structure base-pairs are color-coded according to RNAz conventions (Washietl et al., 2005). Yellow indicates that the base pair is supported by 2 types of pairs. Red indicates that only a single pair-type supports the interaction. The intensity of the color is proportional to the number of sequences having that base pair.
PA2291.1 – This predicted ncRNA candidate is approximately 140nt long, located within a 341bp IGR with no transposon insertions. It is predicted to be in the same direction as and just 20 bases upstream of PA2291 encoding a probable glucose-sensitive porin, and could perhaps be a regulatory structure. RNAz predicted this loci with a z-score of -3.33, and QRNA with a sigmoidal RNA score of 10.25, with multiple sites of covariation. The orthologous sequence in *P. aeruginosa* PA14 does not contain any transposon insertion, and is also located upstream of the corresponding glucose-sensitive porin ortholog. Phylogenetically, this ncRNA candidate is conserved only within *P. aeruginosa*, namely PA14, PA7 and LES strains.
Figure 4-13: Multiple sequence alignment and predicted secondary structure of PA2291.1, a novel putative ncRNA. Alignment columns and secondary structure base-pairs are color-coded according to RNAz conventions (Washietl et al., 2005). Yellow indicates that the base pair is supported by 2 types of pairs. Red indicates that only a single pair-type supports the interaction. The intensity of the color is proportional to the number of sequences having that base pair.
PA3919.1 – This predicted ncRNA candidate is approximately 120nt long, located within a relatively large transposon gap of 478bp (p=0.10), with a mean RNAz z-score of -3.07 and QRNA sigmoidal RNA score of 25.54. RNA signal is stronger in the minus strand, consistent with the transcription direction of the most proximal gene; it is located just 60nt from the start of PA3919, encoding a conserved hypothetical protein. Further downstream lies the moaCDE operon, encoding genes involved in molybdenum cofactor biosynthesis. Phylogenetically, it is widely distributed within the Pseudomonas genus. In P. aeruginosa PA14, the orthologous sequence does not contain any transposon insertion.
Figure 4-14: Multiple sequence alignment and predicted secondary structure of PA3913.1, a novel putative ncRNA. Alignment columns and secondary structure base-pairs are color-coded according to RNAz conventions (Washietl et al., 2005). Yellow indicates that the base pair is supported by 2 types of pairs. Red indicates that only a single pair-type supports the interaction. The intensity of the color is proportional to the number of sequences having that base pair.
4.3.9 Limitations of novel ncRNA prediction

Computationally identifying ncRNAs still remains a difficult and open problem in computational biology (Aksay et al., 2007), as ncRNAs lack many signals and constraints such as clear start and stop sites like in protein-coding genes. As such, it is therefore very difficult to determine the exact start and end of a predicted ncRNA. Though the exact positions are hard to determine computationally, using a smaller offset in the sliding window would aid in narrowing down the range.

Furthermore, it is also difficult to predict the exact strand of the ncRNA, as signals arising from RNA secondary structure would in many cases be similar in either strand (a CAG-CUG stacked pair remains the same stacked pair in either strand), except for cases where G-U wobble pairs are present (where C-A base pair would not be possible in the complementary strand). Analyzing the ncRNA in its genomic context (e.g. proximity to neighboring genes) could aid in determining the correct strand, in particular cis-regulatory ncRNAs.

In this study, both RNAz and QRNA were used to predict ncRNAs based solely on the assumption of the presence of a stable and conserved RNA secondary structure. Therefore, any non-ncRNA genomic feature that contains a conserved secondary structure are potentially predicted, and indeed were, such as rRNA spacers, terminators and insertion elements. Furthermore, we have not distinguished between different types of ncRNAs, such as cis-regulatory RNAs, such as riboswitches which are transcribed together with its target mRNA, from small RNAs (sRNAs), which express their own transcript.
A comparative genomics approach was used in this study to predict ncRNAs, relying on sequence similarity (BLAST) as a preliminary step to find potential orthologous sequences. This is a limitation, because orthologous ncRNAs tend to conserve in secondary structure but the primary sequence could be very dissimilar due to synonymous substitutions. We attempted to overcome this problem in part by using a smaller word size for the BLAST search, however more dissimilar sequences would still be missed. NcRNAs which are not phylogenetically distributed will also not be detected by the comparative genomics approach.

It was observed that long ncRNAs were generally predicted as multiple loci. This could due to the use of a sliding window. Long ncRNAs are sometimes composed of multiple stable substructures (e.g. stem loops), which can be detected if the window is long enough to encompass it, however when the sliding window spans portions of separate substructures, no significant structure signal will then be detected. If these “transition regions” are long enough to span multiple windows, then multiple loci will be predicted. A potential solution is to use a sliding window of larger size, however the disadvantage is that smaller ncRNAs would then yield a weaker signal.

4.4 Concluding remarks

In this chapter I have described methodology to identify non-coding RNAs within intergenic sequences that may encode an essential gene. I have identified five unannotated ncRNAs in *P. aeruginosa* PAO1 belonging to known RNA families, and so are the most likely candidates for being true novel essential
ncRNAs. I also used two RNA prediction programs, RNAz and QRNA, coupled with additional filters for probable false positives, to predict 120 sequences that have features of candidate ncRNA loci. Due to a notable false positive rate (though the false negative rate was reasonable) the most significant five of these candidates were described. Combining predictions from both RNAz and QRNA programs reduced the rate of false positives (increased specificity), according to analyses of known test datasets, while maintaining the rate of true positives detected (sensitivity) observed for each program separately. Similar to the analysis of protein-coding genes, cross-referencing ncRNA candidates in *P. aeruginosa* PAO1 with the transposon mutant library in *P. aeruginosa* PA14 permitted more confidence regarding the likelihood that these putative regions were essential.. However, notably, for those ncRNA candidates examined further that were predicted by RNAz and QRNA, the highest scoring candidates were all limited to the *Pseudomonas* lineage. A filter for those that are found in diverse species may be warranted, both to reduce false positives in terms of true ncRNA gene prediction, as well as to increase the likelihood that the gene is truly essential (based on the analysis of essential genes in Chapter 5).

Three riboswitches have so far been transcriptionally tested by Northern analysis in the laboratory and confirmed to be expressed. Transcriptional verification of other ncRNA candidates are ongoing. The fact that any new ncRNAs were identified though is notable. Regardless, the accurate prediction and functional characterization of ncRNAs remains a difficult problem which warrants further study.
CHAPTER 5  PHYLOGENETIC DISTRIBUTION OF ESSENTIAL GENES

5.1 Summary

Using *P. aeruginosa* PAO1 as the test organism, I performed an analysis on the phylogenetic distribution of essential protein-coding genes and non-coding RNAs that had already been annotated in the PAO1 genome, with the hypothesis that essential genes are phylogenetically distributed across more diverse taxa than non-essential ones. Statistical observations provide evidence supporting this hypothesis, and suggest that a comparative genomics approach is appropriate for identifying essential protein-coding genes and ncRNAs. I also performed analysis on a region in the *P. aeruginosa* PAO1 genome with a notably low transposon insertion rate, and suggested possible explanations in light of the phylogenetic distribution of genes in that region.

5.2 Materials and Methods

Gene annotations for *P. aeruginosa* PAO1 were obtained from the *Pseudomonas* Genome Database (http://www.pseudomonas.com), as of January 10th, 2008.

All annotated protein-coding genes were extracted that were larger than 1kb. I primarily focused on larger genes because they are more likely to be true genes, and shorter genes suffer from false positives from gene prediction programs which may contaminate the data set.
This dataset has been divided into two subsets, one as "putatively essential", the other "non-essential". "Putatively essential" genes are here defined as those genes absent of any transposon insertion for its entire length, whereas "non-essential" genes are defined as those genes with at least one transposon insertion in the first half of the gene. Genes with insertions in the second half have been discarded in order to retain only the most likely candidates to have had a functional disruption.

5.2.1 Cross-referencing gene essentiality evidence in other organisms

In order to obtain added evidence that the putative essential protein-coding genes in *P. aeruginosa* PAO1 were truly essential, two sources of data were referenced as additional evidence of gene essentiality besides the *P. aeruginosa* PAO1 transposon mutant library: the *P. aeruginosa* PA14 transposon mutant library (Liberati et al., 2006), and the Profiling of the *E. coli* Chromosome database (Kato and Hashimoto, 2007).

*P. aeruginosa* PA14 orthologs, if any, were first obtained from http://www.pseudomonas.com for every essential gene in *P. aeruginosa* PAO1 from the previous step. These orthologs were likewise divided into two subsets, "essential" and "non-essential", using the *P. aeruginosa* PA14 transposon mutant library. "Essential" genes were those absent of any transposon insertion for its entire length, while "non-essential" genes were those with at least one transposon insertion in the first half of the gene. Genes that were essential in *P. aeruginosa* PAO1, whose ortholog in *P. aeruginosa* PA14 was also "essential", was thus more likely to be truly essential.
Two BLASTP searches were also performed for each essential gene in *P. aeruginosa* PA01, one against the “essential protein” dataset, the other against the “non-essential protein” dataset, in the PEC database, in order to determine whether orthologs in *Escherichia coli* K-12 were also essential. A stringent e-value cutoff of 1E-20 was used to define orthology. These results represent an additional source of evidence regarding the essentiality of *P. aeruginosa* PA01 genes.

### 5.2.2 Calculating the phylogenetic distribution of essential protein-coding genes

A BLASTP search was performed for each putatively essential gene in *P. aeruginosa* PA01 against the NCBI 'nr' database of proteins, with a stringent e-value cutoff of 1E-30 for precise detection of homologs. I also set the limit of the number of returned results from the default 100, to 1000, in order to obtain a much bigger sample of hits, which may aid in the analysis of the distribution of taxa associated with these BLAST hits.

Organism names were extracted from the returned BLAST hits, and were queried using NCBI Efetch utility for the organism lineage according to the NCBI Taxonomy database. The total number of Superkingdoms, Phyla, Classes, Orders, Families and Genera that were represented by the BLAST hits were counted and analyzed.
5.2.3 Calculating the phylogenetic distribution of essential non-coding RNAs

All annotated non-coding RNAs, including rRNAs, tRNAs, tmRNAs and sRNAs, were extracted from the *P. aeruginosa* PA01 genome annotation. These were filtered for the presence of transposon insertions, and those which were free of any transposon insertion for its entire length were labelled "essential ncRNAs".

A BLASTN search was performed for each essential ncRNA, against the NCBI 'nr' nucleotide database, using an e-value cutoff of 1E-3 and setting the limit of the number of returned results from the default 100 to 1000.

As with the protein-coding regions analysis, organism names were extracted from the returned BLAST hits, and were queried using NCBI Efetch utility for the organism lineage according to the NCBI Taxonomy database. The total number of Superkingdoms, Phyla, Classes, Orders, Families and Genera that were represented by the BLAST hits were counted and analyzed.

5.3 Results and Discussion

A total of 2190 protein-coding genes larger than 1kb in the *P. aeruginosa* PAO1 genome were extracted. Out of these, 155 were deemed "putatively essential", and 1086 "non-essential" based on the criteria described in Materials and Methods in section 5.2.1. The remaining 949 genes were discarded.
5.3.1 Gene essentiality in *P. aeruginosa* PA14 and *E. coli* K-12 orthologs

Figure 5-1 shows the results of cross-referencing the 155 putatively essential genes in *P. aeruginosa* PA01 with the essentiality status of corresponding orthologs in *P. aeruginosa* PA14 and *E. coli* K-12. Out of the 155 putatively essential genes, a total of 97 were also essential in *P. aeruginosa* PA14. Of these, 70 were also essential in *E. coli*, and 11 had no ortholog in *E. coli*. These 81 genes represent a set of genes most likely to be truly essential.

16 genes essential in both strains of *P. aeruginosa* were non-essential in *E. coli*. On the other hand, of the 155 putatively essential genes, 50 of them were non-essential in *P. aeruginosa* PA14, of which 39 were also non-essential or did not have any ortholog in *E. coli*, representing a set of genes likely to be non-essential. However, 11 were essential in *P. aeruginosa* PA01 and *E. coli* but not in *P. aeruginosa* PA14.

Cross-referencing gene essentiality studies in orthologs in other organisms provides a powerful means to increase the confidence of a given gene being truly essential.
5.3.2 Phylogenetic distribution of essential vs. non-essential protein-coding genes in *Pseudomonas aeruginosa* PA01

Table 5-1 shows the mean, median and maximum numbers of taxa represented by orthologs of essential genes compared to non-essential genes. Essential genes are on average represented in higher numbers of taxa than non-essential genes across all levels of taxa, with the former being almost double as taxonomically distributed as the latter for lower levels of taxa. The median shows an even more pronounced disparity, suggesting that essential genes are disproportionately more taxonomically distributed than non-essential genes. The
maximum number of taxa shows no significant difference between the two data sets, suggesting that certain non-essential genes are evolutionarily widely distributed.

Figures 5-2 through 5-7 show histograms comparing the phylogenetic distribution of essential genes vs. non-essential genes across various levels of taxa (Superkingdom, Phylum, Class, Order, Family and Genus). Consistently across all levels of taxa, a larger proportion of essential genes is represented in more numbers of taxa than non-essential genes. For instance, at the Superkingdom level, a larger proportion of non-essential genes is represented in only one Superkingdom (Bacteria), while essential genes have larger proportions represented in Archaea and Eukaryota as well. This disparity is more pronounced for lower levels of taxa, particular at the Order, Family and Genus levels.
Table 5-1: Mean, median and maximum number of taxa represented by orthologs of essential genes vs. non-essential genes.

<table>
<thead>
<tr>
<th></th>
<th>Superkingdoms</th>
<th>Phyla</th>
<th>Classes</th>
<th>Order</th>
<th>Families</th>
<th>Genera</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Essential</td>
<td>2.07</td>
<td>14.91</td>
<td>20.06</td>
<td>58.66</td>
<td>99.43</td>
<td>190.47</td>
</tr>
<tr>
<td>Non-essential</td>
<td>1.79</td>
<td>8.73</td>
<td>12.66</td>
<td>33.54</td>
<td>54.17</td>
<td>97.92</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Essential</td>
<td>2</td>
<td>16</td>
<td>17</td>
<td>63</td>
<td>110</td>
<td>223</td>
</tr>
<tr>
<td>Non-essential</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>22</td>
<td>37</td>
<td>65</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Essential</td>
<td>3</td>
<td>33</td>
<td>59</td>
<td>167</td>
<td>234</td>
<td>427</td>
</tr>
<tr>
<td>Non-essential</td>
<td>3</td>
<td>36</td>
<td>55</td>
<td>137</td>
<td>249</td>
<td>588</td>
</tr>
</tbody>
</table>

Figure 5-2: Histogram of the phylogenetic distribution of essential genes vs. non-essential genes at the Superkingdom level.
Figure 5-3: Histogram of the phylogenetic distribution of essential genes vs. non-essential genes at the Phylum level.

Figure 5-4: Histogram of the phylogenetic distribution of essential genes vs. non-essential genes at the Class level.
Figure 5-5: Histogram of the phylogenetic distribution of essential genes vs. non-essential genes at the Order level

Figure 5-6: Histogram of the phylogenetic distribution of essential genes vs. non-essential genes at the Family level
5.3.3 Essential genes in a region of low transposon insertion density

The taxonomic distribution of the 155 essential genes were plotted along the chromosome, as shown in Figure 5-8. A notable observation is the region roughly between genes PA1914 and PA2492, which coincides with the chromosome region between 2Mbp and 3Mbp, where transposon insertion density is notably low (Jacobs et al., 2003). The cause for this was unknown, however the relatively little phylogenetic distribution of genes in this region compared to other essential genes might shed some insight into the cause.
Figure 5-8: Phylogenetic distribution of essential genes along the chromosome of *P. aeruginosa* PA01.

A more detailed look at the BLASTP reports shows that these genes are conserved across a narrow range of taxa, mostly only within the *Pseudomonas* and *Burkholderia* genera.

That the genes in this region are not as phylogenetically distributed as other essential genes might be an indication that many of them are in fact not essential. In order to test whether this is the case, 23 genes in this region (PA1914 through PA2492) were cross-referenced with orthologs in *P. aeruginosa* PA14 and *E. coli* with regards to their essentiality in these organisms (see Materials and Methods in section 5.2.1).

Among the 23 genes, 17 of them were non-essential in *P. aeruginosa* PA14, five were essential, and one did not have an ortholog. In *E. coli*, 10 were non-essential, and 11 did not have any orthologs; only one had an essential ortholog in *E. coli*. This simple analysis suggests that many of these 23 genes are in fact not essential, and thus provides evidence that the low transposon
insertion density in this region is unlikely to be due to the presence of a cluster of essential genes. The real reason is still unknown, but is perhaps likely to be due to a bias in the transposon insertion process, or in the library construction. Since this is not observed in PA14 these results indicate that this region of low number of transposon insertions may actually be due to a technical error, rather than reflect a true biologically correct result.

5.3.4 Essential but phylogenetically non-distributed protein-coding genes

Essential protein-coding genes tend to be phylogenetically more distributed than non-essential genes, since many of these are involved in absolutely vital processes that sustain cellular life, such as DNA repair, transcription and protein synthesis.

Therefore, genes that appear to be essential in *P. aeruginosa* PA01 but not phylogenetically distributed would represent cases which would be interesting to investigate further. They may reflect genes that have more recently evolved to become essential. In order to add confidence to the essentiality status of these genes, they were cross-referenced with orthologs in *P. aeruginosa* PA14 and *E. coli* with regards to their essentiality in these organisms (see Materials and Methods in section 5.2.1).

Figure 5-9 shows a Venn diagram of 97 genes which are essential in both *P. aeruginosa* PA01 and *P. aeruginosa* PA14. Of these 97, fourteen are interesting in that their orthologs are present in less than 50 genera (the mean is 190.47, see Table 5-1). These could represent genus-specific or species-specific
essential genes, genes undergoing unusually rapid divergence, or they could represent genes that have undergone gene loss in some species (because though they are essential in *P. aeruginosa*, they may not be essential in other species).

![Diagram](image)

**Figure 5-9**: Essential genes that are not phylogenetically distributed. Fourteen genes within the yellow triangle are essential in both *P. aeruginosa* PA01 and PA14, but are distributed in less than 50 different genera.

### 5.3.5 Phylogenetic distribution of essential ncRNAs

A total of 98 non-coding RNAs were extracted from *P. aeruginosa* PA01 for analysis, ranging from rRNAs, tRNAs, tmRNAs and sRNAs. Of these, 69 (70.4% of total) did not contain any transposon insertion and were labelled as
“essential”, while 29 (29.6%) had at least one transposon insertion and were labelled as “non-essential”.

Homologs of each ncRNA were defined as BLAST hits above the e-value cutoff of 1E-3. Table 5-2 shows the mean, median and maximum number of taxa represented by homologs of essential ncRNAs compared to non-essential ncRNAs. Across all levels of taxa, essential ncRNAs have consistently the same value or higher values mean, median and maximum values, with a widening disparity as we move down from Superkingdoms to genera.

Table 5-2: Mean, median and maximum number of taxa represented by homologs of essential ncRNAs vs. non-essential ncRNAs.

<table>
<thead>
<tr>
<th></th>
<th>Superkingdoms</th>
<th>Phyla</th>
<th>Classes</th>
<th>Order</th>
<th>Families</th>
<th>Genera</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Essential</td>
<td>1.65</td>
<td>5.99</td>
<td>8.52</td>
<td>27.77</td>
<td>44.13</td>
</tr>
<tr>
<td></td>
<td>Non-essential</td>
<td>1.45</td>
<td>4.10</td>
<td>6.10</td>
<td>19.28</td>
<td>29.90</td>
</tr>
<tr>
<td>Median</td>
<td>Essential</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>24</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Non-essential</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>Max</td>
<td>Essential</td>
<td>3</td>
<td>23</td>
<td>34</td>
<td>91</td>
<td>139</td>
</tr>
<tr>
<td></td>
<td>Non-essential</td>
<td>3</td>
<td>14</td>
<td>34</td>
<td>91</td>
<td>126</td>
</tr>
</tbody>
</table>

Figures 5-10 and 5-11 show histograms of the phylogenetic distribution of essential ncRNAs vs. non-essential ncRNAs at the Superkingdom level and Genus level, respectively. At the Superkingdom level, it can be observed that a larger proportion of non-essential ncRNAs is represented in only one
Superkingdom, while a larger proportion of essential ncRNAs is represented in more than 2 Superkingdoms. At the Genus level, more than twice as many non-essential ncRNAs are represented in less than 30 genera, while 71% of essential ncRNAs are represented more than 30 genera as opposed to 41% for non-essential genes.

Figure 5-10: Histogram of phylogenetic distribution of essential ncRNAs vs. non-essential ncRNAs at the Superkingdom level
In contrast to protein-coding genes, an analysis on the phylogenetic distribution of essential ncRNAs vs. non-essential ncRNAs is limited by many factors. One main limitation is the availability of only a very small sample of ncRNAs present in \textit{P. aeruginosa} PAO1, which makes it difficult to assess the statistical significance of the difference in phylogenetic distribution. Moreover, this sample of ncRNAs should be overrepresented in tRNAs and rRNAs, adding a certain level of bias. \textit{Cis}-regulatory RNAs that have not been annotated are also not represented.

Since ncRNAs are usually very short in size, the likelihood that a transposon insertion is present is relatively small. According to statistical modelling of transposon insertions in \textit{P. aeruginosa} PAO1 described in Chapter 156.
2, a transposon gap larger than 144bp has a 50% chance of occurring. Thus many ncRNAs which have been labelled as "essential" could in fact be non-essential.

Finding ncRNA homologs is also still an open problem, and not as straightforward as in protein-coding genes, since ncRNA homologs have a conserved RNA secondary structure, but are not necessarily similar in sequence. Thus our simple BLAST analysis above can only capture a certain proportion of homologs with high sequence similarity but is not comprehensive. Yet, even given all these notable limitations, it is interesting that we still detect a difference between these putative essential and non-essential ncRNAs, with respect to their phylogenetic diversity.

5.4 Conclusion

In this chapter I have investigated the phylogenetic distribution of essential protein-coding genes and ncRNAs in *P. aeruginosa* PAO1 compared to their non-essential counterparts. Observations are consistent with the hypothesis that essential protein-coding genes are in general evolutionarily retained in a larger number of taxa, across all levels of taxa. A previous study drew similar conclusions, however analysis had only been done in a selection of 33 organisms (Gerdes et al., 2003).

Exceptions exist, in that fourteen protein-coding genes were identified to be essential in both PAO1 and PA14 strains of *P. aeruginosa*, yet are
phylogenetically distributed in a relatively low number of taxa. These may represent cases of species- or genus-specific essential genes.

Observations also suggest that essential ncRNAs might be more phylogenetically distributed than non-essential ncRNAs, however this analysis of ncRNAs has been severely limited by the small sample size, difficulty in determining essentiality, and the general problem of finding RNA homologs. Still, it is notable that a difference in the taxonomic diversity of homologs to essential ncRNAs, versus non-essential ncRNAs, is detected.

In conclusion, this analysis suggests that a comparative genomics approach is suitable for the identification of putatively essential protein-coding genes and ncRNAs. It also suggests that those putative novel essential genes identified in my screen of strain PAO1 may not be truly essential if they do not have taxonomically diverse homologs. However, it must be appreciated that my analysis may be disproportionately identifying genes that are essential that have low taxonomic diversity, since such genes would be hypothesized to be more likely to not be identified. Still, novel genes have been identified in my screen that are taxonomically diverse and so potentially they are the ones that warrant the most study.
6.1 Summary

The approaches developed to identify protein-coding genes and non-coding RNAs within putatively essential intergenic sequences in *P. aeruginosa* PAO1 were implemented into a computational pipeline as a single software package. To reflect the general applicability of this pipeline, it was used to predict putative novel protein-coding genes and ncRNAs in the significant pathogen, and cause of tuberculosis, *Mycobacterium tuberculosis* H37Rv. This TB causative agent is a major cause of disease worldwide, killing 1-2 million persons per year (WHO, 2008). It is noted for its developing multiple antibiotic resistance and the costly and difficult treatment that is currently recommended for a cure (WHO, 2008). There is a critical need for new antimicrobials active against this TB bacteria. As part of the Bioinformatics for Combating for Infectious Diseases project, we were interested in applying my method to the potential identification of new drug targets in this pathogen. I therefore applied my method to the analysis of the one *M. tuberculosis* genome (strain H37Rv) that has a transposon mutant library associated with it. Unfortunately, the library did not have a high enough coverage to enable accurate predictions of essential genes in intergenic regions, however I was able to use my analysis of *M. tuberculosis* to demonstrate an even more broader applicability of the method – identifying novel genes (not
necessarily essential) in such important pathogen genomes. My analysis identified new genes within previously annotated intergenic regions – some of which pathogen-specific and may represent drug targets worth pursuing further.

6.2 Materials and Methods

6.2.1 Development of automated pipeline

The computational pipeline was written in the Perl language as one single software package. A schematic diagram of the pipeline is illustrated in Figure 6-1. Four input files are required: 1) The nucleotide sequence of the genome interest in FASTA format; 2) Gene annotations in GFF format; 3) A transposon mutant library file containing coordinates of transposon insertion locations; and 4) A configuration file with user-defined parameters of programs used in the pipeline such as BLAST, RNAz and QRNA.

Putatively essential intergenic sequences are first extracted based on an estimated transposon insertion rate and a statistical significance cutoff, similar to the procedure described in Chapter 2. This set of putatively essential IGRs will go through three separate analysis routes. First, protein-coding genes are identified in the procedures described in Chapter 3. A remote BLASTX search is performed against the NCBI protein ‘nr’ database, and significant hits to protein-coding genes are identified from BLAST reports. Identified candidates are queried remotely against protein motif databases (described earlier in section 3.2.3) for functional characterization. Subcellular localization is predicted using
PSORTb as previously described. The result is a list of protein-coding genes in a spreadsheet format, which can be further manually prioritized and reviewed.

The second analysis route involves querying RFAM for identifying unannotated ncRNAs in known RNA families, as described in section 4.2.1. The third analysis route involves searching for novel ncRNAs using the methodologies described in sections 4.2.3 through 4.2.6. First, a BLASTN search is performed against the NCBI nucleotide 'nr' database to obtain sets of orthologous sequences. These are multiply aligned using MAFFT to obtain sets of native multiple sequence alignments, and also subsequently shuffled to obtain sets of shuffled multiple sequence alignments. RNAz and QRNA are run on both native and shuffled alignments, and score distribution graphs generated using R. Predictions from QRNA and RNAz are combined to produce a set of overlapping predictions, output as spreadsheet format, which can be further manually sorted and filtered in consideration of the score distribution graphs.
6.2.2 Data preparation - analysis of an example genome from *M. tuberculosis* H37Rv

The nucleotide genomic sequence and gene annotations of *M. tuberculosis* H37Rv were downloaded from the NCBI Microbial Genomes Resources (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). The transposon mutant insertion coordinates were obtained from the Supplementary Data of the corresponding transposon mutagenesis study (McAdam et al., 2002). The
pipeline used the default configuration settings of a BLASTX e-value cutoff of 1E-10; BLASTN e-value cutoff of 0.001; QRNA/RNAz sliding windows size of 120 columns, and an offset of 20 columns.

6.3 Results - analysis of an example genome from *M. tuberculosis* H37Rv

6.3.1 Extraction of putatively essential IGRs

The transposon library constructed by MacAdam et al. had a total of 1329 unique transposon insertions (MacAdam et al., 2002), over the entire *M. tuberculosis* H37Rv genome of size 4,411,532 bp. This signifies an average insertion rate of 0.30 insertions per kb, or approximately one insertion every 3333bp, representing a very low saturation rate. Statistical modelling of essentiality was therefore of limited use in this case, thus I opted to use a cutoff of 350bp (consistent with a rough size cutoff for an essentiality p-value of P < 0.2 for the *P. aeruginosa* PAO1 study) for defining IGRs of interest, and decided to use this dataset to see if I could simply identify potentially novel genes within the intergenic regions, rather than identifying essential genes. This illustrates the importance of having a transposon library of adequate size. Certainly, more libraries are planned in the future, so my methods are a little bit ready before their time. Still, again, I may be able to determine whether my method can identify any novel genes in the TB bacterium – a discovery which is still of notable interest.

A total of 209 IGRs were therefore extracted using the size cutoff and fed into later stages of the pipeline for analysis.
6.3.2 Novel protein-coding gene candidates

Out of the 209 extracted IGRs, 41 did not return any significant BLASTX hits, and 63 did not contain any ORFs. An additional two contained frameshifts, and another two had premature stop codons with respect to BLAST hits, and were discarded. The remaining 101 sequences contained ORFs that ranged from 105bp to 627bp. We expect many of the shortest genes to be false positives, but since putative novel genes were found to be transcribed in the range of 200-300bp in the *P. aeruginosa* PAO1 study, I have used 200bp as a cutoff, and discarded all ORFs shorter than this. As such, 44 such ORFs were discarded, leaving 57 ORFs.

In order to identify those most likely to be true genes, I have used phylogenetic distribution as an additional filtering criteria. I chose a stringent e-value cutoff of 1E-40 for the top BLAST hit, and a minimum number of 10 returned BLAST hits. This narrowed the final list of putatively novel protein-coding genes to four.

One of these is a predicted 285bp gene encoding a putative secreted antigen, which is widely distributed within the *Mycobacterium* genus. The transcript was predicted in the minus strand, consistent with both flanking genes. It is located 102bp upstream of Rv0946c encoding a glucose-6-phosphate isomerise, and 234bp downstream of an annotated probable mycolyl transferase pseudogene.

The second predicted gene of 333bp encodes a hypothetical protein which is distributed in at least 10 other *Mycobacterium* species, as well as in *Norcardia*
farcinica, another Actinobacteria. It is predicted to be transcribed in the minus strand, in the opposite direction from both flanking genes. It is located 117bp from the gene upstream encoding a probable monophosphatase, and 10bp from the gene downstream encoding a protein of the PPE family.

The third predicted gene is a 297bp gene encoding a hypothetical protein, located 238bp downstream of an hypothetical protein, and 289bp upstream of another hypothetical protein. All three genes are transcribed in the same forward strand. Phylogenetically, it is distributed amongst multiple strains of \textit{M. tuberculosis}. The subcellular localization of its deduced protein product is predicted to be in the cytoplasm.

The fourth gene is the longest amongst the four novel genes, with a predicted transcript of size 531bp, encoding an hypothetical protein. Predicted to be transcribed on the minus strand, it is flanked by nicT, a gene on the forward strand encoding a possible nickel-transport integral membrane protein just 17bp from the stop codon, and by the RV2857c gene on the minus strand encoding a short chain dehydrogenase, 235bp upstream of the start codon. Phylogenetically, it is distributed only among \textit{M. tuberculosis}.

These four predicted genes represent novel protein-coding gene candidates that may warrant further study. One may be concerned that many of these genes are not very taxonomically well spread, however, it is notable that this TB bacterium is relatively unique in organismal structure and function, and has a larger number of genes that are specific to its genera, than normal (Cole et al., 1998).
6.3.3 Novel putative ncRNA candidates with known RNA families

Out of the 209 IGRs extracted for analysis, four returned significant hits to RNA families in RFAM.

A cobalamin riboswitch (RFAM ID: RF00174) was identified just upstream of metE, a 5-methyltetrahydropteroylglutamate-homocysteine methyltransferase involved in methionine metabolism. Some follow-up literature search revealed evidence demonstrating that the metE gene is indeed regulated by this riboswitch via cobalamin-mediated growth inhibition, in *M. tuberculosis* strains H37Rv and CDC1551 (Warner et al., 2007).

A ydaO/yuaA element (RFAM ID: RF00379) was also identified directly upstream of rpfA, a gene encoding a possible resuscitation-promoting factor involved in regulating mycobacterial growth in disease reactivation (Tufariello et al., 2006). YdaO/yuaA leaders are cis-regulatory RNA structures that act as "off" switches, found upstream of *B. subtilis* genes, including ydaO, an amino acid transporter, and the KtrAB operon encoding a K⁺ transporter. *B. subtilis* defective in KtrAB are sensitive to sudden osmotic shock, suggesting the ydaO/yuaA element could possibly be responsive to compounds in the surrounding environment (Barrick et al., 2004). It is plausible that the ydaO/yuaA element could be involved in the regulation of the resuscitation-promoting factor in *M. tuberculosis*. 
A ykoK element (RFAM ID: RF00380) was identified directly upstream of Rv1535, an hypothetical protein. In various bacterial species, the ykoK element is a cis-regulatory Mg$^{2+}$-sensing RNA structure found upstream of major families of Mg$^{2+}$ transporters, and controls the expression of magnesium ion transport proteins (Dann et al., 2007).

Finally, two directly adjacent copies of the glycine riboswitch was also found directly upstream of glyA2, a probable serine hydroxymethyltransferase involved in glycine, serine and threonine metabolism. The presence of two copies is consistent with evidence that glycine riboswitches contain two ligand-binding domains that function cooperatively as a two-state switch (Mandal et al., 2004).

### 6.3.4 Novel putative ncRNA candidates predicted by RNAz/QRNA

Out of the 209 IGRs submitted to BLASTN, 36 did not return any significant orthologs. The remaining 173 were grouped into sets of orthologous sequences, and were multiple-aligned using MAFFT to produce 173 sets of native multiple sequence alignments. These were then also shuffled to produce 173 additional shuffled alignments. Multiple sequence alignments were broken into pairwise alignments in preparation for input to QRNA.

RNAz and QRNA were run on the 346 native and shuffled alignments broken into sliding windows of 120 columns with a 20-column offset. RNAz was thus run on 5892 native windows and 5916 shuffled windows, and QRNA was run on 45253 native windows and 45180 shuffled windows.
Figure 6-2: Distribution of (A) RNAz z-scores and (B) QRNA sigmoidal RNA scores in native windows vs. shuffled windows.
Windows predicted as RNA were clustered into loci, producing 139 candidate ncRNA loci predicted by RNAz, and 198 candidate ncRNA loci predicted by QRNA. 55 loci were predicted only by RNAz but not QRNA, and 114 loci were predicted only by QRNA but not RNAz. These 169 loci were discarded, leaving 84 candidate loci that were predicted by both RNAz and QRNA.

Of these 84 candidate loci, three coincided with predictions via RFAM, namely the ydaO/yuaA element, ykoK element, and the glycine riboswitch (predicted as one single locus). The cobalamin riboswitch could only be predicted by RNAz. These were removed, leaving 81 candidate loci. None of the four identified protein-coding genes were amongst the 81 candidate loci.

Clustered regularly interspaced short palindromic repeats (CRISPRs) (Grissa et al., 2007) were found in four loci, and tandem repeats from the Tandem Repeats Database (Dencœud and Vergnaud, 2004) were found in 3 loci. These 7 loci were discarded, leaving 74 candidate loci.

To retain only the most significant hits, I selected those with an RNAz z-score less than -2.0, and a QRNA sigmoidal RNA score greater than 40. Six loci met this criteria.
Table 6-1: Six predicted ncRNA candidates in *Mycobacterium tuberculosis* H37Rv with RNAz z-score < -2.0, and QRNA sigmoidal RNA score > 40.

<table>
<thead>
<tr>
<th>Left gene (strand)</th>
<th>Right gene (strand)</th>
<th>ncRNA predicted size</th>
<th>ncRNA predicted strand</th>
<th>Distance from left gene</th>
<th>Distance from right gene</th>
<th>Mean RNAz z-score</th>
<th>Mean QRNA sigmoidal score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rv1179c – hypothetical protein (-)</td>
<td>Rv1180 - probable polyketide beta-ketoacyl synthase (+)</td>
<td>120nt</td>
<td>-</td>
<td>40bp</td>
<td>267bp</td>
<td>-5.68</td>
<td>87.61</td>
</tr>
<tr>
<td>Rv2964- probable formyltetrahydrofolate deformylase (+)</td>
<td>Rv2965c- phosphopantetheine adenylyltransferase (-)</td>
<td>136nt</td>
<td>++</td>
<td>60bp</td>
<td>674bp</td>
<td>-3.73</td>
<td>46.49</td>
</tr>
<tr>
<td>Rv3197A- probable transcriptional regulatory protein (-)</td>
<td>Rv3198c- probable ATP-dependent DNA helicase II (-)</td>
<td>170nt</td>
<td>++</td>
<td>20bp</td>
<td>241bp</td>
<td>-2.23</td>
<td>44.14</td>
</tr>
<tr>
<td>Rv3666c- probable periplasmic dipeptide-binding lipoprotein (-)</td>
<td>Rv3667- acetyl-CoA synthetase (+)</td>
<td>121nt</td>
<td>++</td>
<td>280bp</td>
<td>308bp</td>
<td>-2.09</td>
<td>42.40</td>
</tr>
<tr>
<td>Rv3709c- aspartate kinase (-)</td>
<td>Rv3710- 2-isopropylmalate synthase (+)</td>
<td>178nt</td>
<td>++</td>
<td>0bp</td>
<td>199bp</td>
<td>-3.28</td>
<td>61.42</td>
</tr>
<tr>
<td>Rv3843c- probable conserved transmembrane protein (-)</td>
<td>Rv3844- possible transposase (+)</td>
<td>146nt</td>
<td>++</td>
<td>1306bp</td>
<td>728bp</td>
<td>-2.46</td>
<td>70.66</td>
</tr>
</tbody>
</table>
6.4 Discussion and Conclusion

The main drawback in this study was the extremely low saturation rate of the transposon mutagenesis study, which made any statistical assessment of essentiality of limited use. Out of the 1329 total unique transposon insertions, 1189 (89.5% of total number of insertions) were within ORFs, while only 351 (10.5%) fell within intergenic regions. This was due in part to the authors’ intention to focus on identifying essential genes amongst annotated genes, while playing less attention to intergenic regions (McAdam et al., 2002). As such, the transposon mutant library lacked comprehensiveness especially for intergenic regions, and also was non-random. As a result, the essentiality of putative novel protein-coding genes and ncRNAs that were identified within intergenic regions could not be determined and thus of limited value for identifying potential antibacterial targets. This highlights the important value of a comprehensive transposon mutant library for these types of studies, such as those for *P. aeruginosa* PAO1 (Jacobs et al., 2003) and PA14 (Liberati et al., 2006).

Nonetheless, given a list of intergenic sequences (regardless of whether they are essential), here we have demonstrated the applicability of the automated computational pipeline to predict novel protein-coding genes and non-coding RNAs to a genome other than *P. aeruginosa*. Given the availability of a comprehensive transposon library amenable to statistical assessment of essentiality, this automated pipeline can be applied to any bacterial genome in general to identify novel and putatively essential protein-coding genes and ncRNAs. In the absence of such a library, it can still also be used to identify
potentially novel and unannotated protein-coding genes and ncRNAs in any bacterial genome.
CHAPTER 7  CONCLUDING COMMENTS AND FUTURE WORK

To expand the list of available possible antibacterial targets, we have developed a computational approach to identify novel putative essential protein-coding genes and non-coding RNAs within intergenic regions in *P. aeruginosa* PAO1. Using this approach, we have predicted novel protein-coding genes, of which at least three have been found to be transcribed to date, and have other evidence suggesting that they are truly essential genes. Moreover, we have identified five unannotated and putatively essential ncRNAs in known RNA families, including three riboswitches, and predict an additional five novel ncRNAs. The three riboswitches were confirmed to be transcribed. Riboswitches have been proposed as a viable antibacterial drug targets (Blount and Breaker, 2006). In general, these putative essential protein-coding genes and ncRNAs represent an expanded list of potential antibacterial targets that should be studied further.

This computational approach was implemented as an automated pipeline, which, given the availability of comprehensive transposon mutant libraries, can be used to predict novel and putative essential protein-coding genes and ncRNAs in any bacterial genome in general. Its general applicability was reflected by its use in predicting novel protein-coding genes and ncRNAs in *M. tuberculosis* H37Rv. In the future, we would expect it to be applied to other
bacterial genomes as well to expand the list of available potential antibacterial targets.

The ability of identify truly essential genes depends heavily on the availability and comprehensiveness of transposon mutant library data, which we anticipate will increase in the future. The availability of transposon mutant library data for related species and strains also provide a valuable means to cross-reference gene essentiality across multiple species or strains. This approach could be potentially incorporated into the current pipeline in the future.

To better prioritize novel protein-coding genes for lab verification, and to distinguish novel protein-coding genes from false positives, better composition-based scoring methods could also be developed in the future.

Non-coding RNA prediction is currently still an active area of research, and we expect improvements to be made by researchers over the years to come in terms of the sensitivity and accuracy of ncRNA prediction, as well as RNA secondary structure prediction and RNA alignments. Future work would include the incorporation of these improved software into the pipeline for making better ncRNA predictions. In fact, MAFFT has recently just released version 6.5 (Katoh and Toh, 2008), which has incorporated RNA secondary structure information to generate more accurate RNA alignments. Future work would also include the assessment of its performance and potentially incorporating into the current pipeline.

In closing, this work aimed to investigate whether new putative essential genes could be identified in previously annotated bacterial genomes. Using P.
*aeruginosa* as a model, my analysis indicates that even a genome like *P. aeruginosa* PAO1, which was annotated almost a decade ago, is well annotated with regard to essential genes. This is likely in part due to the apparent high conservation of essential genes across species. Still, some putative novel essential genes were identified that warrant further study and if this method is applied in an automated fashion to other genomes, potentially many more novel essential genes, or simply novel genes, could be identified. Such novel essential genes represent important drug targets worth exploring. In the future, this approach could also be coupled though with a screen for pathogen-specific genes or virulence factors, to identify putative novel anti-infective drug targets. Regardless, it is clear that antimicrobial resistance is on the rise, necessitating the identification of new drug targets and new approaches, including genomic approaches, to identify such novel targets and associated therapeutics.
REFERENCE LIST


APPENDIX A: TABLE OF 34 PUTATIVE NOVEL ESSENTIAL PROTEIN-CODING GENE PREDICTIONS IN P. AERUGINOSA PAO1
<table>
<thead>
<tr>
<th>LocusID</th>
<th>Predicted gene size (bp)</th>
<th>PEIS size (bp)</th>
<th>Essentiality p-value</th>
<th>Left gene locus</th>
<th>Left gene product</th>
<th>Left gene strand</th>
<th>Right gene locus</th>
<th>Right gene product</th>
<th>Top hit evalue</th>
<th>Subcellular localization</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA0041.2</td>
<td>195</td>
<td>694</td>
<td>0.04</td>
<td>PA0041</td>
<td>probable hemagglutinin</td>
<td>1</td>
<td>PA0042</td>
<td>hypothetical protein</td>
<td>2.02E-29</td>
<td>Unknown</td>
</tr>
<tr>
<td>PA0050.1</td>
<td>138</td>
<td>674</td>
<td>0.04</td>
<td>PA0050</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA0051</td>
<td>potential phenazine-modifying enzyme</td>
<td>4.07E-16</td>
<td>Unknown</td>
</tr>
<tr>
<td>PA0194.1</td>
<td>144</td>
<td>646</td>
<td>0.04</td>
<td>PA0194</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA0195</td>
<td>putative NAD(P) transhydrogenase subunit alpha 1</td>
<td>4E-18</td>
<td>Unknown</td>
</tr>
<tr>
<td>PA0442.1</td>
<td>129</td>
<td>334</td>
<td>0.20</td>
<td>PA0442</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA0443</td>
<td>probable transporter</td>
<td>9.93E-17</td>
<td>Unknown</td>
</tr>
<tr>
<td>PA0457.1</td>
<td>375</td>
<td>369</td>
<td>0.17</td>
<td>PA0457</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA0458</td>
<td>probable major facilitator superfamily (MFS) transporter</td>
<td>3.03E-45</td>
<td>Cytoplasmic Membrane</td>
</tr>
<tr>
<td>PA0613.1</td>
<td>225</td>
<td>363</td>
<td>0.17</td>
<td>PA0613</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA0614</td>
<td>hypothetical protein</td>
<td>1.02E-37</td>
<td>Cytoplasmic</td>
</tr>
<tr>
<td>PA0716.1</td>
<td>267</td>
<td>417</td>
<td>0.13</td>
<td>PA0716</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA0717</td>
<td>hypothetical protein of bacteriophage Pf1</td>
<td>2.19E-19</td>
<td>Unknown</td>
</tr>
<tr>
<td>PA0728.1</td>
<td>252</td>
<td>309</td>
<td>0.23</td>
<td>PA0728</td>
<td>probable bacteriophage integrase</td>
<td>1</td>
<td>PA0729</td>
<td>hypothetical protein</td>
<td>2.05E-30</td>
<td>Unknown</td>
</tr>
<tr>
<td>PA0836.2</td>
<td>201</td>
<td>306</td>
<td>0.23</td>
<td>PA0836</td>
<td>probable acetate kinase</td>
<td>-1</td>
<td>PA0837</td>
<td>peptidyl-prolyl cis-trans isomerase SlyD</td>
<td>1.63E-35</td>
<td>Cytoplasmic</td>
</tr>
<tr>
<td>PA1152.1</td>
<td>273</td>
<td>545</td>
<td>0.07</td>
<td>PA1152</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA1153</td>
<td>hypothetical protein</td>
<td>1.05E-43</td>
<td>Cytoplasmic</td>
</tr>
<tr>
<td>Gene</td>
<td>1st</td>
<td>2nd</td>
<td>3rd</td>
<td>4th</td>
<td>5th</td>
<td>6th</td>
<td>7th</td>
<td>Position</td>
<td>Protein Description</td>
<td>Location</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>--------------</td>
<td>-----------------------</td>
<td>------------------------</td>
<td>----------</td>
<td>---------------------------------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>PA1263.1</td>
<td>165</td>
<td>296</td>
<td>0.24</td>
<td>PA1263</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA1264 probable transcriptional regulator</td>
<td>-1</td>
<td>1.6E-22 Unknown</td>
<td></td>
</tr>
<tr>
<td>PA1366.1</td>
<td>279</td>
<td>423</td>
<td>0.13</td>
<td>PA1366</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA1367 hypothetical protein</td>
<td>-1</td>
<td>1.82E-42 Unknown</td>
<td></td>
</tr>
<tr>
<td>PA1499.1</td>
<td>222</td>
<td>296</td>
<td>0.24</td>
<td>PA1499</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA1500 probable oxidoreductase</td>
<td>-1</td>
<td>1.22E-30 Unknown</td>
<td></td>
</tr>
<tr>
<td>PA1507.1</td>
<td>168</td>
<td>570</td>
<td>0.06</td>
<td>PA1507</td>
<td>probable transporter</td>
<td>-1</td>
<td>PA1508 hypothetical protein</td>
<td>-1</td>
<td>5.59E-25 Unknown</td>
<td></td>
</tr>
<tr>
<td>PA1566.1</td>
<td>189</td>
<td>370</td>
<td>0.17</td>
<td>PA1566</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA1567 hypothetical protein</td>
<td>-1</td>
<td>1.01E-24 Unknown</td>
<td></td>
</tr>
<tr>
<td>PA1796.5</td>
<td>222</td>
<td>438</td>
<td>0.12</td>
<td>tRNA-His</td>
<td>1</td>
<td>PA1797 hypothetical protein</td>
<td>-1</td>
<td>6.81E-26 Cytoplasmic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1838.1</td>
<td>249</td>
<td>460</td>
<td>0.11</td>
<td>PA1838</td>
<td>sulfite reductase</td>
<td>-1</td>
<td>PA1839 hypothetical protein</td>
<td>-1</td>
<td>9.51E-28 Unknown</td>
<td></td>
</tr>
<tr>
<td>PA2075.1</td>
<td>210</td>
<td>831</td>
<td>0.02</td>
<td>PA2075</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA2076 probable transcriptional regulator</td>
<td>1</td>
<td>3.34E-38 Cytoplasmic</td>
<td></td>
</tr>
<tr>
<td>PA2118.1</td>
<td>243</td>
<td>534</td>
<td>0.08</td>
<td>PA2118</td>
<td>O6-methylguanine-DNA methyltransferase</td>
<td>1</td>
<td>PA2119 alcohol dehydrogenase (Zn-independent)</td>
<td>-1</td>
<td>1.44E-34 Unknown</td>
<td></td>
</tr>
<tr>
<td>PA2145.1</td>
<td>165</td>
<td>498</td>
<td>0.09</td>
<td>PA2145</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA2146 hypothetical protein</td>
<td>1</td>
<td>2.82E-20 Unknown</td>
<td></td>
</tr>
<tr>
<td>PA2173.1</td>
<td>222</td>
<td>308</td>
<td>0.23</td>
<td>PA2173</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA2174 hypothetical protein</td>
<td>-1</td>
<td>3.28E-36 Cytoplasmic</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Start</td>
<td>End</td>
<td>E-value</td>
<td>Protein</td>
<td>Description</td>
<td>Query</td>
<td>Score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>------</td>
<td>---------</td>
<td>-----------</td>
<td>--------------------------------------------------</td>
<td>-------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2566.1</td>
<td>210</td>
<td>658</td>
<td>0.04</td>
<td>PA2566</td>
<td>hypothetical protein -1</td>
<td>PA2567</td>
<td>3.08E-37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2747.1</td>
<td>234</td>
<td>485</td>
<td>0.10</td>
<td>PA2747</td>
<td>hypothetical protein -1</td>
<td>PA2748</td>
<td>2.4E-26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2754.0</td>
<td>1</td>
<td>198</td>
<td>0.11</td>
<td>PA2754</td>
<td>hypothetical protein 1</td>
<td>PA2755</td>
<td>7.35E-28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2754.1</td>
<td>1</td>
<td>198</td>
<td>0.11</td>
<td>PA2754</td>
<td>histidine transport protein HisP</td>
<td>PA2927</td>
<td>4.82E-27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2926.1</td>
<td>183</td>
<td>321</td>
<td>0.21</td>
<td>PA2926</td>
<td>histidine transport protein HisP</td>
<td>PA2927</td>
<td>4.82E-27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA3414.1</td>
<td>459</td>
<td>492</td>
<td>0.09</td>
<td>PA3414</td>
<td>hypothetical protein 1</td>
<td>PA3415</td>
<td>6.6E-35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA3762.1</td>
<td>279</td>
<td>464</td>
<td>0.11</td>
<td>PA3762</td>
<td>hypothetical protein 1</td>
<td>PA3763</td>
<td>1.72E-40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA3865.1</td>
<td>339</td>
<td>548</td>
<td>0.07</td>
<td>PA3865</td>
<td>probable amino acid binding protein</td>
<td>PA3866</td>
<td>1.04E-46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA3887.1</td>
<td>225</td>
<td>361</td>
<td>0.18</td>
<td>PA3887</td>
<td>Na+/H+ antiporter NhaP</td>
<td>PA3888</td>
<td>1.03E-13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA4100.0</td>
<td>1</td>
<td>405</td>
<td>0.05</td>
<td>PA4100</td>
<td>probable dehydrogenase 1</td>
<td>PA4101</td>
<td>3.02E-49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA4108.1</td>
<td>222</td>
<td>392</td>
<td>0.15</td>
<td>PA4108</td>
<td>hypothetical protein 1</td>
<td>PA4109</td>
<td>1.57E-22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA4696.1</td>
<td>132</td>
<td>550</td>
<td>0.07</td>
<td>PA4696</td>
<td>acetylactate synthase 3 catalytic subunit</td>
<td>PA4697</td>
<td>1.34E-17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>Start</td>
<td>End</td>
<td>Annotation</td>
<td>Description</td>
<td>Start</td>
<td>End</td>
<td>Annotation</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>-----</td>
<td>------------</td>
<td>-------------</td>
<td>-------</td>
<td>-----</td>
<td>------------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA4973</td>
<td>192</td>
<td>420</td>
<td>thiamine biosynthesis protein ThiC</td>
<td>-1 PA4973</td>
<td>1</td>
<td>2.58E-12</td>
<td>unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAS141</td>
<td>261</td>
<td>303</td>
<td>1-(5-phosphoribosyl)-5-(5-phosphoribosylamino)</td>
<td>-1 PAS142</td>
<td>4.44E-41</td>
<td>cytoplasmic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX B: TABLE OF OVERLAPPING RNAZ/QRNA NOVEL NCRNA PREDICTIONS IN P. AERUGINOSA PAO1
<table>
<thead>
<tr>
<th>PEIS ID</th>
<th>Predicted size</th>
<th>PEIS size</th>
<th>Essentiality pvalue</th>
<th>left gene locus</th>
<th>left gene product</th>
<th>left gene strand</th>
<th>right gene locus</th>
<th>right gene product</th>
<th>right gene strand</th>
<th>Mean consensus MFE</th>
<th>Mean z-</th>
<th>Mean RNA</th>
<th>Mean RNA mean sigmoid</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA0044-PA0045_1</td>
<td>121</td>
<td>406</td>
<td>0.14</td>
<td>PA0044</td>
<td>exoenzyme T</td>
<td>1</td>
<td>PA0045</td>
<td>hypothetical protein</td>
<td>1</td>
<td>-39.15</td>
<td>-2.09</td>
<td>0.92</td>
<td>2.68</td>
</tr>
<tr>
<td>PA0052-PA0053_1</td>
<td>120</td>
<td>384</td>
<td>0.16</td>
<td>PA0052</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA0053</td>
<td>hypothetical protein</td>
<td>1</td>
<td>-37.80</td>
<td>-2.56</td>
<td>0.97</td>
<td>18.48</td>
</tr>
<tr>
<td>PA0101-PA0102_1</td>
<td>161</td>
<td>342</td>
<td>0.19</td>
<td>PA0101</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA0102</td>
<td>probable carbonic anhydrase</td>
<td>1</td>
<td>-39.98</td>
<td>-2.14</td>
<td>0.95</td>
<td>2.88</td>
</tr>
<tr>
<td>PA0186-PA0187_1</td>
<td>180</td>
<td>326</td>
<td>0.21</td>
<td>PA0186</td>
<td>probable binding protein component of ABC transporter</td>
<td>1</td>
<td>PA0187</td>
<td>hypothetical protein</td>
<td>1</td>
<td>-44.40</td>
<td>-2.38</td>
<td>0.97</td>
<td>7.15</td>
</tr>
<tr>
<td>PA0194-PA0195_1</td>
<td>281</td>
<td>646</td>
<td>0.04</td>
<td>PA0194</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA0195</td>
<td>putative NAD(P) transhydrogenase subunit alpha 1</td>
<td>1</td>
<td>-48.92</td>
<td>-3.24</td>
<td>0.95</td>
<td>1.96</td>
</tr>
<tr>
<td>PA0241-PA0242_1</td>
<td>240</td>
<td>542</td>
<td>0.07</td>
<td>PA0241</td>
<td>probable major facilitator superfamily (MFS) transporter</td>
<td>-1</td>
<td>PA0242</td>
<td>hypothetical protein</td>
<td>1</td>
<td>-37.97</td>
<td>-2.35</td>
<td>0.96</td>
<td>33.69</td>
</tr>
<tr>
<td>PA0263-PA0263_1_1</td>
<td>201</td>
<td>332</td>
<td>0.20</td>
<td>PA0263</td>
<td>secreted protein Hcp</td>
<td>-1</td>
<td>PA0263_1</td>
<td>tRNA-Arg</td>
<td>-1</td>
<td>-28.24</td>
<td>-2.44</td>
<td>0.99</td>
<td>13.13</td>
</tr>
<tr>
<td>PA0263-PA0263_1_2</td>
<td>440</td>
<td>460</td>
<td>0.11</td>
<td>PA0263</td>
<td>secreted protein Hcp</td>
<td>-1</td>
<td>PA0263_1</td>
<td>tRNA-Arg</td>
<td>-1</td>
<td>-45.53</td>
<td>-3.16</td>
<td>0.97</td>
<td>0.73</td>
</tr>
<tr>
<td>PA0290-PA0291_1</td>
<td>359</td>
<td>612</td>
<td>0.05</td>
<td>PA0290</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA0291</td>
<td>Anaerobically-induced outer membrane porin</td>
<td>1</td>
<td>-36.74</td>
<td>-1.55</td>
<td>0.95</td>
<td>13.05</td>
</tr>
<tr>
<td>PA0432-PA0433_1</td>
<td>124</td>
<td>288</td>
<td>0.25</td>
<td>PA0432</td>
<td>S-adenosyl-L-homocysteine hydrolase</td>
<td>-1</td>
<td>PA0433</td>
<td>hypothetical protein</td>
<td>1</td>
<td>-35.76</td>
<td>-1.91</td>
<td>0.97</td>
<td>25.60</td>
</tr>
<tr>
<td>ID</td>
<td>Accession</td>
<td>Score</td>
<td>% Id</td>
<td>Description</td>
<td>Type</td>
<td>ID</td>
<td>E-value</td>
<td>Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-----------</td>
<td>-------</td>
<td>------</td>
<td>---------------------------</td>
<td>---------------</td>
<td>-----</td>
<td>---------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0442-PA0443_1</td>
<td>258 334 0.20</td>
<td>PA0442</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA0443</td>
<td>probable transporter</td>
<td>1</td>
<td>-51.95</td>
<td>-2.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0444-PA0445_1</td>
<td>152 400 0.15</td>
<td>PA0444</td>
<td>allantocate amidohydrolase</td>
<td>1</td>
<td>PA0445</td>
<td>probable transposase</td>
<td>-1</td>
<td>-36.18</td>
<td>-2.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0484-PA0485_1</td>
<td>121 637 0.05</td>
<td>PA0484</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA0485</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>-29.57</td>
<td>-2.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0499-PA0500_1</td>
<td>201 363 0.17</td>
<td>PA0499</td>
<td>probable pili assembly chaperone</td>
<td>-1</td>
<td>PA0500</td>
<td>biotin synthase</td>
<td>1</td>
<td>-30.39</td>
<td>-2.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0609-PA0611_1</td>
<td>140 290 0.25</td>
<td>PA0609</td>
<td>anthranilate synthetase component I</td>
<td>1</td>
<td>PA0611</td>
<td>transcriptional regulator PrtR</td>
<td>-1</td>
<td>-55.43</td>
<td>-4.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0687-PA0688_1</td>
<td>226 513 0.09</td>
<td>PA0687</td>
<td>probable type II secretion system protein</td>
<td>1</td>
<td>PA0688</td>
<td>probable binding protein component of type IV</td>
<td>1</td>
<td>-33.05</td>
<td>-2.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0701-PA0702_1</td>
<td>201 442 0.12</td>
<td>PA0701</td>
<td>probable transcriptional regulator</td>
<td>1</td>
<td>PA0702</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>-58.07</td>
<td>-2.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0730-PA0731_1</td>
<td>198 451 0.11</td>
<td>PA0730</td>
<td>probable transferase</td>
<td>-1</td>
<td>PA0731</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>-34.32</td>
<td>-2.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0736-PA0737_1</td>
<td>121 307 0.23</td>
<td>PA0736</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA0737</td>
<td>hypothetical protein</td>
<td>1</td>
<td>-46.35</td>
<td>-2.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0761-PA0762_1</td>
<td>179 408 0.14</td>
<td>PA0761</td>
<td>L-aspartate oxidase</td>
<td>-1</td>
<td>PA0762</td>
<td>RNA polymerase sigma factor AlgU</td>
<td>1</td>
<td>-48.02</td>
<td>-1.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0805-PA0806_1</td>
<td>121 510 0.09</td>
<td>PA0805</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA0806</td>
<td>hypothetical protein</td>
<td>1</td>
<td>-30.85</td>
<td>-2.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>Start</td>
<td>End</td>
<td>Log2</td>
<td>Strain</td>
<td>Description</td>
<td>Accession</td>
<td>Start</td>
<td>End</td>
<td>Log2</td>
<td>Strain</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>--------</td>
<td>----------------------------------</td>
<td>-----------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>--------</td>
<td>----------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0844</td>
<td>199</td>
<td>341</td>
<td>0.19</td>
<td>PA0844</td>
<td>hemolytic phospholipase C precursor</td>
<td>PA0845</td>
<td>-1</td>
<td></td>
<td>-31.55</td>
<td>0.97</td>
<td>13.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0981</td>
<td>280</td>
<td>345</td>
<td>0.19</td>
<td>PA0981</td>
<td>hypothetical protein</td>
<td>PA0982</td>
<td>1</td>
<td></td>
<td>-42.75</td>
<td>1.00</td>
<td>5.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA0991</td>
<td>121</td>
<td>362</td>
<td>0.18</td>
<td>PA0991</td>
<td>hypothetical protein</td>
<td>PA0992</td>
<td>1</td>
<td></td>
<td>-35.65</td>
<td>0.99</td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1025</td>
<td>121</td>
<td>291</td>
<td>0.25</td>
<td>PA1025</td>
<td>probable porin</td>
<td>PA1026</td>
<td>1</td>
<td></td>
<td>-71.70</td>
<td>0.92</td>
<td>8.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1030</td>
<td>199</td>
<td>558</td>
<td>0.07</td>
<td>PA1030</td>
<td>hypothetical protein</td>
<td>PA1031</td>
<td>1</td>
<td></td>
<td>-33.48</td>
<td>0.94</td>
<td>12.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1091</td>
<td>214</td>
<td>353</td>
<td>0.18</td>
<td>PA1091</td>
<td>hypothetical protein</td>
<td>PA1092</td>
<td>1</td>
<td></td>
<td>-28.28</td>
<td>0.96</td>
<td>12.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1112</td>
<td>121</td>
<td>356</td>
<td>0.18</td>
<td>PA1112</td>
<td>hypothetical protein</td>
<td>PA1113</td>
<td>1</td>
<td></td>
<td>-51.24</td>
<td>0.94</td>
<td>13.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1152</td>
<td>141</td>
<td>545</td>
<td>0.07</td>
<td>PA1152</td>
<td>hypothetical protein</td>
<td>PA1153</td>
<td>1</td>
<td></td>
<td>-26.57</td>
<td>0.99</td>
<td>4.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1182</td>
<td>121</td>
<td>335</td>
<td>0.20</td>
<td>PA1182</td>
<td>probable transcriptional regulator</td>
<td>PA1183</td>
<td>1</td>
<td></td>
<td>-38.91</td>
<td>0.99</td>
<td>5.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1221</td>
<td>120</td>
<td>683</td>
<td>0.04</td>
<td>PA1221</td>
<td>hypothetical protein</td>
<td>PA1222</td>
<td>-1</td>
<td></td>
<td>-51.53</td>
<td>0.94</td>
<td>1.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1221</td>
<td>121</td>
<td>683</td>
<td>0.04</td>
<td>PA1221</td>
<td>hypothetical protein</td>
<td>PA1222</td>
<td>-1</td>
<td></td>
<td>-39.67</td>
<td>0.91</td>
<td>7.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>Description</td>
<td>Accession</td>
<td>Identity</td>
<td>Mismatch</td>
<td>CDR</td>
<td>Fold Change</td>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td>-----------</td>
<td>----------</td>
<td>----------</td>
<td>-----</td>
<td>-------------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1248</td>
<td>Alkaline protease</td>
<td>PA1249</td>
<td>1</td>
<td>30.53</td>
<td>-2.27</td>
<td>15.32</td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1342</td>
<td>Probable binding protein</td>
<td>PA1343</td>
<td>1</td>
<td>38.69</td>
<td>3.32</td>
<td>2.16</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1347</td>
<td>Probable protein component of ABC transporter</td>
<td>PA1348</td>
<td>1</td>
<td>56.57</td>
<td>-2.16</td>
<td>1.57</td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1427</td>
<td>Probable hypotetical protein</td>
<td>PA1428</td>
<td>1</td>
<td>38.67</td>
<td>0.91</td>
<td>7.00</td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1437</td>
<td>Probable protein</td>
<td>PA1438</td>
<td>1</td>
<td>60.63</td>
<td>1.00</td>
<td>15.18</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1519</td>
<td>Probable protein</td>
<td>PA1520</td>
<td>1</td>
<td>47.96</td>
<td>-3.27</td>
<td>2.70</td>
<td>0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1587</td>
<td>Probable protein</td>
<td>PA1588</td>
<td>1</td>
<td>47.96</td>
<td>-3.27</td>
<td>2.70</td>
<td>0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1589</td>
<td>Probable protein</td>
<td>PA1590</td>
<td>1</td>
<td>51.70</td>
<td>-3.26</td>
<td>10.75</td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1655</td>
<td>Probable protein</td>
<td>PA1656</td>
<td>1</td>
<td>51.70</td>
<td>-3.26</td>
<td>10.75</td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1768</td>
<td>Probable protein</td>
<td>PA1769</td>
<td>1</td>
<td>61.36</td>
<td>0.99</td>
<td>2.01</td>
<td>0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1897</td>
<td>Quorum-sensing control repressor</td>
<td>PA1898</td>
<td>1</td>
<td>38.51</td>
<td>-2.98</td>
<td>1.54</td>
<td>0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1899</td>
<td>Probable protein</td>
<td>PA1900</td>
<td>1</td>
<td>353</td>
<td>0.18</td>
<td>15.32</td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein ID</td>
<td>Start</td>
<td>End</td>
<td>Score</td>
<td>Function Description</td>
<td>Protein ID</td>
<td>Start</td>
<td>End</td>
<td>Score</td>
<td>Function Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td>----------------------</td>
<td>------------</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td>----------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1898</td>
<td>160</td>
<td>472</td>
<td>0.10</td>
<td>quorum-sensing control repressor</td>
<td>1</td>
<td>PA1899</td>
<td>1</td>
<td>-62.36</td>
<td>phenazine biosynthesis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1898</td>
<td>121</td>
<td>472</td>
<td>0.10</td>
<td>quorum-sensing control repressor</td>
<td>1</td>
<td>PA1899</td>
<td>1</td>
<td>-36.45</td>
<td>phenazine biosynthesis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA1914</td>
<td>121</td>
<td>567</td>
<td>0.07</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA1915</td>
<td>1</td>
<td>-51.62</td>
<td>hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2035</td>
<td>121</td>
<td>541</td>
<td>0.07</td>
<td>acetalactate synthase 2 catalytic subunit</td>
<td>-1</td>
<td>PA2036</td>
<td>1</td>
<td>-32.70</td>
<td>hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2041</td>
<td>220</td>
<td>487</td>
<td>0.10</td>
<td>probable amino acid permease</td>
<td>1</td>
<td>PA2042</td>
<td>1</td>
<td>-43.58</td>
<td>probable transporter (membrane)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2047</td>
<td>161</td>
<td>312</td>
<td>0.22</td>
<td>probable transcriptional regulator</td>
<td>-1</td>
<td>PA2048</td>
<td>1</td>
<td>-43.33</td>
<td>hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2049</td>
<td>201</td>
<td>570</td>
<td>0.06</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA2050</td>
<td>1</td>
<td>-42.81</td>
<td>RNA polymerase sigma factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2056</td>
<td>140</td>
<td>599</td>
<td>0.06</td>
<td>probable transcriptional regulator</td>
<td>1</td>
<td>PA2057</td>
<td>1</td>
<td>-45.59</td>
<td>hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2056</td>
<td>108</td>
<td>599</td>
<td>0.06</td>
<td>probable transcriptional regulator</td>
<td>1</td>
<td>PA2057</td>
<td>1</td>
<td>-40.86</td>
<td>hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2119</td>
<td>176</td>
<td>332</td>
<td>0.20</td>
<td>alcohol dehydrogenase (Zn-dependent)</td>
<td>-1</td>
<td>PA2120</td>
<td>1</td>
<td>-47.19</td>
<td>hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2127</td>
<td>140</td>
<td>499</td>
<td>0.09</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA2128</td>
<td>1</td>
<td>-38.47</td>
<td>fimbrial subunit CupA1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Start</td>
<td>Stop</td>
<td>E-value</td>
<td>Gene</td>
<td>Function</td>
<td>Start</td>
<td>Stop</td>
<td>E-value</td>
<td>Start</td>
<td>Stop</td>
<td>E-value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-------</td>
<td>-------</td>
<td>---------</td>
<td>-------</td>
<td>-----------------------------------</td>
<td>-------</td>
<td>-------</td>
<td>---------</td>
<td>-------</td>
<td>-------</td>
<td>---------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2127</td>
<td>161</td>
<td>499</td>
<td>0.09</td>
<td>PA2127</td>
<td>hypothetical protein</td>
<td>-1</td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
<td>-44.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>fimbrial subunit CupA1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2204</td>
<td>200</td>
<td>298</td>
<td>0.24</td>
<td>PA2204</td>
<td>probable binding protein component of ABC transporter</td>
<td>-1</td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
<td>-48.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2258</td>
<td>120</td>
<td>363</td>
<td>0.24</td>
<td>PA2258</td>
<td>transcriptional regulator PtXR</td>
<td>-1</td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
<td>-39.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>transcriptional regulator PtX5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-1.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2291</td>
<td>141</td>
<td>341</td>
<td>0.19</td>
<td>PA2291</td>
<td>probable glucose-sensitive porin</td>
<td>-1</td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
<td>-39.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-3.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2310</td>
<td>120</td>
<td>289</td>
<td>0.25</td>
<td>PA2310</td>
<td>hypothetical protein</td>
<td>-1</td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
<td>-52.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2318</td>
<td>156</td>
<td>289</td>
<td>0.25</td>
<td>PA2318</td>
<td>hypothetical protein</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>-30.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>probable transposase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>31.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2345</td>
<td>141</td>
<td>412</td>
<td>0.14</td>
<td>PA2345</td>
<td>hypothetical protein</td>
<td>-1</td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
<td>-49.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2372</td>
<td>121</td>
<td>347</td>
<td>0.19</td>
<td>PA2372</td>
<td>hypothetical protein</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>-41.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2377</td>
<td>219</td>
<td>414</td>
<td>0.14</td>
<td>PA2377</td>
<td>hypothetical protein</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>-47.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>probable aldehyde dehydrogenase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2463</td>
<td>133</td>
<td>312</td>
<td>0.22</td>
<td>PA2463</td>
<td>hypothetical protein</td>
<td>-1</td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
<td>-49.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA2505</td>
<td>120</td>
<td>362</td>
<td>0.18</td>
<td>PA2505</td>
<td>probable porin</td>
<td>-1</td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
<td>-50.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.82</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

206
<table>
<thead>
<tr>
<th>PA2505-PA2506_1</th>
<th>121</th>
<th>362</th>
<th>0.18</th>
<th>PA2505</th>
<th>probable porin</th>
<th>-1</th>
<th>PA2506</th>
<th>hypothetical protein</th>
<th>1</th>
<th>-38.18</th>
<th>-3.34</th>
<th>0.95</th>
<th>1.07</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA2633-PA2634_1</td>
<td>120</td>
<td>594</td>
<td>0.06</td>
<td>PA2633</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA2634</td>
<td>isocitrate lyase</td>
<td>1</td>
<td>-20.01</td>
<td>-1.93</td>
<td>0.97</td>
<td>10.18</td>
</tr>
<tr>
<td>PA2678-PA2679_1</td>
<td>161</td>
<td>339</td>
<td>0.20</td>
<td>PA2678</td>
<td>probable permease of ABC-2 transporter</td>
<td>-1</td>
<td>PA2679</td>
<td>hypothetical protein</td>
<td>1</td>
<td>-47.80</td>
<td>-2.41</td>
<td>0.93</td>
<td>1.26</td>
</tr>
<tr>
<td>PA2683-PA2684_1</td>
<td>119</td>
<td>324</td>
<td>0.21</td>
<td>PA2683</td>
<td>serine/threonine dehydratase</td>
<td>-1</td>
<td>PA2684</td>
<td>hypothetical protein</td>
<td>1</td>
<td>-52.97</td>
<td>-2.14</td>
<td>0.93</td>
<td>13.51</td>
</tr>
<tr>
<td>PA2689-PA2690_1</td>
<td>140</td>
<td>333</td>
<td>0.20</td>
<td>PA2689</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA2690</td>
<td>probable transposase</td>
<td>-1</td>
<td>-33.70</td>
<td>-2.15</td>
<td>0.99</td>
<td>11.41</td>
</tr>
<tr>
<td>PA2689-PA2690_1</td>
<td>140</td>
<td>333</td>
<td>0.20</td>
<td>PA2689</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA2690</td>
<td>probable transposase</td>
<td>-1</td>
<td>-42.98</td>
<td>-1.25</td>
<td>0.97</td>
<td>3.05</td>
</tr>
<tr>
<td>PA2713-PA2714_1</td>
<td>121</td>
<td>353</td>
<td>0.18</td>
<td>PA2713</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA2714</td>
<td>probable molybdopterin oxidoreductase</td>
<td>1</td>
<td>-42.30</td>
<td>-2.44</td>
<td>0.94</td>
<td>1.51</td>
</tr>
<tr>
<td>PA2713-PA2714_1</td>
<td>121</td>
<td>353</td>
<td>0.18</td>
<td>PA2713</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA2714</td>
<td>probable molybdopterin oxidoreductase</td>
<td>1</td>
<td>-42.30</td>
<td>-2.44</td>
<td>0.94</td>
<td>1.51</td>
</tr>
<tr>
<td>PA2733-PA2734_1</td>
<td>185</td>
<td>304</td>
<td>0.23</td>
<td>PA2733</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA2734</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>-21.09</td>
<td>-3.37</td>
<td>0.99</td>
<td>39.54</td>
</tr>
<tr>
<td>PA2746-PA2747_1</td>
<td>161</td>
<td>422</td>
<td>0.13</td>
<td>PA2746</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA2747</td>
<td>hypothetical protein</td>
<td>1</td>
<td>-38.58</td>
<td>-2.34</td>
<td>0.95</td>
<td>1.25</td>
</tr>
<tr>
<td>PA2750-PA2751_1</td>
<td>129</td>
<td>368</td>
<td>0.17</td>
<td>PA2750</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA2751</td>
<td>hypothetical protein</td>
<td>1</td>
<td>-38.63</td>
<td>-2.39</td>
<td>0.96</td>
<td>3.67</td>
</tr>
<tr>
<td>PA2751-PA2752_1</td>
<td>161</td>
<td>395</td>
<td>0.15</td>
<td>PA2751</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA2752</td>
<td>hypothetical protein</td>
<td>1</td>
<td>-67.35</td>
<td>-2.14</td>
<td>0.95</td>
<td>3.26</td>
</tr>
</tbody>
</table>

207
<table>
<thead>
<tr>
<th>Gene1</th>
<th>Gene2</th>
<th>similarity</th>
<th>Gene1 description</th>
<th>Gene2 description</th>
<th>OPLS-DA Score1</th>
<th>OPLS-DA Score2</th>
<th>OPLS-DA Score3</th>
<th>OPLS-DA Score4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA2770-PA2771_1</td>
<td>183</td>
<td>289</td>
<td>0.25</td>
<td>PA2770 hypothetical protein</td>
<td>-1</td>
<td>PA2771 hypothetical protein</td>
<td>1</td>
<td>-48.95</td>
</tr>
<tr>
<td>PA2896-PA2897_1</td>
<td>198</td>
<td>297</td>
<td>0.24</td>
<td>PA2896 RNA polymerase sigma factor</td>
<td>-1</td>
<td>PA2897 probable transcriptional regulator</td>
<td>-1</td>
<td>-48.38</td>
</tr>
<tr>
<td>PA3261-PA3262_1</td>
<td>141</td>
<td>585</td>
<td>0.06</td>
<td>PA3261 hypothetical protein</td>
<td>1</td>
<td>PA3262 probable peptidyl-prolyl cis-trans</td>
<td>-1</td>
<td>-52.37</td>
</tr>
<tr>
<td>PA3325-PA3326_1</td>
<td>138</td>
<td>343</td>
<td>0.19</td>
<td>PA3325 hypothetical protein</td>
<td>1</td>
<td>PA3326 ATP-dependent Clp protease proteolytic</td>
<td>-1</td>
<td>-45.72</td>
</tr>
<tr>
<td>PA3325-PA3326_1</td>
<td>146</td>
<td>343</td>
<td>0.19</td>
<td>PA3325 hypothetical protein</td>
<td>1</td>
<td>PA3326 ATP-dependent Clp protease proteolytic</td>
<td>-1</td>
<td>-31.77</td>
</tr>
<tr>
<td>PA3539-PA3540_1</td>
<td>456</td>
<td>709</td>
<td>0.03</td>
<td>PA3539 hypothetical protein</td>
<td>1</td>
<td>PA3540 GDP-mannose 6-dehydrogenase AlgD</td>
<td>1</td>
<td>-32.42</td>
</tr>
<tr>
<td>PA3919-PA3920_1</td>
<td>119</td>
<td>478</td>
<td>0.10</td>
<td>PA3919 hypothetical protein</td>
<td>-1</td>
<td>PA3920 probable metal transporting P-type ATPase</td>
<td>1</td>
<td>-55.60</td>
</tr>
<tr>
<td>PA3939-PA3940_1</td>
<td>120</td>
<td>368</td>
<td>0.17</td>
<td>PA3939 hypothetical protein</td>
<td>-1</td>
<td>PA3940 probable DNA binding protein</td>
<td>1</td>
<td>-34.54</td>
</tr>
<tr>
<td>PA3948-PA3949_1</td>
<td>200</td>
<td>328</td>
<td>0.21</td>
<td>PA3948 probable two-component response regulator</td>
<td>1</td>
<td>PA3949 hypothetical protein</td>
<td>1</td>
<td>-41.68</td>
</tr>
<tr>
<td>PA3993-PA3994_1</td>
<td>251</td>
<td>310</td>
<td>0.23</td>
<td>PA3993 probable transposase</td>
<td>1</td>
<td>PA3994 probable epoxide hydrolase</td>
<td>-1</td>
<td>-36.31</td>
</tr>
<tr>
<td>PA4033-PA4034_1</td>
<td>201</td>
<td>346</td>
<td>0.19</td>
<td>PA4033 hypothetical protein</td>
<td>1</td>
<td>PA4034 aquaporin Z</td>
<td>1</td>
<td>-50.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>PA4138-</td>
<td>120</td>
<td>315</td>
<td>0.22</td>
<td>PA4138</td>
<td>tyrosyl-tRNA synthetase</td>
<td>-1</td>
<td>PA4139</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>PA4138-</td>
<td>138</td>
<td>297</td>
<td>0.24</td>
<td>PA4138</td>
<td>tyrosyl-tRNA synthetase</td>
<td>-1</td>
<td>PA4139</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>PA4154-</td>
<td>201</td>
<td>435</td>
<td>0.12</td>
<td>PA4154</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA4155</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>PA4204-</td>
<td>115</td>
<td>547</td>
<td>0.07</td>
<td>PA4204</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA4205</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>PA4209-</td>
<td>237</td>
<td>307</td>
<td>0.23</td>
<td>PA4209</td>
<td>probable phenazine-specific methyltransferase</td>
<td>-1</td>
<td>PA4210</td>
<td>probable phenazine biosynthesis</td>
</tr>
<tr>
<td>PA4360-</td>
<td>121</td>
<td>468</td>
<td>0.11</td>
<td>PA4360</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA4361</td>
<td>probable oxidoreductase</td>
</tr>
<tr>
<td>PA4385-</td>
<td>200</td>
<td>538</td>
<td>0.08</td>
<td>PA4385</td>
<td>chaperonin GroEL</td>
<td>-1</td>
<td>PA4387</td>
<td>FxsA</td>
</tr>
<tr>
<td>PA4570-</td>
<td>177</td>
<td>314</td>
<td>0.22</td>
<td>PA4570</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA4571</td>
<td>probable cytochrome c</td>
</tr>
<tr>
<td>PA4613-</td>
<td>141</td>
<td>316</td>
<td>0.22</td>
<td>PA4613</td>
<td>catalase</td>
<td>-1</td>
<td>PA4614</td>
<td>large-conductance mechanosensitive</td>
</tr>
<tr>
<td>PA4623-</td>
<td>141</td>
<td>300</td>
<td>0.24</td>
<td>PA4623</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA4624</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>PA4796-</td>
<td>121</td>
<td>337</td>
<td>0.20</td>
<td>PA4796</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA4797</td>
<td>probable transposase</td>
</tr>
<tr>
<td>Gene1</td>
<td>Gene2</td>
<td>BLAST Score</td>
<td>Gene1 Function</td>
<td>Gene2 Function</td>
<td>Distance</td>
<td>Score1</td>
<td>Score2</td>
<td>Score3</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>PA4879</td>
<td>PA4880</td>
<td>141</td>
<td>295</td>
<td>0.24</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA4880</td>
<td>probable bacterioferritin</td>
</tr>
<tr>
<td>PA4913</td>
<td>PA4914</td>
<td>140</td>
<td>337</td>
<td>0.20</td>
<td>probable binding protein component of ABC transporter</td>
<td>-1</td>
<td>PA4914</td>
<td>probable transcriptional regulator</td>
</tr>
<tr>
<td>PA5019</td>
<td>PA5020</td>
<td>120</td>
<td>302</td>
<td>0.23</td>
<td>hypothetical protein</td>
<td>-1</td>
<td>PA5020</td>
<td>probable acyl-CoA dehydrogenase</td>
</tr>
<tr>
<td>PA5023</td>
<td>PA5024</td>
<td>121</td>
<td>383</td>
<td>0.16</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA5024</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>PA5100</td>
<td>PA5101</td>
<td>141</td>
<td>325</td>
<td>0.21</td>
<td>urocanate hydratase</td>
<td>-1</td>
<td>PA5101</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>PA5118</td>
<td>PA5119</td>
<td>272</td>
<td>311</td>
<td>0.22</td>
<td>thiamine biosynthesis protein Thi</td>
<td>-1</td>
<td>PA5119</td>
<td>glutamine synthetase</td>
</tr>
<tr>
<td>PA5125</td>
<td>PA5126</td>
<td>141</td>
<td>424</td>
<td>0.13</td>
<td>two-component response regulator NtrC</td>
<td>1</td>
<td>PA5126</td>
<td>hypothetical protein</td>
</tr>
<tr>
<td>PA5156</td>
<td>PA5157</td>
<td>128</td>
<td>367</td>
<td>0.17</td>
<td>hypothetical protein</td>
<td>1</td>
<td>PA5157</td>
<td>probable transcriptional regulator</td>
</tr>
<tr>
<td>PA5169</td>
<td>PA5170</td>
<td>197</td>
<td>577</td>
<td>0.06</td>
<td>probable C4-dicarboxylate transporter</td>
<td>1</td>
<td>PA5170</td>
<td>arginine/ornithine antiporter</td>
</tr>
<tr>
<td>PA5252</td>
<td>PA5253</td>
<td>137</td>
<td>296</td>
<td>0.24</td>
<td>probable ATP-binding component of ABC transporter</td>
<td>-1</td>
<td>PA5253</td>
<td>alginate regulatory protein AlgP</td>
</tr>
<tr>
<td>Gene ID 1</td>
<td>Gene ID 2</td>
<td>Gene ID 3</td>
<td>Gene ID 4</td>
<td>Protein/Enzyme</td>
<td>Function</td>
<td>Log2 Fold Change</td>
<td>Log2 Fold Change</td>
<td>p-value</td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>----------------</td>
<td>-----------------------------------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>---------</td>
</tr>
<tr>
<td>PA5288-</td>
<td>PA5289_1</td>
<td>241</td>
<td>374</td>
<td>0.17</td>
<td>PA5288, nitrogen regulatory protein P-II 2</td>
<td>-1</td>
<td>PA5289 protein</td>
<td>1</td>
</tr>
<tr>
<td>PA5304-</td>
<td>PA5305_1</td>
<td>356</td>
<td>356</td>
<td>0.18</td>
<td>PA5304, D-amino acid dehydrogenase small subunit</td>
<td>-1</td>
<td>PA5305 protein</td>
<td>-1</td>
</tr>
<tr>
<td>PA5321-</td>
<td>PA5322_1</td>
<td>121</td>
<td>335</td>
<td>0.20</td>
<td>PA5321, deoxyuridine 5'-triphosphate nucleotidohydrolase</td>
<td>1</td>
<td>PA5322 protein</td>
<td>1</td>
</tr>
<tr>
<td>PA5379-</td>
<td>PA5380_1</td>
<td>151</td>
<td>390</td>
<td>0.15</td>
<td>PA5379, L-serine dehydratase</td>
<td>1</td>
<td>PA5380 protein</td>
<td>1</td>
</tr>
<tr>
<td>PA5445-</td>
<td>PA5446_1</td>
<td>140</td>
<td>377</td>
<td>0.16</td>
<td>PA5445, probable coenzyme A transferase</td>
<td>1</td>
<td>PA5446 protein</td>
<td>1</td>
</tr>
<tr>
<td>PA5565-</td>
<td>PA5566_1</td>
<td>121</td>
<td>473</td>
<td>0.10</td>
<td>PA5565, tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA</td>
<td>-1</td>
<td>PA5566 protein</td>
<td>-1</td>
</tr>
</tbody>
</table>