COMPUTATIONAL PREDICTION AND CHARACTERIZATION OF GENOMIC ISLANDS: INSIGHTS INTO BACTERIAL PATHOGENICITY

by

Morgan Gavel Ira Langille
B.Sc. University of New Brunswick, 2004
B.CS University of New Brunswick, 2004

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

DOCTOR OF PHILOSOPHY

In the
Department of Molecular Biology and Biochemistry

SIMON FRASER UNIVERSITY

Summer 2009

APPROVAL

Name: **Morgan Gavel Ira Langille**

Degree: **Doctor of Philosophy**

Title of Thesis: **Computational prediction and characterization of genomic islands: insights into bacterial pathogenicity**

Examining Committee:

Chair: **Dr. Paul C.H. Li**
Associate Professor, Department of Chemistry

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

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

Dr. Frederic F. Pio
Supervisor
Assistant Professor of Molecular Biology and Biochemistry

Dr. Jack Chen
Internal Examiner
Associate Professor of Molecular Biology and Biochemistry

Dr. Steven Hallam
External Examiner
Assistant Professor of Microbiology and Immunology, University of British Columbia

Date Defended/Approved: Thursday April 16, 2009

Declaration of
Partial Copyright Licence

ABSTRACT

Genomic islands (GIs), including pathogenicity islands, are commonly defined as clusters of genes in prokaryotic genomes that have probable horizontal origins. These genetic elements have been associated with rapid adaptations in prokaryotes that are of medical, economical or environmental importance, such as pathogen virulence, antibiotic resistance, symbiotic interactions, and notable secondary metabolic capabilities. As the number of genomic sequences increases, the impact of GIs in prokaryotic evolution has become more apparent and detecting these regions using bioinformatics approaches has become an integral part of studying microbial evolution and function. In this dissertation, I describe a novel comparative genomics approach for identifying GIs, called IslandPick, and the application of this method to construct robust datasets that were used to test the accuracy of several previously published GI prediction programs. In addition, I will discuss the features of a new GI web resource, called IslandViewer, which integrates the most accurate GI predictors currently available. Further, the role of several GI and prophage regions and their involvement in virulence in an epidemic *Pseudomonas aeruginosa* strain that infects cystic fibrosis patients will be described; as well as an observation that recently discovered phage defence elements, CRISPRs, are over-represented within GIs.

Keywords:
bioinformatics; genomic islands; horizontal gene transfer; phylogenomics; comparative genomics; evolution; bacteria; archaea; pathogenesis; phage

In memory of my uncle,

Dr. Stephen Kerr.

I wish you were here to read this.

ACKNOWLEDGEMENTS

I would like to express my sincerest thanks to my supervisor, Dr. Fiona Brinkman, for her support, guidance, and great insight. In addition, I would like to thank my committee members Dr. David Baillie and Dr. Frederic Pio for their positive suggestions and guidance. I would like to acknowledge all of the collaborators from the LES project, including, Drs. Craig Winstanley, Roger Levesque, Bob Hancock, and Nicholas Thomson. Furthermore, I would like to thank all of the members of the Brinkman Lab especially Dr. William Hsiao for sharing his knowledge on genomic islands. In addition, I would like to thank both the supervisors and students of the Bioinformatics Training Program with a special acknowledgment to Benjamin Good for the many discussions and honest opinions on academic life and science.

Lastly, I would like to thank my parents and brothers for always supporting me, and my wonderful wife Sylvia and son Gavin, who made this journey a truly enjoyable adventure.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

Accuracy	(TP+TN)/(TP+FP+TN+FN)
BLAST	basic local alignment search tool
bp	base pair (i.e. a nucleotide)
CF	Cystic fibrosis
COG	cluster of orthologous groups
FASTA	FAST-All; nucleotide and protein sequence alignment program
FASTA format	a common sequence format: definition line followed by sequence
FN	false negative
FP	false positive
GI	genomic island
GNU GPL	GNU general public license (open source software license)
HGT	horizontal gene transfer
HMM	Hidden Markov Model
IS	insertion sequence
kb	kilobase (1000 bp)
Mb	megabase (1,000,000 bp)
LES	Liverpool Epidemic Strain
ORF	open reading frame
PAI	pathogenicity island

Precision	TP/(TP+FP); usually used equivalently to specificity
Recall	TP/(TP+FN); usually used equivalently to sensitivity
Specificity	TP/(TP+FP)
Sensitivity	TP/(TP+FN);
STM	signature tagged mutagenesis
TN	true negative
TP	true positive
VFDB	virulence factors database

CHAPTER 1 INTRODUCTION

Portions of this chapter have been previously published in the book chapter "Mobile genetic elements and their prediction", co-authored by M.G.I. Langille, F. Zhou, A. Fedynak, W.W.L. Hsiao, Y. Xu, and F.S.L. Brinkman In Y. Xu and J.P. Gogarten (eds.), "Computational Methods for Understanding Bacterial and Archaeal Genomes", Series on Advances in Bioinformatics and Computational Biology, Vol. 7. Imperial College Press, London, 2008 ©2008 Imperial College Press

1.1 Horizontal gene transfer

Bacteria are the most abundant Domain of life that exists on earth (based on biomass) (Suttle, 2005). The species we see today are highly diverse, reflecting adaptations to a wide range of environments over billions of years. One of the major sources of adaptability for bacteria is the ability to obtain genes horizontally from other sources, including other prokaryotes, viruses, and even eukaryotes (Ochman, et al., 2000). Horizontal gene transfer (HGT) can occur by one of three major mechanisms: transformation, conjugation, and transduction.

Transformation is the process by which bacteria uptake naked DNA from their environment (Griffiths, 1928). This transfer method has been shown to be naturally present across various taxa from both the Bacteria and Archaea Domains of life (Lorenz and Wackernagel, 1994). Any cell that is able to uptake naked DNA is considered "competent". This competence state is often an inducible phenotype in response to an environmental stimulus, while some strains exhibit constant competence such as *Neisseria gonorrhoeae* and *Haemophilus influenza* (Dubnau, 1999). The process of transformation starts with

double stranded DNA binding to sites on the cell surface. The DNA is then translocated in single strand form into the cell by a series of proteins, many of which are related to the type IV pili and type II secretion systems (Chen and Dubnau, 2004).

Conjugation is the process by which a donor cell physically joins with a recipient cell and passes DNA through a cell to cell bridge or mating pilus (Lederberg and Tatum, 1946). The DNA substrate that is passed by conjugation is typically a plasmid, but can also be a transposon (see section 1.2.3 below). Those elements that encode the conjugation machinery are referred to as self-transmissible, while those that depend on externally encoded conjugation systems are called mobilizable. The process starts with the extension of the sex pilus from the donor cell to the recipient cell, which has recently been shown to occur at considerable distances (Babic, et al., 2008). The substrate, typically single stranded DNA from a replicating rolling circle, is transferred by a type IV-like secretion system into the recipient cell (Christie, 2001).

Transduction is the movement of DNA by a virus that infects prokaryotic cells, known as a bacteriophage or simply phage. Phage can be divided to into two general groups depending on whether they possess the ability to become dormant, called temperate phage, or if upon infection of the host their only choice is to enter a lytic cycle (the production of phage progeny), called virulent phage (Lwoff, 1953). The dormant phage, upon invading the bacterial cell, will often integrate its own DNA into the bacterium's genome becoming a prophage (Freifelder and Meselson, 1970) and will be replicated for numerous generations

along with the bacterial genome (see section 1.2.1 below). Induction provokes dormant prophage to enter a complete lytic cycle, and this may happen spontaneously or because of change in the bacteria's environmental conditions. Prokaryotic DNA can be horizontally transferred by either generalized or specialized transduction. Generalized transduction occurs when random host DNA fragments mistakenly become packaged into the phage particle during the lytic cycle. Specialized transduction occurs when the host DNA flanking an integrated phage is replicated during phage induction and becomes integrated into the phage particle.

1.2 Mobile genetic elements

Mobile genetic elements (MGEs), such as transposons, integrons, prophage, insertion sequence (IS) elements, and genomic islands (GIs), are regions of DNA that are able to move themselves throughout the genome of a single organism or between organisms. These elements all share three common hurdles to their proliferation. First, the genetic element must be excised from the host genome into either an RNA or DNA molecule. Second, that element must be transmitted between organisms via HGT or within an organism and be ready for integration as a DNA molecule. Third, the element must then be integrated into a replicon in a new location. These elements form the basis of important mechanisms of evolution that result in the transfer, rearrangement or deletion of genes. In addition, many of these elements result in non-Darwinian evolution by allowing genes to be exchanged through HGT and question whether the "Tree of Life" would be better represented as a network (Doolittle and Bapteste, 2007).

1.2.1 Prophage

A prophage is the latent form of a prokaryotic virus known as a phage and the movement of DNA between prokaryotic cells via a phage is referred to as transduction (see section 1.1 above). These integrated prophage account for a large portion of the variation seen between bacterial strains (Ohnishi, et al., 2001) and can represent as much as 10-20% of the genes in a bacterial genome (Casjens, 2003; Casjens, et al., 2000). Furthermore, virulence factors that contribute to a bacterium's pathogenicity, such as cholera toxin in *Vibrio cholerae*, can be mobilized by phage and are seen as a key factor in the evolution of new pathogens (Boyd and Brussow, 2002).

Prophage regions typically contain an integrase and several phage associated genes. However, they can often carry other genes that are not associated with the proliferation of the phage. Similarly to GIs (see below), the presence of a tRNA or a flanking direct repeat is supportive evidence that phage integration may have occurred in a region since these are often common integration sites for phage.

1.2.2 Integrons

Integrons are genetic elements that utilize site-specific recombination to capture and direct expression of exogenous open reading frames (ORFs). They were first identified in the late 1980's for their important role in the capture and spread of antibiotic resistance genes (Stokes and Hall, 1989). Bacteria harbouring integrons possess the ability to incorporate and express genes with potentially adaptive functions, including antibiotic resistance genes, and therefore

pose a major problem for treatment of infectious diseases (Rowe-Magnus, et al., 2002). Furthermore, some bacteria become resistant to multiple antibiotics by harbouring integrons that have captured multiple antibiotic resistance genes and, potentially, genes encoding other traits that give the bacteria an adaptive advantage. Additionally, integrons are often linked with other MGEs, such as plasmids and transposons, leading to rapid dissemination of such traits within a population. In 2007, it was estimated that approximately 10% of the partially or completely sequenced genomes in the Bacteria domain contained integrons (Boucher, et al., 2007), making them an important player in acquisition and spread of adaptive traits and antibiotic resistance in bacterial populations.

Integrons consist of three key elements necessary for the capture and expression of exogenous ORFs: An integrase gene (*intI*) and recombination site (*attI*) are necessary for acquisition of genes, and a promoter (*Pc*) ensures their expression. *IntI*, *attI* and *Pc* comprise the 5' conserved segment (5'CS), and the 3' conserved segment (3'CS) contains known genes that confer resistance to various compounds or provide additional metabolic function (Figure 1.1). *IntI* catalyzes the recombination between *attI* and a recombination site at the 3' end of the gene called *attC* or the 59-base element (59-be). The 59-be consists of a variable region spanning 45-128 nucleotides in length flanked by imperfect inverted repeats at the ends designated R' (GTTRRRY) and R'' (RYYYAAC), where R is a purine and Y a pyrimidine. The recombination site in the 59-be recognized by *intI* is between the G and T bases of R'. An ORF and its associated 59-be is termed a gene cassette. These gene cassettes have been

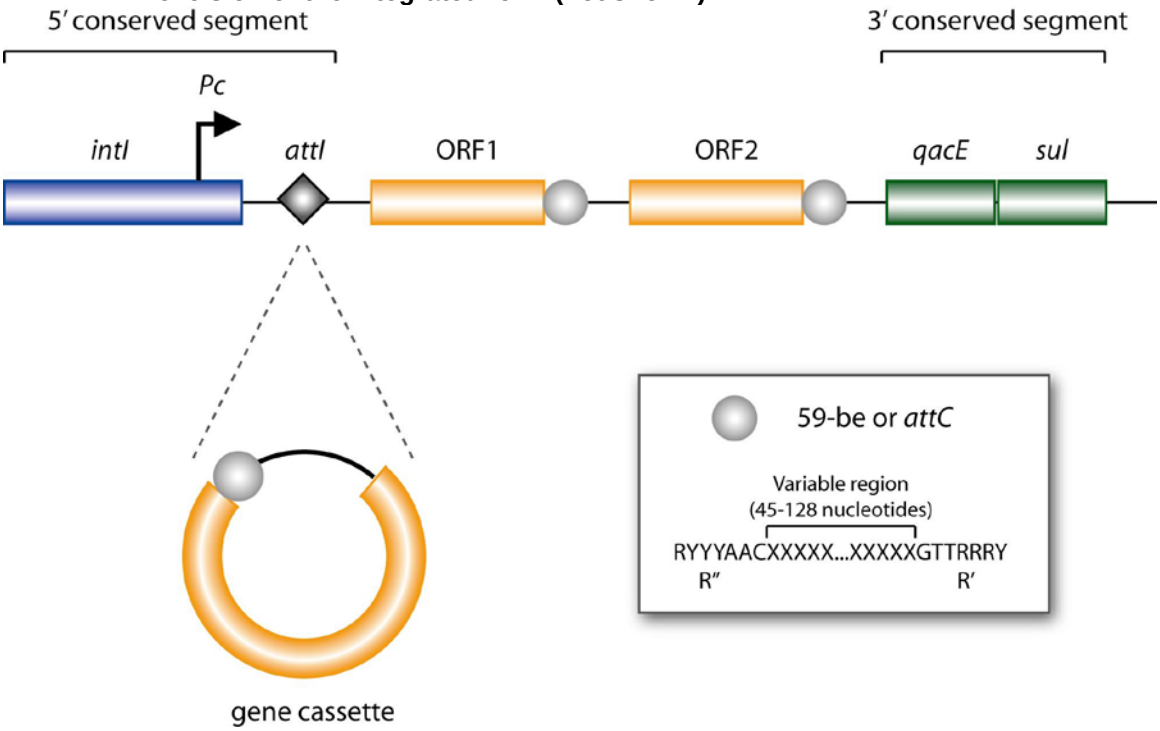
shown to be excised as covalently closed circles that may contain more than one gene cassette linked together (Collis and Hall, 1992).

All integrons characterized to date are classified as either integrons or superintegrons. Integrons are defined as gene cassettes associated with MGEs such as insertion sequences, transposons, and conjugative plasmids, which serve to disseminate genes through mechanisms of HGT. Five classes of integrons have been described, classified based on sequence homology of their integrase genes (Mazel, 2006). Class 1 integrons are the most clinically relevant, isolated frequently from patients with bacterial infections. Bacteria with class 1 integrons often confer multi-antibiotic resistance and possess gene cassettes resistant to a wide variety of antibiotics, including all known β-lactam antibiotics (Mazel, 2006). One such class 1 integron was identified in *E. coli* that contains 8 different antibiotic resistance cassettes including a broad-spectrum β-lactamase gene of clinical importance (Naas, et al., 2001). Association with MGEs can lead to rapid dissemination of integrons and their associated gene cassettes through both intraspecies and interspecies transfer. In support of this, extensive reports have identified integrons in diverse Gram-negative bacteria and in some Gram-positives (Hall, et al., 1999; Mazel, 2006).

Superintegrons differ from integrons in that they are chromosomally located and not linked to MGEs. They also differ in that their cassette arrays can be quite large; one unique superintegron identified in *Vibrio cholerae* harbours over 170 cassettes (Mazel, et al., 1998; Rowe-Magnus, et al., 1999).

In addition to antibiotic resistance genes, integron and superintegron gene cassettes have been shown to encode proteins involved in other adaptive functions, including virulence factors, metabolic genes, and restriction enzymes (Ogawa and Takeda, 1993; Rowe-Magnus, et al., 2001; Vaisvila, et al., 2001). However, a recent study reported that 78% of cassette-encoded genes are uncharacterized or have no known homologs to date (Boucher, et al., 2007).

Figure 1.1 Schematic representation of a class 1 integron.
IntI, integrase gene; attI , integration site; Pc, promoter for expression of integrated gene cassettes; 59-be (attC), site adjacent to ORF recognized by intI; sul, sulphonamide resistance; qacE, quaternary ammonium compound resistance; 59-be, 59 base element. Note that the circular cassette comes from excision of the integrated form (not shown).



1.2.3 Transposons and IS elements

Barbara McClintock was the first to have observed recurring chromosomal breakages in the same region caused by a genetic element, *Ds* (Dissociation), in maize in early 1940s (McClintock, 1941). She later found another element, *Ac* (Activator), in maize that must be present for the *Ds* element to exert chromosomal breakage. These two elements were later proposed to be the autonomous (*Ac*) and non-autonomous (*Ds*) members of the same transposon family (Fedoroff, et al., 1983). More generally, transposons are DNA elements having lengths ranging from a few hundred base pairs (bps) to more than 65,000 bps, that proliferate in the host genome and have been observed in all three domains of life; bacteria, archaea and eukaryotes.

Each group of transposons may consist of autonomous and non-autonomous members. An autonomous transposon encodes transposition catalyzing enzymes, called transposases, and is able to transpose itself. A non-autonomous transposon does not encode such proteins and relies on its autonomous counterparts with similar *cis* signals to transpose it. Movement of transposons is usually limited to within a single cell, but they are often contained within other MGEs such as GIs and prophages that allow for cell-to-cell transfer. Of course, as with any genomic region, transposons could also be transferred between naturally competent cells via transformation. In addition, some transposons called conjugative transposons can move via conjugation and I will discuss these at the end of this section.

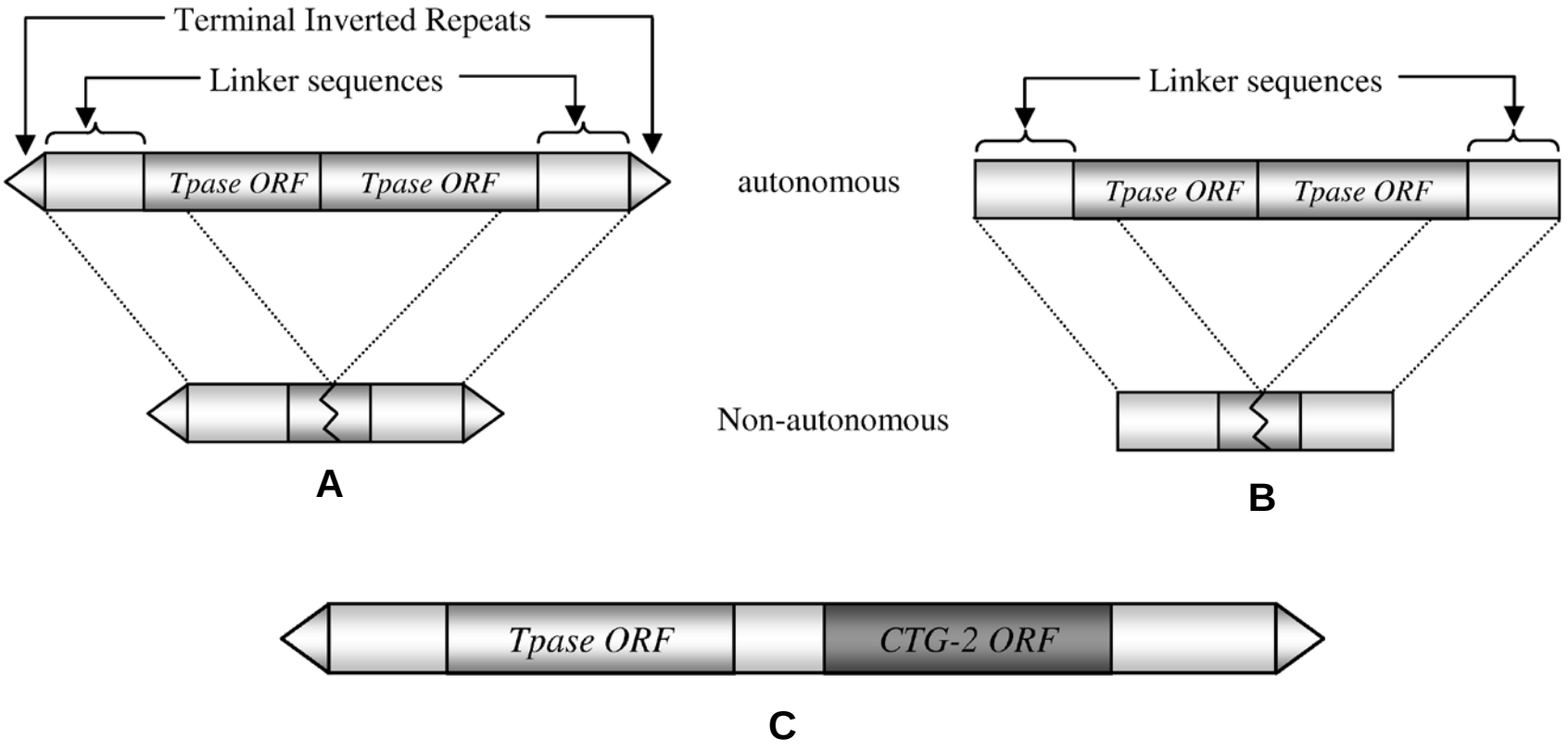
Insertion Sequences (IS elements) are similar to autonomous DNA transposons, in that they encode a transposase, but unlike transposons they do not encode any genes contributing to the phenotype of the host and are typically much smaller than transposons (Adhya and Shapiro, 1969; Shapiro, 1969; Shapiro and Adhya, 1969). As of today, more than 1,500 IS elements have been identified and they are classified into 20 families, with some families being subdivided into groups, based on their genetic structures and the sequence similarities of the encoded transposases (Siguier, et al., 2006). Recent studies suggest that ~99 % of known IS elements in prokaryotes have fewer than 100 copies in their host genomes (Siguier, et al., 2006).

A transposon consists of one or more overlapping genes, one of which may be a transposase (Chandler and Mahillon, 2002; Mahillon and Chandler, 1998; Siguier, et al., 2006), as shown in Figure 1.2. Additional genes may follow, which may alter the host phenotype such as antibiotic resistance genes (Stokes, et al., 2007). Most transposons carry a pair of *terminal inverted repeats* (TIRs) (shorter than 50 bps) at the two termini, and they are termed TIR transposons (Figure 1.2A) while a non-TIR transposon (Figure 1.2B) does not harbour such TIR signals at the termini. Linker sequences are located between each terminal signal and the ORF region.

The relocation of transposons could be deleterious to the host as they may disrupt host genes by inserting into them and may alter the expression of the neighbouring genes with their endogenous promoters (Chandler and Mahillon, 2002; Mahillon and Chandler, 1998). Also, homologous recombination

between two transposons contributes to reorganization and deletion of chromosomal regions in the host genome (Toussaint and Merlin, 2002). After transposons were initially found many studies suggested that transposons were able to introduce beneficial mutations to the host genome through insertion and recombination (Blot, 1994). For example, several studies have shown that transposons can give a selective advantage to the host in specific environments by introducing novel gene mutants in *E. coli* (Lenski, 2004; Naas, et al., 1994; Zambrano, et al., 1993). By taking advantage of such mutagenesis capabilities, transposons have been extensively used in genetic engineering to mediate global insertional mutagenesis of bacteria (Berg, et al., 1984; Ely and Croft, 1982; Rella, et al., 1985; Zink, et al., 1984).

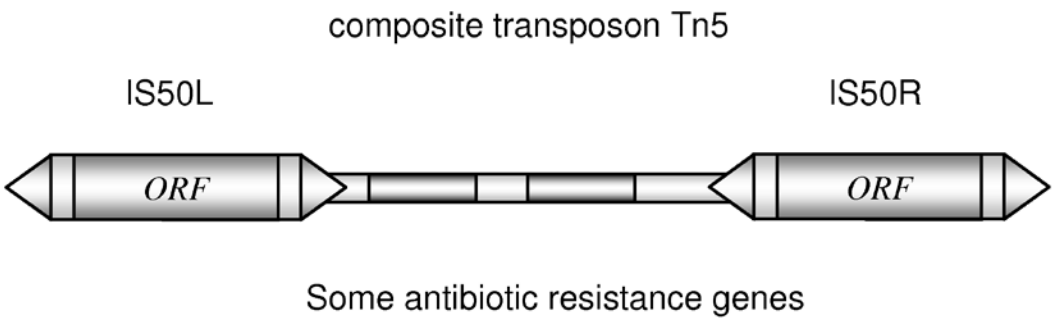
Figure 1.2 Structures of two types of transposons in prokaryotes.
A) TIR (terminal inverted repeat) transposon and B) non-TIR transposon. Both of them have autonomous and non-autonomous members. C) A transposon may also encode proteins other than a transposase.



Two adjacent IS elements, plus intervening DNA sequence, can form a composite transposon as shown in Figure 1.3, which may carry its own protein-encoding genes within the linking DNA sequence, e.g. the antibiotic genes in Tn*5* (Berg, 1989; Reznikof, 2002) and Tn*10* (Haniford, 2002). Several more transposons with much more complex structures, e.g. Tn*3* (Haniford, 2002) and Tn*7* (Craig, 2002), have also been characterized in prokaryotes.

Conjugative transposons (CTns) are MGEs that have features of transposons, plasmids and phage (Clewell and Flannagan, 1993; Scott and Churchward, 1995). As with transposons, conjugative transposons excise and integrate themselves into the genome and are traditionally named under the nomenclature of transposons, e.g. Tn*916* (Franke and Clewell, 1981) and Tn*1545* (Buu-Hoi and Horodniceanu, 1980; Courvalin and Carlier, 1987). However, conjugative transposons are similar to plasmids in that they have a covalently closed circular transfer intermediate that can be transferred by conjugation. This allows conjugative transposons to be integrated within the same cell or between organisms. Contrary to plasmids, conjugative transposons in their circular form cannot autonomously replicate and must become integrated into a prokaryotic genome to maintain their survival (Rice and Carias, 1994; Scott, et al., 1988). Some conjugative transposons have site-specific integration and have integrases that are highly similar to lambdoid phages (Poyart-Salmeron, et al., 1989; Poyart-Salmeron, et al., 1990), but do not form viral particles and therefore are not transferred by transduction.

Figure 1.3 Structure of a composite transposons, Tn5.
The Tn5 composite transposon contains two IS elements, IS50L and IS50R, both of which have terminal inverted repeats at their ends (denoted by triangles). The boxes between these IS elements represent genes that can be carried by the composite transposon and in some cases are antibiotic resistance genes.

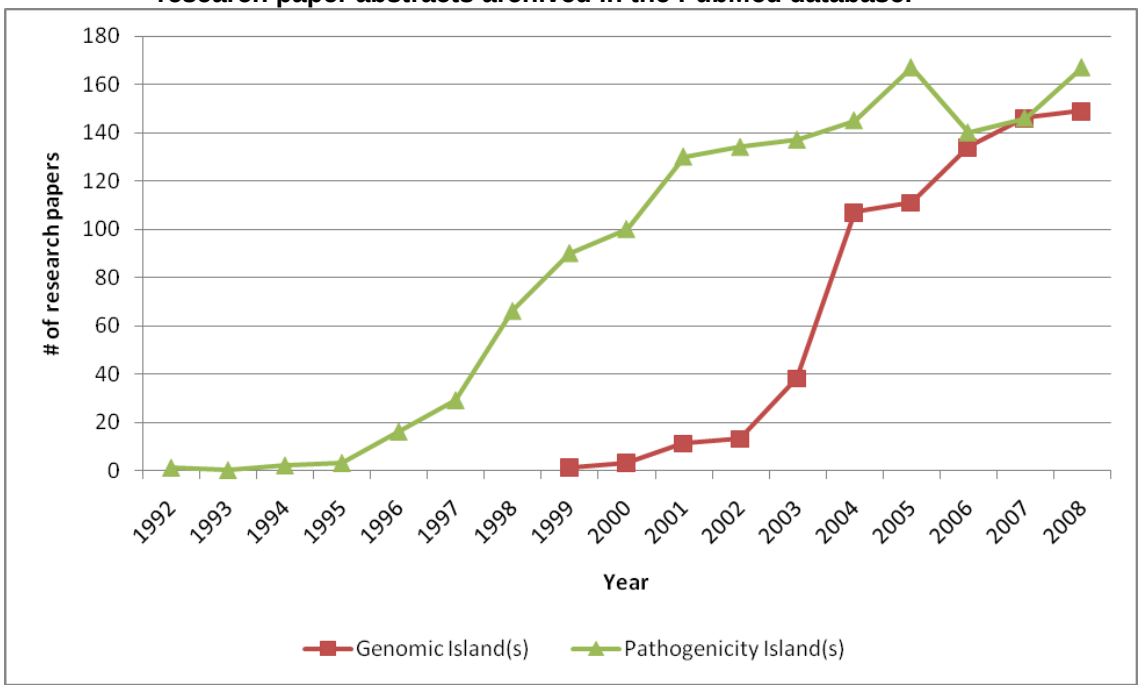


1.2.4 Genomic islands

In 1990, researchers identified many virulence genes clustered together on the chromosome of several *E. coli* strains that were not present in others (Hacker, et al., 1990). These clusters of genes were thought to have been horizontally transferred and based on their association with or presence of virulence determinants were referred to as pathogenicity islands (PAIs). Later studies suggested that other types of islands, besides PAIs, could exist with genes related to other functions such as “secretion islands”, antimicrobial “resistance islands” and “metabolic islands” (Hacker, et al., 1997). GI was then used as a more general term that referred to any cluster of genes, typically 10-200 kilobases in length, with horizontal origins (Hacker and Kaper, 2000). An increase in the use of the terms “pathogenicity islands” and “genomic islands” has continued since these terms were first used (Figure 1.4). This definition of

GIs is broad enough that other mobile genetic elements (MGEs) such as prophage, integrons, conjugative transposons, and integrative conjugative elements have overlapping classifications (Figure 1.5). Typically, many of these other MGEs may be classified as GIs, until further inspection of their mode of integration, site of integration, method of transfer, possible origins, and stability are determined; then a more specific definition can be applied. However, in many cases the transmission mechanism of these genetic elements is not obvious, due to mutations that have obfuscated or destroyed the transmission or integration mechanisms. Therefore, the use of GIs as a generic term is valuable for describing clusters of genes of putative horizontal origin that meet some or all of the criteria listed in Table 1.1, but without a clear mode of transfer or potential for transfer. In understanding this definition, it must therefore be appreciated that GI predictors are usually predicting such generalized regions, which may include chromosomally integrated MGEs such as prophages that have overlapping features.

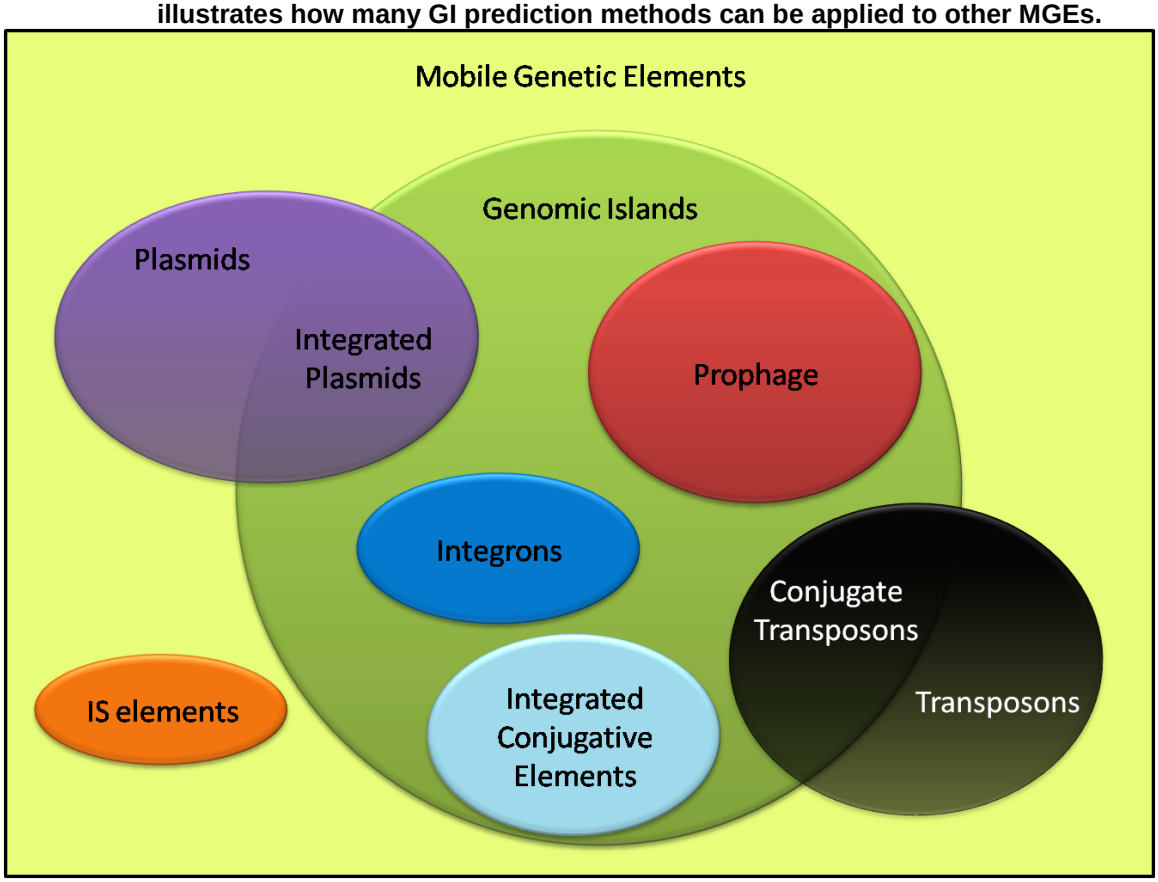
Figure 1.4 Popularity of the terms "genomic islands" and "pathogenicity islands" in research paper abstracts archived in the PubMed database.



The importance of GI prediction should not be underestimated. In this genomic era, where the number of completely sequenced bacteria genomes is increasing rapidly, the identification of GIs in newly sequenced genomes is becoming a common first step in gaining insight into causes of phenotypic differences between species or strains. Links between newly acquired GIs and pathogenic properties continue to be identified since pathogenicity islands were first identified in *E. coli* (Dobrindt, et al., 2004; Gal-Mor and Finlay, 2006; Hacker and Kaper, 2000). In addition, GIs have been found to encode iron uptake functions, type III secretion systems, toxins, and adhesins that augment a pathogen's ability to survive and cause diseases in their host (Dobrindt, et al., 2004; Gal-Mor and Finlay, 2006). Research within Dr. Fiona Brinkman's lab, has additionally recently quantified that, among the genomes sequenced to date,

known virulence factors are over-represented within GIs (unpublished results). New studies emerging also indicate that selective loss/regain of islands may provide an additional means to modulate pathogenicity (Lawrence, 2005; Manson and Gilmore, 2006). Spontaneous excisions of PAIs have been observed in various pathogens resulting in distinct pathogenic phenotypes compared to wild type (Bueno, et al., 2004; Middendorf, et al., 2004). In the case of *Salmonella enterica* serovar Typhi pathogenicity island 7, called SPI7, deletion of this GI is associated with more rapid invasion *in-vitro* and reduced resistance to complement attack (Bueno, et al., 2004). As the genetic requirements for initiation of infection and long-term infection can be quite different, the capability to lose or alter certain genes, such as surface antigens, after the initial infection has been postulated as a means to establish long term colonization and avoid immune detection (Finlay and Falkow, 1997; Gogol, et al., 2007). In addition to this link to virulence, GIs appear to confer many other adaptations of interest to bacteria, including metal resistance, antimicrobial resistance, and secondary metabolic properties of environmental or industrial interest (Dobrindt, et al., 2004). So, the targeted identification of such GI regions in prokaryotic genomes has become of increasing interest.

Figure 1.5 **A general schematic of the class structure of MGE definitions. The fairly broad definition of GIs (large genomic regions with probable horizontal origins) allows several other MGEs to be grouped within GIs and illustrates how many GI prediction methods can be applied to other MGEs.**



GIs share several sequence and structural features that help to distinguish them from the rest of a given prokaryotic genome (Table 1.1, Figure 1.6).

One of the most pronounced features is that their phyletic patterns differ from their host genome, resulting in GIs being sporadically distributed (i.e. only found in some isolates from a given species or strain). Even within a specific strain, there have been several reports showing that GIs are unstable and have the ability to sporadically excise (Hochhut, et al., 2001; Middendorf, et al., 2004). Sequence similarity tools such as BLAST (Altschul, et al., 1997) can be used to search for genomic regions that are present in one particular species/strain, while

being absent in several related species, as a relatively simplistic method for identifying GIs. In addition, whole genome sequence alignment tools such as Mauve (Darling, et al., 2004) can be used to observe conserved genomic regions (based on alignment of multiple related sequences) surrounding apparent newly inserted regions, providing some confidence that a particular region is likely a GI.

Because of the different genome sequence compositions (such as G+C content) that different species lineages or bacteria may exhibit, GIs will often have a sequence composition that is significantly different from their new host genome. Sequence composition-based GI predictors heavily depend on this to identify islands. The simplest measure of sequence composition bias is G+C content (%G+C), but oligo-nucleotides of varying lengths (typically 2-9 nucleotides) are being increasingly used (Karlin, 2001; Karlin, et al., 1998; Sandberg, et al., 2001; Tsirigos and Rigoutsos, 2005; Vernikos and Parkhill, 2006). These measurements are often compared against the average composition of the entire genome and various methods utilize this feature to identify HGT and GIs. However, using only sequence composition bias to identify GIs has several well known flaws. First, highly expressed genes, such as those within ribosomal protein operons, often have a sequence composition that is significantly different from the rest of the genome (Karlin, 2001), resulting in false positive predictions of GIs. Second, any GIs that originated from species with a similar sequence composition as their current host bacterial genome will not be easily detectable. Third, mutational pressure acting on a foreign gene may cause it to adapt to the host genome signature over time in a process termed

"amelioration" (Lawrence and Ochman, 1997), limiting the ability of sequence composition to detect more ancient GI insertions. These problems with sequence composition–based methods for GI prediction can be augmented through the incorporation of other GI features into predictive methods.

Table 1.1 List of features associated with genomic islands

Feature associated with GIs	Method(s) for detection	Benefits and pitfalls when used for GI prediction
-Sporadic distribution -Unstable and can excise spontaneously	-Comparative genomics to identify unique, versus shared regions between genomes	-Requires multiple closely related sequenced genomes for comparison
-Sequence composition bias	-Various tools have been developed to detect bias (see Table 1.2)	- False positive predictions due to highly expressed genes - False negative predictions from gene amelioration
-Adjacent to tRNA	-Detect full or partial tRNAs using BLAST or tRNAscan-SE	- Not all GIs are inserted within tRNAs
-Usually relatively large (>8 kb)	-Comparative genomics to identity large insertions or identifying features such as sequence composition bias over a large region	-Large horizontally acquired regions are easier to predict over single HGTs
-Certain classes of genes are over-represented such as virulence factors, phage-related genes, and genes of unknown functions	-Compare to functional databases such as COG (Clusters of Orthologous Groups) -Not commonly used in GI predictors.	-May allow further sub-classification of GIs into other MGEs such as prophage or integrated plasmids
-Contain mobility genes or elements such as integrases, transposases, IS elements, etc.	-Similarity search of mobility genes using Hidden Markov Models (HMMs) or BLAST	- Can be used as supporting evidence for GI prediction. - Some GIs may not contain a mobility gene. -Annotation of mobility genes may not be available
-Flanked by direct repeats (DRs)	-Use repeat finders such as REPuter(Kurtz and Schleiermacher, 1999)	- Not all GIs are flanked by DRs -Identification of relevant DRs can be difficult due to their size

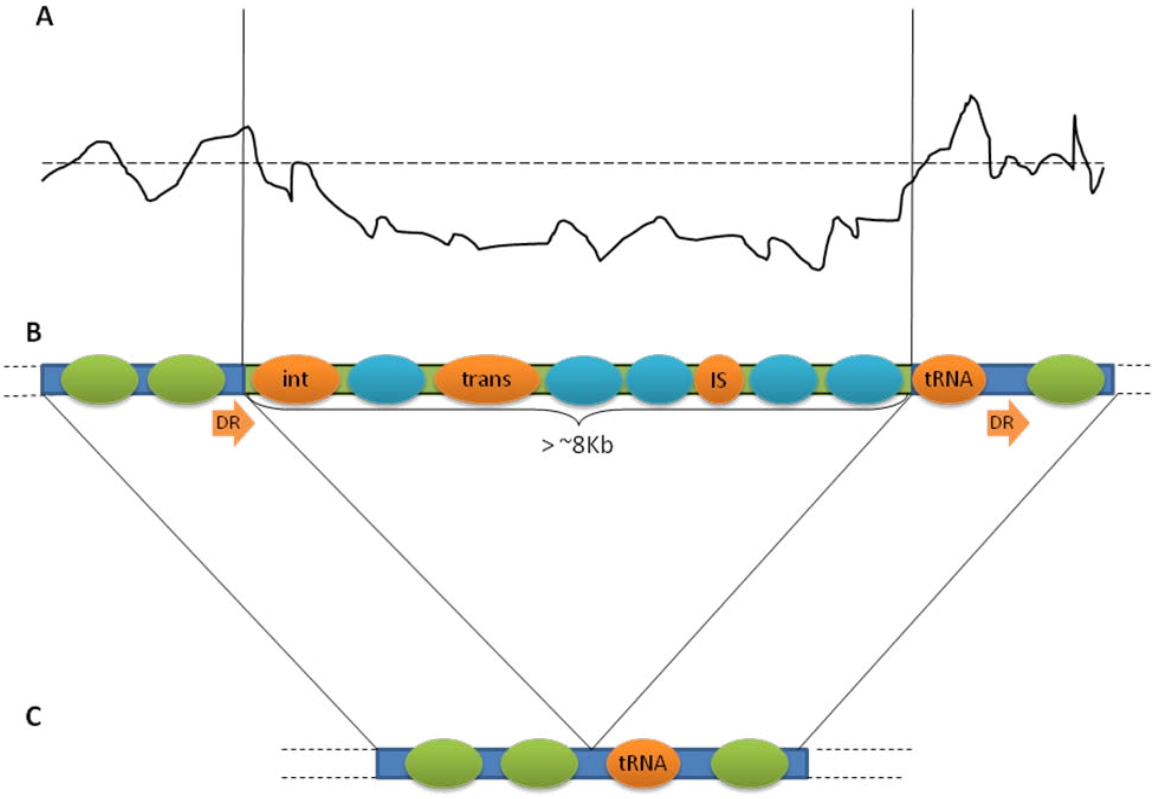
Classically, elements associated with the insertion of some GIs such as tRNA genes and flanking direct repeats can be used as supporting evidence for a GI (Hacker, et al., 1997). tRNA genes are known phage integration sites and direct repeats are often a result of a phage being inserted into a tRNA (Reiter, et al., 1989; Williams, 2002). However, it has been shown that a significant proportion of GIs do not have flanking tRNAs and so using such features to identify GIs only identifies a small subset of them (Hsiao, et al., 2005; Vernikos and Parkhill, 2008).

Certain types of protein coding genes are also associated with GIs. Many of these are related to the mobility of MGEs such as integrases and transposases. These “mobility genes” may indicate that a GI is autonomous or they could reflect remnants of other embedded MGEs such as IS elements that are frequently found in GIs (Hacker, et al., 1997). General functional classes of genes have also been shown to be over-represented within GIs, in particular cell surface proteins, host-interaction proteins, DNA-binding proteins, and (related to the mobility genes mentioned above) phage related proteins (Nakamura, et al., 2004; Vernikos and Parkhill, 2008; Waack, et al., 2006). The most prevalent bias is that GIs disproportionately contain genes with no known homologs or with unknown function (Hsiao, et al., 2005). This latter observation is thought to reflect the phage origin of many GIs and the fact that the gene pool predicted for phage is much larger than for bacteria (Hsiao, et al., 2005; Suttle, 2005).

While it is not necessary for every feature to be present in a region that is named a GI, the simultaneous presence of a subset of these features is generally

viewed as strong evidence for the region's horizontal origin. These regions are often thought to be a single HGT event, but many may be the result of several individual HGT events in the same genomic region. Recent studies have suggested that, apart from phylogeny-based methods, sequence composition bias, phage related genes, GI size, and integrase genes are the most important features for GI prediction, while flanking direct repeats, tRNAs, and gene density were informative but not as important (Vernikos and Parkhill, 2008).

Figure 1.6 **Graphical representation of several genomic features associated with GIs. A) Line plot representing nucleotide composition bias (e.g. %G+C or dinucleotide bias) that deviates significantly from the genome average (dotted line), B) presence of sequence elements (orange) such as mobility genes (transposases, integrases, etc.), IS elements, direct repeats (DR), and tRNAs, C) absence of large region in closely related isolates, strains, or species with conserved flanking genes (green) are all supporting indicators that a particular region is a GI.**



1.3 Detection of genomic islands

There are essentially two main bioinformatics approaches for identifying GIs: sequence composition-based and comparative genomics-based methods (Table 1.2).

1.3.1 Sequenced composition based methods

Sequence composition based approaches for detecting horizontally acquired genetic material have been shown to be a capable and versatile tool for detecting GIs, even considering the problems with false positives and false negatives alluded to above (i.e. missing ameliorated islands or identifying clusters of highly expressed genes as islands). Composition-based methods are desirable because they only require the single genome/sequence being analyzed, whereas comparative genomics-based approaches require that other similar genomes be available for comparison – and the latter is not always available. All of the methods described below essentially calculate the frequency of nucleotide sequences of a certain length, referred to as a k-mer, where k is the length of the sequence (usually from 1 to 9), for a sub-region of a genome and compare these results with the expected frequencies from that genome. Deviation from the genome frequencies is scored and if the score is above a certain cut-off, these regions are marked as putative GIs.

1.3.1.1 SIGI-HMM

SIGI-HMM uses codon usage (frequency of a tri-nucleotide normalized by synonymous codons) as a genome sequence composition signature (Merkl,

2004; Waack, et al., 2006). The codon usage frequency table of an organism is derived either from its whole genome, if available, or from its species' entry in the CUTG codon usage database (Nakamura, et al., 1999). For each gene, the multiplicative-product of the codon usage frequency from each codon in the gene is determined using the organism's own codon frequency table (the host table). The same multiplicative-product is also calculated using the same gene sequence but instead of using the organism's own table, other organisms' frequency tables are used (the donor tables). Lastly, a score in the form of a normalized odds-ratio is calculated from each pair-wise comparison between the product derived from the host table and that derived from a donor table. The score value can be used to decide whether the codon usage of a gene resembles more to the codon prevalence of the host species or to that of another (putative donor) species. In the cases where the sequence resembles another donor more, the gene, if it meets a custom cut-off, is marked as a putative foreign gene. Non-contiguous clusters of putative foreign genes are combined to form a putative GI using a Hidden Markov Model (HMM) (Eddy, 2004). The HMM incorporates an alternative probabilistic model based on randomly generated nucleotide sequences using the same amino acid sequence as the real gene product. This alternative model provides a baseline measure for the random noise in the sample. Moreover, an additional filter to remove potentially highly expressed genes was also incorporated into the HMM using the codon usage of ribosomal proteins as a reference. Using a path-generating algorithm of HMM, a final list of GIs is predicted. All genes assigned to a putative foreign state (i.e.

more similar to a donor frequency table) are considered in GIs and these regions are further combined if there are less than four native (not foreign) genes between them.

1.3.1.2 IslandPath-DINUC and IslandPath-DIMOB

IslandPath-DINUC and IslandPath-DIMOB were two methods that were used to construct datasets using dinucleotide bias, and dinucleotide bias plus the presence of at least a single mobility gene, respectively (Hsiao, et al., 2005). Mobility genes are identified, by searching the genome annotation for terms commonly used to describe mobility genes and by HMMER search of each predicted gene against PFAM mobility gene profiles.

1.3.1.3 PAI-IDA

PAI-IDA uses iterative discriminant analysis on three different genome signatures: %G+C, dinucleotide frequency, and codon usage(Tu and Ding, 2003). Initial window size of 20kb and step size of 5kb were used to calculate the DNA signature of a window compared to the whole genome. A small list of known PAIs from seven genomes was used as the initial training data to generate the parameters used in the linear functions to discriminate anomaly regions from the rest of the genome. Then through iteration, the discriminant function is improved by considering additional (predicted) anomaly regions. The iteration ends if the status of each region stops changing.

1.3.1.4 Centroid

Centroid allows the user to measure sequence composition using various options including k-mer size (1-8) and window size (Rajan, et al., 2007). The authors use a default k-mer size of five and found that by artificially inserting large genomic regions their method was more sensitive than %G+C and dinucleotide methods; especially at larger insert sizes of 20Kb and 40Kb.

1.3.1.5 AlienHunter

AlienHunter uses “Interpolated Variable Order Motifs” (IVOMs) which generates variable length k-mers and prefers longer k-mers over shorter k-mers as long as there is enough information (Vernikos and Parkhill, 2006). The length k is set from one to eight. The program assigns a weight to each k-mer based on its length in order to linearly combine all the k-mer frequencies as a score. The weights are necessary because shorter k-mers are more likely to appear than longer k-mers, but longer k-mers contain more information and are more specific. An HMM is used to refine the boundaries of the HGT regions. The advantage of this approach is in its ability to incorporate variable length k-mers, and based on the developer’s own analyses, longer k-mers provide better sensitivity and specificity than shorter k-mers alone (Vernikos and Parkhill, 2006).

1.3.2 Comparative genomics methods

Comparative genomics based approaches compare multiple genome sequences to detect GIs. While there are not always multiple genomes available for comparison, limiting the use of this method, comparative genomics-based

evidence of an island is preferred when possible. This preference over sequence composition-based methods is due to the problems, alluded to above, with false positive and false negative predictions. This comparative genomics-based approach will likely become more frequently used, as the number and diversity of genome sequences available increases.

Table 1.2 GI prediction programs.

Program Name	Description	Program Availability
SIGI-HMM	Measures codon usage and removes ribosomal regions	Downloadable Graphical Program
IslandPath-DIMOB	Measures dinucleotide bias and presence of at least one mobility gene	Command line program
PAI-IDA	Measures GC, dinucleotide, and codon usage	Command line program
Centroid	Allows various options, but 5-mers are the default	Command line program
Alien Hunter/ IVOM	Uses variable length k-mers	Command line program
Mobilome FINDER	Uses tRNA locations and whole genome alignments to identify GIs	Web Resource

The main premise of the comparative genomics approach is to identify clusters of genes in one genome that are not present in several closely related genomes (Figure 1.6). These regions can be often identified using whole genome alignment methods such as Mauve (Darling, et al., 2004) and Mummer (Delcher, et al., 2002). Genome regions that are aligned across multiple genomes are conserved regions that are unlikely to have horizontal origins (relative to each other), while regions that are unique to a genome (not aligned) can be considered putative GIs for the genome that they reside in. Any comparative genomics based method will rely heavily on the query genome and the available

comparison genomes that are used in the analysis. For example, the inclusion of very distant genomes (with extensive rearrangements) in the comparison could make alignment of genomes difficult and lead to false positive predictions. Using at least one genome that has more recently diverged (i.e. large regions of conserved synteny) may result in more robust predictions of GIs, however, if the genomes are too closely related then GIs that have inserted before the divergence of the genomes will not be predicted. Again, a comparative genomics based approach depends on the availability of several related genomes being already sequenced, and for some genomes, there are no closely related genomes yet available to perform this comparison. However, with the rapid increase of sequenced genomes this limitation would continue to diminish and a comparative genomics approach would likely increase in utility.

1.3.2.1 MobilomeFINDER

When I started my thesis project in 2004 no comparative genomics based GI prediction method had yet to be published. However, since then one other method called MobilomeFINDER has been published (Ou, et al., 2007). MobilomeFINDER focuses on identifying those islands that are associated with tRNAs, a site that GIs often use as integration points. The method starts by identifying shared tRNAs among several related genomes and then uses Mauve to search for GIs within the up- and downstream regions of these orthologous tRNAs (Ou, et al., 2006). The extra requirement of having a tRNA nearby the predicted GI makes this method quite robust; however, not all GIs use tRNAs as insertion sites and so this results in many GIs being missed (Hsiao, et al., 2005).

In addition to this limitation, MobilomeFINDER requires the manual selection of both the query and the comparative genomes as input, which may result in inconsistent selection criteria due to the unfamiliarity of different phylogenetic distances within genera.

1.3.3 Databases and other computational resources

In addition to the GI prediction programs listed above, there are several other computational resources that can be useful in GI research (Table 1.3).

1.3.3.1 IslandPath

IslandPath provides a visual interface to aid researchers in the detection of GIs (Hsiao, et al., 2003). Each gene in the genome is represented as a small circle that has a colour assigned to it based on whether it has significant deviation from the G+C content of the genome. Genes that have unusual dinucleotide bias are also marked with a strikethrough. In addition, any mobility genes and tRNAs are marked with additional shapes. The result is a clickable whole genome graphical view that highlights features that associated with GIs and aids manual identification of putative GIs.

1.3.3.2 MOSAIC

MOSAIC is a database that contains pre-computed whole genome alignments for several bacteria species (Chiapello, et al., 2005). Users can browse and download conserved and “variable” regions for genomes within the database, with the variable regions being potentially GIs.

1.3.3.3 Islander

Islander is a database of 84 GIs and their tRNA integration sites for 106 genomes (Mantri and Williams, 2004). GI predictions were made by using tRNAs and tmRNAs predicted by tRNAscan-SE (Lowe and Eddy, 1997) and BRUCE (Laslett, et al., 2002) in a BLAST search and filtering out regions that do not contain an integrase genes. GIs can be browsed by GI name, organism name, or integration site (e.g. all GIs inserted in leucine tRNAs).

1.3.3.4 PAIDB

PAIDB is a database that provides GI information for those regions that are homologous to previously described pathogenicity islands (PAIs) (Yoon, et al., 2006). They call these regions PAI-like and any of these regions that also show sequence composition bias using %G+C are labelled as candidate PAIs (cPAIs). PAIs can be browsed by species, text searched, or searched with BLAST.

1.3.3.5 VFDB

VFDB **(**Virulence Factor Database) contains curated lists of virulence factors and pathogenicity islands for several species (Yang, et al., 2008). In addition, a larger number of virulence factor related genes are listed based on similarity to known virulence factors. These can be browsed by species, text searched, or searched with BLAST and PSI-BLAST.

Table 1.3 GI databases and other computational resources

Resource Name	Description	Query and Download Options
IslandPath	Aids in GI detection by visualizating multiple features of GIs (dinucleotide bias, mobility genes, tRNAs, etc.) in a single genomic view.	Whole genome graphical view is clickable and provides browsing of gene annotations
MOSAIC	Contains pre-computed whole genome alignments for several bacteria species	Conserved and variable (potential GIs) can be browsed and downloaded
Islander	Database of GIs within tRNA and tmRNA integration sites	GIs can be browsed by organism, GI name, or integration site
PAIDB	Contains GIs (identified by %G+C bias) that are homologous to previously described PAIs	Predictions can be browsed by species or by searching with text or BLAST
VFDB	Contains curated and putative (similarity based) virulence factors as well as a small list of curated PAIs	BLAST and PSI-BLAST can be used to search for virulence factors. PAIs can be found using a text search or by browsing species

1.4 Goal of present research

At the onset of my project, there were approximately 200 completely sequenced prokaryotic genomes and no method that used comparative genomics to identify GIs had yet to be developed. In addition, several sequence composition based GI prediction methods had been published but, surprisingly, a thorough comparison of their accuracy had not been conducted. In addition, many of these methods were not user friendly or easily accessible by the researchers needing to use them for new genome sequencing projects. Lastly, several studies had previously shown that pathogenic strains of bacteria often contained GIs that were not present in their non-pathogenic relatives, but few studies had shown direct evidence that these GIs provided an *in-vivo* competitive advantage in the infected host organism.

To meet these goals, I created a new method for GI prediction, called IslandPick that used comparative genomics and used this tool to produce robust datasets of GIs and non-GIs (conserved regions). These datasets, that were independent of the sequence composition based approaches used by previous methods, were used to assess and compare the accuracy of several previously published GI prediction programs (Langille, et al., 2008). After determining the most accurate GI predictors, I integrated the top three methods into a single integrated web resource, called IslandViewer, providing researchers with a user-friendly and informative web site for predicting, viewing and downloading GIs (Langille and Brinkman, 2009). Developing these resources allowed me to identify several GIs and prophage regions within a newly sequenced epidemic strain in cystic fibrosis patients (Winstanley, et al., 2008). In collaboration with other researchers, several of these GIs were found to harbour genes that provided an *in-vivo* competitive advantage in an infection model. Lastly, the robust prediction of GIs across hundreds of genomes allowed for a newly identified association between GIs and recently discovered phage defence elements, CRISPRs.

CHAPTER 2 ISLANDPICK: A COMPARATIVE GENOMICS APPROACH FOR GENOMIC ISLAND IDENTIFICATION

Portions of this chapter have been previously published in the article "Evaluation of genomic island predictors using a comparative genomics approach", co-authored by M.G.I. Langille, W.W.L. Hsiao, and F.S.L. Brinkman in BMC Bioinformatics, Volume 9 © 2008 Langille et al; licensee BioMed Central Ltd.

2.1 Introduction

An alternative approach that is independent from sequence composition-based approaches for GI identification is to use comparative genomics to identify genomic regions that have a clear phyletic pattern of non-vertical inheritance. In these methods, putative GIs are often defined as clusters of genes in one genome that are not present in a related genome. They are based on the observation that GIs are sporadically distributed among closely related species or strains and can sometimes be found between very distantly related species as judged by the degrees of sequence divergence in 16S rRNAs or other orthologs (Ragan, 2001). Until recently, this approach could only be performed manually for a few species that had enough similar sequenced genomes (Beres, et al., 2002; Hayashi, et al., 2001; Karaolis, et al., 1998; McClelland, et al., 2001; Parkhill, et al., 2001; Perna, et al., 2001).

In this chapter, I describe the development of "IslandPick", the first completely automated comparative genomics approach to identify GIs. Starting with all sequenced bacterial genomes as input (gathered using a new in-house

resource I developed called MicrobeDB), I use stringent but potentially flexible criteria, with distance cut-offs, to select query genomes that have a sufficient number of suitably related species or strains to conduct an analysis of GIs. IslandPick is then used to identify robust datasets of GIs from several genomes for the primary purpose of creating a benchmark that can be used for evaluating previously published sequence composition based GI prediction methods. As additional genome sequences become available, IslandPick will become increasingly useful for GI prediction and can be applied to those genomes to expand the benchmark dataset in a consistent and automated fashion.

2.2 MicrobeDB

A new in-house database that stores all completely sequenced genomes from National Center for Biotechnology Information (NCBI) and an application programming interface (API) for access to this data, called MicrobeDB, was constructed to aid in the analysis of large scale bacterial genomic studies. All sequenced genomes are downloaded monthly from the NCBI FTP server (ftp://ftp.ncbi.nih.gov/genomes/Bacteria) and stored locally. Information at the genome project, replicon (chromosome or plasmid), and gene level are parsed from these files and stored in a MySQL database (see database schema in Appendix File 2.1). Each monthly download is given a new version number so that experiments can be conducted on a stable “snap-shot” of the currently available genomes and annotations at a given time. Information within the database can be accessed directly by MySQL queries, or through a novel Perl API that allows easier access to all of the data.

Additional tables have been added to store GI predictions from multiple methods in a single resource. MicrobeDB is constructed so that additional analyses on microbial genomes such as ortholog prediction, protein subcellular prediction, etc. can be easily incorporated.

2.3 Identifying genomic islands using a comparative genomics approach

The overall approach that is used to predict GIs in IslandPick is to identify those regions that are present in only one genome (called the “query genome”) and are absent in several other related genomes (called “comparison genomes”). These unique regions are presumed to have been horizontally transferred (see section 2.7, for alternative theories), and are considered putative GIs. To identify these unique regions, whole genome alignments were constructed using the command line program mauveAligner from the Mauve 1.2.3 software package (Figure 2.1) (Darling, et al., 2004). Mauve allows for genomic insertions, deletions, inversions, and rearrangements and has been used by several researchers for prokaryote genome alignment (Glasner, et al., 2008; Glasner, et al., 2006; Greene, et al., 2007). After the comparison genomes are selected (see next section), the query genome is aligned pair wise against each of the comparison genomes in the dataset. Although Mauve can create a single multiple genome alignment for all genomes, pair wise alignments were used instead. This is due to a limitation in Mauve 1.2.3 that would not allow alignment of regions from a subset of genomes. In addition, this allowed the pair-wise alignments to be parallelized on the Brinkman computer cluster resulting in a

faster implementation. For each pair wise alignment, regions longer than 8000 nucleotides were extracted from the query genome that could not be aligned. Regions of the query genome that were not aligned in any of the pair wise genome alignments were retained for additional filtering as described below.

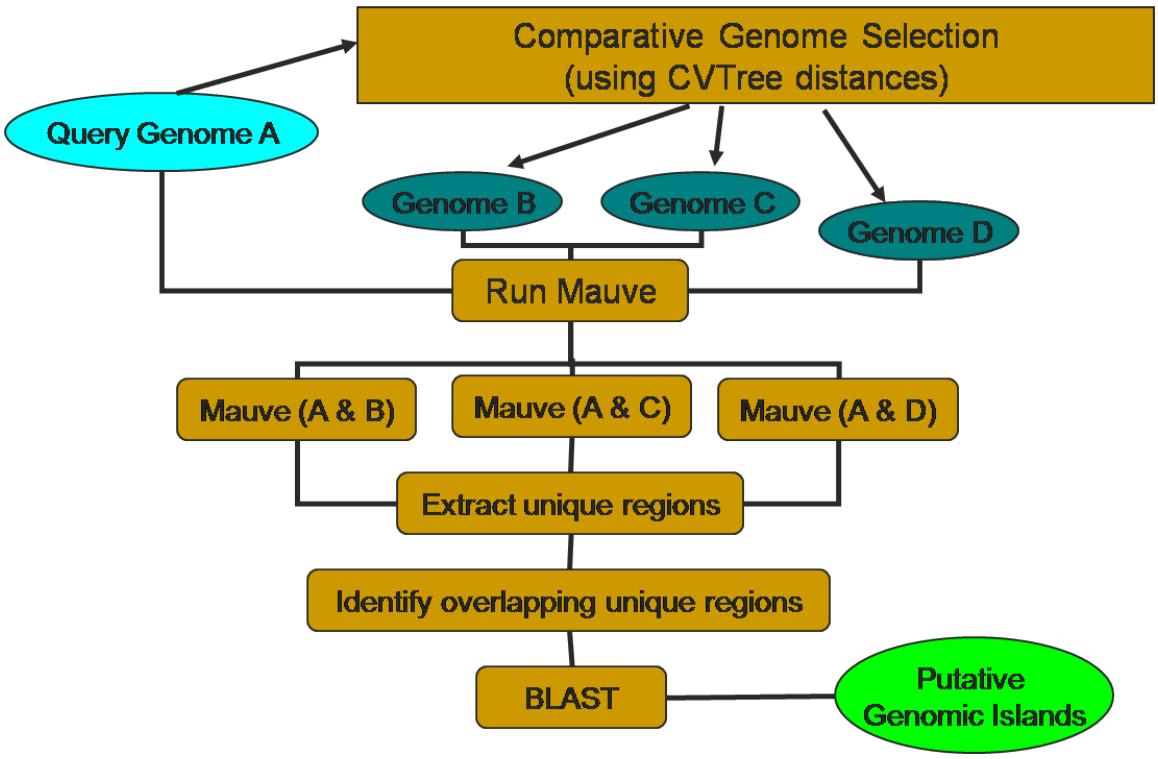
One caveat of Mauve is that it enforces a one-to-one alignment so if a duplication event occurs in one of the input genomes only one of the copies will be aligned. These possible genome duplications were excluded and ensured that the putative genomics islands were truly unique to only the query genome, with respect to the comparison genomes, by using BLAST similarity search as an additional filter. Each “unique” region was searched against the query genome and all comparison genomes using BLASTn (Altschul, et al., 1997), with default parameters except for an e-value of 1 (instead of 10). All similarity search matches (hits) less than 700 nucleotides were discarded while remaining hits were clustered together if they were less than 200 base pairs a part. Any unique regions that contained clustered hits that covered more than half of the minimum island size (8kb) were removed and the remaining regions were considered putative GIs. This additional BLAST filter may limit the prediction of GIs that contain genes with homologs in other regions of the genome, but the filter is needed to increase the precision of GI predictions in IslandPick.

2.4 Automated selection of comparison genomes

The application of my GI prediction method (see previous section), depends heavily on the genomes that are selected for comparison. In the past, when the number of sequenced genomes was limited, the selection of

comparison genomes was usually based on the genomes that were available at that time. In addition, most analyses were often conducted on a set of species that a particular researcher was extremely familiar with. Currently, due to the rapid increase of fully sequenced genomes, there are often numerous choices for comparison genomes. In addition, experiments that require analysis of hundreds of genomes makes manual selection not only time consuming but also increase the chance of bias resulting from personal choices. My goal was to produce a method that could “pick” comparison genomes automatically, in a uniform manner, with as little bias as possible. This would allow IslandPick to make predictions for any genome (if enough suitable genomes existed) without human intervention (Figure 2.1).

Figure 2.1 Pipeline of the IslandPick prediction program.
Comparison genomes are automatically selected for the given query genome using a pre-computed genome distance matrix calculated with CVTree. If enough suitable comparison genomes exist, GIs are predicted by taking each query genome and aligning it pair-wise with each comparison genome. Then all unaligned overlapping regions found in the query genome from the pair wise alignments are filtered using BLAST to ensure that they are truly unique to the query genome.

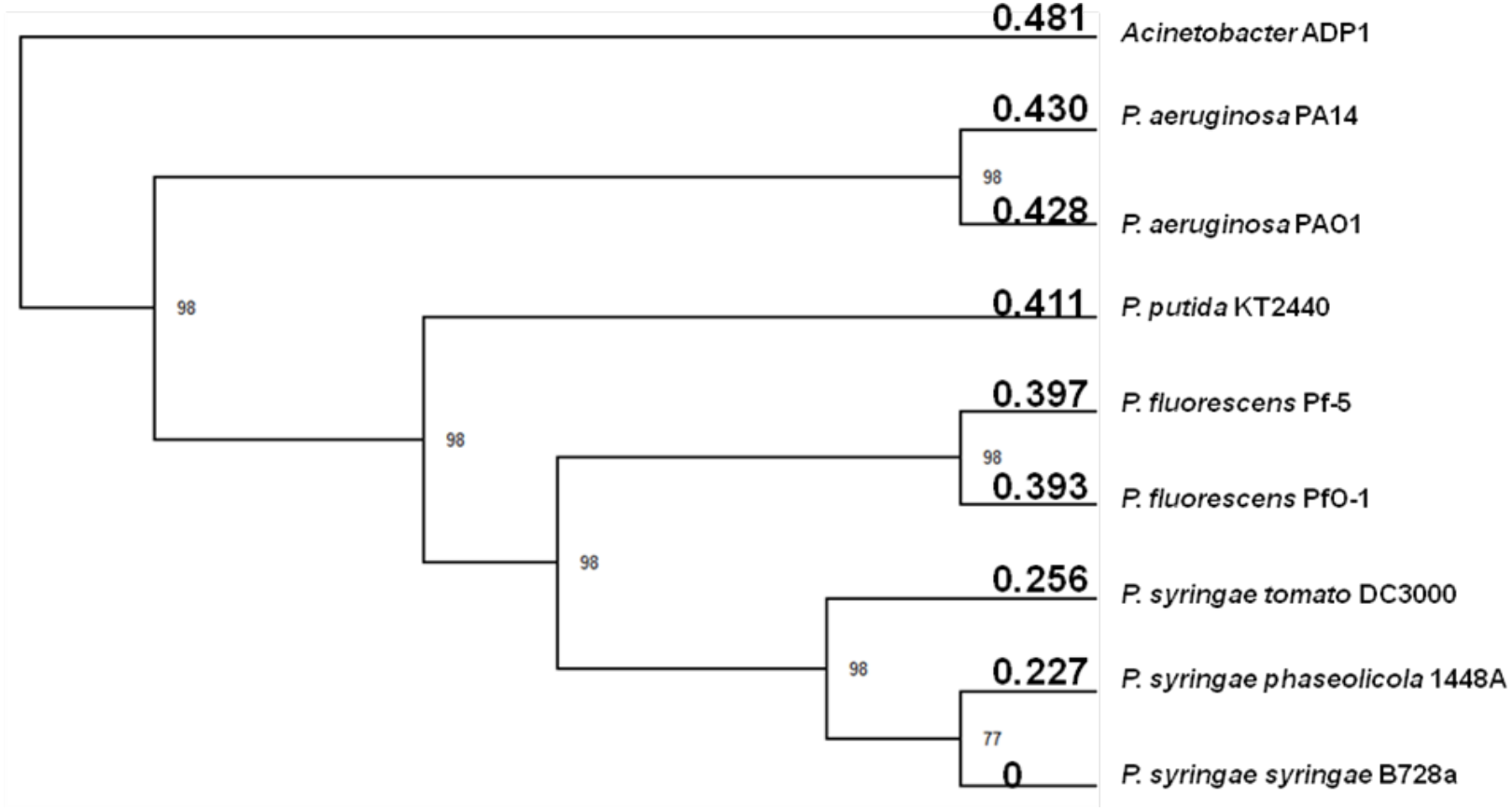


2.4.1 Calculating genome distances

I used an external application called CVTree, which infers evolutionary relatedness based on oligo-peptide content of complete predicted proteomes, to establish relative-phylogenetic distances between organisms. The input for CVTree is the translated protein sequence files for each genome. The algorithm then finds matching protein sequences of length 5 between the genomes, removes background noise using a Markov model, and finds the cosine between each genome composition vector (20^5). The resulting distance measure output by CVTree ranges from 0 (identical) to 0.5 (no similarity). The source code for

CVTree (Qi, et al., 2004) was obtained and used to calculate the 270,480 distances between every pair of the 736 currently available bacterial chromosomes; requiring approximately 526 hours (or ~7 seconds per calculation) of computation time (based on a single Intel Xeon 2.8 GHz machine) or ~4 hours on a 130 node cluster. The distances outputted by CVTree are on a scale and range between 0 and 0.5. To ensure that CVTree was behaving suitably, I compared these distance calculations to those produced by more conventional phylogenetic distance measures using PHYLIP (http://evolution.genetics.washington.edu/phylip.html), using *carB* and *omp85* genes as comparison sequences, as was previously used for phylogenetic analysis of species (Lawson, et al., 1996). Other approaches were tested to calculate evolutionary distances, such as SHOT (Korbel, et al., 2002), however CVTree was found to give the most consistent results across various genera including *Pseudomonas* (Figure 2.2).

Figure 2.2 A *Pseudomonas* species tree with overlaid CVTree distances.
The species tree was constructed using the conserved genes carB and Omp85 using maximum parsimony. Boot strap values are shown on the inner nodes of the tree. The CVTree distances (shown on the leaves of the tree) are pair-wise distances to *P. syringae* B728a. This is only one example of several trees that were inspected to confirm that CVTree was calculating suitable species distances.



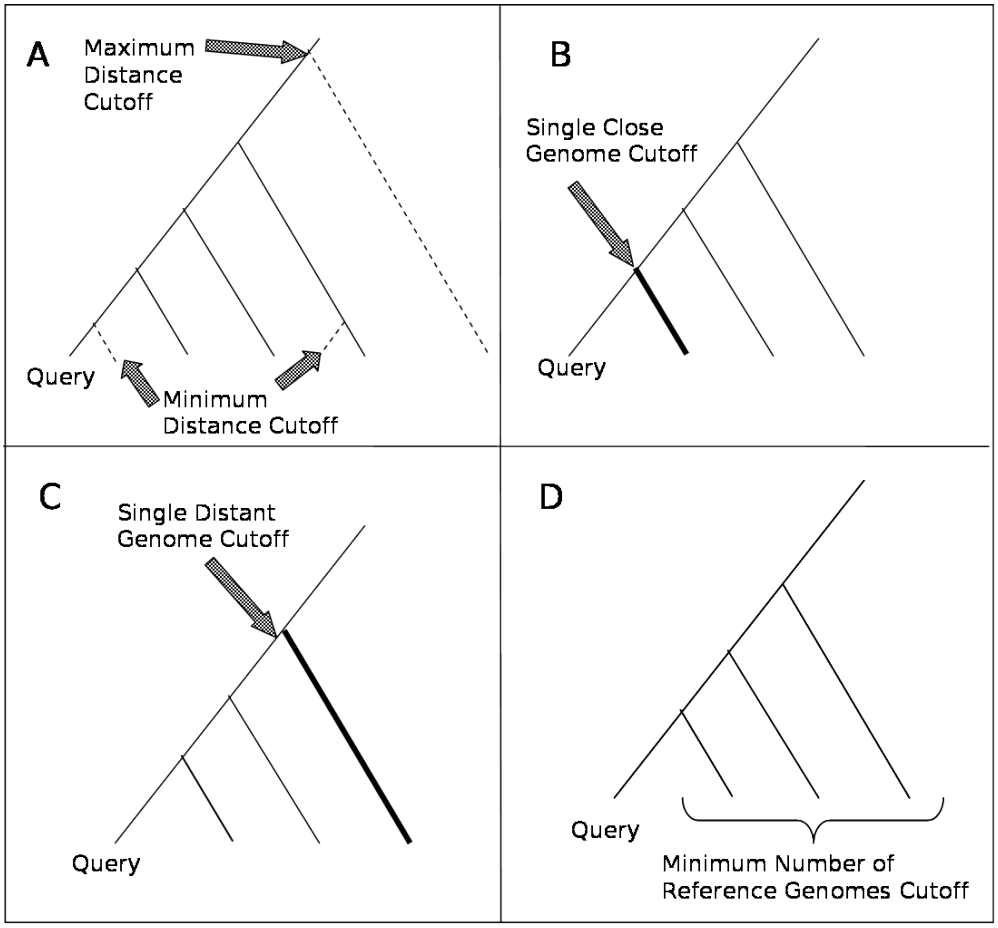
2.4.2 Genome selection parameters

Several CVTree distance cut-offs were formulated to ensure that appropriate genomes were selected for comparison to the query genome (Figure 2.1, Figure 2.3). These parameters were selected using known groups of strains that are within the proper distance for comparative genomics (e.g. *Pseudomonas aeruginosa* and *Escherichia coli* strains) (Figure 2.2). In addition, the alignments were inspected to ensure that the alignment results were not too sparse or too similar to gain any useful information. A “Maximum Distance Cutoff” of 0.42 was used to remove any genomes that were too distant from the query genome. It was observed that at such cutoff, often less than 5% of the genomes could be aligned. A “Minimum Distance Cutoff” of 0.1 was used to remove very closely related strains that would not provide any additional information and may have prevented the identification of some notable islands that were shared between such closely related strains (Figure 2.3A). In addition, by allowing for a larger span of insertion time, this parameter ensures that IslandPick is not limited to identifying only very recently inserted GIs. At least one comparison genome must have a distance less than 0.30 from the query genome to ensure that identification of GIs are all within a similar evolutionary time (Figure 2.3B). Decreasing this parameter would restrict GI identification to only recent insertions, while increasing it would allow prediction of more ancient insertions (see section 2.7 below for more details). In addition, to ensure that the stable backbone regions identified are ancient enough to be reliable (see section 2.6 below), at least one comparison genome must have a distance greater than 0.34

from the query genome (Figure 2.3C). Lastly, it is required that there be at least three suitable comparison genomes for each query genome to be used for further analysis (Figure 2.3D).

These entire set of cut-offs can be changed to permit prediction of GIs acquired from different time frames. For example, by increasing the "Minimum Distance Cutoff" and the "Single Close Genome Cutoff" the period of time that GI acquisitions are detected is changed by choosing more divergent genomes for the analysis. Overall, the parameters, in particular the default parameters, were selected to ensure high precision and confidence in the resulting predictions, so that they could be used to fairly evaluate the accuracy of several sequence composition based GI prediction tools (see Chapter 3). The parameters were not changed to maximize the accuracy scores of any GI prediction tools that were evaluated; however, default parameters resulted in the highest apparent accuracy when GI datasets were compared with a curated, literature-based dataset (see section 3.3 below).

Figure 2.3 **Effect of IslandPick comparison genome cut-offs on a sample genome tree. First, for each query genome any genomes that are too distant to the query genome or too closely related to each other are removed (dotted lines) A). Second, to ensure that the identified GIs were inserted from similar time frames and were not biased by the genomes that are currently available, at least one genome (bold line) must be close enough to the query genome B). Also, it is required that at least one genome is a certain distance away from the query genome (bold line) to ensure that the backbone sequences identified were not inserted recently C). Finally, there must be a minimum of three comparison genomes that have met all other criteria D). The comparison genomes that have passed all the cut-offs are used for comparison with the query genome.**



2.5 Genomic island predictions using IslandPick

Of the 675 completely sequenced microbial genomes, 736 chromosomes were initially used as queries in the IslandPick pipeline (Figure 2.1). Three hundred and seventy seven of these did not have at least three related species/strains while many others did not meet the stringent criteria to do a

comparative analysis (see section 2.4.2 above). One hundred and seventy three chromosomes met the requirements for the prediction of GIs and a subset of 134 chromosomes contained GIs while IslandPick did not detect GIs in the other 39 chromosomes (see genome list in Appendix File 2.2). Many of these 39 genomes may contain GIs that are smaller than 8kb or have other cases of HGT that were not being targeted by IslandPick. The dataset was further reduced to 118 chromosomes, because a negative dataset could not be predicted for 14 chromosomes and the GI prediction tool SIGI-HMM gave errors on another two chromosomes (see section 2.6 and Chapter 3). In total, I identified 771 GIs, comprising 12.4Mb and ranging in size from 8-105kb, within 118 chromosomes from 117 different strains and 22 genera (see Appendix File 2.3). These putative GIs contained a total of 11,404 annotated genes with an average of 14.8 genes/GI and 97.5 genes/strain (see Appendix File 2.4).

2.6 Developing a negative dataset of GIs

In order to evaluate the accuracy of several previously developed GI predictors (see Chapter 3), a dataset of genomic regions that were not likely to contain GIs was constructed (negative dataset). The IslandPick pipeline was adapted to identify large genomic regions that were conserved in several genomes. These regions are likely to form the stable backbone of the genomes and are unlikely to be acquired by HGT among the strains considered. A multiple genome alignment of each query genome and all comparison genomes previously selected (see section 2.4 above) was performed using Mauve with minimum backbone length and maximum gap size parameters set to 8000 and

300, respectively. The regions that were conserved across all genomes were extracted from Mauve's backbone output file. These conserved genomic regions were identified for the same 134 query genomes that were used for prediction of GIs. Conserved regions larger than 8000 base pairs could not be identified for 14 of these chromosomes and so these were removed from both the positive and negative datasets. The resulting negative dataset was about 4 times larger than the IslandPick dataset of GIs, containing approximately 50.6 Mb over 3770 separate genomic regions (see Appendix File 3.1). The size difference between the negative and positive datasets was expected since the proportion of HGT versus conserved backbone in a genome is normally much smaller (Daubin and Ochman, 2004; Vernikos, et al., 2007; Waack, et al., 2006).

2.7 Discussion

I have introduced and outlined, IslandPick, a novel automated method for predicting GIs using comparative genomics. To date, this is the first attempt at trying to automate genome selection for comparative genomics. I have used IslandPick, with its stringent default criteria, to generate datasets of GIs and non-GI regions that can be used to evaluate the accuracy of multiple sequence composition based GI predictors (see Chapter 3).

Of course, there are some limitations to predicting GIs using comparative genomics. The choice of genomes for comparison to each query genome can result in differences in the predicted GIs. IslandPick's genome selection criterion uses several distance cutoffs to minimize this bias as much as possible (example given in the next paragraph). GIs could be present in the negative dataset if a GI

inserted before the divergence of all genomes examined. To minimize these in my datasets, IslandPick requires that at least three comparison genomes be used for each query genome and that at least one comparison genome is at least some minimal distance away from the query genome. The number of false positive GI predictions is minimized by requiring that any putative GI is present only in the query genome when compared to all comparison genomes. Therefore, a deletion of the same genomic region would need to occur in three or more strains for it to be mis-predicted as a GI in my analysis. Similarly, a GI that inserted into multiple genomes would have to be conserved in all of the diverse genomes studied, to be improperly placed in the negative dataset. Although using several rules in the genome selection process results in very stringent datasets of GI and non-GI regions, it does limit the number of organisms that can be used by IslandPick. Relaxing the genome selection process by the removal of some of these cut-offs would allow IslandPick to be applicable to more genomes. It should be emphasized that IslandPick was not developed to be a GI prediction tool that would replace sequence-based composition tools, which can be used on any genome without the requirement of having several other comparative genomes; rather, the IslandPick approach allows the testing of these tools and in certain cases can also be used for GI prediction (cases that should increase notably in the future, as more and more genomes are sequenced).

As an example of the IslandPick approach, when *Salmonella enterica* Typhi CT18 is used as a query genome to identify islands using the default cutoffs, very closely related genomes including *S. enterica* Typhi Ty2 and *S.*

enterica Paratyphi A str. ATCC 9150 were excluded from comparison. Therefore, IslandPick identifies GIs that have inserted after the divergence of *S. enterica* Typhi CT18 and the next most related genome that has been sequenced, which is *S. typhimurium* LT2. Islands that inserted before the divergence of CT18 and LT2 would also not be included in the positive dataset, using these stringent cutoffs. However, IslandPick requires that at least one genome be a certain distance from the query genome (*Shigella dysenteriae* Sd197 in this example), so that these more ancient GIs are not improperly included in the negative dataset. It is assumed that any sequences shared between the query genome (e.g. *Salmonella enterica* Typhi CT18) and the comparative genomes including those that meet the single distant genome cutoff (e.g. *S. dysenteriae* Sd197) are sufficiently stable and can be considered as the conserved genome backbone. Again, distance cutoffs can be modified in IslandPick to detect islands that are more ancient or those acquired more recently.

In many instances, IslandPick tends to split large islands into smaller ones, which is probably the result of a few similar genes being identified in one or more of the comparison genomes. Considering that as an island gets bigger there is a greater chance of detecting some similarity between the genomic regions being compared, one would expect that very large GIs might be split into smaller ones. As indicated in recent research, this limitation could be improved in the future by spanning together islands that are interrupted by only small regions of low similarity (Azad and Lawrence, 2007).

Similar to other GI prediction tools IslandPick does not try to identify the origins or the methods of horizontal transfer for these GIs. Indeed future research on many of these large regions of HGT will likely allow them to be sub-classified into known mobile elements such as conjugative transposons, integrated plasmids, integrons, and prophage; and will depend on robust prediction tools and knowledge of their strengths and weaknesses. Comparative genomics studies like this one, will aid in these areas by providing an independent method for GI prediction. As more genomes are sequenced, the distance cutoffs used in this method should be re-evaluated, but this overall approach should only increase in utility as the number of completely sequenced microbial genomes increases into the thousands.

CHAPTER 3 EVALUATING SEQUENCE COMPOSITION BASED GENOMIC ISLAND PREDICTION METHODS

Portions of this chapter have been previously published in the article "Evaluation of genomic island predictors using a comparative genomics approach", co-authored by M.G.I. Langille, W.W.L. Hsiao, and F.S.L. Brinkman in BMC Bioinformatics, Volume 9 © 2008 Langille et al; licensee BioMed Central Ltd.

3.1 Introduction

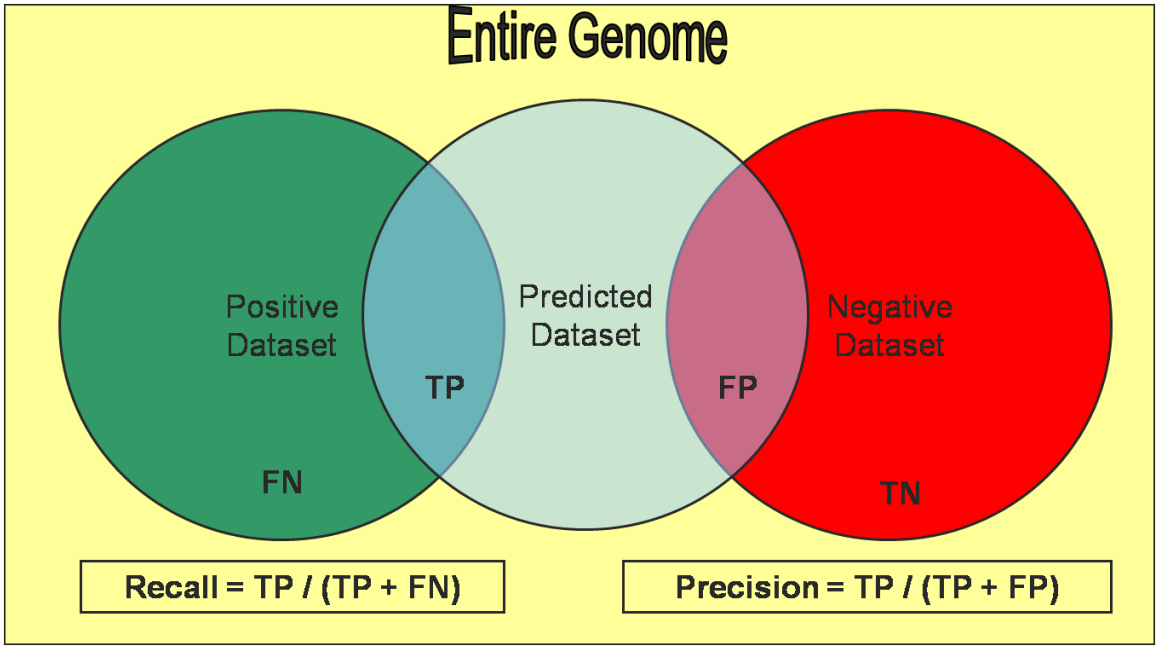
Using the positive and negative datasets of GIs developed in Chapter 2, I evaluated how well these datasets agreed with those predicted using previously published sequence composition-based GI tools since IslandPick's comparative genomics based method is independent of sequence composition-based methods. Analyses of GI/HGT prediction tools have been previously published, but have used either artificial datasets or real data from only a few species (Azad and Lawrence, 2007; Vernikos and Parkhill, 2006). In addition, I evaluated how well these datasets agreed with GIs reported in a small literature-based dataset of GIs. The accuracy of these sequence composition based GI prediction methods are compared and suggestions for their use are discussed.

3.2 Comparison with sequence composition based GI prediction methods

The GIs and the non-GIs identified from my comparative genomics approach were used as positive and negative datasets, respectively, to evaluate the accuracy of several previously published sequence based GI prediction tools:

PAI_IDA (Tu and Ding, 2003), AlienHunter (Vernikos and Parkhill, 2006; Waack, et al., 2006), SIGI-HMM (as part of the Colombo package) (Waack, et al., 2006), Centroid (Rajan, et al., 2007), and IslandPath (included both DIMOB and DINUC methods) (Hsiao, et al., 2003; Hsiao, et al., 2005) (Table 3.1). The tools were run using their default parameters on the same 118 chromosomes and any overlapping regions with the negative dataset were considered false positives (FP) while overlapping regions with the positive dataset were considered true positives (TP) (see Appendix File 3.2). False negatives (FN) and true negatives (TN) were those regions in the positive dataset and negative dataset, respectively, which were not predicted by the method being evaluated. Precision or specificity was calculated using the standard formula TP/ (TP +FP) and recall or sensitivity was calculated using TP/ (TP+FN). The equation, (TP + TN)/ (TP+TN+FP+FN) was used to measure the overall accuracy (Figure 3.1).

Figure 3.1 Accuracy calculations using IslandPick derived positive and negative datasets.



The following accuracy calculations were measured using the number of overlapping nucleotides (Appendix File 3.3); although results were not significantly different when counting only GIs with greater than 50% overlap (data not shown). I found that the precision and recall for the tools evaluated varied considerably (Table 3.1). SIGI-HMM performed the best with 92% precision (though only 33% recall) whereas AlienHunter had the best recall at 77% (though only 38% precision). SIGI-HMM and the IslandPath/DIMOB tool had comparable overall highest accuracy of 86% with IslandPath-DIMOB more suitable for analyses requiring a slightly higher recall (precision of 86% with a recall of 36%). All of the tools had similar overall accuracies ranging from 82-86% (but with differing emphasis on precision versus recall) except for AlienHunter, which had an accuracy of only 71%. This appeared to be primarily due to the large number of predictions being made by AlienHunter (1264.8 kb of GI/genome) versus the other methods (163.2 to 444.2 kb of GI/genome).

For completeness, I also calculated the accuracy of each tool using every other tool as the benchmark (Appendix File 3.4). The average accuracy measurements over all benchmarks for each tool were very similar to those calculated using only my datasets, indicating that the datasets generated using IslandPick may be an appropriate reference dataset for future use. These positive and negative GI datasets, and the source code for development of these datasets, are available at www.pathogenomics.sfu.ca/islandpick_GI_datasets.

Table 3.1 Average number of GI predictions and accuracy measurements of several GI prediction tools.

Tool	Average number of nucleotides in GIs per genome (kb)	Precision	Recall	Overall Accuracy
SIGI-HMM	232.7	92.3	33.0	86.3
IslandPath/ DIMOB	170.7	85.8	35.6	86.2
PAI IDA	163.2	68.0	32.2	83.7
Centroid	171.3	61.3	27.6	82.4
IslandPath/ DINUC	444.4	54.8	53.3	82.2
Alien Hunter	1264.8	38.0	77.0	70.8
Literature	639.4	100	87.0	96.3

3.3 Comparison with previously published genomic islands

Although, there is no gold standard dataset of GIs, I wanted to examine how previously published GIs overlapped with my datasets. Five strains from the list of 118 had published GIs (Beres, et al., 2002; Hayashi, et al., 2001; McClelland, et al., 2001; Parkhill, et al., 2001; Perna, et al., 2001). As with the analysis of the sequence composition based GI predictors, I calculated the overlap of the published GIs against the positive and negative dataset. I found, potentially due in part to the similar manual comparative genomics methods sometimes used to identify GIs in the literature dataset, that the literature GIs had the most agreement with my datasets (versus the GI predictors evaluated below). Literature GIs had the highest precision, recall, and overall accuracy of 100,

87%, and 96%, respectively, when using IslandPick-predicted islands as the text dataset (Table 3.1).

3.4 Comparison of sequence composition based approaches using additional GI datasets constructed with more relaxed criteria.

IslandPick's parameters can be modified to allow the prediction of GIs with more ancient origins. Although the inclusion of more ancient GIs could lead to a more comprehensive dataset, it may result in an increase in false positives since the proper identification of older evolutionary events can be easily mistaken. However, I did use two additional "relaxed" sets of parameters to determine the effect on GI prediction of changing the default parameters. These relaxed parameters should identify GIs with origins that are more ancient. The first relaxed set used the same default parameters, except that the "Minimum Distance Cutoff" was changed to 0.15 and the "Single Close Genome Cutoff" changed to 0.34. The second set of parameters was even more relaxed by increasing the "Single Close Genome Cutoff" to 0.20, with all other parameters being the same as the first relaxed set.

The first "relaxed" dataset had approximately 46% more GIs predicted per genome, while as expected the negative datasets stayed about the same size with a 3% increase in the relaxed dataset. Notably, accuracy relative to the literature dataset went down slightly (see Appendix File 3.5 and Appendix File 3.6), indicating that the IslandPick defaults do most accurately reflect literature-based GI data. The sequence composition-based tools also all had a relative

decrease in accuracy using this more relaxed dataset: Accuracy decreased between 4.5 and 6.6% for all methods, with the exception of Alien Hunter (the method with highest recall but lowest precision) which showed the smallest decrease of 0.6% (see Appendix File 3.5 and Appendix File 3.6). Using a second more relaxed dataset of parameters resulted in yet another decrease in predicted accuracy of the GI tools and the accuracy relative to the literature-based dataset also decreased further (data not shown). While the use of more relaxed criteria for GI prediction may still have its uses, the results indicate that the default settings of the IslandPick method are most appropriate for predicting islands that most closely resemble what is reported in the literature. In addition, the sequence composition-based methods appear to perform best when using the default IslandPick-predicted GI datasets for evaluation.

3.5 Discussion

I have used IslandPick, with its stringent default criteria, to generate test datasets of GIs and non-GI regions that are used to evaluate the accuracy of multiple sequence composition based GI predictors. This represents the first evaluation of GI predictors based on real (non-artificial) GI data from several different strains of bacteria (Azad and Lawrence, 2007; Vernikos and Parkhill, 2006). By developing separate negative and positive datasets that were independent of sequence composition based approaches, I was able to assess the accuracy of several GI predictors.

According to this analysis, SIGI-HMM has the highest precision and shares comparable overall accuracy with IslandPath-DIMOB, which has higher

recall at the expense of precision. SIGI-HMM is the only tool tested that measures codon usage and notably it also identifies codon usage associated with highly expressed genes and then discards such genes from the analysis. While more study is needed, this suggests that regions displaying codon usage bias of a pattern that is not associated with highly expressed genes are more likely to be GIs. Consistent with this, the IslandPath/DIMOB method that requires both a dinucleotide bias and the presence of a mobility gene for a GI prediction does much better than the IslandPath/DINUC method, which measures only dinucleotide bias. The latter can result in false positives from highly expressed genes but higher predictive recall/sensitivity. AlienHunter had the lowest precision (38%); however, it had by far the highest recall value (77%) with more than twice as many predictions as any other tool.

Based on the results, the use of SIGI-HMM is suggested for making very precise predictions where a high confidence dataset of GIs is preferred while AlienHunter can be used as a first-pass tool to capture most GIs for further refinement. If suitable comparative genomes are available, IslandPick would be a top choice for GI prediction. If comparative genomes are not available, the results generally suggest that by combining multiple features of GIs, as in the IslandPath/DIMOB dataset, and accounting for highly expressed genes, which SIGI-HMM does and IslandPath/DIMOB does indirectly, a better overall predictor could be created. Considering that sequence composition based methods often make non-overlapping predictions, the use of more than one method may result in improved prediction accuracy. For example, I tested the accuracy of combining

the predictions from IslandPath-DIMOB and SIGI-HMM and found that there was a large increase in recall/sensitivity to 48% (from IslandPath-DIMOB (36%); SIGI-HMM (33%)) and overall accuracy 88% (IslandPath-DIMOB (86%); SIGI-HMM (86%)) while maintaining roughly the same precision/specificity 86% (IslandPath-DIMOB (86%); SIGI-HMM (92%)) (data not shown). More analysis of the differences in sequence composition between true positives and false positives in this analysis could be insightful.

The results show that all GI predictors had a decrease in overall accuracy when trying to predict more ancient islands. Considering that sequence composition based predictors would have trouble detecting significant signals in older GIs due to amelioration to the host genome, it was not surprising that the overall accuracy for all tools decreased (Lawrence and Ochman, 1997). Alien Hunter had the lowest decrease in overall accuracy however, it still maintained the lowest precision and overall accuracy for the prediction of this dataset and SIGI-HMM still out performed the other sequence composition-based tools for predicting these more divergent islands. It is possible that the accuracy of some of these sequence composition-based tools could be improved by optimizing their parameters. However, out of all the tools, SIGI-HMM and Centroid were the only ones with a clearly defined sensitivity/statistical parameter and even for these there were no recommend suggestions besides the default. Although default parameters for all tools are presumably maximized to result in the best overall accuracy, some fine-tuning may improve their results.

It must also be appreciated that the GI regions identified with IslandPick represent a set of GIs that were acquired within a particular window of divergence of the strains being examined. Any genomic regions that did not have clear evidence of GI or non-GI status were not included in either of the datasets so that tools that predicted such possible/uncertain GIs were not penalized. This would include GIs that have inserted into multiple strains or those that have partial similarity with other genomic regions. Rather, my methodology penalizes tools that falsely predict GIs in highly conserved backbone regions that very likely do not contain true GIs, and my method penalizes tools that do not predict a subset of GIs that are very likely true positives. When compared to all of the sequence composition based methods tested in this study, IslandPick produced the smallest dataset of GIs compared to all of the methods (see Appendix File 3.2) and the proportion of the genomes that are covered in both of the positive and negative datasets combined, ranges from 10%-30% per genome. Therefore, IslandPick does not make predictions for the majority (70%-90%) of the genome, reflecting the high accuracy of the positive and negative datasets. In addition, the comparative genomics-based GI datasets had the highest agreement with the smaller curated, literature-based dataset.

This analysis of the accuracy of composition-based GI predictors should aid both development and use of such predictors, which are becoming of increasing importance as the critical role of GIs in microbial evolution becomes more apparent. My analyses of the accuracy of GI predictors should aid researchers in formulating an appropriate approach to identify GIs, based on

whether they prefer high recall/sensitivity or precision/specificity. Such GI predictors are likely to become of increasing importance in bacterial genome analysis, as appreciation grows of their significant role in adaptations of medical and environmental importance.

CHAPTER 4 ISLANDVIEWER: AN INTEGRATED INTERFACE FOR COMPUTATIONAL IDENTIFICATION AND VISUALIZATION OF GENOMIC ISLANDS

Portions of this chapter have been previously published in the article "IslandViewer: an integrated interface for computational identification and visualization of genomic islands", co-authored by M.G.I. Langille and F.S.L. Brinkman in Bioinformatics, Volume 25, Issue 5 ©2009 The Author(s)

4.1 Introduction

After developing IslandPick (Chapter 2) and conducting an analysis of the accuracy of several sequence composition based GI prediction methods (Chapter 3), I saw the need for a user friendly web resource that would integrate the most accurate GI prediction programs. In this chapter, I present IslandViewer (http://www.pathogenomics.sfu.ca/islandviewer/), the first web accessible interface that facilitates viewing and downloading of GI datasets predicted from user-submitted sequences, or based on pre-computed analyses, using the sequence composition based approaches SIGI-HMM and IslandPath-DIMOB, and the comparative genomics approach IslandPick.

4.2 Implementation

GI predictions are pre-computed using SIGI-HMM, IslandPath-DIMOB, and IslandPick (see section 4.3 below) for all completed genomes and are stored in a local MySQL database called MicrobeDB (see section 2.2 above). All methods are run in parallel for each genome so that automatic monthly updates

are quickly performed on a computer cluster, while all dynamic web pages are implemented using PHP.

4.3 Selection and integration of genomic island prediction methods

The inclusion of particular GI prediction methods into IslandViewer were based on several factors. The most obvious is that I could only consider using methods that had obtainable software and could be run without manual intervention. Therefore, many GI resources that are simply a database and have no downloadable software such as Islander (Mantri and Williams, 2004) could not be included into IslandViewer. In addition, I did not consider the inclusion of MobilomeFINDER (Ou, et al., 2007), a tool that uses a comparative genomics based approach similar to IslandPick because it requires the manual selection of comparison genomes (making pre-computed results for all genomes impossible). However, all of these methods are listed on IslandViewer's "Resources" page and users are recommended to visit their respective websites if interested.

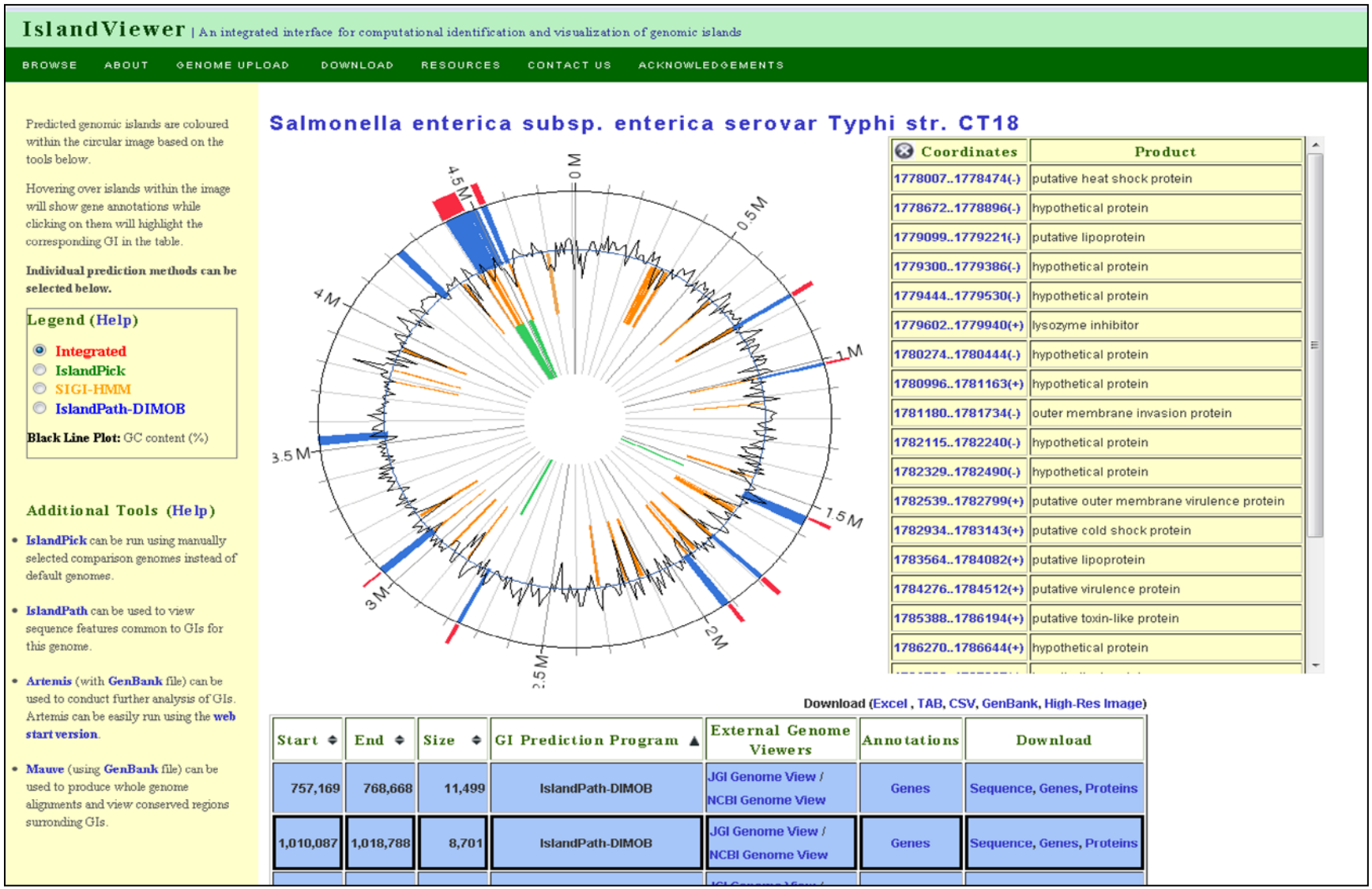
For those tools that did have their software freely available, IslandPath-DIMOB (Hsiao, et al., 2005) and SIGI-HMM (Waack, et al., 2006) were included because they were shown to have the highest specificity (86-92%) and overall accuracy (86%) (Chapter 3). In addition, the automated comparative genomics method, IslandPick, was included since it provides predictions that are not based on sequence composition and showed the most agreement with a manual curated dataset of literature based GIs. These three methods sometimes predict the same GIs, but often give slightly different results suggesting that they

complement each other well without being redundant. Methods that had lower specificity (some as low as 38% precision), which would result in a large number of false predictions in IslandViewer were avoided. Finally, none of the methods included in IslandViewer had been previously available as a web resource; therefore, giving new user-friendly access to three different GI prediction methods.

4.4 Features and design of IslandViewer

IslandViewer allows the viewing of all GI predictions for the above predictors through a single integrated interface (Figure 4.1). Predictions are pre-computed for all published GIs and are updated on a monthly basis, while users with newly sequenced unpublished genomes can submit their genome for analysis and receive an email notification when finished. These user-submitted genomes are not viewable by other IslandViewer users and are accessible for at least one month. IslandPick automatically selects comparison genomes for use using default distance parameters, but since researchers may have particular insights into a particular species, they can choose to run IslandPick with their own manually selected comparison genomes and have the option of being notified by email when the results are available.

Figure 4.1 A screenshot of the IslandViewer interface.



Once the genome of interest is selected it is presented as a circular genome image with each predicted GI highlighted (different colours for different tools in the IslandViewer) and is also available as a high-resolution image suitable for publication. In addition to the predicted GIs for each tool, IslandViewer highlights any GIs that have been predicted by two or more methods. The annotations for genes within each GI can be quickly viewed by hovering over the GI of interest within the image. Clicking on an island jumps to the corresponding row in a table below the genome image and gives information such as GI coordinates, links to tables showing genes and annotations within the GI region, links to external genome viewers at NCBI and Joint Genome Institute (JGI), and links to IslandPath to allow further examination of GI related features in the genome of choice. GI predictions may be downloaded in various formats including Excel, tab-delimited, comma-delimited, FASTA, and GenBank (allowing easy input into the genome browser and annotation tool Artemis (Rutherford, et al., 2000)). All datasets and source code are available for download under a GNU GPL license.

4.5 Discussion

GI identification is becoming a first critical step in the characterization of a bacterial genome, due to the growing appreciation for the role of GIs in important adaptations of interest. Recent research has therefore focused on developing new computational methods for their prediction. However, these methods tend to use different approaches and identify different features of GIs. The result is that the most accurate methods each have high precision, but low recall, leading to

slightly different regions being predicted. Previously, researchers could either pick a single method or try to manually integrate the results from multiple methods themselves. In addition, many of these tools did not have their own web interfaces and often required that the user download and run the program on their computer. IslandViewer alleviates these concerns by providing a web interface for three accurate GI prediction methods that were not previously available through a web interface. By pre-computing GI datasets for all completed genomes and providing a single submission process for new user genomes IslandViewer allows researchers access to a user-friendly resource that can be used as the first step in GI analysis of bacterial genomes. It would be expected that researchers would manually inspect any GI predictions shown in IslandViewer to determine their validity and make more accurate predictions of their boundaries. IslandViewer helps aid further analysis of GI predictions by providing data in various formats that can be used in other bioinformatic tools such as Artemis, and by providing numerous links to other GI resources. IslandViewer should be a useful resource for any researcher studying GIs and microbial genomes.

CHAPTER 5 THE ROLE OF GENOMIC ISLANDS IN THE VIRULENT *PSEUDOMONAS AERUGINOSA* LIVERPOOL EPIDEMIC STRAIN

Portions of this chapter have been previously published in the article "Newly introduced genomic prophage islands are critical determinants of in-vivo competitiveness in the Liverpool Epidemic Strain of Pseudomonas aeruginosa", co-authored by C. Winstanley, M.G.I. Langille, J.L. Fothergill, I. Kukavica-Ibrulj, C. Paradis-Bleau, F. Sanschagrin, N. R. Thomson, G.L. Winsor, M.A. Quail, N. Lennard, A. Bignell, L. Clarke, K. Seeger, D. Saunders, D. Harris, J. Parkhill, R. E.W. Hancock, F.S.L. Brinkman, and R.C. Levesque in Genome Research, Volume 19, Issue 1 ©2009 by Cold Spring Harbor Laboratory Press

5.1 Introduction

Pseudomonas aeruginosa is a ubiquitous organism distributed widely in the environment, including the soil and water and in association with various living host organisms. It is one of the most prevalent causes of opportunistic infections in humans and is the most common cause of eventually fatal, persistent respiratory infections in cystic fibrosis (CF) patients. It has been assumed to owe its versatility to its genetic complexity. Sequencing of four strains (Lee, et al., 2006; Mathee, et al., 2008; Stover, et al., 2000), and molecular genetic analysis of others, has revealed an approximately 6-7 Mb genome with around 5,500 ORFs. Based on comparisons of the first two *P. aeruginosa* genomes sequenced, those of strains PA01 (Stover, et al., 2000) and PA14 (Lee, et al., 2006) [the latter of which is the most common genotype encountered in diverse habitats in one study of 240 isolates (Wiehlmann, et al., 2007)], it was revealed that there is a quite highly conserved core genome

representing up to 90% of the total genomic sequence; subsequent studies have revealed an extraordinary similarity of the core genome with an average nucleotide divergence of around 0.5% (1 in 200 nucleotides). Other changes that can occur include the loss of core genes through deletion or loss of expression through mutation [e.g. with the pyoverdine and O-antigen biosynthesis genes; (Spencer, et al., 2003)].

In addition to this core genome, there are variable accessory genes, which are largely associated with GIs that are subject to what is termed diversifying selection, or rapid change that is presumed to be due to certain selective pressures. Some of these GIs have been well described including a 108-kb pathogenicity island PAPI-1 (Qiu, et al., 2006) that, in strain PA14, carries several regulatory genes, including *pvrR* that regulates antibiotic resistance and biofilm formation, a smaller (11-kb) PA14 pathogenicity island PAPI-2 encoding the exotoxin ExoU, a 14 gene island of PAK that encodes the flagellin glycosylation machinery (Arora, et al., 2001), two tandem defective phage (pyocin) islands in PA01 (but widely distributed) that are determinants of fluoroquinolone susceptibility (Brazas and Hancock, 2005), and a 103kb mobile GI pKLC102 from clone C isolates that appears to comprise a hybrid of plasmid and phage features (Klockgether, et al., 2004). While these specific instances have been studied and general features of the diversifying GIs are well understood, there is still considerable debate as to what are the forces that shape genomic diversity among *P. aeruginosa* isolates and in particular what selective advantages are provided by the variable accessory genes. The

discovery of epidemic strains from the lungs of patients with CF provided an unprecedented opportunity to address this issue.

The widespread assumption that CF patients acquire only unique strains of *P. aeruginosa* from the environment was challenged when molecular typing was used to demonstrate the spread of a β-lactam-resistant isolate, now known as the Liverpool Epidemic Strain (LES), at a children's CF unit in Liverpool, UK (Cheng, et al., 1996). Subsequent identification of other CF epidemic strains in the UK (Lewis, et al., 2005; Scott and Pitt, 2004) and Australia (Armstrong, et al., 2003; O'Carroll, et al., 2004) indicate that transmissible *P. aeruginosa* strains make a significant contribution to the infection of patients in some CF centres. LES is the most frequent clone isolated from CF patients in England and Wales (Scott and Pitt, 2004) and has also been reported in Scotland (Edenborough, et al., 2004). In addition, LES can cause superinfection (McCallum, et al., 2001), exhibits enhanced survival on dry surfaces (Panagea, et al., 2005), and is associated with greater patient morbidity than other *P. aeruginosa* strains (Al-Aloul, et al., 2004). In two unusual cases, transmission of an LES strain occurred from a CF patient to both non-CF parents, causing significant morbidity and infections that have persisted (McCallum, et al., 2002), and from a CF patient to a pet cat (Mohan, et al., 2008). LES isolates, including isolate LESB58, exhibit an unusual phenotype, characterised by early (in the growth curve) over-expression of the cell-density-dependent quorum sensing regulon, including virulence-related secreted factors such as LasA, elastase and pyocyanin (Fothergill, et al., 2007; Salunkhe, et al., 2005). Furthermore, LESB58 is known to be a biofilm

hyperproducer (Kukavica-Ibrulj, et al., 2008). Hence, LES is a successful and aggressive clone that is particularly well adapted to the CF lung. While all *P. aeruginosa* isolates are intrinsically resistant to antimicrobials, like other CF isolates that cause chronic infections and are treated over time with antibiotics, LES can readily mutate to resistance to the common antibiotics utilized in therapy (although LESB58 does not have a mutator phenotype like many other mature CF isolates, including other LES isolates). Indeed LES was first identified because of the widespread occurrence of *P. aeruginosa* isolates exhibiting ceftazidime resistance in a clinic where ceftazidime monotherapy was in routine use (Cheng, et al., 1996). A survey of multiple LES isolates demonstrated that the strain can also acquire resistance to meropenem, aztreonam, tobramycin and ciprofloxacin (Fothergill, et al., 2008).

The *P. aeruginosa* strains PA01 and PA14 were previously compared with LES isolate LESB58 to assess *in-vivo* growth, infection kinetics, bacterial persistence and localization within tissues in a rat model of chronic lung infection (Kukavica-Ibrulj, et al., 2008). The three *P. aeruginosa* strains demonstrated similar growth curves *in-vivo* but differences in lung tissue distribution and in virulence in a competitive *in-vivo* assay. The LESB58 strain persisted in the agarose beads used to deliver bacteria into the bronchial lumen, while PA01 and PA14 strains were found to disseminate into the alveolar regions and grew as macrocolonies after 14 days post-infection.

To learn about the forces that have shaped the development of this very important epidemic strain, a collaboration of researchers including myself, set out

to sequence and analyse the genome of the earliest archived LES isolate, LESB58. LESB58 was obtained from a Liverpool CF patient in 1988, eight years prior to the first published study on the LES (Cheng, et al., 1996). The LESB58 genome was sequenced by the Pathogen Production team at the Sanger Institute and I led the genome annotation; including the identification of many large GIs including five prophage clusters, one defective (pyocin) prophage cluster and five non-phage islands. In addition, Roger Levesque's research group performed an unbiased signature tagged mutagenesis (STM) study, and screening in a chronic rat lung infection model. I mapped these STM primer sequence reads to determine the genes implicated in the pathogenesis of LES. This study revealed genes from the prophage clusters that strongly impacted on competitiveness in this chronic infection model, indicating that acquisition of these prophage genes contributed to the success of the LES strain.

5.2 Genome annotation

I annotated the genome of LESB58, depicted in Figure 5.1 and with statistics available in Table 5.1, using a combination of automated methods and manual curation (see next paragraph). The genome is available through the *Pseudomonas* Genome Database at www.*Pseudomonas*.com, which represents a repository for all completed *Pseudomonas* genome sequences released publicly to date (Winsor, et al., 2009).

Coding sequences (CDS) within LES were predicted using Glimmer3 (Delcher, et al., 2007) and were assigned LES locus identifiers consisting of a "PLES_" prefix followed by five digits that are incremented in multiples of 10 to

allow for additional CDSs or non-coding RNAs. Orthologs in PA14 and PAO1 were identified for each LES CDS or non-coding RNA using a reciprocal best BLAST approach coupled with synteny and Ortholuge (Fulton, et al., 2006) analysis: In particular, each LES CDS was used as the query input for a FASTA search with either PA14 or PAO1, using an identity cutoff of 30% that covered at least 80% of the query and hit. The relaxed 30% cutoff was used to capture possible cases of substantial gene divergence and the following methods were used to eliminate cases of non-orthologous homologs. If the original LES CDS was identified as the top hit using the same search as for the PA14 or PAO1 top hit, then the top hits were considered probable orthologs. In cases where multiple top hits with the same score were identified, gene synteny, from whole genome alignments obtained with the program Mauve (Darling, et al., 2004), was used to identify the most probable ortholog. Orthologs were additionally characterized using Ortholuge (Fulton, et al., 2006). LES genes with identified orthologs in either PAO1 or PA14, with the most recent annotations from www.*Pseudomonas*.com (Winsor, et al., 2009), were transferred automatically. Gene annotations from PAO1 were selected for transfer over PA14 in cases where LES genes had orthologs in both, due to the higher level of updated manual curation of the PAO1 genome. LES CDSs that did not have an identified ortholog in PA14 or PAO1 were manually annotated based on significant BLAST matches from the NCBI nr database. Protein subcellular localization and COGs were predicted for each LES CDS using PSORTb 2.0 (Gardy, et al., 2005) and RPS-BLAST (Marchler-Bauer, et al., 2002), respectively.

Table 5.1 *P. aeruginosa* LESB58 genome statistics

Feature		Characteristics
Genome Size		6,601,757 base pairs
Total Number of Genes		6027
Protein Coding Genes		5931
RNA Genes		96
Pseudogenes		34
Genomic Islands (genes)		5 (214)
Prophage (genes)		6 (210)
PALES genes with no orthologs in[a]:	PAO1	574
	PA14	528
	PA7	825
	Any *P. aeruginosa* strains	350

[a] Orthologs were determined using a combination of reciprocal best BLAST hits and gene synteny analysis, with some validation by Ortholuge.

5.2.1 Virulence genes

The LESB58 genome carries virtually all of the reported virulence genes of *P. aeruginosa*. Of the 265 *P. aeruginosa* virulence factor CDSs described for strain PA01 (Wolfgang, et al., 2003), all but two are present in the LESB58 genome. Clearly orthologous CDSs to PA01 PA2399 (*pvdD*) and PA1392 were not present. PA2399 is a putative non-ribosomal peptide synthetase within the type I pyoverdine synthesis gene cluster. Instead, the LESB58 genome carries genes for the synthesis of a type III pyoverdine, which include a type-specific, divergent *pvdD* (Smith, et al., 2005). Notably, there are novel duplications of pyoverdine-associated genes in the genome of strain LESB58, which carries three identical copies of the *fpvAIII* gene (encoding the type III pyoverdine receptor) and the adjacent gene *pvdE* (encoding an ABC transporter). Two

additional but truncated versions of *pvdF* are also present. PA1392 is a hypothetical protein of unknown function. Some virulence-related LESB58 CDSs were divergent from strain PA01, including those matching PA1695 (*pscP*) and PA2525-7 (*pilABC*). The LES genome contains the type III secretion gene *pscP*, but with a ten residue deletion (5'-PTPTPTPTPT-3'; position 108-117) in the predicted protein in comparison to the strain PA01 predicted protein. Further analysis of virulence was performed in the signature tagged mutagenesis study described further below.

5.2.2 Motility organelles

Variations in the type IV pilin *pil* locus are not uncommon (Kus, et al., 2004). The LESB58 genome contained PilB and PilC CDSs sharing 88% and 84% identity with PAO1 orthologs respectively, but both matched *P. aeruginosa* strain 2192 orthologs with 99% identity. The LES putative PilA was identical to a previously reported unusual PilA (GenBank AAC63060; (Pasloske, et al., 1988)), but shared only 32% identity with the PA01 ortholog. Most important however in this regard were the experiments performed by Dr. Craig Winstanley's research group that showed the parental strain and tested clonal derivatives were completely devoid of any form of motility, including flagellin-dependent swimming motility, pilus dependent twitching motility and viscosity-regulated swarming motility (Table 5.2). This is consistent with the observation that unlike strains PA01 and PA14, LES tends to remain tightly associated with the agar beads utilized in the rat chronic lung model (Kukavica-Ibrulj, et al., 2008). Dr. Winstanley's research group also used electron microscopy to detect that neither

flagella nor pili were on the surface of LESB58, explaining the loss of motility. A whole *P. aeruginosa* PA14 genomic mutant library screen for deficiencies in swarming motility revealed that PA1628 mutants were less motile (E. Torfs and R.E.W. Hancock, unpublished data) and the equivalent gene in LESB58 was a pseudogene (PLES_36981/91; see Table 5.3 for a listing of all pseudogenes identified). Similarly other genes, that were adjacent to the homologs of other pseudogenes (PA2023, PA2026, PA2399, PA4688, PA5454, PA5655), led to loss of swarming motility when mutated in *P. aeruginosa* PA14.

Table 5.2 Motility defect in LES isolates.
Average zone diameters measured in militmeters (mm) from three replicates exhibiting about 5-15% standard deviation. Experiment performed by Dr. Winstanley's research group.

Strain	Swimming Zone (mm)	Twitching Zone (mm)	Swarming Zone (mm)
WT	23	50	29
H1024	2	8	4
H1025	4	8	4
H1026	3	5	5
H1027	7	15	4
H1028	2	8	3
H1029	3	5	6
H1030	3	8	7
H1031	2	3	7
H1032	3	7	7
H1033	8	11	8

5.2.3 Phenazine biosynthesis

Phenazine compounds produced by fluorescent *Pseudomonas* species are metabolites that function in microbial competitiveness, and appear to play a role in virulence in *P. aeruginosa*. As with other *P. aeruginosa* genomes, the genome of LESB58 contained two clusters of genes encoding putative phenazine biosynthesis pathways. One cluster matched that of strain PA01 *phzA2-phzG2* (PA1899-1905) but contained a *phzB* gene sharing greater identity to PA01 *phzB1* (PA4211). The second cluster began with orthologs to the strain PA01 *phzA1-phzB1* (PA4210-4211) but the downstream genes shared greater identity with PA01 *phzC2-phzG2*.

5.2.4 Lipopolysaccharide (LPS)

The genome of LESB58 carries a cluster of LPS O-antigen serotype O6 genes (Raymond, et al., 2002). O6 is a common serotype (Pirnay, et al., 2002) shared by the second most prevalent clone amongst the UK CF population, the Midlands 1 strain (Smart, et al., 2006). However as for many mature CF isolates (Hancock, et al., 1983), LES strains are non-typable and thus probably contain rough LPS lacking O-antigen. One likely reason for this is mutation to a pseudogene of the GDP-mannose 4,6-dehydratase gene (*rmd,* a homolog of PA5453), which is within the LPS biosynthesis gene cluster. It has been demonstrated that *rmd* knockout mutants are deficient in A-band LPS biosynthesis (Rocchetta, et al., 1998).

Table 5.3 Predicted pseudogenes in *P. aeruginosa* LESB58.

Pseudo-gene	Start	End	PLES ID	Product	PAO1 Locus ID	Homolog accession
	68630	68884	00541	Conserved hypothetical protein	PA0054	AAG03444
	68957	69178	00551	Conserved hypothetical protein	PA0054	AAG03444
	983975	984271	9011	Hypothetical protein	PA4075	AAG07462
	1064301	1064510	09841	Hypothetical protein	PA3991	AAG07378
wspE	1390018	1391562	12781	Probable chemotaxis sensor/effector	PA3704	AAG07091
wspF	1391589	1391954	12791	Probable methylesterase	PA3703	AAG07090
wspF	1391979	1392566	12801	Probable methylesterase	PA3703	AAG07090
	2542804	2543022	23691	Phage minor tail protein L		YP001347786
	2549148	2551028	23761	Hypothetical protein	PA0978	EAZ52070
pltB	2757842	2763775	25821	Polyketide synthase type I		AAQ90173
	2801226	2801822	26081	4-hydroxyphenylpyruvate dioxygenase		EAV77455
	2801863	2802039	26091	Transcriptional regulator, asnc family		ABE46104
	2802409	2802636	26111	Glutaredoxin		ABF54143
	2814258	2815349	26201	Outer membrane efflux protein		YP973578
	2833295	2834071	26331	Major facilitator superfamily MFS_1		YP001372980
	2881389	2881904	26861	Tn3 family transposase		ABI20725
mexF	3015431	3017683	28011	RND multidrug efflux transporter	PA2494	AAG05882
mexF	3017882	3018619	28021	RND multidrug efflux transporter	PA2494	AAG05882
mexT	3020021	3021151	28041	Transcriptional regulator mext	PA2492	AAG05880
pvdF	3181636	3181848	29001	Pyoverdine synthetase F	PA2396	AAG05784
pvdF	3186857	3187069	29031	Pyoverdine synthetase F	PA2396	AAG05784
gor	3647780	3648310	32971	Glutathione reductase	PA2025	AAG05413
gor	3648307	3649134	32981	Glutathione reductase	PA2025	AAG05413
	3652654	3653145	33031	Probable transcriptional regulator	PA2020	AAG05408

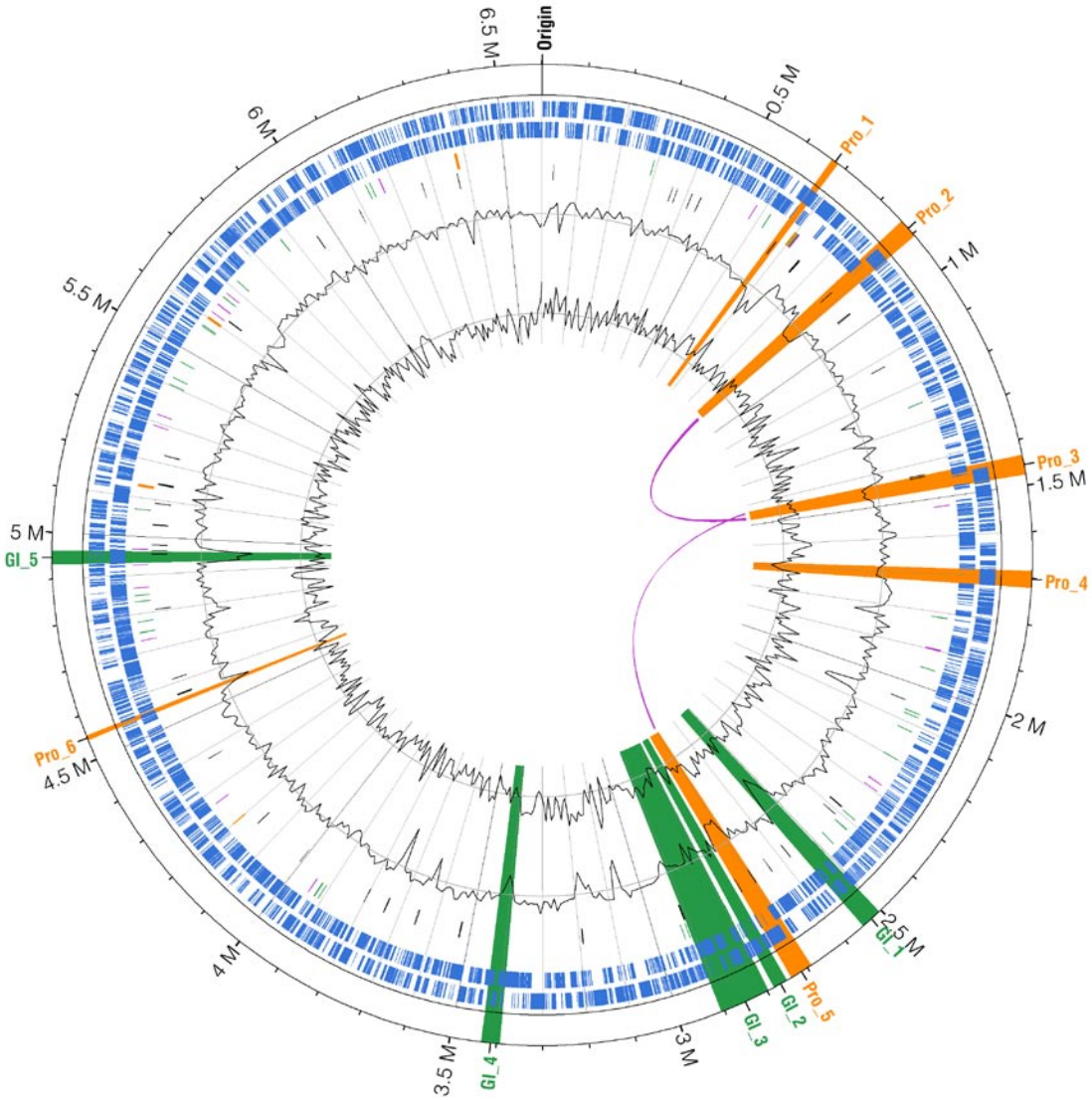
Pseudo-gene	Start	End	PLES ID	Product	PAO1 Locus ID	Homolog accession
	3977401	3977685	35801	Hypothetical protein	PA1749	AAG05138
	4095534	4096457	36981	Prob. 3-hydroxyacyl-coA-ehydrogenase	PA1628	AAG05017
	4096599	4097063	36991	Prob. 3-hydroxyacyl-coa dehydrogenase	PA1628	AAG05017
	5050925	5051860	45951	Still frameshift probable transcriptional regulator	PA0748	AAG04137
	5062180	5062368	46051	Hypothetical protein		ABJ09889
	5541211	5542146	50241	Hypothetical protein	PA4638	AAG08026
hitA	5592840	5594360	50731	Ferric iron-binding periplasmic protein	PA4687	AAG08074
	5708468	5709400	51711	Probable short-chain dehydrogenase	PA4786	AAG08172
gmd	6479023	6479541	58481	GDP-mannose 4,6-dehydratase	PA5453	AAG08838
	6597036	6597260	59621	Hypothetical protein	PA5566	AAG08951

5.2.5 Antibiotic Resistance

The original LESB58 isolate did not demonstrate remarkable antibiotic resistance, although like other *P. aeruginosa* isolates that infect the lungs of individuals with CF it is virtually impossible to eradicate once it becomes established (Hancock and Speert, 2000). In such cases, initial infections are suppressed by antibiotic treatment but over time antibiotics become increasingly less effective and resistance becomes established to one antibiotic after another. While many CF isolates acquire hyper mutator capabilities, e.g. by mutations in their *mutT* or *mutS* genes, LESB58 is not hypermutable, although subsequent isolates of this epidemic strain had acquired such status (Fothergill, et al., 2007). Nevertheless, the seeds for resistance development as observed in subsequent isolates are indeed present in the chromosome. The major cause of β-lactam

resistance is derepression of the class-C chromosomal β-lactamase (PA4110), and its homolog and those of all of the accessory regulatory genes are present in the genome. Another major cause of multidrug resistance is derepression of the expression of particular efflux pumps of which *P. aeruginosa* has a wide variety. Mutations in certain efflux pump genes were observed. For example the positive regulator of MexEFOprN, *mexT* (PA2492 homolog), was a pseudogene in LESB58, while the *mexF* (PA2494) gene is present but mutated suggesting that the MexEFOprN efflux system was minimally operative and perhaps not derepressible in the LES. Similarly, MexZ (PA2020) was also a pseudogene. However, the major efflux pump contributing to intrinsic and mutational resistance MexABOprM, and the ancillary system MexCDOprJ were intact. In other LES isolates exhibiting greater antimicrobial resistances, depression of AmpC and mutations in *mexR* and *mexZ*, implicated in up-regulation of the MexAB-OprM and MexXY efflux pumps respectively, have been identified (Salunkhe, et al., 2005). Of the 31 PAO1 CDSs annotated as functional class “antibiotic resistance and susceptibility” in the *Pseudomonas* Genome Database, only PA2818 (*arr*), a putative aminoglycoside response regulator, was absent from the genome of the LES.

Figure 5.1 Circular map of the *P. aeruginosa* LES genome.
Starting from outermost circle going inwards: major (500kb) and minor tick (100kb) measurements of the genome with estimated location of the origin; prophage (orange) and GIs (green) are highlighted across all tracks; protein coding genes (blue) on plus (outer) and minus strand (inner); tRNAs (green), rRNAs (orange), and all other non-coding RNA genes (purple); Signature Tagged Mutants (black); GC content (outer black line plot) with GC content average (grey line) and GC skew (inner black line plot) were calculated using a 10kb non-overlapping window. The location of two highly similar genomic regions of length 7.5 kb and 13.5 kb within the prophages are marked with looping purple lines, between their locations on the innermost circle. The identified prophage and GIs are distributed around the genome, but there is one notable cluster of LESGI-1, LESGI-2, and LESGI-3, reflecting the non-random nature of GI insertion in *P. aeruginosa* (Wiehlmann, et al., 2007). Significant sequence composition bias in 7 of the 9 regions was computationally identified (Table 5.4), while GC content deviating from the average can be observed for these regions in the figure.



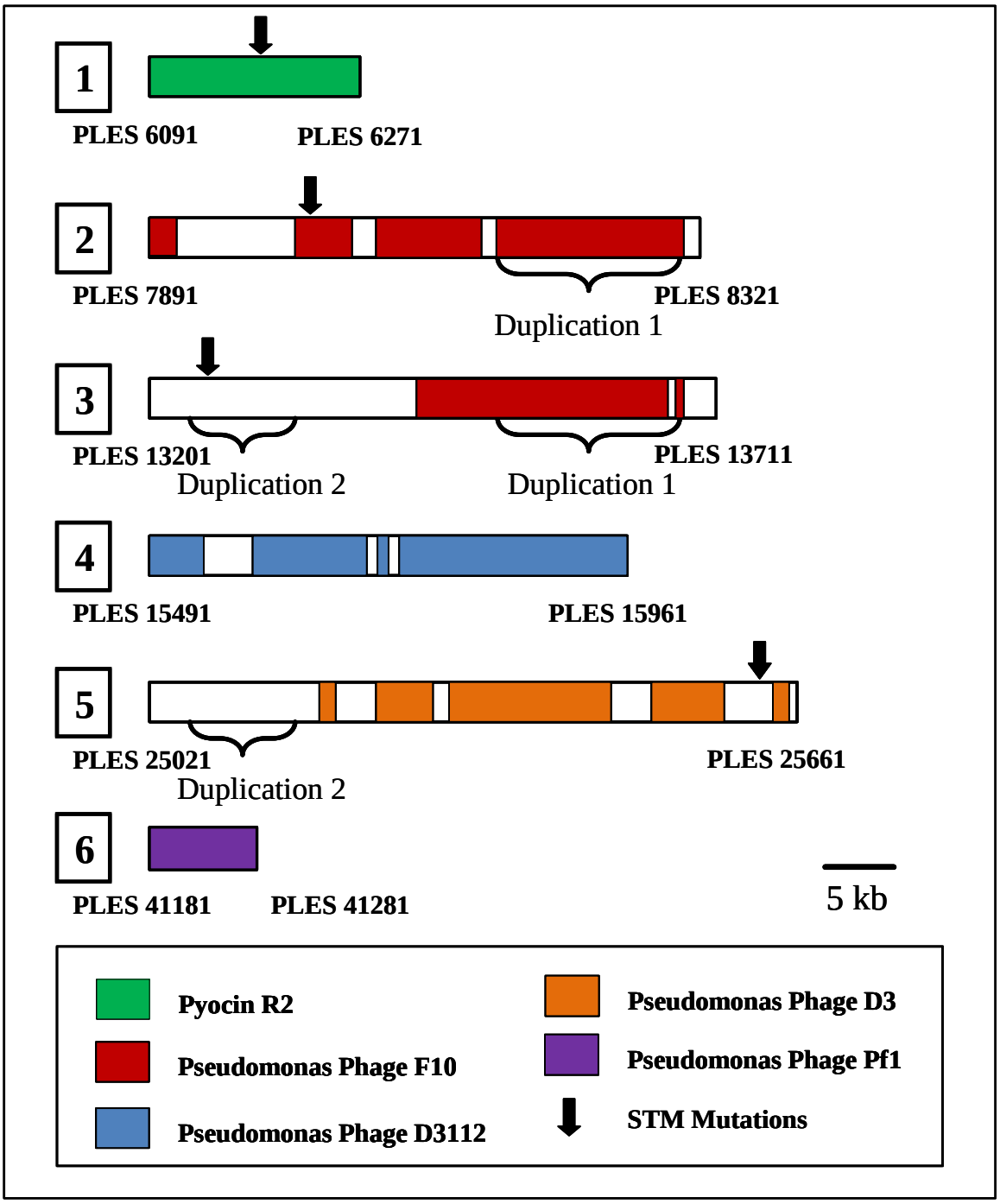
5.3 Identification of prophage and genomic islands within LES

Prior to the sequencing of LESB58, previous studies used subtractive hybridization to identify several regions that were not present in PAO1 and further quantified the prevalence of these regions amongst LES and non-LES CF isolates (Smart, et al., 2006). I refined these novel regions further and identified several new GIs and prophage regions using IslandPick (see Chapter 2). The exact boundaries of several of these regions were determined by Craig Winstanley's research group, by designing PCR primers reading out from each terminal region, and sequencing the resultant amplicons (Table 5.4).

5.3.1 LES bacteriophage gene clusters

Isolate LESB58 contained six prophage gene clusters, termed here prophages 1-6 (Table 5.4; Figure 5.2; Appendix File 5.1), of which four are absent from strain PA01. The LES prophage 1 gene cluster was a defective prophage predicted to encode pyocin R2. In strain PA01, two gene clusters in tandem encode pyocin R2 and F2, both of which are predicted to be evolved from phage tail genes. It has been demonstrated that either can be present or absent in *P. aeruginosa* (Ernst, et al., 2003; Nakayama, et al., 2000). The LES genome carried the pyocin R2 (P2 phage homolog) cluster (PLES06091-PLES06271) but not the pyocin F2 (phage λ homolog) cluster. It also carried pyocin S2 (PLES41691).

Figure 5.2 Phage clusters identified in LESB58 with significant similarities and positioning of STM mutants after *in-vivo* screening.



The LES prophage 2 gene cluster is 42.1 kb long and includes 44 CDSs of which 32 are homologous to the sequenced bacteriophage F10 (Kwan, et al., 2006), a member of the Siphoviridae family. Where orthologs were detected,

synteny was maintained between the two phage genomes, but matching regions were interspersed with non-matching CDSs (Appendix File 5.1).

The LES prophage 3 gene cluster was 42.8 kb and included 53 CDSs. A 13.6 kb region of this prophage, comprising 16 CDSs, shared 82.2% identity with a region of prophage 2 with homology to bacteriophage F10. Much of the rest of LES prophage 3 was similar to a region of the *P. aeruginosa* strain 2192 genome. However, LES prophage 3 also contained a 7.5 kb region (11 CDSs) with 99.8% identity to a region of LES prophage 5. LES prophage 4 shared a high level of similarity with the transposable phage D3112 (Wang, et al., 2004) but with some variation, especially at one terminus. LES prophage 5 had considerable similarity to bacteriophage D3 (Kropinski, 2000), although there was evidence of substantial genetic rearrangements (Figure 5.2).

The LES prophage 6 gene cluster was similar to the genome of bacteriophage Pf1 (Hill, et al., 1991). It has been suggested that Pf1 genes might be important in CF infections, in that Pf1 genes are up-regulated under conditions of reduced oxygen supply (Platt, et al., 2008), implicated in the augmentation of the antimicrobial efficacy of antibiotics (Hagens, et al., 2006), and play an active role in the activity and adaptation of *P. aeruginosa* populations biofilms (Mooij, et al., 2007; Sauer, et al., 2004; Webb, et al., 2004; Webb, et al., 2003). However, since most clinical isolates carry Pf1-like phages, these activities are not restricted to successful CF strains such as the LES (Finnan, et al., 2004).

Table 5.4 Identified genomic islands and prophage regions.

Region Name	Integration Site Relative To PAO1	Approximate start position		Number of Genes	Characteristics	
		Start[a]	End[a]		Sequence Composition Bias[b]	Mobility Gene(s) Present
Prophage1	PA0611 - PA0649	665561	680385	19	No	None
Prophage 2	PA4138 - PA4139	863875	906018	44	Yes	Integrase
Prophage 3	PA3663 - PA3664	1433756	1476547	53	Yes	Integrase
Prophage 4	PA3463 - PA3464	1684045	1720850	48	No	Transposase
LESGI-1	PA2727 - PA2737	2504700	2551100	31	Yes	Transposases & Integrases
Prophage 5	PA2603 - PA2604	2690450	2740350	65	Yes	Integrase
LESGI-2	PA2593 - PA2594	2751800	2783500	18	No	None
LESGI-3	PA2583 - PA2584	2796836	2907406	107	Yes	Integrase
LESGI-4	PA2217 - PA2229	3392800	3432228	32	Yes	None
Prophage 6	PA1191 - PA1192	4545190	4552788	12	Yes	Integrase
LESGI-5	PA0831 - PA0832	4931528	4960941	26	Yes	Integrase

[a] The approximate start and end positions are given for those regions without PCR analysis, except for Prophages 2 and 3 and LESGI-5.

[b] Sequence composition bias is indicated if the majority of the region was found to have sequence bias by either Alien Hunter (Vernikos et al., 2006) or the IslandPick-DIMOB (Hsiao et al., 2005) method.

5.3.2 LES genomic islands

The observed five LESB58 GIs are summarized in Table 5.4, depicted in Figure 5.3, and described in greater detail in Appendix File 5.1.

Many GIs have been identified in *P. aeruginosa* strains in previous studies; including, PAGI-1 (Liang, et al., 2001), PAGI-2 and PAGI-3 (Larbig, et al., 2002), PAGI-4 (Klockgether, et al., 2004), PAGI-5 (Battle, et al., 2008), PAGI-6 to PAGI-11(Battle, et al., 2009), PAPI-1 and PAPI-2 (He, et al., 2004), and pKLC102 (Klockgether, et al., 2004). Only two of the five GIs identified within the LES strain showed similarity to any previously identified *P. aeruginosa* island, with the last 67 kb of the 110 kb LESGI-3 island showing similarity to PAGI-2, PAGI-3, PAGI-5 and PAPI-1 (Figure 5.4), while LESGI-4 shared 46% identity with PAGI-1 over its entire length. As previously noted, pKLC102 and the related PAPI-1 were not found within the LES strain (Wurdemann and Tummler, 2007). In addition, PAGI-4 and PAGI-6 to PAGI-11 showed no significant homologs in the LESB58 genome.

LESGI-1 is inserted at a tRNA locus, and contains phage- and transposon-related CDSs. However, it also contained several CDSs sharing similarity with predicted proteins from non-pseudomonads such as the thermophilic anaerobe *Clostridium thermocellum* and the marine bacteria *Marinobacter sp*. Although mostly matching hypothetical proteins of no known function, the island included homologs of regulatory proteins, restriction-modification proteins, an ATPase and a sensor-kinase. This island included PALES23591, which contains the LES-F9 marker, although it is not unique to LES isolates (Smart, et al., 2006).

Figure 5.3 GIs identified in LESB58 with significant similarities and positioning of STM mutants after *in-vivo* screening.

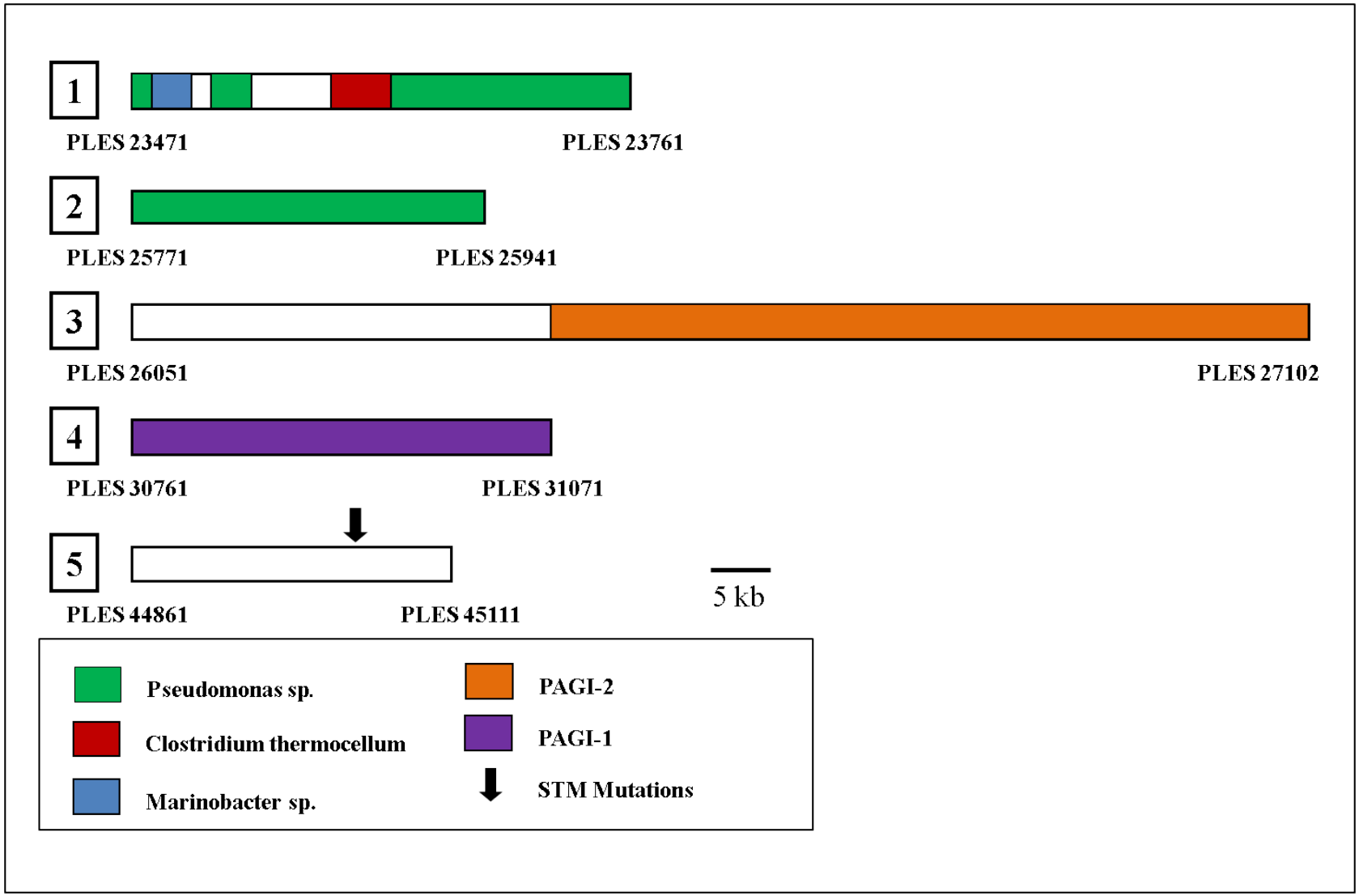
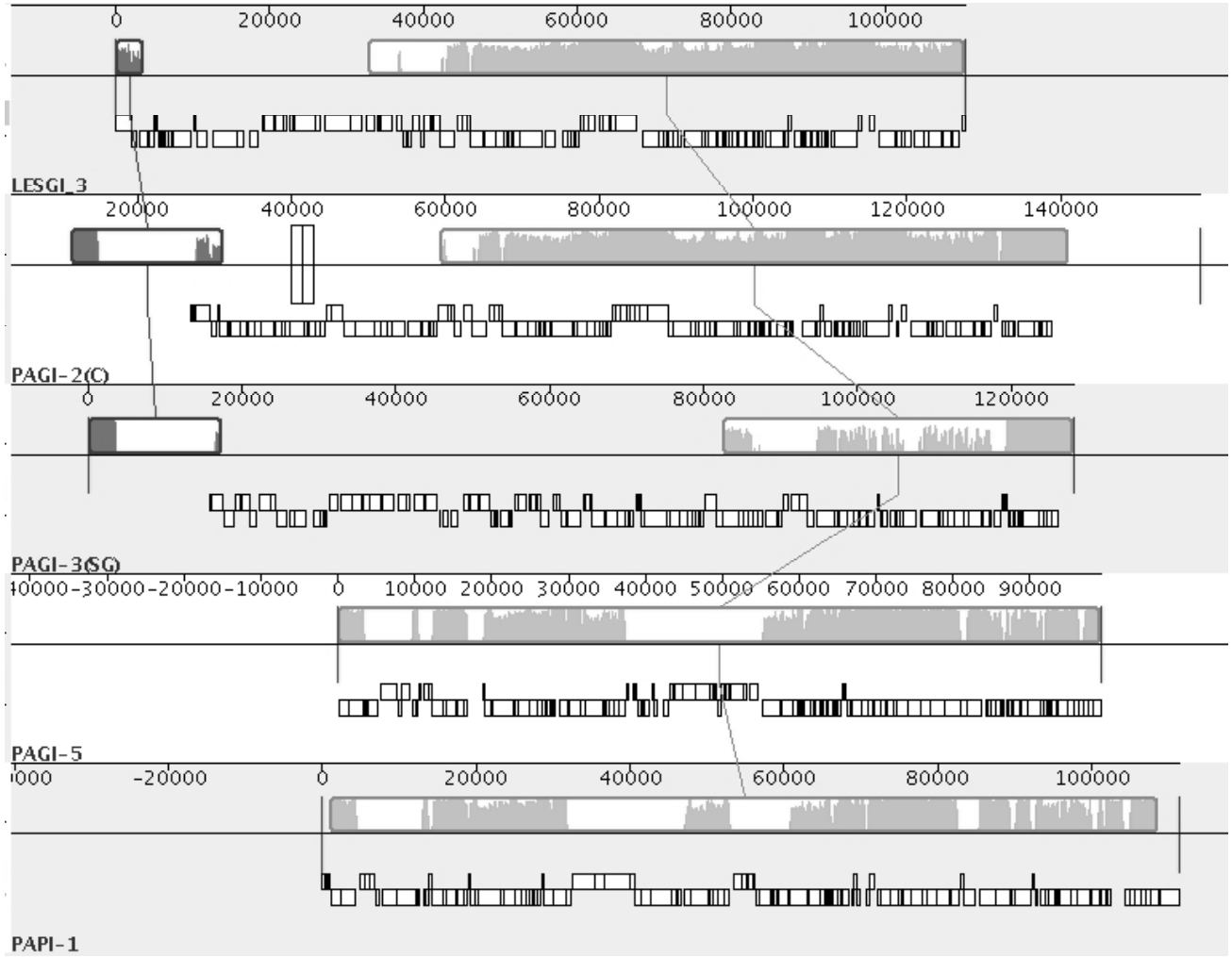


Figure 5.4 Alignment of LESGI-3 and four other previously published GIs in *P. aeruginosa*.
The similar regions are shown with sequence similarity for LESGI-3, PAGI-2 (AF440523), PAGI-3 (AF440524), PAGI-5(EF611301) and PAPI-1 (AY273869) (Battle, et al., 2008; He, et al., 2004; Larbig, et al., 2002). Genes within each region are shown as black boxes. Alignment was created using Mauve (Darling, et al., 2004).



LESGI-2 contained a pyoluteorin biosynthesis gene cluster (pltMRLABCDEFGZHIJKNO) sharing 99% nucleotide sequence identity with a cluster from *Pseudomonas* sp. M18 (AY394844), but containing a frameshift mutation in pltB. Pyoluteorin has antifungal activities (Bender, et al., 1999) and may play an important role in the ability of plant associated pseudomonads, such as *P. fluorescens*, to suppress a variety of plant diseases (Nowak-Thompson, et al., 1999). Interestingly, in LESB58, as previously found in the genome of *Pseudomonas* sp. M18, the island was adjacent to a PA2593-like CDS.

LESGI-3 was related to the PAGI-2 GI of Clone C (Klockgether, et al., 2004; Larbig, et al., 2002) with an alternative cargo region containing multiple putative transport proteins. LESGI-4 was related to the GI PAGI-1 (Liang, et al., 2001).

LES GI-5 was a novel island containing genes that largely match those of organisms other than *P. aeruginosa*, and including a putative phage integrase and plasmid replication genes (Appendix File 5.1). Aside from those associated with mobile elements, most predicted protein BLASTP matches shared <50% identity.

5.4 Signature tagged mutagenesis of LESB58

Signature tagged mutagenesis (STM) is a well defined method for determining, in a relatively unbiased manner, the importance of specific genes in *in-vivo* growth, through the relative ability of mutants to survive in animal models of infection. Since LES is an extremely robust epidemic isolate in CF and since it

previously demonstrated a competitive advantage over other *P. aeruginosa* strains in relevant animal models of infection (Kukavica-Ibrulj, et al., 2008), a STM analysis was performed on LESB58 by Dr. Roger Levesque's lab.

Of the 60 LESB58 STM mutants that were attenuated in lung infection, I was able to map 47 of them to an unambiguous sequence location (Table 5.5). Six of these genes were also found in a previous STM screening using strain PA01 (Table 5.5). DNA sequencing revealed insertions in most known functional gene classes. These included insertions in genes encoding products or processes previously implicated in pathogenesis of *P. aeruginosa*, such as the type III secretion protein PscH, a haem iron uptake receptor PhuR, TolA, the fimbrial usher CupA3, the alginate biosynthesis protein MucD, and two transcriptional regulators PLES27111 and PLES33031. Insertions in genes involved in the biosynthesis of type III pyoverdine (*pvdE*) and pyochelin (PLES07011) were identified, emphasizing the importance of both siderophores.

Table 5.5 List of 47 LESB58 virulence associated genes. Identified by PCR-based screening of 9216 STM mutants after passage through the chronic rat lung agar bead infection model.

STM Mutants	Insertion Site in LES genome	PAO1[a] ortholog	Putative function / comments
L103T13G	PLES00271	PA0028	Hypothetical protein
L28T5G	PLES03211	PA0325	Putative permease of ABC transporter
L70T18G	PLES03331	PA0336	Nudix hydrolase YgdP
L64T24G	PLES03721	PA0375	Cell division ABC transporter, permease protein FtsX
L52T19T	PLES04001	**PA0402**	PyrB Aspartate carbamoyltransferase
L114T20G	PLES06181	PA0622	Put. phage tail sheath protein/pyocin R2 (LES prophage 1)
L15T13G	PLES07011	PA4226	Dihydroaeruginoic acid synthetase
L124T1G	PLES08021	None	DNA replication protein DnaC (LES prophage 2)
L114T14G	PLES08731	PA4100	Probable dehydrogenase
L6T19G	PLES08751	PA4098	Probable short-chain dehydrogenase
L113T14T	PLES10401	PA3936	Probable permease of ABC taurine transporter
L124T11G	PLES13181	PA3666	Tetrahydrodipicolinate succinylase
L94T20G	PLES13261	None	Hypothetical protein (LES prophage 3)[b]
L111T2G	PLES19021	PA3166	Chorismate mutase
L14T10G	PLES22061	PA2858	Putative ABC transporter, permease protein
L111T13T	PLES22341	PA2831	Putative zinc carboxypeptidase
L106T24G	PLES23991	PA2705	Hypothetical protein
L52T24G	PLES23991	PA2705	Hypothetical protein
L52T5T	PLES23991	PA2705	Hypothetical protein
L14T9G	PLES24551	PA2650	Putative methyltransferase
L58T23G	PLES25621	None	Putative lytic enzyme (LES prophage 5)[c]
L19T13G	PLES27111	PA2583	Probable sensor /response regulator hybrid
L70T1G	PLES29051	None	PvdE; component of type III pyoverdine locus
L113T14G	PLES31971	PA2130	CupA3, fimbrial usher protein
L110T9G	PLES33001	PA2023	UTP-glucose-1-phosphate uridylyltransferase
L110T14G	PLES33031	PA2020	Probable transcriptional regulator
L13T13G	PLES33821	PA1941	Hypothetical protein
L124T10G	PLES33821	PA1941	Hypothetical protein
L82T13G	PLES34271	PA1897	Putative desaturase
L13T19G	PLES36081	PA1721	Type III export protein PscH

STM Mutants	Insertion Site in LES genome	PAO1[a] ortholog	Putative function / comments
L109T23T	PLES37591	**PA1569**	Prob major facilitator superfamily (MFS) transporter
L25T11T	PLES39641	**PA1449**	Flagellar biosynthetic protein FlhB[d]
L106T19G	PLES41401	**PA1181**	Conserved hypothetical protein
L54T20T	PLES41751	PA1144	Probable major facilitator superfamily (MFS) transporter
L54T13T	PLES43701	PA0945	PurM, phosphoribosylaminoimidazole synthetase
L57T4G	PLES45041	None	Hypothetical protein (LES GI-5)
L65T15G	PLES45141	PA0829	Probable hydrolase
L19T14G	PLES45311	PA0811	Probable major facilitator superfamily (MFS) transporter
L22T17G	PLES45771	**PA0766**	Serine protease MucD precursor
L121T13G	PLES46381	PA0692	Hypothetical protein
L64T1G	PLES46641	PA4284	Exodeoxyribonuclease V beta chain
L10T7G	PLES47381	PA4360	Putative chromosome segregation ATPase
L14T13G	PLES50951	PA4710	Putative haem uptake outer membrane receptor PhuR
L20T20G	PLES53911	PA5002	Hypothetical protein
L21T13G	PLES55011	**PA5111**	Lactoylglutathijne lyase
L61T13G	PLES56651	PA5271	Hypothetical protein
L127T13G	PLES57621	PA5367	ABC phosphate transporter membrane component

[a]Genes previously identified by STM screening of *P. aeruginosa* strain PAO1 (or present in the same operon as previously identified genes) are indicated in bold.

[b]This location was tentatively identified as it is within a duplicated region shared by LES prophage 5

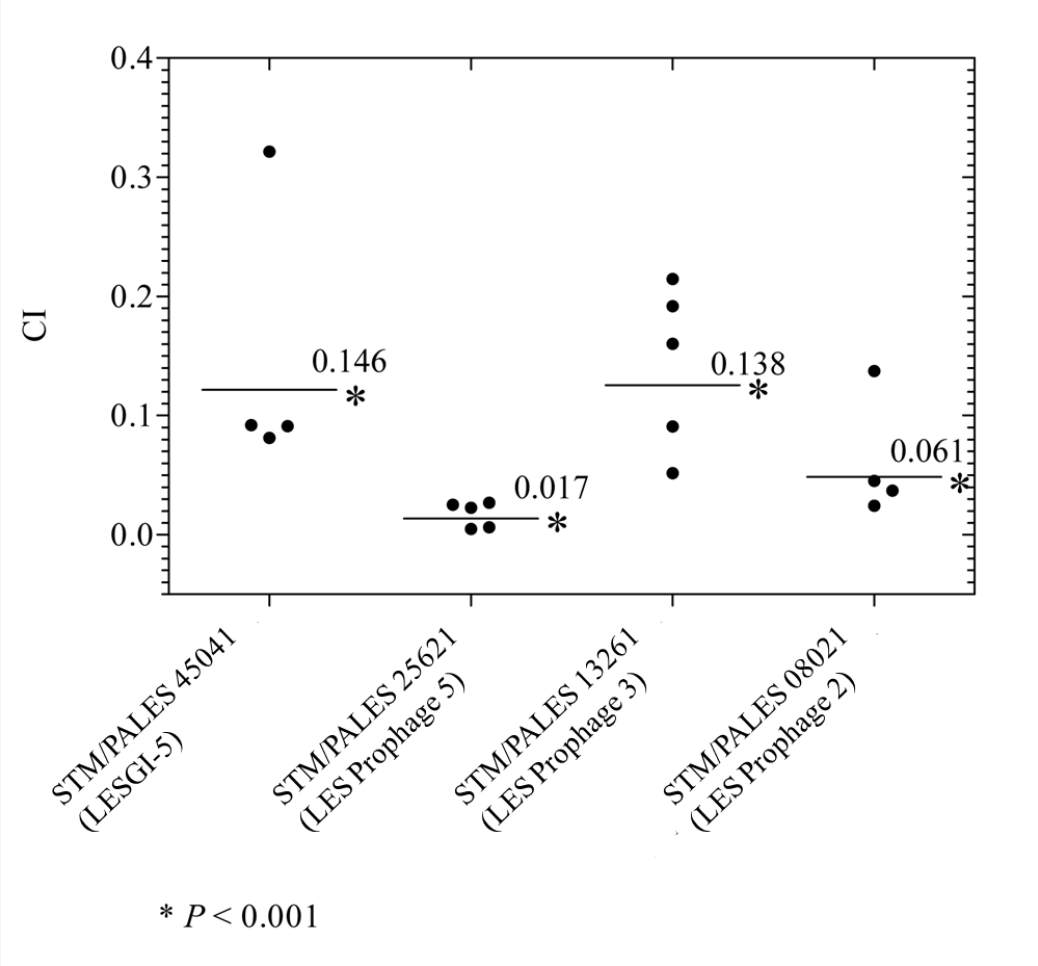
[c]Since it is likely that gene PLES25621 would not be expressed in a lysogen, it seems probable that the insertion in gene PLES25621 had a polar effect on downstream genes, affecting the expression of PLES25631, PLES26641and PLES25651, which are known to be part of LES prophage 5.

[d]Since the parent strain LES5B is relatively deficient in swimming motility which depends of flagella function (Table 5.2), it is hypothesized that the observation of this mutation within the characterized STM mutants reflects either an importance for the residual motility function, an alternative function for FlhB (e.g. in a Type III-like secretion event or adherence) or polar effects on one of the downstream genes.

5.4.1 *In-vivo* analysis of STM mutants having insertions in prophage and genomic islands

To assist in understanding the basis for the successful colonization of the LES in CF patients, the level of attenuation *in-vivo* was determined by Dr. Levesque's research group for 3 STM mutants having insertions in LES prophages -2, -3 and -5 and one STM mutant in the unique LES GI, LESGI-5 (Table 5.4). *In-vitro* growth was assessed for each of these STM mutants in mixed cultures with the wild-type (*in-vitro* competitive index [CI]) to confirm that these mutants did not affect *in-vitro* growth, and were not out-competed *in-vitro* by the wild-type LESB58 strain, yielding an *in-vitro* competitive index of around 1.0 after 18 hr in BHI broth. This contrasted with the results when competition was assessed *in-vivo*, for which the mutants were mixed with the wild-type strain LESB58 and grown in the rat lung infection model for 7 days. As depicted in Figure 5.5, mutants with insertions in both Prophages 2 and 5 caused a severe defect in growth and maintenance *in-vivo* which gave a significant 16- to 58- fold decrease of CFUs in rat lung tissues with competitive index values of 0.061 and 0.017, respectively. Mutants in Prophage 3 and LESGI-5 could be partially maintained in lung tissues with approximately 7-fold decreases in growth *in-vivo*.

Figure 5.5 ***In-vivo*** **competitive index (CI) of four STMs within** ***P. aeruginosa*** **LESB58. STM PALES_45041 (within LESGI-5), PALES_25621 (within LES Prophage 5), PALES_13261 (within LES Prophage 3), and PALES_08021 (within LES Prophage 2) grown for 7 days in the rat lung in competition with the wild-type LESB58 strain. Each circle represents the CI for a single animal in each group. A CI of less than 1 indicates an attenuation of virulence. The geometric mean of the CIs for all rats is shown as a solid line and statistically significant p value is indicated with an asterisk (* P < 0.001 with the Mann-Whitney sum test). This experiment was performed by Dr. Levesque's research group.**



5.5 Conclusions

The genome of *P. aeruginosa* exhibits a mosaic structure (Ernst, et al., 2003) and is composed of a "core genome" (approximately 90%) and an "accessory genome" (approximately 10%). The latter includes gene clusters involved in determining O-serotype (Raymond, et al., 2002), flagellin type (Arora,

et al., 2001), type IV pili (Kus, et al., 2004), siderophore production (Spencer, et al., 2003) as well as genomic/pathogenicity islands (Gal-Mor and Finlay, 2006; He, et al., 2004; Klockgether, et al., 2004; Larbig, et al., 2002; Liang, et al., 2001) and prophages. Although many of the known virulence genes are carried within the core genome of *P. aeruginosa* (Wolfgang, et al., 2003), genes from the accessory genome can contribute to pathogenicity. The genome of LESB58, like those sequenced previously, carries the core genome, including the vast majority of recognized virulence genes of *P. aeruginosa*. The genomic variations lie largely within five prophages and one defective prophage, and five large GIs, a few of which are related to those found in other strains of *P. aeruginosa.*

Extensive genome plasticity has been reported for *P. aeruginosa* clinical isolates, with phage sequences making a significant contribution to HGT leading to sequence diversity (Shen, et al., 2006). Indeed, it has been suggested that integrase-driven instability plays an important role in bacterial genomic evolution (Manson and Gilmore, 2006). Furthermore it has been demonstrated that phages can drive diversification of *P. aeruginosa* (Brockhurst, et al., 2005). More than 60 temperate phages have been isolated from *P. aeruginosa* (Akhverdian, et al., 1984; Wang, et al., 2004), and many have been genome sequenced.

It is well known that *P. aeruginosa* pathogenesis involves a variety of well known core genome functions (e.g. Type II and III secretion, iron transport, etc) and as well as other functionally important “accessory” gene clusters determining O-serotype, flagellin type, type IV pili and siderophore production [although these are only named accessory genes because of their sequence divergence and it is

arguable that these are really core functions]. This chapter has shown for one of the few well characterized “epidemic” strains of *P. aeruginosa* that the success of this organism, permitting it to be retained in a infection model relevant to CF, requires genetic information encoded on three prophages and one GI. This sheds some light on the crucial nature of the flow of genetic information through the accessory genome in such critical functions as the ability of an organism to grow successfully in a host possessing multiple mechanisms for impeding bacterial survival.

It has been demonstrated that *P. aeruginosa* virulence is combinatorial (Lee, et al., 2006). The studies described here indicate an ability to successfully establish colonization in what is usually a protected niche, the lung, indicate that this too involves a combinatorial process and involves both the core genome and key prophage and GI genes from the “accessory” genome to increase competitiveness.

CHAPTER 6 CRISPRS AND THEIR ASSOCIATION WITH GENOMIC ISLANDS

6.1 Introduction

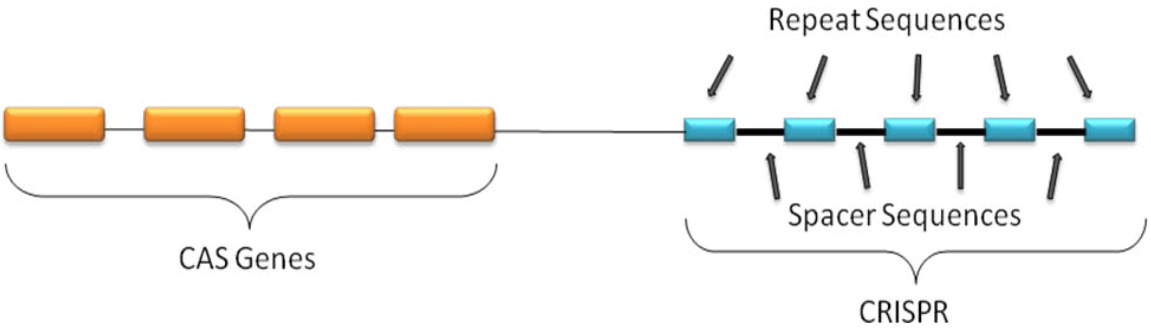
Clustered, regularly interspaced short palindromic repeats (CRISPRs) are genetic elements that have been identified in approximately 40% and 90% of Bacteria and Archaea genomes, respectively (Grissa, et al., 2007). A CRISPR consists of several identical repeats, separated by non-identical spacer sequences (Figure 6.1)(Sorek, et al., 2008). These repeat and spacer sequences typically range in size from 25-40 bps long, while the number of repeats in a single CRISPR varies widely from 2 to 250 (Grissa, et al., 2007).

Initially, CRISPRs were thought to be simple repetitive elements with no known function; however, recent research has shown that these elements along with CRISPR associated (CAS) genes are involved in a silencing mechanism that can provide protection against phage (Barrangou, et al., 2007). In this study, the authors showed that phage-resistant *Streptococcus thermophilus* could be produced when infected with phage. These phage resistant mutants were shown to have newly acquired spacer sequences that matched with 100% identity to the genome of the challenging phage. Barrangou, et al. verified that these spacer sequences were the cause of the newly acquired phage resistance by introducing these phage related spacer sequences into a phage-sensitive *S. thermophilus* and showing gain of phage-resistance. Several studies have shown

that the CRISPR region is expressed as a single RNA molecule that is then processed into small RNAs (sRNAs) (Tang, et al., 2002; Tang, et al., 2005). The CRISPR system was initially thought to target mRNAs and be analogous to the well described RNAi system in eukaryotes (Hannon, 2002), but a recent study showed that the CRISPR system targeted DNA and could block HGT of a plasmid by conjugation (Marraffini and Sontheimer, 2008).

Previous studies that have analysed the phylogenetic profiles of CAS genes suggest that CRISPR systems could be primarily transferred by HGT (Godde and Bickerton, 2006; Haft, et al., 2005). Although CRISPRs have been identified on 10 megaplasmids (Godde and Bickerton, 2006) and within two prophage in *Clostridium difficile* (Sebaihia, et al., 2006), a large scale analysis of CRISPRs and GIs has not been conducted.

Figure 6.1 Typical structure of a CRISPR system.



6.2 Over representation of CRISPRs within GIs

Predicted CRISPRs were obtained from the CRISPRdb (http://crispr.u-psud.fr/crispr/CRISPRHomePage.php). The entire database of CRISPRs was not available through the web interface, so the complete list was sent by email from

Ibtissem Grissa on October 29th, 2008. This database contained 1043 confirmed CRISPRs for 355 species (306 Bacteria and 49 Archaea). The coordinates of these CRISPRs were searched among the 5172 GIs in these 255 species that had been predicted by any of the GI prediction methods: IslandPick, SIGI-HMM, or IslandPath-DIMOB. In total, 128 CRISPRs in 76 organisms were found to be within these GIs and based on the proportion of sequence within GIs was twice as many as expected (see Table 6.1 and Appendix File 6.1). This over-representation of CRISPRs within GIs was found to be statistically significant (p value = $1.6x\ 10^{-16}$) using a chi-squared test.

Considering that CRISPRs have been identified in a larger proportion of Archaea genomes versus Bacteria genomes, the over-representation of CRISPRs in GIs was tested separately on Archaea and Bacteria datasets. While the Bacteria dataset still showed a significant over-representation of CRISPRs ($p = 8.1x10^{-18}$), the Archaea dataset was not statistically significant ($p = 0.02$) even though a similar trend was observed. This lack of CRISPR over-representation in GIs in Archaea could have biological significance in terms of Archaea obtaining CRISPRs through MGEs other than GIs. However, it is possible that it could be simply the result of not having enough sequenced Archaea genomes and that Archaea has less GIs (3.4% of genomes) than Bacteria (6.4% of genomes); therefore, limiting the statistical strength of the association calculation.

Table 6.1 Over-representation of CRISPRs in GIs.

Domain of Life	Number of Genomes	Number of GIs	Proportion of Genome in GIs	Total Number of CRISPRs	Expected Number of CRISPRs in GIs	Observed Number of CRISPRs in GIs	Significance (Chi-square Test)*
Archaea	49	298	3.7%	206	7.7	14	0.020
Bacteria	306	4874	6.4%	837	53.3	114	8.1×10^{-18}
Archaea and Bacteria	355	5172	6.1%	1043	64.0	128	1.6×10^{-16}

*x^2 test includes number of observed and expected CRISPRs outside of islands (data not shown).

6.3 GIs and CRISPRs have more phage genes

To approximate the contribution of phage to all GIs, the frequency of genes in GIs with 'phage' occurring in the annotation (referred to as 'phage genes' from now on) was enumerated and compared to the number of phage genes outside of GIs. As expected, GIs disproportionately contained a large number of genes with a phage annotation (6990 observed; 1264.2 expected; p =~0), indicating that a large number of GIs are likely prophage regions (Table 6.2). This over-representation of phage genes is seen in both Archaea (p= 4.5 x 10^{-20}) and Bacteria (p=~0). However, the proportion of GIs that contain at least one phage gene is much less in Archaea (18/355 =5.1%) compared to Bacteria (2095/11875=17.6%), which is representive of the smaller proportion of phage genes in general seen within Archaea (0.10%) versus Bacteria (0.79%).

GIs that contained CRISPRs showed the same over-representation of phage genes (p= 5.7 x10^{-5}, Table 6.2) when compared to genomic regions outside of GIs. The number of phage genes within GIs with CRISPRs was not significantly different when compared to the number of phage genes within GIs not containing CRISPRs (p=0.54, Table 6.2). A comparison between Archaea and Bacteria GIs containing CRISPRs and phage genes could not be conducted due to the small sample sizes for these categories.

Table 6.2 Over-representation of genes with 'phage' annotation in CRISPRs and GIs.

Genomic Regions	Number of 'phage genes'		Total number of genes in region	Chi-square test (x^2)
	Observed	Expected[3]		
Inside GIs[1]	6990	1264.22	165784	~0
Outside GIs[1]	12868	18593.78	2438303	
GIs containing CRISPR(s)[2]	13	4.5	1500	5.7×10^{-5}
Outside GIs[2]	812	820.5	274073	
GIs containing CRISPR(s)[2]	13	22.9	1500	0.54
GIs not containing CRISPR(s)[2]	267	257.1	16825	

[1]Total of 12230 GIs in 853 organisms

[2]Total of 5172 GIs in 355 organisms

[3]Expected = Total number of genes in region * Observed number of phage genes in both regions / Total number of genes in both regions

6.4 Conclusions

In this chapter, I have provided supporting evidence that CRISPRs are over-represented within GIs and therefore are likely being horizontally transferred. In addition, it has been shown that some of these GIs containing CRISPRs are likely to be prophage hinting that some phage are carrying these CRISPRs within their genome. Typically, it has been thought that CRISPRs are mainly beneficial to bacteria to defend against viral infections. However, it could be that some phage have started to take advantage of this system to possibly eliminate competing phage strains. In addition, I have identified differences between Archaea and Bacteria with respect to prophage, GIs, and CRISPRs. Upon reflection of the data it appears that Archaea have less phage genes within their genomes resulting in a lower proportion of GIs containing phage genes, and

presumably causing a lower proportion of GIs in Archaea when compared to Bacteria. This lack of phage genes in Archaea could be because approximately 90% of all Archaea genomes contain CRISPRs and provide a stronger defense against phage. Of course, most of this data is speculative, but it does provide some initial findings suggesting that the CRISPR system is widely spread and has broad functionality. Understanding the association of CRISPRs with islands is important, given the association of GIs with virulence and other microbial adaptations of medical and industrial importance.

CHAPTER 7 CONCLUDING REMARKS

When I started this project in 2005, the genome sequencing era was well under way with approximately 200 completely sequenced microbial genomes and rumours of the next generation sequencing machines that could sequence a microbial genome in an afternoon were just starting to surface. Researchers were quickly realizing that analyzing the genes from a single organism without reference to previously published genomes had limited conclusions. "Comparative genomics" was the new field that allowed comparisons between several related species (often with varying phenotypes) and provided insight into possible function of those new or missing genes. Now in 2009, the number of completed genomes will likely pass the 1000 mark and in coming years will continue to grow rapidly. Any genomics project today would not be considered without a large comparative genomics component. During this time span, I designed a computational method to identify GIs using a comparative genomics approach. Initially, I used this approach to identify GIs in a few classical organisms that had several closely related sequenced genomes (e.g. *Escherichia*, *Salmonella*, and *Pseudomonas*), but realized that selecting comparison genomes for all genomes would be increasingly time consuming as more genomes were sequenced. Therefore, I set out to develop a method that would automatically select comparison genomes in a relatively unbiased fashion and allow GI prediction without manual intervention. Although this sub-project

was challenging and had never been tackled previously, the outcome resulted in IslandPick, a GI prediction program that would become more useful as genome sequencing continued instead of becoming obsolete.

IslandPick, along with stringent parameters, was used to generate a robust dataset of GIs and with some slight adaptations, a dataset of conserved regions that were considered non-GIs. Considering that these datasets were derived from a non-sequence composition based method and had high agreement with a smaller dataset of previously published GIs, they were used to compare the accuracy of several previously published GI prediction programs. This was a much needed analysis that showed sequence based GI prediction programs had varying strengths and weaknesses with respect to precision, recall, and overall accuracy. Overall, two methods, SIGI-HMM and IslandPath-DIMOB, had the highest precision and accuracy.

Considering that these two methods (SIGI-HMM and IslandPath-DIMOB) were not available through a web interface, I decided to integrate them along with IslandPick into a new web-based GI prediction resource called IslandViewer. This is the first web resource that integrates multiple GI prediction methods applied to all sequenced genomes, in an automated, continually updated fashion, and allows users to submit their own newly sequenced genomes for analysis. This research has improved access to predictions of GIs, and evaluation of predictive methods, which should benefit a broad range of researchers interested in GIs, prokaryotic evolution, and comparative genomics.

I was then able to apply my GI prediction research in two different ways. First, I used the GI prediction programs to help identify GIs and prophage regions within the newly sequenced *P. aeruginosa* Liverpool Epidemic Strain, which had been shown previously to have increased pathogenicity in CF patients. Several genes within these variable regions were revealed to provide a competitive advantage over mutant strains in an *in-vivo* rat lung infection model; confirming the role of these GIs in bacteria pathogenicity. Second, the thousands of predicted GIs in IslandViewer were able to provide evidence that CRISPRs are over-represented within GIs and are often associated with prophage regions. This large-scale association could not have been conducted without the GI datasets from hundreds of organisms, and provides possible insight into global mechanisms of bacterial evolution that should be further studied.

Future study of GIs depends on robust computational methods that can help identify these regions accurately. My research has provided a major step forward in this direction, providing our first understanding of the accuracy of GI predictors, and providing tools that will facilitate more in depth analysis of GIs. There is still much we do not understand, but clearly GIs need to be studied further, given their apparent important role in prokaryotic evolution and medically important adaptations in pathogens.

APPENDIX

The CD-ROM attached forms a part of this work.

Spreadsheet files (.xls) can be opened with Microsoft Excel or Open Office, while other documents (.pdf) can be opened with any PDF reader.

Chapter 2:

Appendix File 2.1 MicrobeDB Schema.pdf	173KB
Appendix File 2.2 IslandPick Query Genomes.xls	83KB
Appendix File 2.3 IslandPick GI Predicitons.xls	190KB
Appendix File 2.4 Genomic Island Genes.xls	1424KB

Chapter 3:

Appendix File 3.1 Negative Dataset.xls	704KB
Appendix File 3.2 All GI method predictions.xls	5572KB
Appendix File 3.3 GI accuracy measurements.xls	126KB
Appendix File 3.4 Summary of GI accuracy.xls	21KB
Appendix File 3.5 Summary of GI accuracy using relaxed dataset.pdf	17KB
Appendix File 3.6 GI accuracy using relaxed dataset.xls	122KB

Chapter 5:

Appendix File 5.1 LES genes within prophage and genomic islands.xls	84KB

Chapter 6:

Appendix File 6.1 Genomic island containing CRISPRs.xls	78KB

REFERENCE LIST

Adhya, S.L. and Shapiro, J.A. (1969) The galactose operon of *E. coli* K-12. I. Structural and pleiotropic mutations of the operon, *Genetics*, **62**, 231-247.

Akhverdian, V.Z., Khrenova, E.A., Bogush, V.G., Gerasimova, T.V. and Kirsanov, N.B. (1984) [Wide distribution of transposable phages in natural *Pseudomonas aeruginosa* populations], *Genetika*, **20**, 1612-1619.

Al-Aloul, M., Crawley, J., Winstanley, C., Hart, C.A., Ledson, M.J. and Walshaw, M.J. (2004) Increased morbidity associated with chronic infection by an epidemic *Pseudomonas aeruginosa* strain in CF patients, *Thorax*, **59**, 334-336.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.

Armstrong, D., Bell, S., Robinson, M., Bye, P., Rose, B., Harbour, C., Lee, C., Service, H., Nissen, M., Syrmis, M. and Wainwright, C. (2003) Evidence for spread of a clonal strain of *Pseudomonas aeruginosa* among cystic fibrosis clinics, *J Clin Microbiol*, **41**, 2266-2267.

Arora, S.K., Bangera, M., Lory, S. and Ramphal, R. (2001) A genomic island in *Pseudomonas aeruginosa* carries the determinants of flagellin glycosylation, *Proc Natl Acad Sci U S A*, **98**, 9342-9347.

Azad, R.K. and Lawrence, J.G. (2007) Detecting laterally transferred genes: use of entropic clustering methods and genome position, *Nucleic Acids Res*, **35**, 4629-4639.

Babic, A., Lindner, A., Vulic, M., Stewart, E. and Radman, M. (2008) Direct visualization of horizontal gene transfer, *Science*, **319**, 1533 - 1536.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A. and Horvath, P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes, *Science*, **315**, 1709-1712.

Battle, S.E., Meyer, F., Rello, J., Kung, V.L. and Hauser, A.R. (2008) Hybrid pathogenicity island PAGI-5 contributes to the highly virulent phenotype of a *Pseudomonas aeruginosa* isolate in mammals, *J Bacteriol*, **190**, 7130-7140.

Battle, S.E., Rello, J. and Hauser, A.R. (2009) Genomic islands of *Pseudomonas aeruginosa*, *FEMS Microbiol Lett*, **290**, 70-78.

Bender, C.L., Rangaswamy, V. and Loper, J. (1999) Polyketide production by plant-associated pseudomonads, *Annu Rev Phytopathol*, **37**, 175-196.

Beres, S.B., Sylva, G.L., Barbian, K.D., Lei, B., Hoff, J.S., Mammarella, N.D., Liu, M.Y., Smoot, J.C., Porcella, S.F., Parkins, L.D., Campbell, D.S., Smith, T.M., McCormick, J.K., Leung, D.Y., Schlievert, P.M. and Musser, J.M. (2002) Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence, *Proc Natl Acad Sci U S A*, **99**, 10078-10083.

Berg, D.E. (1989) Transposon Tn5. In Berg, D.E. and Howe, M.M. (eds), *Mobile DNA*. ASM, Washington, DC, 185-210.

Berg, D.E., Berg, C.M. and Sasakawa, C. (1984) Bacterial transposon Tn5: evolutionary inferences, *Mol Biol Evol*, **1**, 411-422.

Blot, M. (1994) Transposable elements and adaptation of host bacteria, *Genetica*, **93**, 5-12.

Boucher, Y., Labbate, M., Koenig, J.E. and Stokes, H.W. (2007) Integrons: mobilizable platforms that promote genetic diversity in bacteria, *Trends Microbiol*, **15**, 301-309.

Boyd, E.F. and Brussow, H. (2002) Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved, *Trends Microbiol*, **10**, 521-529.

Brazas, M.D. and Hancock, R.E. (2005) Ciprofloxacin induction of a susceptibility determinant in *Pseudomonas aeruginosa*, *Antimicrob Agents Chemother*, **49**, 3222-3227.

Brockhurst, M.A., Buckling, A. and Rainey, P.B. (2005) The effect of a bacteriophage on diversification of the opportunistic bacterial pathogen, *Pseudomonas aeruginosa*, *Proc Biol Sci*, **272**, 1385-1391.

Bueno, S.M., Santiviago, C.A., Murillo, A.A., Fuentes, J.A., Trombert, A.N., Rodas, P.I., Youderian, P. and Mora, G.C. (2004) Precise excision of the large pathogenicity island, SPI7, in *Salmonella enterica* serovar Typhi, *J Bacteriol*, **186**, 3202-3213.

Buu-Hoi, A. and Horodniceanu, T. (1980) Conjugative transfer of multiple antibiotic resistance markers in *Streptococcus pneumoniae*, *J Bacteriol*, **143**, 313-320.

Casjens, S. (2003) Prophages and bacterial genomics: what have we learned so far?, *Molecular microbiology*, **49**, 277-300.

Casjens, S., Palmer, N., van Vugt, R., Huang, W.M., Stevenson, B., Rosa, P., Lathigra, R., Sutton, G., Peterson, J., Dodson, R.J., Haft, D., Hickey, E., Gwinn, M., White, O. and Fraser, C.M. (2000) A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*, *Mol Microbiol*, **35**, 490-516.

Chandler, M. and Mahillon, J. (2002) Insertion sequences revisited. In Lambowitz, A.M. (ed), *Mobile DNA*. American Society for Microbiology, Washington, DC., 631-662.

Chen, I. and Dubnau, D. (2004) DNA uptake during bacterial transformation, *Nat Rev Microbiol*, **2**, 241-249.

Cheng, K., Smyth, R.L., Govan, J.R., Doherty, C., Winstanley, C., Denning, N., Heaf, D.P., van Saene, H. and Hart, C.A. (1996) Spread of beta-lactam-resistant *Pseudomonas aeruginosa* in a cystic fibrosis clinic, *Lancet*, **348**, 639-642.

Chiapello, H., Bourgait, I., Sourivong, F., Heuclin, G., Gendrault-Jacquemard, A., Petit, M.A. and El Karoui, M. (2005) Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops, *BMC Bioinformatics*, **6**, 171.

Christie, P.J. (2001) Type IV secretion: intercellular transfer of macromolecules by systems ancestrally related to conjugation machines, *Mol Microbiol*, **40**, 294-305.

Clewell, D.B. and Flannagan, S.E. (1993) The conjugative transposons of Gram-positive bacteria. In Clewell, D.B. (ed), *Bacterial Conjugation*. Plenum, New York, 369-393.

Collis, C.M. and Hall, R.M. (1992) Gene cassettes from the insert region of integrons are excised as covalently closed circles, *Mol Microbiol*, **6**, 2875-2885.

Courvalin, P. and Carlier, C. (1987) Tn1545: a conjugative shuttle transposon, *Mol Gen Genet*, **206**, 259-264.

Craig, N.L. (2002) Tn7. In Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. (eds), *Mobile DNA II*. ASM Press, Washington, D.C., 423-456.

Darling, A.C.E., Mau, B., Blattner, F.R. and Perna, N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements, *Genome research*, **14**, 1394-1403.

Daubin, V. and Ochman, H. (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*, *Genome Res*, **14**, 1036-1042.

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

Delcher, A.L., Phillippy, A., Carlton, J. and Salzberg, S.L. (2002) Fast algorithms for large-scale genome alignment and comparison, *Nucleic Acids Res*, **30**, 2478-2483.

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

Doolittle, W.F. and Bapteste, E. (2007) Pattern pluralism and the Tree of Life hypothesis, *Proc Natl Acad Sci U S A*, **104**, 2043-2049.

Dubnau, D. (1999) DNA uptake in bacteria, *Annu Rev Microbiol*, **53**, 217-244.

Eddy, S.R. (2004) What is a hidden Markov model?, *Nat Biotechnol*, **22**, 1315-1316.

Edenborough, F.P., Stone, H.R., Kelly, S.J., Zadik, P., Doherty, C.J. and Govan, J.R. (2004) Genotyping of *Pseudomonas aeruginosa* in cystic fibrosis suggests need for segregation, *J Cyst Fibros*, **3**, 37-44.

Ely, B. and Croft, R.H. (1982) Transposon mutagenesis in *Caulobacter crescentus*, *J Bacteriol*, **149**, 620-625.

Ernst, R.K., D'Argenio, D.A., Ichikawa, J.K., Bangera, M.G., Selgrade, S., Burns, J.L., Hiatt, P., McCoy, K., Brittnacher, M., Kas, A., Spencer, D.H., Olson, M.V., Ramsey, B.W., Lory, S. and Miller, S.I. (2003) Genome mosaicism is conserved but not unique in *Pseudomonas aeruginosa* isolates from the airways of young children with cystic fibrosis, *Environ.Microbiol.*, **5**, 1341-1349.

Fedoroff, N., Wessler, S. and Shure, M. (1983) Isolation of the transposable maize controlling elements Ac and Ds, *Cell*, **35**, 235-242.

Finlay, B.B. and Falkow, S. (1997) Common themes in microbial pathogenicity revisited, *Microbiol Mol Biol Rev*, **61**, 136-169.

Finnan, S., Morrissey, J.P., O'Gara, F. and Boyd, E.F. (2004) Genome diversity of *Pseudomonas aeruginosa* isolates from cystic fibrosis patients and the hospital environment, *J Clin Microbiol*, **42**, 5783-5792.

Fothergill, J.L., Panagea, S., Hart, C.A., Walshaw, M.J., Pitt, T.L. and Winstanley, C. (2007) Widespread pyocyanin over-production among isolates of a cystic fibrosis epidemic strain, *BMC Microbiol*, **7**, 45.

Fothergill, J.L., Upton, A.L., Pitt, T.L., Hart, C.A. and Winstanley, C. (2008) Diagnostic multiplex PCR assay for the identification of the Liverpool, Midlands 1 and Manchester CF epidemic strains of *Pseudomonas aeruginosa*, *J Cyst Fibros*, **7**, 258-261.

Franke, A.E. and Clewell, D.B. (1981) Evidence for a chromosome-borne resistance transposon (Tn916) in *Streptococcus faecalis* that is capable of "conjugal" transfer in the absence of a conjugative plasmid, *J Bacteriol*, **145**, 494-502.

Freifelder, D. and Meselson, M. (1970) Topological relationship of prophage lambda to the bacterial chromosome in lysogenic cells, *Proc Natl Acad Sci U S A*, **65**, 200-205.

Fulton, D.L., Li, Y.Y., Laird, M.R., Horsman, B.G., Roche, F.M. and Brinkman, F.S. (2006) Improving the specificity of high-throughput ortholog prediction, *BMC Bioinformatics*, **7**, 270.

Gal-Mor, O. and Finlay, B.B. (2006) Pathogenicity islands: a molecular toolbox for bacterial virulence, *Cell Microbiol*, **8**, 1707-1719.

Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M. and Brinkman, F.S. (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis, *Bioinformatics*, **21**, 617-623.

Glasner, J.D., Plunkett, G., 3rd, Anderson, B.D., Baumler, D.J., Biehl, B.S., Burland, V., Cabot, E.L., Darling, A.E., Mau, B., Neeno-Eckwall, E.C., Pot, D., Qiu, Y., Rissman, A.I., Worzella, S., Zaremba, S., Fedorko, J., Hampton, T., Liss, P., Rusch, M., Shaker, M., Shaull, L., Shetty, P., Thotakura, S., Whitmore, J., Blattner, F.R., Greene, J.M. and Perna, N.T. (2008) Enteropathogen Resource Integration Center (ERIC): bioinformatics support for research on biodefense-relevant enterobacteria, *Nucleic Acids Res*, **36**, D519-523.

Glasner, J.D., Rusch, M., Liss, P., Plunkett, G., 3rd, Cabot, E.L., Darling, A., Anderson, B.D., Infield-Harm, P., Gilson, M.C. and Perna, N.T. (2006) ASAP: a resource for annotating, curating, comparing, and disseminating genomic data, *Nucleic Acids Res*, **34**, D41-45.

Godde, J.S. and Bickerton, A. (2006) The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes, *J Mol Evol*, **62**, 718-729.

Gogol, E.B., Cummings, C.A., Burns, R.C. and Relman, D.A. (2007) Phase variation and microevolution at homopolymeric tracts in Bordetella pertussis, *BMC Genomics*, **8**, 122.

Greene, J.M., Perna, N. and Blattner, F. (2007) ERIC-Comprehensive bioinformatics resources for enteropathogens, *Microbe*, **2**, 322-323.

Griffiths, F. (1928) The significance of pneumococcal types, *Journal of Hygiene*, **27**, 113-159.

Grissa, I., Vergnaud, G. and Pourcel, C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats, *BMC Bioinformatics*, **8**, 172.

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R. and Goebel, W. (1990) Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal *Escherichia coli* isolates, *Microb Pathog*, **8**, 213-225.

Hacker, J., Blum-Oehler, G., Muhldorfer, I. and Tschape, H. (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution, *Molecular Microbiology*, **23**, 1089-1097.

Hacker, J. and Kaper, J.B. (2000) Pathogenicity islands and the evolution of microbes., *Annual Review of Microbiology*, **54**, 641-679.

Haft, D.H., Selengut, J., Mongodin, E.F. and Nelson, K.E. (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes, *PLoS Comput Biol*, **1**, e60.

Hagens, S., Habel, A. and Blasi, U. (2006) Augmentation of the antimicrobial efficacy of antibiotics by filamentous phage, *Microb Drug Resist*, **12**, 164-168.

Hall, R.M., Collis, C.M., Kim, M.J., Partridge, S.R., Recchia, G.D. and Stokes, H.W. (1999) Mobile gene cassettes and integrons in evolution, *Ann.N.Y.Acad.Sci.*, **870**, 68-80.

Hancock, R. and Speert, D. (2000) Antibiotic resistance in *Pseudomonas aeruginosa*: mechanisms and impact on treatment, *Drug Resist Updat*, **3**, 247-255.

Hancock, R.E., Mutharia, L.M., Chan, L., Darveau, R.P., Speert, D.P. and Pier, G.B. (1983) *Pseudomonas aeruginosa* isolates from patients with cystic fibrosis: a class of serum-sensitive, nontypable strains deficient in lipopolysaccharide O side chains, *Infect Immun*, **42**, 170-177.

Haniford, D.B. (2002) Transposon Tn10. In Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. (eds), *Mobile DNA II*. ASM Press, Washington, D.C., 457-483.

Hannon, G.J. (2002) RNA interference, *Nature*, **418**, 244-251.

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.G., Ohtsubo, E., Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C., Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori, M. and Shinagawa, H. (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12, *DNA Res*, **8**, 11-22.

He, J., Baldini, R.L., Deziel, E., Saucier, M., Zhang, Q., Liberati, N.T., Lee, D., Urbach, J., Goodman, H.M. and Rahme, L.G. (2004) The broad host range pathogen *Pseudomonas aeruginosa* strain PA14 carries two pathogenicity islands harboring plant and animal virulence genes, *Proc Natl Acad Sci U S A*, **101**, 2530-2535.

Hill, D.F., Short, N.J., Perham, R.N. and Petersen, G.B. (1991) DNA sequence of the filamentous bacteriophage Pf1, *J Mol Biol*, **218**, 349-364.

Hochhut, B., Lotfi, Y., Mazel, D., Faruque, S.M., Woodgate, R. and Waldor, M.K. (2001) Molecular analysis of antibiotic resistance gene clusters in *Vibrio cholerae* O139 and O1 SXT constins, *Antimicrob Agents Chemother*, **45**, 2991-3000.

Hsiao, W., Wan, I., Jones, S.J. and Brinkman, F.S. (2003) IslandPath: aiding detection of genomic islands in prokaryotes, *Bioinformatics*, **19**, 418-420.

Hsiao, W.W., Ung, K., Aeschliman, D., Bryan, J., Finlay, B.B. and Brinkman, F.S. (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands, *PLoS Genet*, **1**, e62.

Karaolis, D.K., Johnson, J.A., Bailey, C.C., Boedeker, E.C., Kaper, J.B. and Reeves, P.R. (1998) A *Vibrio cholerae* pathogenicity island associated with epidemic and pandemic strains, *Proc Natl Acad Sci U S A*, **95**, 3134-3139.

Karlin, S. (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes., *Trends In microbiology*, **9**, 335-343.

Karlin, S., Mrazek, J. and Campbell, A.M. (1998) Codon usages in different gene classes of the *Escherichia coli* genome, *Mol Microbiol*, **29**, 1341-1355.

Klockgether, J., Reva, O., Larbig, K. and Tummler, B. (2004) Sequence analysis of the mobile genome island pKLC102 of *Pseudomonas aeruginosa* C, *J Bacteriol*, **186**, 518-534.

Korbel, J.O., Snel, B., Huynen, M.A. and Bork, P. (2002) SHOT: a web server for the construction of genome phylogenies., *Trends in genetics : TIG.*, **18**, 158-162.

Kropinski, A.M. (2000) Sequence of the genome of the temperate, serotype-converting, *Pseudomonas aeruginosa* bacteriophage D3, *J Bacteriol*, **182**, 6066-6074.

Kukavica-Ibrulj, I., Bragonzi, A., Paroni, M., Winstanley, C., Sanschagrin, F., O'Toole, G.A. and Levesque, R.C. (2008) In vivo growth of *Pseudomonas aeruginosa* strains PAO1 and PA14 and the hypervirulent strain LESB58 in a rat model of chronic lung infection, *J Bacteriol*, **190**, 2804-2813.

Kurtz, S. and Schleiermacher, C. (1999) REPuter: fast computation of maximal repeats in complete genomes, *Bioinformatics*, **15**, 426-427.

Kus, J.V., Tullis, E., Cvitkovitch, D.G. and Burrows, L.L. (2004) Significant differences in type IV pilin allele distribution among *Pseudomonas aeruginosa* isolates from cystic fibrosis (CF) versus non-CF patients, *Microbiology*, **150**, 1315-1326.

Kwan, T., Liu, J., Dubow, M., Gros, P. and Pelletier, J. (2006) Comparative genomic analysis of 18 *Pseudomonas aeruginosa* bacteriophages, *J Bacteriol*, **188**, 1184-1187.

Langille, M.G. and Brinkman, F.S. (2009) IslandViewer: an integrated interface for computational identification and visualization of genomic islands, *Bioinformatics*, **25**, 664-665.

Langille, M.G.I., Hsiao, W.W.L. and Brinkman, F.S.L. (2008) Evaluation of genomic island predictors using a comparative genomics approach, *BMC Bioinformatics*, **9**, 329.

Larbig, K.D., Christmann, A., Johann, A., Klockgether, J., Hartsch, T., Merkl, R., Wiehlmann, L., Fritz, H.-J. and Tummler, B. (2002) Gene islands integrated into tRNA(Gly) genes confer genome diversity on a *Pseudomonas aeruginosa* clone, *J Bacteriol*, **184**, 6665-6680.

Laslett, D., Canback, B. and Andersson, S. (2002) BRUCE: a program for the detection of transfer-messenger RNA genes in nucleotide sequences, *Nucleic Acids Res*, **30**, 3449-3453.

Lawrence, J.G. (2005) Common themes in the genome strategies of pathogens, *Curr Opin Genet Dev*, **15**, 584-588.

Lawrence, J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange, *J Mol Evol*, **44**, 383-397.

Lawson, F.S., Charlebois, R.L. and Dillon, J.A. (1996) Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life, *Mol Biol Evol*, **13**, 970-977.

Lederberg, J. and Tatum, E.L. (1946) Novel genotypes in mixed cultures of biochemical mutants of bacteia, *Cold Spring Harbor Symposia on Quantitative Biology*, **11**, 113-114.

Lee, D., Urbach, J., Wu, G., Liberati, N., Feinbaum, R., Miyata, S., Diggins, L., He, J., Saucier, M., Deziel, E., Friedman, L., Li, L., Grills, G., Montgomery, K., Kucherlapati, R., Rahme, L. and Ausubel, A. (2006) Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial, *Genome Biology*, **7**, 90.

Lenski, R.E. (2004) Phenotypic and genomic evolution during a 20000 generation experiment with the bacterium *Escherichia coli*, *Plant Breeding Reviews*, **24**, 225-265.

Lewis, D.A., Jones, A., Parkhill, J., Speert, D.P., Govan, J.R., Lipuma, J.J., Lory, S., Webb, A.K. and Mahenthiralingam, E. (2005) Identification of DNA markers for a transmissible *Pseudomonas aeruginosa* cystic fibrosis strain, *Am J Respir Cell Mol Biol*, **33**, 56-64.

Liang, X., Pham, X.Q., Olson, M.V. and Lory, S. (2001) Identification of a genomic island present in the majority of pathogenic isolates of *Pseudomonas aeruginosa*, *J Bacteriol*, **183**, 843-853.

Lorenz, M.G. and Wackernagel, W. (1994) Bacterial gene transfer by natural genetic transformation in the environment, *Microbiol Rev*, **58**, 563-602.

Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res*, **25**, 955-964.

Lwoff, A. (1953) Lysogeny, *Bacteriol Rev*, **17**, 269-337.

Mahillon, J. and Chandler, M. (1998) Insertion sequences, *Microbiol Mol Biol Rev*, **62**, 725-774.

Manson, J.M. and Gilmore, M.S. (2006) Pathogenicity island integrase cross-talk: a potential new tool for virulence modulation, *Mol Microbiol*, **61**, 555-559.

Mantri, Y. and Williams, K.P. (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities, *Nucleic Acids Research*, **32**, D55-58.

Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y. and Bryant, S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure, *Nucleic Acids Res*, **30**, 281-283.

Marraffini, L.A. and Sontheimer, E.J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA, *Science*, **322**, 1843-1845.

Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J.M., Koehrsen, M., Rokas, A., Yandava, C.N., Engels, R., Zeng, E., Olavarietta, R., Doud, M., Smith, R.S., Montgomery, P., White, J.R., Godfrey, P.A., Kodira, C., Birren, B., Galagan, J.E. and Lory, S. (2008) Dynamics of *Pseudomonas aeruginosa* genome evolution, *Proc Natl Acad Sci U S A*, **105**, 3100-3105.

Mazel, D. (2006) Integrons: agents of bacterial evolution., *Nature reviews. Microbiology*, **4**, 608-620.

Mazel, D., Dychinco, B., Webb, V.A. and Davies, J. (1998) A distinctive class of integron in the *Vibrio cholerae* genome, *Science*, **280**, 605-608.

McCallum, S.J., Corkill, J., Gallagher, M., Ledson, M.J., Hart, C.A. and Walshaw, M.J. (2001) Superinfection with a transmissible strain of *Pseudomonas aeruginosa* in adults with cystic fibrosis chronically colonised by *P aeruginosa*, *Lancet*, **358**, 558-560.

McCallum, S.J., Gallagher, M.J., Corkill, J.E., Hart, C.A., Ledson, M.J. and Walshaw, M.J. (2002) Spread of an epidemic *Pseudomonas aeruginosa* strain from a patient with cystic fibrosis (CF) to non-CF relatives, *Thorax*, **57**, 559-560.

McClelland, M., Sanderson, K.E., Spieth, J., Clifton, S.W., Latreille, P., Courtney, L., Porwollik, S., Ali, J., Dante, M., Du, F., Hou, S., Layman, D., Leonard, S., Nguyen, C., Scott, K., Holmes, A., Grewal, N., Mulvaney, E., Ryan, E., Sun, H., Florea, L., Miller, W., Stoneking, T., Nhan, M., Waterston, R. and Wilson, R.K. (2001) Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2, *Nature*, **413**, 852-856.

McClintock, B. (1941) The Stability of Broken Ends of Chromosomes in Zea Mays, *Genetics*, **26**, 234-282.

Merkl, R. (2004) SIGI: score-based identification of genomic islands, *BMC Bioinformatics*, **5**, 22.

Middendorf, B., Hochhut, B., Leipold, K., Dobrindt, U., Blum-Oehler, G. and Hacker, J. (2004) Instability of pathogenicity islands in uropathogenic *Escherichia coli* 536, *J Bacteriol*, **186**, 3086-3096.

Mohan, K., Fothergill, J.L., Storrar, J., Ledson, M.J., Winstanley, C. and Walshaw, M.J. (2008) Transmission of *Pseudomonas aeruginosa* epidemic strain from a patient with cystic fibrosis to a pet cat, *Thorax*, **63**, 839-840.

Mooij, M.J., Drenkard, E., Llamas, M.A., Vandenbroucke-Grauls, C.M., Savelkoul, P.H., Ausubel, F.M. and Bitter, W. (2007) Characterization of the integrated filamentous phage Pf5 and its involvement in small-colony formation, *Microbiology*, **153**, 1790-1798.

Naas, T., Blot, M., Fitch, W.M. and Arber, W. (1994) Insertion sequence-related genetic variation in resting *Escherichia coli* K-12, *Genetics*, **136**, 721-730.

Naas, T., Mikami, Y., Imai, T., Poirel, L. and Nordmann, P. (2001) Characterization of In53, a class 1 plasmid- and composite transposon-located integron of *Escherichia coli* which carries an unusual array of gene cassettes, *Journal of Bacteriology*, **183**, 235-249.

Nakamura, Y., Gojobori, T. and Ikemura, T. (1999) Codon usage tabulated from the international DNA sequence databases; its status 1999, *Nucleic Acids Res*, **27**, 292.

Nakamura, Y., Itoh, T., Matsuda, H. and Gojobori, T. (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes., *Nature genetics.*, **36**, 760-766.

Nakayama, K., Takashima, K., Ishihara, H., Shinomiya, T., Kageyama, M., Kanaya, S., Ohnishi, M., Murata, T., Mori, H. and Hayashi, T. (2000) The R-type pyocin of *Pseudomonas aeruginosa* is related to P2 phage, and the F-type is related to lambda phage, *Mol Microbiol*, **38**, 213-231.

Nowak-Thompson, B., Chaney, N., Wing, J.S., Gould, S.J. and Loper, J.E. (1999) Characterization of the pyoluteorin biosynthetic gene cluster of *Pseudomonas fluorescens* Pf-5, *J Bacteriol*, **181**, 2166-2174.

O'Carroll, M.R., Syrmis, M.W., Wainwright, C.E., Greer, R.M., Mitchell, P., Coulter, C., Sloots, T.P., Nissen, M.D. and Bell, S.C. (2004) Clonal strains of *Pseudomonas aeruginosa* in paediatric and adult cystic fibrosis units, *Eur Respir J*, **24**, 101-106.

Ochman, H., Lawrence, J.G. and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation., *Nature*, **405**, 299-304.

Ogawa, A. and Takeda, T. (1993) The gene encoding the heat-stable enterotoxin of *Vibrio cholerae* is flanked by 123-base pair direct repeats, *Microbiol.Immunol.*, **37**, 607-616.

Ohnishi, M., Kurokawa, K. and Hayashi, T. (2001) Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors?, *Trends Microbiol*, **9**, 481-485.

Ou, H.Y., Chen, L.L., Lonnen, J., Chaudhuri, R.R., Thani, A.B., Smith, R., Garton, N.J., Hinton, J., Pallen, M., Barer, M.R. and Rajakumar, K. (2006) A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria, *Nucleic Acids Res*, **34**, e3.

Ou, H.Y., He, X., Harrison, E.M., Kulasekara, B.R., Thani, A.B., Kadioglu, A., Lory, S., Hinton, J.C., Barer, M.R., Deng, Z. and Rajakumar, K. (2007) MobilomeFINDER: web-based tools for in silico and experimental discovery of bacterial genomic islands, *Nucleic Acids Res*, **35**, W97-W104.

Panagea, S., Winstanley, C., Walshaw, M.J., Ledson, M.J. and Hart, C.A. (2005) Environmental contamination with an epidemic strain of *Pseudomonas aeruginosa* in a Liverpool cystic fibrosis centre, and study of its survival on dry surfaces, *J Hosp Infect*, **59**, 102-107.

Parkhill, J., Dougan, G., James, K.D., Thomson, N.R., Pickard, D., Wain, J., Churcher, C., Mungall, K.L., Bentley, S.D., Holden, M.T., Sebaihia, M., Baker, S., Basham, D., Brooks, K., Chillingworth, T., Connerton, P., Cronin, A., Davis, P., Davies, R.M., Dowd, L., White, N., Farrar, J., Feltwell, T., Hamlin, N., Haque, A., Hien, T.T., Holroyd, S., Jagels, K., Krogh, A., Larsen, T.S., Leather, S., Moule, S., O'Gaora, P., Parry, C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S. and Barrell, B.G. (2001) Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18, *Nature*, **413**, 848-852.

Pasloske, B.L., Joffe, A.M., Sun, Q., Volpel, K., Paranchych, W., Eftekhar, F. and Speert, D.P. (1988) Serial isolates of *Pseudomonas aeruginosa* from a cystic fibrosis patient have identical pilin sequences, *Infect Immun*, **56**, 665-672.

Perna, N.T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E.J., Davis, N.W., Lim, A., Dimalanta, E.T., Potamousis, K.D., Apodaca, J., Anantharaman, T.S., Lin, J., Yen, G., Schwartz, D.C., Welch, R.A. and Blattner, F.R. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7, *Nature*, **409**, 529-533.

Pirnay, J.-P., De Vos, D., Cochez, C., Bilocq, F., Vanderkelen, A., Zizi, M., Ghysels, B. and Cornelis, P. (2002) *Pseudomonas aeruginosa* displays an epidemic population structure, *Environ Microbiol*, **4**, 898-911.

Platt, M.D., Schurr, M.J., Sauer, K., Vazquez, G., Kukavica-Ibrulj, I., Potvin, E., Levesque, R.C., Fedynak, A., Brinkman, F.S., Schurr, J., Hwang, S.H., Lau, G.W., Limbach, P.A., Rowe, J.J., Lieberman, M.A., Barraud, N., Webb, J., Kjelleberg, S., Hunt, D.F. and Hassett, D.J. (2008) Proteomic, microarray, and signature-tagged mutagenesis analyses of anaerobic *Pseudomonas aeruginosa* at pH 6.5, likely representing chronic, late-stage cystic fibrosis airway conditions, *J Bacteriol*, **190**, 2739-2758.

Poyart-Salmeron, C., Trieu-Cuot, P., Carlier, C. and Courvalin, P. (1989) Molecular characterization of two proteins involved in the excision of the conjugative transposon Tn1545: homologies with other site-specific recombinases, *Embo J*, **8**, 2425-2433.

Poyart-Salmeron, C., Trieu-Cuot, P., Carlier, C. and Courvalin, P. (1990) The integration-excision system of the conjugative transposon Tn 1545 is structurally and functionally related to those of lambdoid phages, *Mol Microbiol*, **4**, 1513-1521.

Qi, J., Luo, H. and Hao, B. (2004) CVTree: a phylogenetic tree reconstruction tool based on whole genomes., *Nucleic Acids Research*, **32**, 45-47.

Qiu, X., Gurkar, A.U. and Lory, S. (2006) Interstrain transfer of the large pathogenicity island (PAPI-1) of *Pseudomonas aeruginosa*, *Proc Natl Acad Sci U S A*, **103**, 19830-19835.

Ragan, M.A. (2001) Detection of lateral gene transfer among microbial genomes, *Curr Opin Genet Dev*, **11**, 620-626.

Rajan, I., Aravamuthan, S. and Mande, S.S. (2007) Identification of compositionally distinct regions in genomes using the centroid method, *Bioinformatics*, **23**, 2672-2677.

Raymond, C.K., Sims, E.H., Kas, A., Spencer, D.H., Kutyavin, T.V., Ivey, R.G., Zhou, Y., Kaul, R., Clendenning, J.B. and Olson, M.V. (2002) Genetic variation at the O-antigen biosynthetic locus in *Pseudomonas aeruginosa*, *J Bacteriol*, **184**, 3614-3622.

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

Rella, M., Mercenier, A. and Haas, D. (1985) Transposon insertion mutagenesis of *Pseudomonas aeruginosa* with a Tn5 derivative: application to physical mapping of the arc gene cluster, *Gene*, **33**, 293-303.

Reznikof, S.W. (2002) Tn5 Transposition. In Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. (eds), *Mobile DNA II*. ASM Press, Washington, D.C., 403-422.

Rice, L.B. and Carias, L.L. (1994) Studies on excision of conjugative transposons in enterococci: evidence for joint sequences composed of strands with unequal numbers of nucleotides, *Plasmid*, **31**, 312-316.

Rocchetta, H.L., Pacan, J.C. and Lam, J.S. (1998) Synthesis of the A-band polysaccharide sugar D-rhamnose requires Rmd and WbpW: identification of multiple AlgA homologues, WbpW and ORF488, in Pseudomonas aeruginosa, *Mol Microbiol*, **29**, 1419-1434.

Rowe-Magnus, D.A., Guerout, A.M. and Mazel, D. (1999) Super-integrons, *Res.Microbiol.*, **150**, 641-651.

Rowe-Magnus, D.A., Guerout, A.M. and Mazel, D. (2002) Bacterial resistance evolution by recruitment of super-integron gene cassettes, *Molecular microbiology*, **43**, 1657-1669.

Rowe-Magnus, D.A., Guerout, A.M., Ploncard, P., Dychinco, B., Davies, J. and Mazel, D. (2001) The evolutionary history of chromosomal super-integrons provides an ancestry for multiresistant integrons, *Proc.Natl.Acad.Sci.U.S.A.*, **98**, 652-657.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. and Barrell, B. (2000) Artemis: sequence visualization and annotation, *Bioinformatics*, **16**, 944-945.

Salunkhe, P., Smart, C.H., Morgan, J.A., Panagea, S., Walshaw, M.J., Hart, C.A., Geffers, R., Tummler, B. and Winstanley, C. (2005) A cystic fibrosis epidemic strain of *Pseudomonas aeruginosa* displays enhanced virulence and antimicrobial resistance, *J Bacteriol*, **187**, 4908-4920.

Sandberg, R., Winberg, G., Branden, C.I., Kaske, A., Ernberg, I. and Coster, J. (2001) Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier, *Genome Res*, **11**, 1404-1409.

Sauer, K., Cullen, M.C., Rickard, A.H., Zeef, L.A., Davies, D.G. and Gilbert, P. (2004) Characterization of nutrient-induced dispersion in *Pseudomonas aeruginosa* PAO1 biofilm, *J Bacteriol*, **186**, 7312-7326.

Scott, F.W. and Pitt, T.L. (2004) Identification and characterization of transmissible Pseudomonas aeruginosa strains in cystic fibrosis patients in England and Wales, *J Med Microbiol*, **53**, 609-615.

Scott, J.R. and Churchward, G.G. (1995) Conjugative transposition, *Annu Rev Microbiol*, **49**, 367-397.

Scott, J.R., Kirchman, P.A. and Caparon, M.G. (1988) An intermediate in transposition of the conjugative transposon Tn916, *Proc Natl Acad Sci U S A*, **85**, 4809-4813.

Sebaihia, M., Wren, B.W., Mullany, P., Fairweather, N.F., Minton, N., Stabler, R., Thomson, N.R., Roberts, A.P., Cerdeno-Tarraga, A.M., Wang, H., Holden, M.T., Wright, A., Churcher, C., Quail, M.A., Baker, S., Bason, N., Brooks, K., Chillingworth, T., Cronin, A., Davis, P., Dowd, L., Fraser, A., Feltwell, T., Hance, Z., Holroyd, S., Jagels, K., Moule, S., Mungall, K., Price, C., Rabbinowitsch, E., Sharp, S., Simmonds, M., Stevens, K., Unwin, L., Whithead, S., Dupuy, B., Dougan, G., Barrell, B. and Parkhill, J. (2006) The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome, *Nat Genet*, **38**, 779-786.

Shapiro, J.A. (1969) Mutations caused by the insertion of genetic material into the galactose operon of *Escherichia coli*, *J Mol Biol*, **40**, 93-105.

Shapiro, J.A. and Adhya, S.L. (1969) The galactose operon of *E. coli* K-12. II. A deletion analysis of operon structure and polarity, *Genetics*, **62**, 249-264.

Shen, K., Sayeed, S., Antalis, P., Gladitz, J., Ahmed, A., Dice, B., Janto, B., Dopico, R., Keefe, R., Hayes, J., Johnson, S., Yu, S., Ehrlich, N., Jocz, J., Kropp, L., Wong, R., Wadowsky, R.M., Slifkin, M., Preston, R.A., Erdos, G., Post, J.C., Ehrlich, G.D. and Hu, F.Z. (2006) Extensive genomic plasticity in *Pseudomonas aeruginosa* revealed by identification and distribution studies of novel genes among clinical isolates, *Infect Immun*, **74**, 5272-5283.

Siguier, P., Filee, J. and Chandler, M. (2006) Insertion sequences in prokaryotic genomes, *Curr Opin Microbiol*, **9**, 526-531.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. and Chandler, M. (2006) ISfinder: the reference centre for bacterial insertion sequences., *Nucleic acids research.*, **34**, 32-36.

Smart, C.H., Walshaw, M.J., Hart, C.A. and Winstanley, C. (2006) Use of suppression subtractive hybridization to examine the accessory genome of the Liverpool cystic fibrosis epidemic strain of *Pseudomonas aeruginosa*, *J Med Microbiol*, **55**, 677-688.

Smart, C.H.M., Scott, F.W., Wright, E.A., Walshaw, M.J., Hart, C.A., Pitt, T.L. and Winstanley, C. (2006) Development of a diagnostic test for the Midlands 1 cystic fibrosis epidemic strain of *Pseudomonas aeruginosa*, *J Med Microbiol*, **55**, 1085-1091.

Smith, E.E., Sims, E.H., Spencer, D.H., Kaul, R. and Olson, M.V. (2005) Evidence for diversifying selection at the pyoverdine locus of *Pseudomonas aeruginosa*, *J Bacteriol*, **187**, 2138-2147.

Sorek, R., Kunin, V. and Hugenholtz, P. (2008) CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea, *Nat Rev Microbiol*, **6**, 181-186.

Spencer, D.H., Kas, A., Smith, E.E., Raymond, C.K., Sims, E.H., Hastings, M., Burns, J.L., Kaul, R. and Olson, M.V. (2003) Whole-genome sequence variation among multiple isolates of *Pseudomonas aeruginosa*, *J Bacteriol*, **185**, 1316-1325.

Stokes, H.W., Elbourne, L.D. and Hall, R.M. (2007) Tn1403, a multiple-antibiotic resistance transposon made up of three distinct transposons, *Antimicrob Agents Chemother*, **51**, 1827-1829.

Stokes, H.W. and Hall, R.M. (1989) A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons, *Molecular microbiology*, **3**, 1669-1683.

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K., Wu, Z., Paulsen, I.T., Reizer, J., Saier, M.H., Hancock, R.E., Lory, S. and Olson, M.V. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen, *Nature*, **406**, 959-964.

Suttle, C.A. (2005) Viruses in the sea, *Nature*, **437**, 356-361.

Tang, T.H., Bachellerie, J.P., Rozhdestvensky, T., Bortolin, M.L., Huber, H., Drungowski, M., Elge, T., Brosius, J. and Huttenhofer, A. (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus, *Proc Natl Acad Sci U S A*, **99**, 7536-7541.

Tang, T.H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., Bachellerie, J.P. and Huttenhofer, A. (2005) Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus, *Mol Microbiol*, **55**, 469-481.

Toussaint, A. and Merlin, C. (2002) Mobile elements as a combination of functional modules, *Plasmid*, **47**, 26-35.

Tsirigos, A. and Rigoutsos, I. (2005) A new computational method for the detection of horizontal gene transfer events, *Nucleic Acids Res*, **33**, 922-933.

Tu, Q. and Ding, D. (2003) Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis., *FEMS Microbiology Letters*, **221**, 269-275.

Vaisvila, R., Morgan, R.D., Posfai, J. and Raleigh, E.A. (2001) Discovery and distribution of super-integrons among pseudomonads, *Molecular microbiology*, **42**, 587-601.

Vernikos, G.S. and Parkhill, J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands, *Bioinformatics*, **22**, 2196-2203.

Vernikos, G.S. and Parkhill, J. (2008) Resolving the structural features of genomic islands: a machine learning approach, *Genome Res*, **18**, 331-342.

Vernikos, G.S., Thomson, N.R. and Parkhill, J. (2007) Genetic flux over time in the *Salmonella* lineage, *Genome Biol*, **8**, R100.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W.F., Surovcik, K., Meinicke, P. and Merkl, R. (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models, *BMC Bioinformatics*, **7**, 142.

Wang, P.W., Chu, L. and Guttman, D.S. (2004) Complete sequence and evolutionary genomic analysis of the *Pseudomonas aeruginosa* transposable bacteriophage D3112, *J Bacteriol*, **186**, 400-410.

Webb, J.S., Lau, M. and Kjelleberg, S. (2004) Bacteriophage and phenotypic variation in *Pseudomonas aeruginosa* biofilm development, *J Bacteriol*, **186**, 8066-8073.

Webb, J.S., Thompson, L.S., James, S., Charlton, T., Tolker-Nielsen, T., Koch, B., Givskov, M. and Kjelleberg, S. (2003) Cell death in *Pseudomonas aeruginosa* biofilm development, *J Bacteriol*, **185**, 4585-4592.

Wiehlmann, L., Wagner, G., Cramer, N., Siebert, B., Gudowius, P., Morales, G., Kohler, T., van Delden, C., Weinel, C., Slickers, P. and Tummler, B. (2007) Population structure of *Pseudomonas aeruginosa*, *Proc Natl Acad Sci U S A*, **104**, 8101-8106.

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

Winsor, G.L., Van Rossum, T., Lo, R., Khaira, B., Whiteside, M.D., Hancock, R.E. and Brinkman, F.S. (2009) Pseudomonas Genome Database: facilitating user-friendly, comprehensive comparisons of microbial genomes, *Nucleic Acids Res*, **37**, D483-488.

Winstanley, C., Langille, M.G., Fothergill, J.L., Kukavica-Ibrulj, I., Paradis-Bleau, C., Sanschagrin, F., Thomson, N.R., Winsor, G.L., Quail, M.A., Lennard, N., Bignell, A., Clarke, L., Seeger, K., Saunders, D., Harris, D., Parkhill, J., Hancock, R.E., Brinkman, F.S. and Levesque, R.C. (2008) Newly introduced genomic prophage islands are critical determinants of *in vivo* competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*, *Genome Res*.

Wolfgang, M.C., Kulasekara, B.R., Liang, X., Boyd, D., Wu, K., Yang, Q., Miyada, C.G. and Lory, S. (2003) Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa*, *Proc Natl Acad Sci U S A*, **100**, 8484-8489.

Wurdemann, D. and Tummler, B. (2007) In silico comparison of pKLC102-like genomic islands of *Pseudomonas aeruginosa*, *FEMS Microbiol Lett*, **275**, 244-249.

Yang, J., Chen, L., Sun, L., Yu, J. and Jin, Q. (2008) VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics, *Nucleic Acids Res*, **36**, D539-542.

Yoon, S., Park, Y.-K., Lee, S., Choi, D., Oh, T., Hur, C.-G. and Kim, A. (2006) Towards pathogenomics: a web-based resource for pathogenicity islands., *Nucleic Acids Res*.

Zambrano, M.M., Siegele, D.A., Almiron, M., Tormo, A. and Kolter, R. (1993) Microbial competition: *Escherichia coli* mutants that take over stationary phase cultures, *Science*, **259**, 1757-1760.

Zink, R.T., Kemble, R.J. and Chatterjee, A.K. (1984) Transposon Tn5 mutagenesis in *Erwinia carotovora* subsp. carotovora and *E. carotovora* subsp. atroseptica, *J Bacteriol*, **157**, 809-814.