

# **INTEGRONS IN PSEUDOMONADS ARE ASSOCIATED WITH HOTSPOTS OF GENOMIC DIVERSITY.**

Neil Wilson BSc. (Hons)

Department of Biological Sciences,  
Macquarie University, Australia.



Thesis submitted: August 2007

A thesis submitted for fulfilment of the requirements of the degree  
of Doctor of Philosophy

# TABLE OF CONTENTS

<b>SYNOPSIS.....</b>	<b>VII</b>
<b>STATEMENT OF CANDIDATE .....</b>	<b>IX</b>
<b>ACKNOWLEDGMENTS .....</b>	<b>X</b>
<b>ABBREVIATIONS .....</b>	<b>XIII</b>
<b>CHAPTER 1 – LITERATURE REVIEW .....</b>	<b>- 1 -</b>
<b>1.1 – Bacterial Diversity .....</b>	<b>- 1 -</b>
<b>1.2 - The nature of genetic organisation .....</b>	<b>- 2 -</b>
<b>1.3 – Origins of Genetic Novelty.....</b>	<b>- 3 -</b>
<b>1.4 – Genetic modularisation .....</b>	<b>- 6 -</b>
<b>1.5 – Operons .....</b>	<b>- 7 -</b>
1.5.1 – Operon Evolution .....	- 8 -
<b>1.6 – Integrons.....</b>	<b>- 11 -</b>
1.6.1 – Background .....	- 11 -
1.6.2 – Core Integron Distribution and Diversity .....	- 16 -
1.6.3 – Integrons and Genomic Context .....	- 27 -
<b>1.7 – The mobile integron paradigm .....</b>	<b>- 31 -</b>
<b>1.8 – The generalised chromosomal integron .....</b>	<b>- 34 -</b>
1.8.1 – Limitations of the generalised CI concept.....	- 38 -
<b>1.9 - The need for model organisms to characterise CIs .....</b>	<b>- 40 -</b>
1.9.1 – <i>Pseudomonas</i> spp. as a model organism .....	- 40 -
<b>1.10 – The genus <i>Pseudomonas</i> .....</b>	<b>- 42 -</b>
1.10.1 – A brief history of genus <i>Pseudomonas</i> .....	- 42 -
1.10.2 – The <i>Ps. stutzeri</i> species complex.....	- 42 -
1.10.3 – Relationships between <i>Ps. stutzeri</i> genomovars .....	- 44 -
1.10.4 – <i>Ps. stutzeri</i> distribution and ecology .....	- 45 -

1.10.5 - By what mechanisms do genomovars diverge?.....	46 -
1.10.6 - A role for integrons in the diversification of <i>Ps. stutzeri</i> genomovars.....	46 -
<b>1.11 – Aims of the present study .....</b>	<b>48 -</b>
<b>CHAPTER 2 – GENERAL MATERIALS AND METHODS.....</b>	<b>50 -</b>
<b>2.1 – Bacterial Strains .....</b>	<b>50 -</b>
<b>2.2 – DNA Extraction.....</b>	<b>52 -</b>
<b>2.3 – PCR.....</b>	<b>52 -</b>
<b>2.4 – Agarose Gel Electrophoresis.....</b>	<b>53 -</b>
<b>2.5 – Southern Hybridisation .....</b>	<b>53 -</b>
<b>2.6 – Preparation of PCR Products and plasmid DNA for Sequencing .....</b>	<b>54 -</b>
<b>2.7 – Sequence Analysis .....</b>	<b>55 -</b>
<b>CHAPTER 3 – CHARACTERISATION OF STRAIN COLLECTION.....</b>	<b>57 -</b>
<b>3.1 – Introduction .....</b>	<b>57 -</b>
<b>3.2 – Materials and methods .....</b>	<b>64 -</b>
3.2.1 – BOX-PCR.....	64 -
3.2.2 – Analysis of BOX-PCR profiles .....	64 -
3.2.3 – 16S rDNA and 16S-23S IGS PCR.....	65 -
3.2.4 – RFLP analysis of 16S rDNA and IGS1 sequences.....	65 -
3.2.5 – Sequence Analysis.....	66 -
<b>3.3 – Results .....</b>	<b>69 -</b>
3.3.1 – BOX-PCR .....	69 -
3.3.2 – Analysis of 16S rRNA gene RFLP profiles.....	72 -
3.3.3 – Analysis of IGS1 RFLP profiles .....	74 -
3.3.4 – Confirmation of strain independence and provenance .....	76 -
3.3.5 – 16S rRNA gene phylogeny .....	78 -
3.3.6 – 16S-23S IGS phylogeny.....	82 -
<b>3.4 – Conclusions .....</b>	<b>86 -</b>
<b>CHAPTER 4 – DISTRIBUTION OF INTEGRONS AND GENE CASSETTES IN <i>PSEUDOMONAS</i> .....</b>	<b>89 -</b>

<b>4.1 - Introduction.....</b>	<b>- 89 -</b>
<b>4.1 - Materials and methods .....</b>	<b>- 95 -</b>
4.2.1 - PCR screening for integrons.....	- 95 -
4.2.2 – Probe design for integron detection .....	- 98 -
4.2.3 – Southern hybridization .....	- 100 -
4.2.4 – Sequence Analysis.....	- 101 -
4.2.5 – Public Database Searches .....	- 101 -
<b>4.3 – Results .....</b>	<b>- 103 -</b>
4.3.1 – PCR detection of integrons .....	- 103 -
4.3.2 – Southern hybridisation integron detection – Core Integron .....	- 106 -
4.3.3 – Southern hybridisation integron detection – 59-be .....	- 110 -
4.3.4 – Analysis of Recovered integrons .....	- 114 -
4.3.4.1 – Analysis of recovered <i>intl</i> genes.....	- 114 -
4.3.4.2 – Analysis of recovered <i>attI</i> regions .....	- 118 -
4.3.4.3 – Putative identification of $P_c$ .....	- 120 -
4.3.4.4 - Putative identification of <i>intl</i> regulatory sequences .....	- 122 -
4.3.4.5 – Analysis of Gene Cassettes.....	- 124 -
4.3.5 – <i>in silico</i> analysis of <i>Pseudomonas</i> family integron distribution .....	- 124 -
<b>4.4 – Conclusions .....</b>	<b>- 126 -</b>
<b>CHAPTER 5 – GENOMIC CONTEXT OF <i>PSEUDOMONAS</i> INTEGRONS .....</b>	<b>- 131 -</b>
<b>5.1 - Introduction.....</b>	<b>- 131 -</b>
<b>5.2 - Materials and methods .....</b>	<b>- 135 -</b>
5.2.1 – Summary of cloning, detection and sequencing strategy.....	- 135 -
5.2.2 - Analysis of gene cassettes .....	- 136 -
5.2.3 - Sequence Analysis.....	- 136 -
5.2.4 - Phylogenetic Analysis .....	- 138 -
<b>5.3 – Results .....</b>	<b>- 139 -</b>
5.3.1 – Recovery of addition integron sequences.....	- 139 -
5.3.1.1 – Extension of existing integron sequences .....	- 139 -
5.3.1.2 – Characterisation of <i>Ps. stutzeri</i> 19SMN4 (4A) and DNSP21 (5A) integrons.....	- 140 -
5.3.1.3 – Characterisation of <i>Ps. stutzeri</i> Gv.2 integrons .....	- 144 -
5.3.1.4 - Characterisation of additional integrons / cassette arrays .....	- 144 -
5.3.1.5 – Analysis of recovered gene cassettes.....	- 148 -
5.3.2 - Recovery and analysis of integron flanking genes .....	- 152 -
5.3.2.1 – Characterisation of integrons at locus 1 .....	- 158 -

5.3.2.2 – Characterisation of integrons at locus 2 .....	- 159 -
5.3.3 - Phylogenetic analysis of integron flanking genes .....	- 164 -
5.3.4 - Phylogenetic analysis of <i>intl</i> relative to genomic context .....	- 166 -
5.3.5 - Synteny and ancestry of integron loci across <i>Pseudomonas</i> spp. ....	- 172 -
5.3.6 - Analysis of integron boundaries .....	- 178 -
5.3.7 - Detection of a CI in the genome sequence of <i>Ps. mendocina</i> YMP .....	- 183 -
<b>5.4 - Conclusions .....</b>	<b>- 192 -</b>
<b>CHAPTER 6 - EVOLUTIONARY ANALYSIS OF <i>PSEUDOMONAS</i> SPP. INTEGRONS-</b>	<b>199 -</b>
<b>6.2 - Materials and methods .....</b>	<b>- 202 -</b>
6.2.1 – Genomic Analyses and Codon Usage Analyses .....	- 202 -
6.2.2 – Statistical analysis of codon usage data .....	- 202 -
<b>6.3 – Results .....</b>	<b>- 203 -</b>
6.3.1 – G + C content analysis .....	- 203 -
6.3.2 – Codon Usage Analysis .....	- 206 -
6.3.2.1 – Analysis of relative codon usage data using PCA and CA .....	- 208 -
6.3.2.2 – Assessment of <i>Ps. aeruginosa</i> as a surrogate for codon usage in <i>Ps. stutzeri</i> .....	- 209 -
6.3.2.3 – Codon usage of <i>Pseudomonas</i> spp. CIs relative to chromosomal genes .....	- 212 -
<b>6.4 – Conclusions .....</b>	<b>- 215 -</b>
<b>CHAPTER 7 – FINAL DISCUSSION .....</b>	<b>- 218 -</b>
<b>7.1 – Integron acquisition and loss in <i>Pseudomonas</i> spp.....</b>	<b>- 219 -</b>
<b>7.2 – Integrons are associated with hotspots for recombination .....</b>	<b>- 222 -</b>
<b>7.3 – Implications for the concept of the generalised CI.....</b>	<b>- 224 -</b>
<b>7.4 – Future Work .....</b>	<b>- 231 -</b>
<b>APPENDIX 1 – FOSMID LIBRARY CONSTRUCTION AND SCREENING .....</b>	<b>- 233 -</b>
<b>A.1 - Introduction .....</b>	<b>- 233 -</b>
<b>A.2 - Materials and methods.....</b>	<b>- 233 -</b>
A.2.1 - Fosmid library construction.....	- 233 -
A.2.2 - Estimation of library coverage.....	- 233 -
A.2.3 - Colony hybridisation.....	- 234 -
A.2.4 - Fosmid purification.....	- 235 -
A.2.5 - Sequencing of fosmid clones .....	- 236 -

A.2.6 - PCR recovery of additional integron boundary sequences .....	- 238 -
A.2.7 - Hybridisation screening for integron flanking genes.....	- 239 -
<b>A.3 - Results.....</b>	<b>- 241 -</b>
A.3.1 - Construction of fosmid libraries and estimation of genomic coverage.....	- 241 -
A.3.2 - Screening fosmid clone libraries for integron sequences.....	- 244 -
A.3.3 - Recovery of complete integrons and flanking sequences .....	- 248 -
A.3.4 - PCR recovery of additional integron boundary regions .....	- 249 -
A.3.4.1 - Recovery of 5' boundary regions .....	- 249 -
A.3.4.2 - Recovery of 3' boundary from <i>Ps. stutzeri</i> BAM17 .....	- 250 -
A.3.5 - Screening strain collection for integron flanking genes .....	- 251 -
A.3.6 – Screening fosmid libraries for integron flanking genes .....	- 254 -
<b>A.4 - Conclusions .....</b>	<b>- 255 -</b>
<b>REFERENCES.....</b>	<b>- 257 -</b>

## SYNOPSIS

Integrans associated with mobile genetic elements have played a central role in the emergence and spread of multiple antibiotic resistance in many pathogenic bacteria. However, the discovery of integrans in the chromosomes of diverse, non-pathogenic bacteria suggests that integrans have a broader role in bacterial evolution. The *Pseudomonas stutzeri* species complex is a well studied model for bacterial diversity. Members of the complex are genetically closely related, but sub-taxa are not able to be defined by exclusively shared sets of phenotypic characters. Rather, on the basis of total DNA:DNA similarity, *Ps. stutzeri* strains have been divided into 17 different groups (termed genomovars). Two *Ps. stutzeri* strains have been found to contain Chromosomal Integrans (CIs). This thesis involved exploration of the hypothesis that a CI was present in the common ancestor of the *Ps. stutzeri* species complex and assessed the impact of integrans on diversity across all Pseudomonads. The history and significance of integrans is discussed in Chapter 1 as part of a literature review, and general materials and methods are provided in Chapter 2. Chapters 3 – 6 comprise the sections in which data generated during my PhD project are presented. A comprehensive analysis of the relationships between the strains being analysed is presented in Chapter 3. In Chapter 4, results of PCR and hybridisation screening for integrans across the strain collection are presented. In Chapter 5 the recovery of additional integrans and in depth sequence analysis of the recovered integrans are described. Finally, Chapter 6 contains statistical analyses of integron-associated genes and Chapter 7 contains a final discussion the most significant findings. Twenty-three *Pseudomonas* spp. strains were screened for the presence of integrans. All but three were found to contain integron-like sequences; however, most integron sequences recovered contained inactivated core integrans.

Despite having a chromosomal locus, integrons in *Pseudomonas* were found to have properties indicative of frequent horizontal transfer. Evidence was also obtained which suggests that integrons have been acquired at the same locus on multiple independent occasions. This has not been observed in other families of chromosomal integrons and suggests that the loci at which integrons in *Pseudomonas* are found are hotspots for recombination.



## **STATEMENT OF CANDIDATE**

This work is original and has not been submitted for a higher degree to any other university or institution. The work of others, when drawn upon, is referenced fully.

Approximately 50% of the DNA sequencing performed in Chapter 5 was performed by another researcher. All analyses of DNA sequence data presented in this thesis were performed by the author.

Signed

Neil Wilson

Date

## **ACKNOWLEDGMENTS**

To my beautiful girlfriend Linh, thanks most of all just for being there for me (and generally putting up with me!). Thanks for your patience while I was writing-up. Thanks for listening and being interested in my nerdy stories and for getting excited when I get a good result, even if it is just a band on a gel! Thanks for sharing the good times and for listening to the whinges in the not so good times. It is great to have a partner that shares my passions. It's been great to be able share my life and my passion for science with you.

To my parents, Frank and Lesley, before anyone else I would not have been able to accomplish this without you both. I can't thank you enough for all the support you have given me. From making sure I received a good education to the unwavering moral and financial support you provided. You have both been great parents and I am eternally grateful. To my Mum Lesley in particular, thank you for putting a roof over my head throughout my PhD and for providing such a great living environment for me. You have always believed so strongly in me; I may not have had the confidence to do this without that. Also, the endless stream of great coffee that you made sure flowed was absolutely crucial to the thesis writing effort. Thanks also for helping in the printing and in proof-reading my thesis for typos; that can't have been much fun and there were plenty of them!

To Kathy and Jane, thanks for being great big sisters. I have always looked up to you both. Thanks for always being more or less nice to a younger brother who was only very occasionally annoying! To Jane in particular, thanks for being interested in my science

stories and theories, and for leaving me alone (most of the time!) when I needed to study for weeks on end. Thanks most for helping to fight for my right to be a cool science nerd!

Thank you to my friends for sticking by me, for understanding why I need to work on weekends at times, for tolerating my warped sense of punctuality, and for finding science-nerdiness amusing! I really am lucky to have such a great bunch of mates.

To my supervisors, Michael Gillings and especially Andrew Holmes, I thank you for the endless guidance you given me throughout my PhD. I have learnt a huge amount from you both and it has been an honour to have each of you as a mentor. To Andy, thank you for always having your office door open to me. Your willingness to help and advice went above and beyond the call of duty. To Nick Coleman, you were a great mentor in the lab throughout my PhD. If you had not been in the lab, I would have had less fun, but more importantly I would have learned far less, generated far less data and not become as accomplished a scientist. Thank you.

Thank you to both Sasha and Claire being my fellow PhD lab buddies. The moral support, good fun and philosophical conversations were invaluable. We will forevermore share a weird bond, a mutual understanding of the pain we have endured! Seriously though, the value of having other people in the same boat should not be undervalued, especially when they are fun to work with too! It wouldn't have been the same without you guys. Good luck finishing up your respective theses, and whatever you do, don't take as long as I did!

Thank you to all the people in the Holmes lab at Sydney University and the EMMA lab at Macquarie University who have assisted me and/or shared a lab-space with me throughout my PhD. It is a pleasure to work each day with so many people who are

passionate about what they do. Thank you to Johannes Sikorski for supplying most of the *Pseudomonas* strains analysed. Thank you to Ruth Hall for helpful guidance throughout both the research and writing-up component of my PhD.

## **ABBREVIATIONS**

59-be – 59-base element

bp – Base pairs

CA – correspondence analysis

CI – chromosomal integron

CTAB – Hexadecyltrimethylammonium bromide

DIG-6-dUTP – digoxigenin-6-dUTP

DNA – Deoxyribonucleic acid

EDTA – Ethylenediaminetetraacetic acid

Gv. – Genomovar

HGT – horizontal gene transfer

IGS1 – 16S-23S rDNA intergenic spacer

IntI – integron integrase

LB broth – Luria-Bertani broth

MI – mobile integron

PCA – principal components analysis

PCR – polymerase chain reaction

RFLP – Restriction fragment length polymorphism

RSCU – relative synonymous codon usage

SSC – Standard Saline Citrate

SDS – Sodium dodecyl sulfate

TBE buffer – Tris borate EDTA buffer

TE buffer – Tris EDTA buffer

# CHAPTER 1 – LITERATURE REVIEW

## 1.1 – Bacterial Diversity

The two domains of prokaryotic microorganisms, Bacteria and Archaea, encompass the vast majority of diversity (in terms of both genotypic and physiological diversity) within the biosphere (Woese, 1987). Such high levels of physiological variation mean that prokaryotes are found in almost every conceivable habitat on earth, from the deepest seas and deep underground to the upper atmosphere, and are able to thrive in extremes of pH, salinity, temperature, nutrient concentrations and atmospheric pressure (Horikoshi and Grant, 1998; Madigan and Marrs, 1997; Rothschild and Mancinelli, 2001). The extreme physiological diversity of bacteria is also reflected at the genomic level. The minimal core set of genes required to sustain a bacterial cell has been estimated to consist of <250 genes (Gil *et al.*, 2004) which comprises only 5-10 % of the typical bacterial genome. The remainder of the genome is comprised of genes which provide the specific traits necessary for persistence of bacterial lineage in a particular ecological niche, and constitutes a massive genetic potential for niche specialisation.

The extraordinary diversity and versatility of bacteria raises intriguing questions in regard to the generation of genetic diversity in these organisms, especially in light of the fact that bacteria are unicellular, asexual and exhibit relatively low levels of recombination. One obvious explanation for this is the long evolutionary history of prokaryotes relative to eukaryotes. However, the ability of bacteria to rapidly adapt to environmental change (eg. spread of multiple antibiotic resistance and degradation of xenobiotic compounds) suggests that bacterial versatility cannot be explained by a

long evolutionary history alone and that specific mechanisms must exist which facilitate the acquisition and exploitation of genetic novelty for adaptive advantage (Poole *et al.*, 2003). The mechanisms by which bacteria acquire and exploit genetic novelty are therefore of particular interest and are critical to understanding how bacteria evolve and adapt to environmental change.

## **1.2 - The nature of genetic organisation**

Genetic diversity may be considered to occur at four different levels: primary diversity – variation at the level of the DNA sequence, secondary diversity – variation in sequence entities (eg. genes), tertiary diversity – diversity of modules of sequence entities (eg. operons), and quaternary diversity – variation in regulation and expression of sequence entities (eg. regulons). Novelty at one level is not necessarily underpinned by novelty at other levels. The reason for this is that different novelty-generating mechanisms may impact one or two levels of diversity, but rarely impact all levels. The generation of diversity at all levels of organisation is critical in allowing populations to respond to environmental change, as diversity at each level impacts the evolution of populations in different ways. Generation of primary and quaternary diversity results in gradual stepwise changes at the DNA sequence level and are important in the fine-tuning of phenotypes in response to gradual environmental changes. In contrast, generation of secondary and, in particular, tertiary diversity can result in the instantaneous acquisition of novel and complex phenotypes, and is thus important in adaptive responses to rapid environmental changes.

### 1.3 – Origins of Genetic Novelty

Genetic novelty arises through changes to DNA sequences encoding RNA, protein or regulatory elements, and natural selection. The end result is sets of genes that are expressed in such a fashion as to enable the stable and sustainable participation of the cell within an ecological community (niche). The need for populations to adapt at a genomic level to environmental changes means that novel genes must be generated or acquired to accommodate and/or exploit these changes, while redundant genetic information is lost. How do bacterial populations acquire genetic novelty?

Diversity in bacteria is generated through a combination of mutation, rearrangement and gene acquisition. Mutation (in the form of point, insertion and deletion mutations) mainly generates stepwise changes in the primary sequence, or the expression patterns, of previously existing genes, and consequently results in the generation of diversity at the primary and quaternary levels of organisation. More rarely, mutation may generate new sequence entities (secondary diversity) through frameshift mutations which result in the disruption or fusion of pre-existing genes or generation of new coding sequences. Mutation does not directly affect tertiary level diversity. A variety of mechanisms is known which can affect mutation rates including error-prone polymerases and DNA repair mechanisms. Genomic rearrangements generate diversity by shuffling gene order, creating novel gene associations and altering gene regulation. Numerous specialised mechanisms are involved in genomic rearrangement, including homologous recombination and site-specific recombination. Several phenomena which are mediated by these mechanisms are known, such as phase variation (Hilton *et al.*, 2006; van der Woude, 2006), gene



inversion (Komano, 1999; Kutsukake *et al.*, 2006) and gene rearrangement. As with mutation, genomic changes arising by rearrangement occur in a stepwise or progressive fashion, and mainly impact diversity at the primary and quaternary levels. Rearrangement may generate diversity at the tertiary level of organisation through the formation of novel gene associations; however, such events are likely to be relatively rare. The gradual nature of genomic changes resulting from mutation and rearrangement means that these mechanisms are not effective as initiators of rapid adaptive responses, except in the context of small environmental changes for large populations.

With the advent of whole genome datasets, Horizontal Gene Transfer (HGT) has received great emphasis as a contributor to the generation of bacterial diversity. The relative contribution of HGT to bacterial evolution with respect to other diversity generating mechanisms is one of the most contentious issues in evolutionary microbiology. Estimates of the frequency of HGT vary massively; different methods have predicted that as few as 2% (Ge *et al.*, 2005) and as many as 90% (Mirkin *et al.*, 2003) of all prokaryotic genes have been affected by HGT. More recently, exhaustive analyses of complete microbial genomes under ultraconservative assumptions have indicated that at least 65% of all gene families have been affected by HGT (Dagan and Martin, 2007). This analysis arguably represents the most robust estimate to date of the impacts of HGT.

Unlike mutation and recombination, HGT results in sudden innovations or jumps in the secondary diversity of populations, and can lead to similar and sudden changes to tertiary and quaternary diversity. Entire gene pathways can be acquired in a single HGT event, and may result in the immediate acquisition of complex and novel

phenotypes. Such events provide an opportunity for evolution to proceed in leaps through rapid adaptation to environmental changes (Lawrence, 2001; Ochman *et al.*, 2000). While the frequency and relative impact of HGT on bacterial evolution remains a matter of contention, several unambiguous examples of HGT events which have resulted in the acquisition of novel multi-gene phenotypes exist and illustrate the potential for HGT to impact bacterial evolution. HGT generates diversity through the acquisition of novel genes which may result in new phenotypes. For example, pathogenicity in many *Shigella* spp. strains can be attributed to single acquisition events (Pupo *et al.*, 2000; Yang *et al.*, 2007), genes encoding xenobiotic degradation pathways are frequently found clustered on plasmids and exhibit properties indicative of acquisition by HGT (Chauhan *et al.*, 1998; Mosqueda and Ramos, 2000), and multiple antibiotic phenotypes have been widely disseminated by HGT through associations with diverse mobile genetic elements (Leverstein-van Hall *et al.*, 2002; Summers, 2006; Wright, 2007).

Most mutation and recombination events are likely to have no discernable impact on phenotype, while some will have a deleterious effect or provide a selective advantage. Similarly, most horizontally transferred genes are unlikely to result in changed phenotypes, as most fail to persist for long periods in the recipient genome and may even interact antagonistically with native genes (Bouma and Lenski, 1988; Schrag and Perrot, 1996; Wright, 1977). Given the apparent instability of acquired genes and the huge potential for HGT to facilitate rapid adaptation, it has been proposed that selection may favour the evolution of genomic features that minimise antagonistic interactions between newly acquired genes and native genes and, therefore, increase the probability of these genes persisting in the recipient genome (Rainey and Cooper, 2004). Modular genetic organisation is one such mechanism. Modularity may

increase the probability of all genes of a particular pathway being horizontally transferred in a single event (spatial modularity) and of transferred genes being successfully integrated into the regulatory machinery of the recipient (regulatory modularity) (Rainey and Cooper, 2004). Modularity may therefore be seen as a mechanism that increases evolvability in bacteria, by reducing potential fitness costs associated with expressing acquired genes and by increasing the likelihood of persistence of newly acquired genes.

## **1.4 – Genetic modularisation**

Modularity is a pervasive feature of genome organisation in bacteria. Genes can be spatially modularised (eg. plasmids, genomic islands, toxin/antitoxin modules and operons), or occur in regulatory networks such as regulons. Much research has been dedicated to characterising both of these forms. While the selective advantages of modular genetic organisation are obvious (eg. concerted regulation of genes of particular pathways), its evolutionary origins are a matter of much contention. It has been hypothesised that regulatory modularisation (eg. signal transduction pathways) occurs because of selection for genome architectures that minimise the pleiotropic effects associated with accommodation of newly acquired DNA sequences by lateral gene transfer (Rainey and Cooper, 2004). Functional modularity of mobile genetic elements increases versatility and means that addition or subtraction of a particular module is less likely to affect the viability of the element (Hendrix *et al.*, 2000; Osborn and Boltner, 2002; Toussaint and Merlin, 2002). Evolution of spatial modularisation of genes is hypothesised to occur through selection for coregulation of genes belonging to particular pathways, and/or to facilitate transmission of gene clusters by HGT (Lawrence, 1999; Price *et al.*, 2005b). Operons are the quintessential genetic

modules, and provide an example of both spatial and regulatory modularisation, with member genes being clustered behind common regulatory machinery.

## 1.5 – Operons

Operons are groups of genes that are transcribed in a single mRNA. Operons are a major form of genomic organisation in prokaryotes, occurring in all known bacterial and archaeal genomes (Price *et al.*, 2005a; Wolf *et al.*, 2001). More than 10% of the genes in a typical bacterial genome are predicted to occur in operons, which is likely an underestimate due to the difficulties associated with prediction of operons from genome sequences (Ermolaeva *et al.*, 2001). Operons often, but not always, encode functionally complementary genes (de Daruvar *et al.*, 2002; Rogozin *et al.*, 2002). Gene order of operons is frequently conserved across species by vertical transmission (Overbeek *et al.*, 1999; Wolf *et al.*, 2001); however, few are structurally conserved among distantly related bacteria (Itoh *et al.*, 1999; Mushegian and Koonin, 1996; Watanabe *et al.*, 1997; Wolf *et al.*, 2001). Some operons are ancient in origin (eg. the histidine operon) (Price *et al.*, 2006), undergoing stable vertical transfer with very little influence of horizontal transfer and high levels of gene order conservation, while others show extensive HGT across large phylogenetic distances, exhibiting a mosaic structure with low levels of gene order conservation (Davies *et al.*, 2002; Sentchilo *et al.*, 2000). These observations indicate that while individual operons appear unstable over long evolutionary periods, the rate of operon creation must be significantly higher than that of operon disruption, as operons remain a principle form of gene organisation in extant bacteria.

The recent explosion of bacterial genome sequences available in public databases has provided many insights into the nature of bacterial operons. The diversity of genes which may be found in operons appears to be limited only by total bacterial genetic diversity (Price *et al.*, 2005a). Further, operons have also been found to be highly variable with respect to a number of other characteristics. Operons are dynamic mosaic structures (Itoh *et al.*, 1999; Xie *et al.*, 2003). However, many operons exhibit high levels of global stability over long evolutionary timescales (Price *et al.*, 2006). Once operon structures have evolved, selection for their maintenance appears to be strong; however, selection may sometimes drive sub-division of an operon into two or more co-transcribed units, and may even lead to complete dissociation of the constituent genes (Price *et al.*, 2005b; Xie *et al.*, 2003).

### 1.5.1 – Operon Evolution

The selective pressures responsible for operon maintenance are generally well agreed upon, and include co-regulation of gene expression and facilitating horizontal transfer of complete gene pathways. In contrast, the mechanisms which drive operon formation are a matter of much contention. The probability of multiple genes simultaneously coming into close proximity and becoming co-regulated by random mechanisms alone is thought to be so low that operons are hypothesised to evolve by gradual processes, and thus models for operon evolution are generally based around this assumption.

Several models have been proposed to account for the evolution of operons. For almost forty years after their discovery, co-regulation from a single promoter was thought to be the main selective pressure driving operon formation (Jacob *et al.*,

1962). Recently, this idea has been challenged by a theory termed the Selfish Operon Model (SOM), where operons evolve gradually to assist their own dispersal by horizontal gene transfer (Lawrence, 1999; Lawrence and Roth, 1996). According to the SOM, capabilities which are only occasionally useful can easily be lost by random deletion of a single gene. The other genes in the pathway will then be quickly lost, as they now serve no function. The capability could then be reacquired by HGT, but only if the genes are in close proximity and can be transferred in a single event. If the genes are not already found within an operon, occasional random rearrangements which bring the genes closer together will increase the probability of all complementary genes being transferred together. After several rounds of rearrangement and transfer, the genes will become tightly clustered and an operon can form by deletion of any intervening DNA still remaining. At this point, both coordinated expression and ease of horizontal transfer serve to maintain operons.

At the time the SOM model was proposed, it was thought that with the exception of a few broadly conserved operons such those involved in ribosome synthesis, operons consisted of mostly non-essential genes (Lawrence, 2001). However, as expanded datasets have become available, it has been found that essential genes are preferentially found within operons, even when broadly conserved operons are excluded (Pal and Hurst, 2004). Furthermore, in an extensive survey of the *E. coli* K12 genome, Price *et al.* (2005b) found that horizontally acquired genes were no more likely to form operons (at the time of acquisition or after acquisition) than essential single-copy genes. These findings cast doubts on the role of the SOM in operon formation, as the loss and regain of genes by HGT is a central requirement of the SOM, and thus precludes the involvement of essential genes which are lethal if lost. However, the possibility that some new operons form selfishly is not excluded.

Their conclusion that HGT cannot account for the evolution of all operons led Price *et al.* (2005b) to reconsider the co-regulation model of operon formation. The selective force driving operon formation under this model is reduction in the amount of regulatory information required to control the expression of operons relative to isolated genes. Optimisation and/or adaptation of expression levels should occur more efficiently when the genes for a pathway are under the control of a single regulator, as any changes in regulatory information will impact all genes in the operon equally, and minimise any detrimental effects associated with variation in expression of genes within particular pathways. Thus, as the amount of regulatory information required increases, the more likely the genes should be found in an operon (Price *et al.*, 2005b).

One limitation of the co-regulation model in relation to the SOM is that genes with similar optimal expression patterns must be placed adjacent immediately for a selective advantage to be conferred, whereas the SOM allows for gradual clustering of genes. By what mechanisms can two genes become adjacent? As gradual deletion of intervening sequences is not plausible under the co-regulation model, mechanisms which can place two genes adjacent in a single step are required. In other words, mechanisms which generate diversity at the tertiary level of organisation are required to drive operon formation. Price *et al.* (2005b) discuss only fortuitous rearrangements or insertion of genes as mechanisms to bring genes into close proximity; however, they do discuss phenomena which may increase the frequency of such events. No mention is made of site-specific recombination mechanisms or integrons. Integrons in particular exhibit several characteristics which make them potentially well suited to a role in the formation of co-regulated gene clusters.

## 1.6 – Integrations

### 1.6.1 – Background

Integrations were discovered as a result of their role in the accumulation and spread of multiple antibiotic resistance among gram-negative pathogenic bacteria. In the relatively short time since the discovery of antibiotics, resistance to virtually all known antibiotics has emerged in pathogenic bacteria (Davies *et al.*, 2002). It is clear that the majority of antibiotic resistance determinants (ie. genes encoding proteins that specifically target antibiotics) have not evolved *de novo* on several occasions since the beginning of the antibiotic era; rather they have been acquired by bacterial lineages via HGT. The development of DNA sequencing technology led to the discovery that resistance genes in proteobacterial pathogens are often clustered and found at conserved genetic loci (Cameron *et al.*, 1986; Hall and Vockler, 1987; Martinez and De La Cruz, 1990; Ouellette *et al.*, 1987; Sundstrom *et al.*, 1991). This observation suggested that a specific mechanism was behind the accumulation of genes at these loci and led to the discovery of integrations (Stokes and Hall, 1989). A considerable amount of research has been dedicated to characterising integrations, and their properties as a site-specific recombination system are well established (Table 1.1). The minimal integration (in terms of the minimal sequence features required for site specific-recombination to occur) consists of a gene, *intI*, that encodes a site-specific recombinase belonging to the tyrosine recombinase family, an adjacent recombination site, *attI*, and a promoter,  $P_{int}$ , which drives expression of *intI* (Figure 1.1). Integrations may contain one or more gene cassettes integrated at *attI*. Gene cassettes are the smallest known genetic elements. A cassette usually consists of a single Open Reading Frame (ORF), an associated recombination site (termed 59-be



or *attC*) that is specifically recognised by IntI, and very little extraneous non-coding DNA (Figure 1.1). Multiple gene cassettes can be assembled into arrays, and rearranged within the integron, through the action of *intI* mediated recombination. Together, the core integron and gene cassettes constitute a system of two independent elements which is unique among other site-specific recombination systems. Most known mobile genetic elements capable of site-specific recombination are autonomous, one-component elements, in which all the components of the element are clustered and move as a single unit. Gene cassettes constitute the mobile component of integrons, while the core integron is the scaffold into which cassettes are integrated and excised. Thus, integrons are non-autonomous mobile genetic elements, as gene cassette mobility is dependent on the non-mobile activity of the core integron.

All integrons characterised to date share a common set of sequence characteristics and thus, integrons may be identified on the basis of sequence information alone. As a group, IntI integrases share all the diagnostic features of, but form a distinct group within, the tyrosine recombinases (Nunes-Duby *et al.*, 1998). IntI integrases also contain an additional motif (which is absent from other tyrosine recombinases) that appears to be a signature for IntI family members and is necessary for IntI activity (Messier and Roy, 2001). Different *attI* sites are defined on the basis of its position between *intI* and the first gene cassette. Apart from the residue adjacent to the *attI* recombination crossover, no diagnostic sequences are known that identify *attI*. Experimental data is therefore required to confirm functionality of different *attI* sites. Across all known examples, 59-be exhibit very low levels of sequence conservation. Different 59-be do, however, exhibit a conserved structure, consisting two-halves of an imperfect inverted repeat between which a variable number of nucleotides (3 – 79

bp) is inserted (Holmes *et al.*, 2003a). The conservation of structure indicates that 59-be sites are orthologous structures and aids the identification of 59-be from diverse sources. However, the lack of sequence conservation between different 59-be complicates the process of 59-be detection using molecular probes. The *intI* gene is the only integron component that can be unambiguously identified in partial and/or divergent integron sequences. Therefore, there are two levels at which to consider integron and gene cassette diversity: IntI diversity – presence of an IntI homologue indicating the presence of a potential integron, and integron diversity – presence of IntI, *attI* and 59-be indicating the presence of an actual integron.

Key Discovery	Reference
<b>Integrans first recognised a genetic element</b>	Stokes and Hall 1989
<b>Integrans encode a site-specific recombination system</b> Intl-mediated site-specific recombination Definition of <i>attI</i> Definition of 59be ( <i>attC</i> ) Definition of P <sub>c</sub> Integration and excision of gene cassettes	Collis <i>et al.</i> 1993; Recchia <i>et al.</i> 1994 Partridge <i>et al.</i> 2000 Hall <i>et al.</i> 1991 Levesque <i>et al.</i> 1994; Collis and Hall 1995 Collis and Hall 1992b
<b>Intl exhibits specificity to cognate <i>attI</i>, but recognises diverse 59be</b>	Collis <i>et al.</i> 2002; Holmes <i>et al.</i> 2003
<b>Cassette encoded ORFs are expressed</b>	Levesque <i>et al.</i> 1994; Collis <i>et al.</i> 1995
<b>Integrans are phylogenetically and environmentally widespread</b> Phylogenetically wide distribution Abundant in diverse environments Recombinationally functional	Rowe-Magnus <i>et al.</i> 2001; Holmes <i>et al.</i> 2003 Neild <i>et al.</i> 2001; Nemergut <i>et al.</i> 2004 Holmes <i>et al.</i> 2003; Leon and Roy 2003; Drouin <i>et al.</i> 2002
<b>Integrans may form long-term associations with bacterial lineages</b>	Rowe-Magnus <i>et al.</i> 2001; Gillings <i>et al.</i> 2005
<b>Gene cassette content is not restricted</b>	Stokes <i>et al.</i> 2001; Holmes <i>et al.</i> 2003; Michael <i>et al.</i> 2004

Table 1.1 - Significant discoveries relating to the function and diversity of the integron gene cassette system.

## Generic integron model

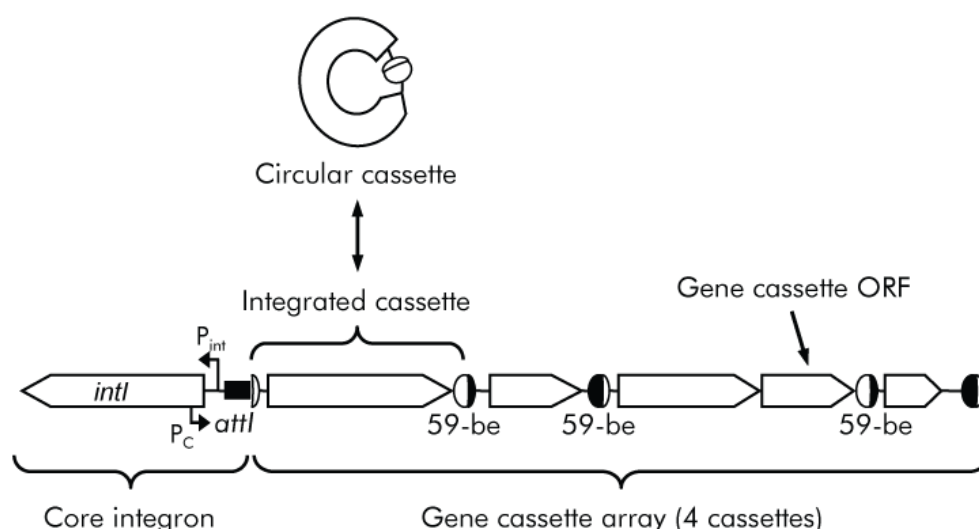


Figure 1.1 – The generic integron model has been formulated as more examples of diverse integrons emerged, and the fundamental components of the system became apparent. Integrons are defined as a two-component system, the core integron and cassette array. The core integron comprises the ‘minimal integron’, and consists of an integrase gene (*intl*), an associated recombination site (*attI*), and a promoter that drives expression of *intl* ( $P_{int}$ ); however, integrons are frequently identified on the basis of an *Intl* homologue alone. The core integron may also be associated with an additional promoter that drives expression gene cassettes ( $P_C$ ), but this is not essential for recombination activity. A sequence is considered to be an integron if an *Intl* homologue is present. An integron is considered to be complete and potentially functional if it has a complete *Intl* homologue, and an attached sequence consistent with an *attI* site. The core integron may be associated with one or more gene cassettes inserted at *attI*. Gene cassettes are excised as circular elements and consist of a recombination site (59-be) and usually a single ORF; however, cassettes may consist of two ORFs (cassette 3), one or more ORFs in reverse orientation, or no detectable ORF. When circular cassettes are integrated at *attI*, the recombination crossover occurs in the right hand region of the 59-be, resulting in the far right hand region of the 59-be being placed at the 5' boundary of the cassette, with the remainder of the 59-be at the 3' boundary of the cassette. Protein-coding genes are represented by arrowed boxes, with the direction of the arrow indicating the direction of transcription.

Note - With the exception of an unusual integron in the genome of *Treponema denticola* ATCC 35405 (Coleman *et al.*, 2004), in which the *intl* gene is transcribed in the opposite direction, all known integrons conform to the generic integron model; however, degenerate integrons and orphan gene cassettes or cassette arrays are also common.

### 1.6.2 – Core Integron Distribution and Diversity

Integrans are divided into classes on the basis of core integron sequence divergence, although all are likely to function in a similar way. A formal classification of what constitutes a class (in terms of sequence identity) has not been made; for the purpose of this thesis, integrans exhibiting >95% Intl amino acid identity and >95% *attI* nucleotide identity were considered to be members of a single class. While members of particular classes share similar core integrans, they may differ in the cassettes they carry and/or the genomic context in which they are found. Integron classes were originally named according to numbering scheme, but have more recently been named according to bacterial species in which it is found (eg. InPstQ for the Integron of P*s. stutzeri* Q; Holmes *et al.*, 2003b). Across known examples, Intl integrases exhibit as little as 33% amino acid identity (Collis *et al.*, 2002b), which reflects a long evolutionary history for this protein family. Relative to Intl, very little sequence conservation is observed in the *attI* sites of diverse core integrans; however, considerable sequence conservation is apparent in the *attI* sites of more closely related core integrans (eg. minimum of 44% nucleotide identity across the *Pseudomonas* integrans listed in Table 1.3). The *attI* region is assumed to have properties which assist Intl-mediated recombination. The identification of binding sites for Intl1 in the *attI1* region of class 1 integrans lends support to this hypothesis (Collis *et al.*, 1998).

For several years after their discovery, all known integrans belonged to one of three types, termed class 1, 2 and 3 integrans. Integrans belonging to these classes are typically associated with antibiotic resistance phenotypes and are contained within mobile genetic elements such as transposons (Table 1.2 and Figure 1.3). Integrans

outside the antibiotic resistance context were first discovered in the *Vibrio cholerae* O1 genome sequence (Clark *et al.*, 1997; Mazel *et al.*, 1998). This integron exhibited several differences to integrons belonging to classes 1-3, including having a chromosomal locus rather than a locus with a mobile element (see Section 1.7). This discovery served as a catalyst for several successful investigations aimed at the specific detection and recovery of integrons (Holmes *et al.*, 2003b; Nemergut *et al.*, 2004; Nield *et al.*, 2001; Rowe-Magnus *et al.*, 2001), and with the ever increasing availability of complete genome sequences, the number of known non-antibiotic resistance integrons has grown steadily since (Figure 1.2). There are currently more than 70 known classes of integron, from a phylogenetically diverse array of bacteria including  $\gamma$ -proteobacteria,  $\beta$ -proteobacteria,  $\delta$ -proteobacteria,  $\epsilon$ -proteobacteria, Planctomycetes, Chlorobia, Chloroflexi, Cyanobacteria and Spirochaetes (Table 1.3). Integrons, or integron remnants, can be detected in approximately 5% of complete bacterial genome sequences. Diverse *intI* genes and gene cassettes can also be amplified directly from environmental DNA samples (Nemergut *et al.*, 2004; Nield *et al.*, 2001). Collectively, these data indicate that integrons which do not appear to confer multiple resistance phenotypes and are most frequently found in the chromosome (as opposed to a mobile genetic element) are relatively common features of the genomes of phylogenetically diverse bacteria.

Core integrons of different classes have also been shown to exhibit functional diversity. Small changes in the sequence of  $P_c$  between various class 1 and class 3 integrons have been shown to impact expression levels of cassette encoded genes (Collis and Hall, 1995; Levesque *et al.*, 1994). *IntI* integrases of different classes also exhibit a preference for their cognate *attI* site (Collis *et al.*, 2002b), and exhibit

variation in activity as site-specific recombinases (Biskri *et al.*, 2005; Collis *et al.*, 2002a). These observations indicate that integrons are functionally variable and raise the possibility that local selective pressures may result in the fine-tuning of particular aspects of integron function.

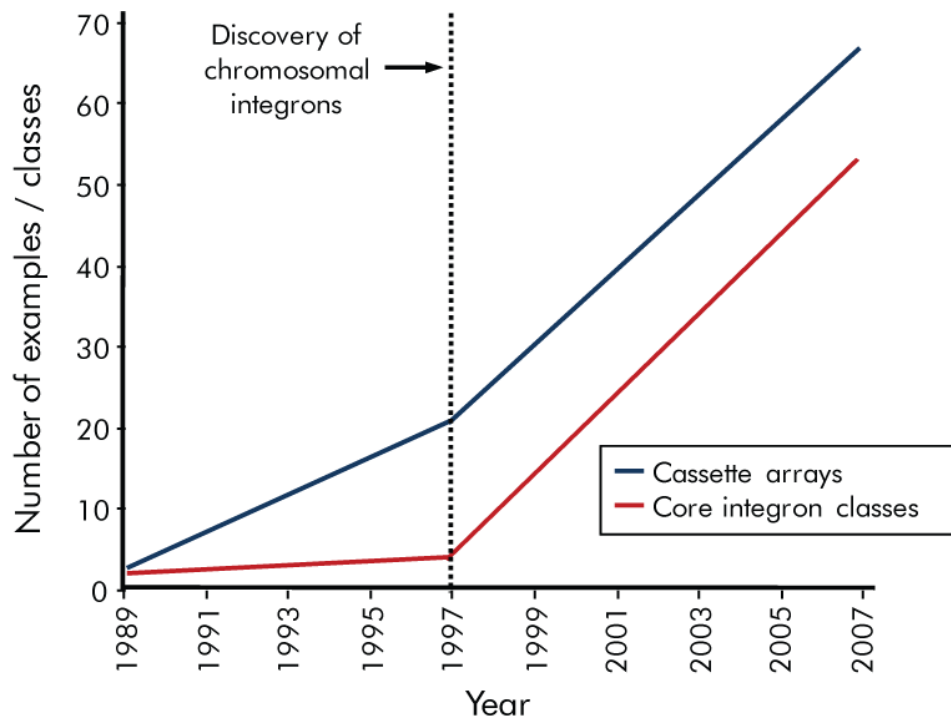


Figure 1.2 – Cumulative number of integron classes and cassette arrays known since the discovery of the integron / gene cassette system. Prior to 1997, all known integrons belonged to one of three classes. These integrons were found to be associated with different gene cassettes in different bacteria and hence the number of known cassette arrays rose sharply over this time. The discovery of the first chromosomal integron coincided with the rapid increase in the availability of genome sequences and resulted in the detection of diverse integrons in different bacteria. Consequently, rate of integron and gene cassette detection increased rapidly after this time. The lines in the graph are not drawn to an accurate scale and are intended as a guide only.





Class	Mobile element	Range of species observed	Activity <sup>1</sup>	Reference
Class 1	Tn402, Tn21	>20 species representing four bacterial divisions	D	Radstrom <i>et. al.</i> , 1994; Brown <i>et. al.</i> , 1996
Class 2	Tn7	3 γ-proteobacterial species	D <sup>2</sup>	Sundstrom <i>et. al.</i> , 1991
Class 3	Transposon	2 γ-proteobacterial species	D	Collis <i>et. al.</i> , 2002
Class 4 <sup>3</sup>	SXT element	<i>Vibrio cholerae</i>	P	Hochhut <i>et. al.</i> , 2001

Table 1.2 - Diversity of known Mobile Integrans (MI) from published literature. A MI is defined as an integron which is directly linked to a mobile element such as a transposon or conjugative plasmid, such that movement of the transposon will result in movement of the integron. All MIs recovered to date have been found to be associated with cassettes encoding resistance to antibiotics, although this need not be the case for all MIs.

- <sup>1</sup> D – Activity demonstrated experimentally, P – Activity predicted on the basis of conserved IntI amino acid sequences, In – Integron predicted to be inactive due to the presence of disrupted or truncated *intI* genes.
- <sup>2</sup> Most class 2 integrons contain an internal stop codon which results in a truncated and non-functional IntI2 integrase. Replacement of the internal stop codon with an amino acid codon has been shown to restore the function of IntI2 (Sundstrom *et al.*, 1991). Recently, several class 2 integrons which contain complete *intI* genes have also been recovered (Barlow and Gobius, 2006).
- <sup>3</sup> The term class 4 integron was originally used to refer to the chromosomal integron of *V. cholera* O1; however, the SXT element integron of *V. cholerae* included in the table above was assigned as class 4 by Mazel (2006).

Table 1.3 (opposite and following 3 pages) - Integrons identified from partial or complete genome sequences, in addition to examples presented in published literature. Included in the table are all *intl* genes detected by performing Blastp and tBlastn searches on published Genbank sequences, including partial and complete genome sequences (search performed on 18th March 2007).

\* - A third *intl* gene not described in the Genbank annotation of *Nitrosomonas europaea* ATCC 19718 was identified using a tBlastn search (corresponding to bases 2096564-2095931 of the complete genome sequence NC\_004757.1 |). The *intl* gene contained at least 4 mutations, resulting in several reading frame changes, and a highly fragmented gene remnant. This remnant was located >300 Kbp from the other *intl* genes in this strain.

- <sup>1</sup> Where available GenPept accession numbers are given as identifiers for each *intl* gene. When GenPept accession numbers were not available, the Genbank accession number is given, in addition to the specific nucleotide positions of the *intl* gene.
- <sup>2</sup> Describes the source of the integron sequence: G - predicted based on conserved sequence information from partial or complete genome sequences; D - detected by molecular methods or specifically described from a genome sequence; MG - genomic information reconstructed from a metagenomic library of environmental DNA.
- <sup>3</sup> Indicates predicted activity of Intl as a site-specific recombinase - described by one of three terms: D - Intl recombination activity demonstrated experimentally; P - Intl activity inferred on the basis of complete and conserved Intl sequences which are predicted to encode functional integrases; In - Intl inferred to be inactive on the basis of deletions/mutations which disrupted the expected Intl sequence.
- <sup>4</sup> Reference - Refers to publications which specifically describe or otherwise characterise particular integron(s). I – Leon and Roy 2003; II – Drouin *et al.*, 2002; III – Rowe-Magnus *et al.*, 2001; IV – Vaisvila *et al.*, 2001; V – Holmes *et al.*, 2003; VI – Rowe-Magnus *et al.*, 2003; VII – Shi *et al.*, 2006; VIII – Mazel *et al.*, 1998; IX – Boucher *et al.*, 2006; X – Gillings *et al.*, 2005; XI – Coleman *et al.*, 2004.

Table 1.3 - Page 1 of 4

Bacterial group	Species/strain	GenPept accession number <sup>1</sup>	Source <sup>2</sup>	Activity <sup>3</sup>	Reference <sup>4</sup>
Chlorobia	<i>Chlorobium phaeobacteroides</i> DSM 266	ABL64322.1	G	P	
	<i>Prosthecochloris aestuarii</i> DSM 271	AAIJ01000003.1 (PaesDRAFT_1746-1747)	G	In	
Chloroflexi	<i>Chloroflexus aggregans</i> DSM 9485	EAV09786.1	G	In	
		EAV09784.1	G	In	
	<i>Roseiflexus castenholzii</i> DSM 13941	EAV25484.1	G	P	
		EAV25499.1	G	P	
		EAV25496.1	G	In	
Cyanobacteria	<i>Synechococcus</i> sp. WH 5701	EAQ75559.1	G	P	
		EAQ76664.1	G	In	
	<i>Synechococcus</i> sp. RS9917	EAQ687111	G	P	
		NZ_AANP01000001 (539362-539820)	G	In	
	<i>Synechococcus</i> sp. JA-2-3B	NC_007776 (939776-939937)	G	In	
	<i>Nostoc</i> sp. PCC 7120	NC_003272.1 (1756531-1756316)	G	In	
Planctomycetes	<i>Kuenenia stuttgartiensis</i>	CAJ73190.1	MG	P	
	<i>Gemmata obscuriglobus</i> UQM 2246	contig:4999 (TIGR*) 37034-36093	G	P	
	<i>Blastopirellula marina</i> DSM 3645	AANZ01000002.1 (DSM 3645_0390)	G	P	
Spirochaetes	<i>Treponema denticola</i> ATCC 35405	AAS12359	G	P	X/
β-Proteobacteria	<i>Acidovorax</i> sp. JS42	ABM43488.1	G	P	
	<i>Methylobium petroleiphilum</i> PM1 (AKA - <i>Rubrivivax gelatinosus</i> PM1)	ABM93658.1	G	P	
	<i>Thiobacillus denitrificans</i> ATCC 25259	AAZ97168.1	G	P	
	<i>Methylobacillus flagellatus</i> KT	ABE48680.1	G	P	
	<i>Nitrosomonas europaea</i> ATCC 19718	CAD86100.1	G	D	/
		CAD843611	G	P	/
		Not annotated*	G	In	
	<i>Nitrosomonas eutropha</i> C91	ABI59614.1	G	P	
	<i>Azoarcus</i> sp. EbN1	CAI06133	G	P	
γ-Proteobacteria	<i>Azoarcus</i> sp. BH72	YP_935226-7	G	In	
	<i>Dechloromonas aromatica</i> RCB	AAZ48156.1	G	P	
	<i>Acidithiobacillus ferrooxidans</i> ATCC 23270	contig:10428 (TIGR*) (2368191-2368647)	G	In	

Table 1.3 - Page 2 of 4

Bacterial group	Species/strain	GenPept accession number <sup>1</sup>	Source <sup>2</sup>	Activity <sup>3</sup>	Reference <sup>4</sup>
γ-Proteobacteria	<i>Alteromonadales bacterium</i> TW-7	EAW25851.1	G	P	
	<i>Alteromonas macleodii</i> 'Deep ecotype'	EAR05568.1	G	P	
	<i>Colwellia psychrerythraea</i> 34H	NC_003910 (4198312-4197959)	G	In	
	<i>Marinobacter aquaeolei</i> VT8	ABM20277.1	G	P	
		ABM18073.1	G	P	
	<i>Pseudoalteromonas haloplanktis</i> TAC125	CAI86536.1	G	P	
	<i>Pseudoalteromonas tunicata</i> D2	EAR28382.1	G	P	
		EAR30315.1	G	P	
	<i>Pseudoalteromonas atlantica</i> T6c	ABG40363	G	P	
	<i>Psychromonas ingrahamii</i> 37	ABM03211.1	G	P	
	<i>Saccharophagus degradans</i> 2-40	YP_525901.1	G	In	
	<i>Shewanella amazonensis</i> SB2B	ABL99562.1	G	P	
		ABL99840.1	G	In	
	<i>Shewanella baltica</i> OS195	EAU26882.1	G	P	
	<i>Shewanella oneidensis</i> MR-1	AAN55084.1	G	D	//
	<i>Shewanella putrefaciens</i> CIP 69.34	AAK01408.1	D	P	///
	<i>Shewanella putrefaciens</i> 200	EAY54292.1	G	P	
		EAY52096.1	G	P	
	<i>Shewanella</i> sp. MR-7	AB143120.1	G	P	
	<i>Shewanella woodyi</i> ATCC 51908	EAV38279.1	G	P	
	<i>Alkalilimnicola ehrlichei</i> MLHE-1	YP_742633.1	G	P	
	<i>Nitrococcus mobilis</i> Nb-231	EAR20185.1	G	P	
	<i>Oceanobacter</i> sp. RED65	EAT13689.1	G	P	
	<i>Oceanospirillum</i> sp. MED92	EAR62174.1	G	P	
	<i>Pseudomonas alcaligenes</i>	AAK73287.1	D	P	/V
	<i>Pseudomonas mendocina</i> ymp	EAV21281.1	G	In	
	<i>Pseudomonas stutzeri</i> BAM	AAN16071.1	D	P	✓
	<i>Pseudomonas stutzeri</i> Q	AAN16061.1	D	D	✓
	<i>Azotobacter vinelandii</i> AvOP	EAM04304.1	G	In	
	<i>Listonella anguillarum</i>	AAM95157.1	D	P	VI
	<i>Listonella pelagia</i>	AAK02082.2	D	P	///
	<i>Vibrio alginolyticus</i> 12G01	EAS76953.1	G	In	
	<i>Vibrio cholerae</i> AM-19226	NZ_AA01000020	G	P	
	<i>Vibrio cholerae</i> B33	EAZ75637.1	G	P	
	<i>Vibrio cholerae</i> MAK 757	EAY38273.1	G	P	
	<i>Vibrio cholerae</i> MO10	ZP_01479120.1	G	P	

Table 1.3 - Page 3 of 4

Bacterial group	Species/strain	GenPept accession number <sup>1</sup>	Source <sup>2</sup>	Activity <sup>3</sup>	Reference <sup>4</sup>
γ-Proteobacteria	<i>Vibrio cholerae</i> NCTC 8457	EAZ74583.1	G	P	VII
	<i>Vibrio cholerae</i> MZO-2	NZ_AAWF01000136	G	P	
	<i>Vibrio cholerae</i> MZO-3	EAY39948.1	G	P	
	<i>Vibrio cholerae</i> O1 biovar eltor	AAK95987.2	G	P	
	<i>Vibrio cholerae</i> O139	BAE71364.1	D	P	
	<i>Vibrio cholerae</i> O395	ZP_00755120.1	G	P	
	<i>Vibrio cholerae</i> V51	ZP_01484802.1	G	P	
	<i>Vibrio cholerae</i> V52	AAC38424	G	P	VIII
	<i>Vibrio cholerae</i> 2740-80	ZP_01675851.1	G	In	
	<i>Vibrio cholerae</i> 569B	AAC38424	D	P	
	<i>Vibrio cholerae</i> 623-39	NZ_AAWG01000020	G	P	III
	<i>Vibrio fischeri</i> CIP 103206	AAK02079	D	P	
	<i>Vibrio fischeri</i> ES114	AAW87733	G	P	
	<i>Vibrio harveyi</i> HY01	NZ_AAWP01000002	G	P	III
	<i>Vibrio metschnikovii</i>	AAK02074.1	D	P	
	<i>Vibrio mimicus</i>	AAD55407.2	D	P	
	<i>Vibrio natriegens</i> CIP 103193	AAO38263.1	D	P	VI
	<i>Vibrio parahaemolyticus</i> CIP 75.2T	AAK02076.1	D	P	
	<i>Vibrio parahaemolyticus</i> RIMD 2210633	NP_798244.1	G	P	
	<i>Vibrio salmonicida</i>	CAC35342.1	D	P	IX
	<i>Vibrio</i> sp. DAT722	ABA55859.1	D	P	
	<i>Vibrio</i> sp. Ex25	ZP_01476288.1	G	P	
	<i>Vibrio</i> sp. MED222	EAQ52311.1	G	In	VI
	<i>Vibrio vulnificus</i> CIP 75.4	AAN33109.1	D	P	
	<i>Vibrio vulnificus</i> CMCP6	AAO10775.1	G	P	
	<i>Vibrio vulnificus</i> YJ016	BAC94705.1	G	P	III
	<i>Xanthomonas</i> sp. CIP 102397	AAK07447.1	D	P	
	<i>X. axonopodis</i> DAR34895	AAX24163.1	D	In	
	<i>X. axonopodis</i> DAR73877	AAX24165.1	D	In	
	<i>X. axonopodis</i> AAX24163	AAX24163	D	In	
	<i>X. axonopodis</i> DAR26930	AAX14926.1	D	P	
	<i>X. campestris</i> DAR69819	AAX24191.1	D	P	
	<i>X. campestris</i> DAR54703	AAX24195.1	D	P	
	<i>X. campestris</i> CIP 100069T	AAK07444.1	D	P	
	<i>X. campestris</i> DAR30538	AAX24185.1	D	P	
	<i>X. campestris</i> 8004	AAV47440.1	G	P	
	<i>X. campestris</i> ATCC 33913	AAM39663.1	G	P	

Table 1.3 - Page 4 of 4

Bacterial group	Species/strain	GenPept accession number <sup>1</sup>	Source <sup>2</sup>	Activity <sup>3</sup>	Reference <sup>4</sup>
$\gamma$ -Proteobacteria	<i>X. campestris</i> DAR61713	AAX149411	D	P	X
	<i>X. campestris</i> 85-10	CAJ24368.1	G	P	
		CAJ21987.1	G	P	
	<i>X. campestris</i> DAR61714	AY928789	D	In	X
	<i>X. oryzae</i> KACC10331	AAW77486.1	G	In	
	<i>X. oryzae</i> MAFF 311018	AAW75733.1	G	In	
		BAE70752.1	G	In	
		BAE69103.1	G	In	
	<i>X. campestris</i> DAR73878	AY928790	D	In	X
	<i>X. axonopodis</i> DAR34876	AY928777	D	In	X
	<i>X. axonopodis</i> DAR26930	AAX14926.1	D	P	X
	<i>X. translucens</i> DAR26929	AY928787	D	In	X
$\delta$ -Proteobacteria	<i>Desulfuromonas acetoxidans</i> DSM 684	EAT14222.1	G	P	
	<i>Desulfotalea psychrophila</i> LSv54	CAG36707	G	P	
	<i>Geobacter lovleyi</i> SZ	EAV89277.1	G	P	
	<i>Geobacter metallireducens</i> GS-15	ABB332211	G	P	
		ABB330211	G	P	
	<i>Geobacter sulfurreducens</i> PCA	NP_953513	G	P	
	<i>Alkaliphilus metalliredigens</i> QYMF	EAO79949.1	G	P	
		EAO80020.1	G	P	
	<i>Pelobacter carbinolicus</i> DSM 2380	ABA88624.1	G	In	
	<i>Pelobacter carbinolicus</i> DSM 2380	ABA87802.1	G	P	
$\epsilon$ -Proteobacteria	<i>Sulfurimonas denitrificans</i> ATCC 33889	YP_393293.1	G	P	

### 1.6.3 – Integrons and Genomic Context

As mentioned previously, integrons are a two-component system in which cassettes are the mobile elements and the core integron is the non-mobile scaffold into which cassettes are integrated and expressed. As integrons may be contained within other genetic elements, the potential exists for different selective pressures to operate on an element which is wholly mobile (when contained within a mobile element), as opposed to an element with a combination of mobile and fixed components (integrons not contained within mobile elements). Several differences between integrons found within mobile elements and many found within bacterial chromosomes lend support to this hypothesis (Table 1.4). It is possible that differences between integrons in different genomic contexts are the result of functional differences; however, the fact that all available data on integron function indicate that genomic context has no discernable effect recombination activity argues against this (Table 1.5) (Biskri *et al.*, 2005; Holmes *et al.*, 2003b). Class 1 integrons are by far the best characterised integron family and have served as a model for the demonstration of integron function. Based on class 1 model, the essential components that constitute a functioning integron have been determined to be a *IntI*-integrase capable of integration and excision of cassettes at *attI* (Table 1.5). The presence of  $P_{cI}$ , while not related to recombination activity is also a fundamental integron component, as the expression of cassette encoded genes gives integrons the opportunity to provide a selective advantage. All but one of these properties has been demonstrated for two different CIs (Table 1.5). This provides strong evidence that the fundamental function of integrons is independent of genomic context. The differences between integrons contained within mobile elements and chromosomes are thus most likely to result from selective constraints which operate on integrons in different



genomic contexts. The implications of this phenomenon are discussed in the following two sections.

Different names have been variously assigned to different integrons, including Multiple Resistance Integrons and Mobile Integrons for those encoding multiple antibiotic resistance, and Super Integrons and Chromosomal Integrons for those found within bacterial chromosomes and not encoding antibiotic resistance. However, specific definitions that would allow all integrons to be unambiguously classified on the basis of genomic context have yet to be proposed. Differences between integrons may occur on the basis of genomic context; integrons found within mobile elements have mobile core integrons and may be subject to HGT frequently, while integrons found within chromosomes have fixed loci and likely rarely undergo HGT. Investigation of these questions requires an unambiguous and meaningful classification scheme that reflects this difference. It is therefore proposed (and will be used for the purpose of this thesis) that the term 'Mobile Integron' (MI) be used to refer to any integron that is found within an element predicted to confer autonomous mobility, and the term 'Chromosomal Integron' (CI) be used to refer to any integron which is found in a bacterial chromosome and is predicted to have a 'framework' locus (ie. is found at a chromosomal locus, has no detectable association with a larger mobile genetic element, and is vertically inherited in that bacterial lineage).

	<b>Class 1</b>	<b>Generalised MI</b>	<b>Generalised CI</b>	<b><i>Vibrio</i> spp. CIs</b>
<b>Immediate Context</b>	Transposon (IR boundaries)	<b>Mobile element</b>	<b>Chromosomal locus</b>	Chromosomal framework genes
<b>Propensity for HGT</b>	High (found in > 10 bacterial genera)	<b>High</b>	<b>Very low</b>	Low**
<b>Vertical Inheritance</b>	Relatively Low	<b>Relatively Low</b>	<b>High</b>	High (Present in 20/24 <i>Vibrio</i> spp. genome sequences)
<b>Array size</b>	0 - 10	<b>Constrained</b>	<b>Unconstrained</b>	Usually >50
<b>Higher order phenotypes</b>	Multiple antibiotic resistance	<b>Common</b> (Immediate adaptive advantage conferred)	<b>Unknown</b>	None observed
<b>59be Sequence Diversity</b>	High (> 100 known from >20 subfamilies)	<b>High</b>	<b>Low (most cassettes &gt;80% identity)</b>	Low (most <i>V. cholerae</i> cassettes >80% identity)

Table 1.4 - Generalised differences between mobile and chromosomal integrons. The differences listed are divided into those relating to the core integron (top-half of table) and gene cassettes (bottom-half of table). The fundamental difference between the integron types is that MIs are located within a mobile element, while CIs are not; all other differences are most likely a direct or indirect consequence of this difference in genomic context.

Integron type	Integron Class	P <sub>c</sub>	<i>attI</i> recognised by IntI	59be subfamilies recognised	SSR integration	SSR excision	Refs
MI	Class 1	✓	✓	>15	✓	✓	1-5
	Class 3	✓	✓	2	✓	✓	6
CI	InVchO1	?	✓	1	✓	✓	7 & 8
	InPstQ	✓	✓	3	✓	?	9 & 10

Table 1.5 - Experimentally demonstrated activities of integrons belonging to different classes and in different genomic contexts. Question marks indicate functions which have not been demonstrated experimentally.

SSR – Site Specific Recombination. ‘SSR integration’ refers to IntI mediated integration of gene cassettes, while ‘SSR excision’ refers to IntI mediated excision of cassettes.

References: 1 - (Martinez and De La Cruz, 1990); 2 – (Hall *et al.*, 1991); 3 – (Collis and Hall, 1992b); 4 – (Collis and Hall, 1992a); 5 – (Collis *et al.*, 1993); 6 – (Collis *et al.*, 2002a); 7 – (Rowe-Magnus *et al.*, 2002); 8 – (Biskri *et al.*, 2005); 9 – (Holmes *et al.*, 2003b); 10 – (Coleman and Holmes, 2005).

## 1.7 – The mobile integron paradigm

Integrans that are contained within mobile genetic elements are predicted to undergo HGT frequently. This assumption forms the basis of the mobile integron paradigm (Figure 1.4a). Under this model, cassette encoded genes lead to the maintenance of integrans on mobile genetic elements under strong selective pressures. Selection favours dissemination of the integron and parent element by HGT. Over the course of several transfer events, the integron recruits new gene cassettes from different hosts/environments and acquires new cassette encoded phenotypes. The ability of MIs to accumulate suites of genes which confer a selective advantage and then readily transfer these genes to other organisms in the same ecosystem or community, means that entire bacterial communities have the potential to respond in a concerted manner to rapid changes in selective pressures. Thus, the ability to undergo horizontal transfer means that MIs are likely to impact bacterial evolution at the community level of organisation (Figure 1.4a).

The potential for MIs to impact bacterial evolution is well illustrated by prevalence of integron encoded multiple antibiotic resistance in pathogenic bacteria. MIs currently consist of integrans belonging classes 1-4 (see Table 1.2), all of which have been found to carry cassettes encoding antibiotic resistance determinants. Of these, class 1 integrans are by far the most abundant and widely distributed MIs (Table 1.2). Class 1 integrans are clearly homologous entities, sharing identical or near-identical core integron regions. More than 100 independent examples of class 1 integrans have been submitted to public databases, including >25 species from 4 bacterial domains. This is unambiguous evidence of extensive HGT of the class 1 core integron across large phylogenetic distances. Class 1 integrans also exhibit mobility between different

parent mobile elements and are also found associated with several different mobile elements; however, most occur in the same immediate genomic context, bounded by inverted repeat sequences and associated with *tni* gene modules or *tni* remnants (Figure 1.3). The high level of diversity of cassette-associated recombination sites (59-bp) indicates that individual cassettes, while encoding common general functions, are recruited from diverse sources, which is again consistent with frequent horizontal transfer of MIs.

All currently known MIs are associated almost exclusively with cassettes encoding antibiotic resistance determinants, such that they have sometimes been referred to as Multiple Resistance Integrations (MRIs). This is likely due to a sampling bias towards clinical isolates and strong selective pressure for the accumulation and dissemination of multiple antibiotic resistance in clinical and agricultural environments. However, there is no logical reason why MIs should be restricted to a multiple antibiotic resistance context and it is likely that MIs outside this context will be found as sequence databases continue to expand.

## Class 1 integron

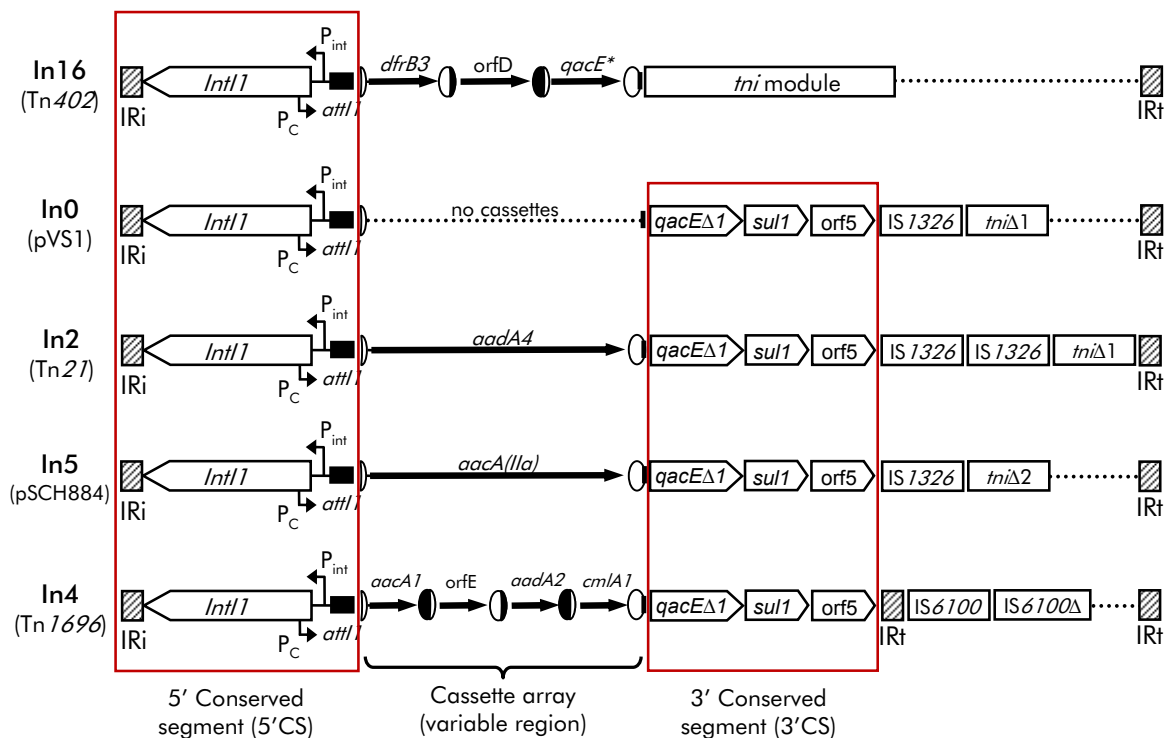


Figure 1.3 – Characteristics of the class 1 integron gene cassette site-specific recombination system. Class 1 integrons are defined on the basis of sequence conservation within the 5' conserved segment. Class 1 integrons typically (but not always, as in the case of In0) contain one or more gene cassettes which encode a variety of antibiotic resistance determinants. All class 1 integrons are flanked by conserved inverted repeat sequences (termed *IRi* and *IRt*). Many class 1 integrons also share conserved sequences between the last gene cassette and *IRt*, and is referred to as the 3' conserved segment (3' CS); however, several changes due to the insertion of IS elements and deletion of existing sequences are also apparent. In16 does not contain the 3' CS associated with other class 1 integrons, but instead contains the complete *tni* gene module which confers a transposition phenotype. This integron does, however, contain a homologue of *qacEΔ1* which is not truncated and is contained in a cassette in the last position of the array (indicated by a \*), which suggests that *qacEΔ1* originated from inactivation of a gene cassette. The structure of In16, containing a complete *tni* module, *qacE* as a gene cassette and the absence of the 3' CS, suggests that it may represent an ancestral form of the other integrons included. Elements within the figure are not drawn to scale.

## 1.8 – The generalised chromosomal integron

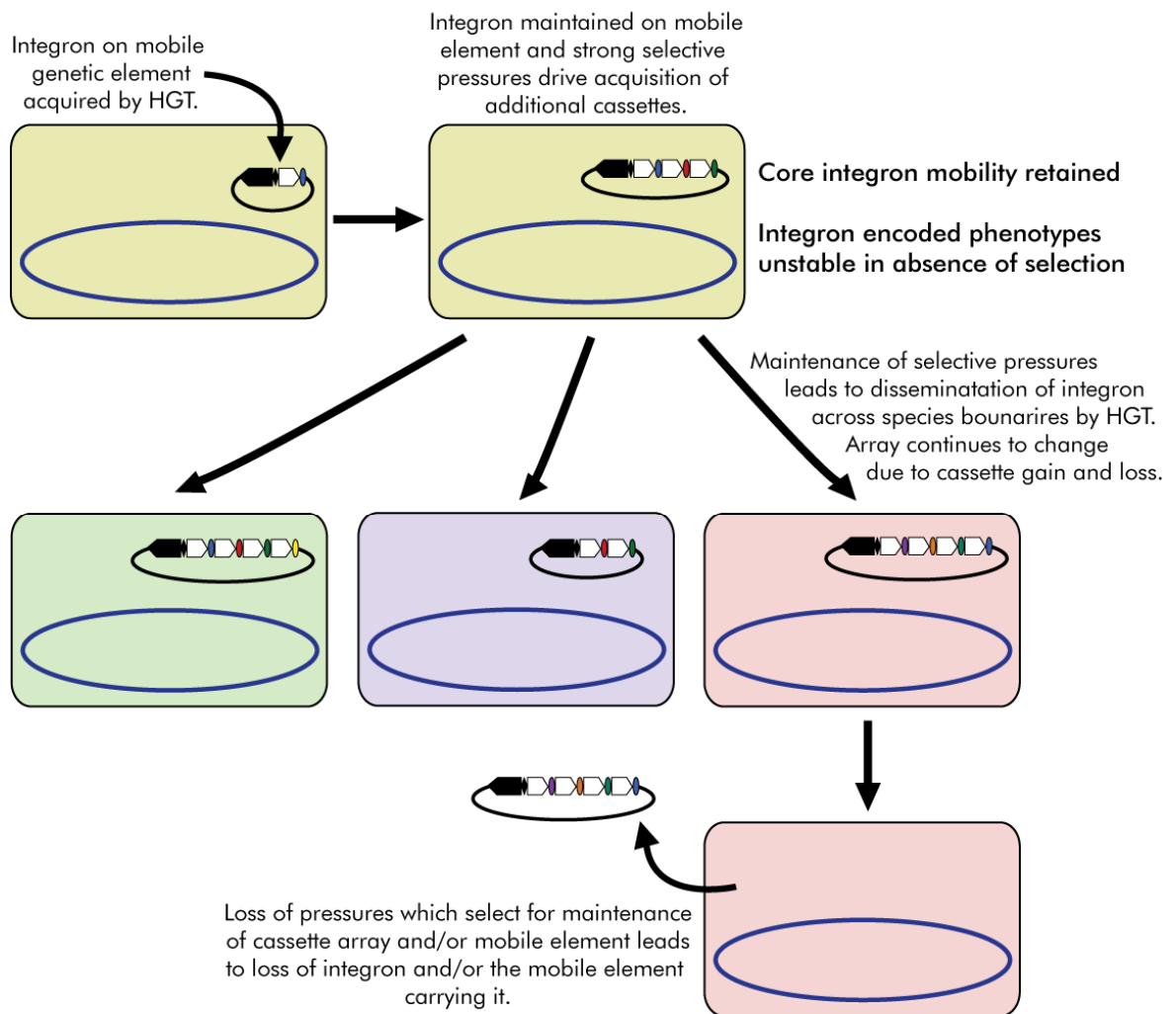
Chromosomal integrons are hypothesised to occur at fixed chromosomal loci and are consequently relatively rarely exposed to HGT. This assumption forms the basis of the chromosomal integron paradigm (Figure 1.3b). The chromosomal integron paradigm was based primarily on patterns of CI diversity in several *Vibrio* strains (Table 1.4) (Rowe-Magnus *et al.*, 2001). However, CIs consistent with the paradigm were subsequently identified in *Xanthomonas* (Gillings *et al.*, 2005) and *Pseudomonas* (Holmes *et al.*, 2003b; Vaisvila *et al.*, 2001). The conformation of CIs from bacterial groups beyond these three genera is yet to be specifically analysed, and thus the dataset upon which the CI paradigm is based remains somewhat limited.

In addition to occurring at conserved chromosomal loci, CIs are predicted to contain *intI* genes with phylogenies that are congruent with chromosomal marker genes, and tend to be associated with larger cassette arrays with similar 59-be sequences (Table 1.4). The 59-be of CIs typically form evolutionarily distinct subgroups on the basis of sequence conservation (>80% nucleotide identity) and/or the presence of heterologous sequence insertions that are unique with respect to other 59-be (Holmes *et al.*, 2003b). The fixed nature of CI loci means that these integrons are predicted to be exposed to HGT rarely, which may limit the size of the cassette pool which can be accessed by the integron. A fixed locus and stable vertical transmission also increases the long-term stability of selective pressures, which in turn provides increased opportunity for the evolution of multi-gene phenotypes. Loss of *IntI* function prevents further rearrangement of cassettes in the array and may serve to fix combinations of genes which provide selectable phenotypes.

The broad-scale significance of CIs remains unclear. Their apparent sedentary nature, large cassette arrays and high 59-be sequence identity suggest that different functional and/or evolutionary constraints operate on CIs relative to MIs. Integrons share many similarities with operons. These similarities include: clustering of constituent genes (gene cassettes in integrons) behind a common promoter, co-transcription of constituent genes, and containing genes which are most often not associated with independent regulatory information. The key difference between integrons and operons is that gene cassettes are mobile and cassette arrays dynamic, whereas gene order in operons is fixed and often stable over long evolutionary timescales. Theoretically, all that needs to occur to fix the genes in a cassette array is a loss of 59-be or Intl recombination activity. As Intl inactivation can occur in a single step, it may be expected that cassette array fixation most often occurs through inactivation of Intl, rather than inactivation of multiple 59-be. It is not difficult to imagine a scenario in which selection would favour immobilisation of a cassette array that encodes a multi-gene phenotype which provides a selective advantage (Figure 1.2). While the probability of accumulation of a suite of functionally complementary genes through the random acquisition and rearrangement of single gene cassettes is extremely low, the long residence time of CIs is likely to result in a stabilisation of selective pressures (relative to an integron frequently exposed to HGT) which may facilitate the accumulation of complementary genes. This is an attractive potential role for chromosomal integrons; however, in the absence of an example of a CI assembled multi-gene phenotype or experimental data on phenotype acquisition from directed evolution experiments, a role for CIs as operon engineers can be inferred only.



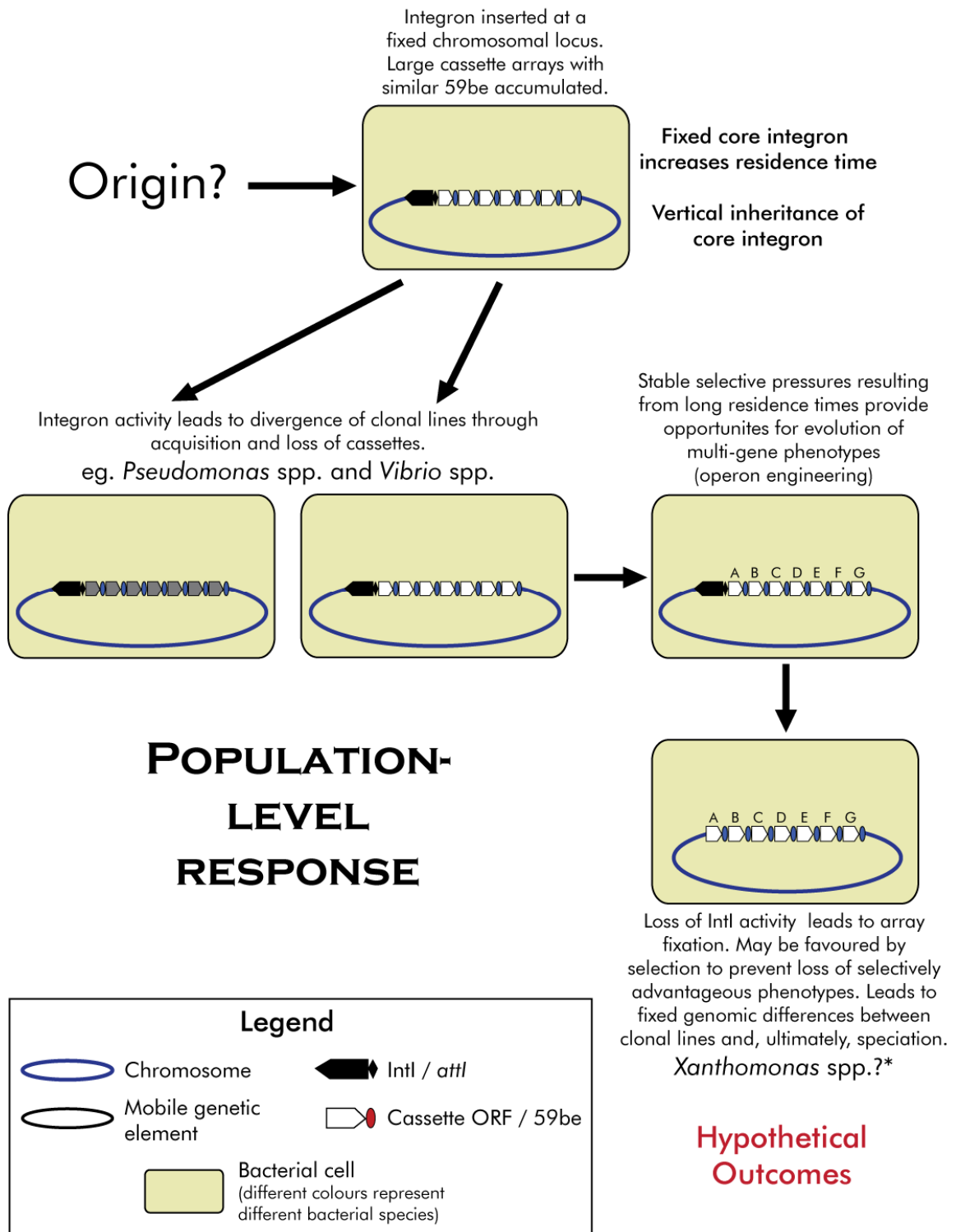
## a. Mobile Integrin Paradigm



## COMMUNITY-LEVEL RESPONSE

Figure 1.4 (above and opposite) – Hypothetical evolutionary consequences of integrin activity in mobile integrons (a.) and chromosomal integrons (b.). The unique, two-component nature of integrons, in which the core integron is not independently mobile, means that the response of integrons to particular selective pressures is likely to vary depending on genomic context. The key difference between MIs and CIs is that MIs are wholly mobile and are therefore exposed to HGT frequently and are maintained for long periods only in the presence of selection, while CIs are fixed in the chromosome and are stably transmitted primarily by vertical inheritance. This phenomenon is hypothesised to result in a dichotomy between integrons in different genomic contexts.

## b. Chromosomal Integrator Paradigm



The chromosomal integron concept was based on a somewhat limited dataset (CIs in several *Vibrio* spp. and a single example of a CI in a Pseudomonad). As more data has accumulated in the form of genome sequences and studies targeted at integron recovery, the number of apparently chromosomal integrons known has grown to more than 100 (Table 1.3). Several patterns emerging from this dataset suggest that while chromosomal integrons may be stably vertically inherited in some lineages, other integrons with no detectable link to mobile elements have been subjected to HGT (see Section 1.7). Specific examination of CIs in additional bacterial lineages is therefore required to assess the universality of the chromosomal integron paradigm.

### **1.8.1 – Limitations of the generalised CI concept**

Members of the *Vibrio* genus are rapidly becoming a model system for investigation of the chromosomal integron/gene cassette system (Biskri *et al.*, 2005; Boucher *et al.*, 2006; Rowe-Magnus *et al.*, 2001; Rowe-Magnus *et al.*, 2002; Rowe-Magnus *et al.*, 2003), and the CI paradigm was formulated primarily on the basis of observations of *Vibrio* CIs. While limited data on the characteristics of CIs in Pseudomonads (Holmes *et al.*, 2003b; Vaisvila *et al.*, 2001) and Xanthomonads (Gillings *et al.*, 2005) is consistent with the CI paradigm, the dataset upon which the CI paradigm is based is very limited with respect to the range of bacterial species which are predicted to contain CIs (Table 1.3). Thus, uncertainty exists about the applicability of the CI paradigm as model for the evolutionary impacts of integrons across the strains predicted to contain CIs in Table 1.3.

Emerging patterns of diversity apparent from integrons in complete genome sequences are revealing several inconsistencies with the CI paradigm. Integrons show

a patchy distribution across phylogenetically diverse bacteria (approximately 5% of complete genome sequences contain integrons, representing 9 bacterial divisions). Multiple copies of identical (or near-identical) core integrons (eg. *Nitrosomonas europaea* ATCC 19718) and multiple divergent core integrons (eg. *Roseiflexus castenholzii* DSM 13941 and *Geobacter metallireducens* GS-15) are commonly found within single genomes. CIs associated with *Shewanella* spp. exhibit many characteristics which are inconsistent with the chromosomal integron paradigm, including multiple chromosomal loci and low levels of 59-bp sequence conservation.

Differences are also apparent between the CIs in *Vibrio* spp. and *Xanthomonas* spp. with respect to those of *Pseudomonas* spp.; CIs or CI remnants are found in most *Vibrio* (20/24) and *Xanthomonas* (6/6) complete genome sequences. In contrast, CIs are detectable in only 1 (*Ps. mendocina* YMP) of the 17 *Pseudomonas* genome sequences available. Further, Vaisvila *et. al.* (2001) failed to detect sequences indicative of CI presence in 5 of 8 *Pseudomonas* spp. screened. Collectively, these observations indicate that the differences between MIs and CIs are not at all clear cut and it is possible that integrons with intermediate characteristics will be more common than integrons that conform in all aspects to the paradigmatic CI or MI. Nonetheless, the presence of consistent stark differences between some MIs and CIs from distantly related bacteria remains convincing evidence of the existence of a dichotomy in at least some integrons, which is most likely caused by differences in genomic context.

## 1.9 - The need for model organisms to characterise CIs

The evolutionary history of CIs in bacterial lineages can be investigated through comparison of integrons from closely related strains. Determination of genomic context, comparative phylogenies of integron-associated genes with phylogenetic marker genes, comparison of 59-bes and comparison of cassette arrays can provide an abundance of information on the history of integron activity in a bacterial lineage. As gene cassette arrays are dynamic structures, very closely related strains are needed to assess rates of cassette turnover. However, as the core integron of CIs is not mobile, and may be an ancient feature of the bacterial lineage, more distantly related strains are needed to assess their evolutionary history. There is a need for model bacterial groups which meet these criteria; a bacterial group in which several closely and distantly related, taxonomically well-characterised strains are available and in which at least some strains contain CI's. Several bacterial genera, including *Vibrio*, *Xanthomonas* and *Pseudomonas* do meet these criteria to differing degrees.

### 1.9.1 – *Pseudomonas* spp. as a model organism

Several strains representing three different species (*Ps. alcaligenes*, *Ps. straminea* and *Ps. stutzeri*) are known to contain integrons (Holmes *et al.*, 2003b; McNichol, 2002; Vaisvila *et al.*, 2001), and indirect evidence suggests their presence in at least one additional member of the genus (Vaisvila *et al.*, 2001). However, the apparent absence of CIs in all but one of the available *Pseudomonas* spp. genome sequences, in addition to the fact that integrons were not detected in 3 of the 8 *Pseudomonas* spp. strains screened by Vaisvila *et al.* (2001), indicate that while CIs may be

ubiquitous features of the genomes of some *Pseudomonas* spp., they appear to be absent in most. This observation is in contrast to the patterns of CI diversity observed in *Vibrio* and *Xanthomonas*, suggesting a different evolutionary history for CIs in this genus. The significant differences apparent between *Pseudomonas* CIs and the CIs of other bacterial lineages warrant attention and make *Pseudomonas* a good candidate model organism for characterising CIs in their own right. In addition, the availability of numerous *Pseudomonas* strains with well defined phylogenetic relationships and representing a spectrum of evolutionary divergence will allow robust inferences to be made on the evolutionary history of CIs in this lineage.

In order to map the evolutionary history of a given sequence, it is desirable to have a phylogenetically well characterized set of strains, so that robust assertions can be made about the phylogeny of the sequence in question. The genus *Pseudomonas* is one of the most diverse, biologically relevant, and extensively characterized bacterial genera known (Spiers *et al.*, 2000). It is a well characterized genome in terms of 16S rRNA phylogeny, having undergone a massive restructuring on its basis, resulting in the reclassification of several *Pseudomonas* spp. as different genera, and the establishment of a coherent, 'core' group of *Pseudomonas* strains, termed *Pseudomonas sensu stricto* (Palleroni *et al.*, 1973). Seven complete and several incomplete genome sequences for members of the genus are available. Of the members of the genus known to contain CIs, the *Ps. stutzeri* species complex to date constitutes one of the most extensively characterized bacterial groups. Several *Ps. stutzeri* strains are known to contain integrons (Holmes *et al.*, 2003b), the phylogenetic structure of the species complex is well defined, and many strains exhibiting varying degrees of relatedness are available. This provides a strong basis from which to investigate the evolutionary history of chromosomal integrons.

Members of the genus *Pseudomonas*, and in particular the *Ps. stutzeri* species complex, are thus good candidate organisms to serve as a model for the significance of CIs to bacterial evolution.

## **1.10 – The genus *Pseudomonas***

### **1.10.1 – A brief history of genus *Pseudomonas***

Members of the genus *Pseudomonas* are highly versatile metabolically, utilising a wide range of simple and complex organic compounds. They are ubiquitous in soil and water and are important pathogens of animals, plants and humans (Palleroni, 1992). The genus *Pseudomonas* was originally established based upon characteristics of cell morphology and metabolism. With the growing use of molecular typing methods over the past 20 years, it became evident that phenotypic data did not always correlate with molecular taxonomies (Woese, 1987). As a result of this, many organisms originally described as species of the genus *Pseudomonas* have been reclassified. *Pseudomonas* species are now grouped on the basis of 16S rRNA-DNA hybridisation studies (Palleroni *et al.*, 1973). Genuine *Pseudomonas* species (*Pseudomonas sensu stricto*) are currently defined as those found within the rRNA group I as defined by Palleroni *et al* (1973).

### **1.10.2 – The *Ps. stutzeri* species complex**

*Ps. stutzeri* was first described as *Bacillus denitrificans* by (Burri and Stutzer, 1895) and was later reclassified as *Ps. stutzeri* (Van Niel and Allen, 1952). It is a non-fluorescent, gram negative bacterium ( $\gamma$ -proteobacteria), mobile by means of a single polar flagellum. All strains are denitrifiers, liberating nitrogen gas from nitrate,

amylase positive and gelatinase negative, and can utilise starch and maltose as sole carbon sources (Bennasar *et al.*, 1998; Sikorski *et al.*, 1999). Most strains have distinctive colony morphology, with a dry and wrinkled appearance and are highly cohesive. *Ps. stutzeri* has a global distribution and has been isolated from diverse natural and anthropogenic environments (Holmes, 1986; Hubalek *et al.*, 1998; Papapetropoulou *et al.*, 1994; Rossello *et al.*, 1991; Vaisanen *et al.*, 1991). Strains have most commonly been isolated from soil and water, however several clinical isolates also exist, as *Ps. stutzeri* is an opportunistic pathogen of humans (Noble and Overman, 1994). *Ps. stutzeri* is of particular interest to microbial ecology studies. Strains are able to degrade a wide variety of environmental pollutants (Baggi *et al.*, 1987; Garcia-Valdes *et al.*, 1989), naphthalene and methylnaphthalenes (one of the most abundant aromatic components of crude oil and potentially toxic components of petroleum) (Rossello-Mora *et al.*, 1994b), and high-molecular weight polyethylene glycols (Obradors and Aguilar, 1991). Several strains have also served as a model species for the study of denitrification (Zumft and Korner, 1997) and natural transformation (Carlson *et al.*, 1983; Lorenz and Sikorski, 2000; Sikorski *et al.*, 1998).

*Ps. stutzeri* has a complex taxonomy, and may be one of the best characterised species within the genus *Pseudomonas*. All strains exhibit phenotypic traits which permit description at the species level. However, *Ps. stutzeri* strains are highly variable with respect to many phenotypic traits and exhibit very high genomic diversity. In terms of allelic diversity (as determined by MLEE and MLST), *Ps. stutzeri* is the most diverse bacterial species yet described (Cladera *et al.*, 2004; Rius *et al.*, 2001). Several taxonomic studies have indicated that relationships among members of the species are complex, and some strains are clearly more closely related to each



other than to other strains. DNA-DNA hybridisation studies have identified 9 distinct genomic groups, termed genomovars (Gv) (Rossello *et al.*, 1991). The DNA-DNA similarity cut-off value used for genomovar circumscription is equivalent to that recommended for species circumscription using polyphasic taxonomy methods, corresponding to approximately 70% DNA-DNA similarity (Rossello *et al.*, 1991; Vandamme *et al.*, 1996). Several studies have indicated that the high level of intraspecific genotypic and phenotypic heterogeneity between strains is sufficient to warrant division of *Ps. stutzeri* into two or more formal species (Gavini *et al.*, 1989; Palleroni *et al.*, 1970; Stainer *et al.*, 1966). However, no set of diagnostic phenotypic traits is currently known which discriminates any genomovar from other genomovars, preventing formal division of the complex into more than one species. Thus, *Ps. stutzeri* genomovars can be considered as 'genomic species', displaying sufficient genomic divergence to be considered different species, but which are unable to be discerned on the basis of phenotype. It is likely that such phenotypic traits will become available as the biochemical diversity of *Ps. stutzeri* becomes better understood. Only one genomovar has thus far been re-classified; members of Gv.6 were reclassified as *Ps. balearica* based on 16S rRNA phylogeny and a set of defining phenotypic characteristics (Bennasar *et al.*, 1996).

### 1.10.3 – Relationships between *Ps. stutzeri* genomovars

A considerable amount of effort has been devoted to clarification of the relationship between *Ps. stutzeri* genomovars. While no diagnostic (unique and universal) within-genomovar phenotypes have been identified, several diversity measures concur with genomovar designations with varying degrees of resolution. 16S rRNA gene sequencing (Bennasar *et al.*, 1996; Sikorski *et al.*, 1999; Sikorski *et al.*, 2005), PCR-

based genomic fingerprinting (Sikorski *et al.*, 1999), macrorestriction fragment length analysis (Ginard *et al.*, 1997; Rainey *et al.*, 1994), 16S-23S ITS sequence analysis (Cladera *et al.*, 2004; Guasp *et al.*, 2000; Sikorski *et al.*, 2005), total fatty acid analysis, total protein analysis, and multi-locus sequence typing (Cladera *et al.*, 2004) all lend support to at least some genomovar designations (and collectively lend support to most genomovar designations). This substantial data on *Ps. stutzeri* diversity indicates that the genomovar concept provides a robust framework which differentiates subsets of closely related *Ps. stutzeri* strains from other subsets within the species complex.

#### **1.10.4 – *Ps. stutzeri* distribution and ecology**

The *Ps. stutzeri* species complex is ubiquitous in an ecological sense. All genomovars are found in diverse habitats and have been isolated from multiple continents (Holmes *et al.*, 2003b; Rossello *et al.*, 1991; Stainer *et al.*, 1966). Members of different *Ps. stutzeri* genomovars can be isolated from a single site, indicating a complex population structure between different genomovars (Sikorski *et al.*, 2002). The wide geographic distribution and high metabolic diversity of *Ps. stutzeri* suggests a large effective population size. Co-existence of multiple genomovars in single habitats suggests that genomovars, to the extent that they represent a species concept, are sympatric at current scales of observation. However, finer-scale analyses of *Ps. stutzeri* populations may reveal allopatry between genomovars with respect to micro-niches. Niche-specific adaptation to varied microenvironments and subsequent divergence of clonal lines (allopatric divergence) has been suggested as an explanation for the extraordinarily high genetic diversity observed in *Ps. stutzeri* (Schmidt *et al.*, 1996).

### 1.10.5 - By what mechanisms do genomovars diverge?

Low levels of assortive recombination have been observed among members of different genomovars, suggesting a strongly clonal population structure (Rius *et al.*, 2001; Sikorski *et al.*, 2002). Thus, despite the fact that several large scale intra-genomic recombination events have been observed among members of different genomovars (Ginard *et al.*, 1997), it appears that mechanisms other than recombination are largely responsible for the divergence of *Ps. stutzeri* strains into genomovars. Mutation is also likely to drive divergence; however, it is difficult to account for the level of genomic divergence observed between different strains via mutation alone. The natural competence of many *Ps. stutzeri* strains (Lorenz and Sikorski, 2000; Sikorski *et al.*, 1999), the presence of varied plasmids, insertion sequences and mosaic gene structures (Coleman and Holmes, 2005; Ginard *et al.*, 1997), and variation in overall genome size (Ginard *et al.*, 1997; Rainey *et al.*, 1994) suggest that LGT may have played a significant role in the diversification of *Ps. stutzeri*.

### 1.10.6 - A role for integrons in the diversification of *Ps. stutzeri* genomovars

IntI recombination can lead to rapid microevolutionary divergence of clonal lines through the divergence of cassette arrays by integration and excision of cassettes. Diversity may also arise through the acquisition and loss of entire integrons. Fixed differences between clonal lines can arise if integron function is lost, as this prevents further IntI-mediated cassette acquisition or loss, and immobilises the cassettes found in the array. Loss of integron function can occur through the loss or mutation of 59-be or through mutations to the core integron which affect the activity of IntI or *attI*. As

core integron inactivation can occur through a single mutation event, while cassette array inactivation involves at least one mutation event for each cassette in the array, it is expected that loss-of-core-integron-function mutations are most often responsible for integron inactivation. If the fixed cassettes contain genes which confer a selective advantage, this genetic differentiation may lead to ecological differentiation and, ultimately, speciation.

### 1.11 – Aims of the present study

The emerging patterns in CI diversity and characteristics raise several questions regarding the universal applicability of the CI paradigm. Do all CIs exhibit long residence times? *Vibrio* is the only bacterial genus which has been shown to contain CIs with a considerably long residence time. CIs in *Xanthomonas* spp. also exhibit characteristics indicative of a long residence time; however, *Xanthomonas* strains known to contain CIs exhibit significantly less divergence than *Vibrio* strains which contain CIs. Where do CIs come from? Addressing this question requires determination of the frequency with which CIs are acquired by HGT and the upper limit for the residence time of a CI. Why do some strains contain multiple integrons, and did they arise by duplication or independent acquisition? How dynamic are cassette arrays? The limited availability of closely related *Vibrio* spp. strains and frequent inactivation of *Xanthomonas* spp. CIs have prevented this question from being thoroughly addressed. Why are so many *intl* genes predicted to encode non-functional integrases? Several explanations for this are possible. It may simply reflect that integrons have a long evolutionary history. Alternatively, the core integron region may be attractive to mutators or *Intl* inactivation may be favoured by selection to fix useful combinations of cassette encoded genes.

The primary aim of this thesis was to extensively test the conformation of *Pseudomonas* CIs to the CI paradigm (as seen in *Vibrio* and *Xanthomonas*), and to determine the contribution of integrons to intra- and inter-genomovar diversity in the *Ps. stutzeri* species complex. Specifically, the possibility that CIs in *Pseudomonas* spp. exhibit properties consistent with ancient acquisition (on a single occasion) and subsequent stable vertical inheritance of the integron platform throughout the

evolutionary radiation of the genus was tested. Based on the differences between CIs in *Pseudomonas* and those found in *Vibrio* and *Xanthomonas*, it was hypothesised that *Pseudomonas* CIs will exhibit several inconsistencies with the CI paradigm. If this is the case, reconsideration of the nature and evolutionary history of CIs within bacterial lineages will be required.

A strain collection consisting of 24 *Pseudomonas* strains representing five different species was screened for the presence of integrons. DNA typing methods and phylogenetic analysis of marker sequences was used to provide a measure of the relatedness of strains being screened and to establish strain provenance (Chapter 3). The abundance of integron-like sequences across the strain collection was assessed using DNA-based molecular detection methods (Chapter 4). Genomic context was determined through cloning of entire integrons and analysis of the sequences flanking different integrons (Chapter 5). Finally, the evolutionary history of CIs in *Pseudomonas* was explored using phylogenetic analysis of IntI sequences in addition to several other measures of genomic divergence, such as G+C content and codon usage analysis (Chapters 5 & 6). The implications of the results obtained are discussed in relation to the applicability of the CI paradigm to bacterial groups other than *Vibrio* and *Xanthomonas*.

## CHAPTER 2 – GENERAL MATERIALS AND METHODS

### 2.1 – Bacterial Strains

A collection of 20 *Ps. stutzeri* strains (representing 6 genomovars), and a single strain each of *Ps. balearica* (formerly *Ps. stutzeri* Gv.6), *Ps. straminea*, *Ps. aeruginosa*, *Ps. fluorescens* and *Ps. putida* were screened for the presence of integrons (Table 2.1). A conjugative plasmid carrying a class 1 integron (R388) and an *E. coli* strain (JM109) which does not contain an integron were also included in most analyses. Details of all strains used in the study are listed in Table 2.1. Isolation of *Ps. stutzeri* strains Q, BAM17, and P1 is described in Holmes *et al.*, (2003). All other *Ps. stutzeri* strains were kindly provided by J. Sikorski (University of Haifa, Israel). *Ps. straminea* KM91 was isolated from soil in the Hawkesbury Region, NSW (McNichol, 2002). *Ps. fluorescens* NCTC 7244 and *Ps. aeruginosa* NCTC 3756 strains were obtained from the University of Sydney strain collection. *Ps. putida* F1 and the conjugative plasmid R388 were kindly supplied by H. W. Stokes (Macquarie University, Australia). Chromosomal DNA from *Xanthomonas campestris* DAR 30538 was kindly supplied by M. Joss (Macquarie University, Australia). All strains were grown in LB broth (10g/L tryptone, 5g/L yeast extract, 5g/L NaCl, pH 7.5) or on LB agar (LB broth supplemented with 1.5%), and were stored at -80°C in LB medium containing 15% v/v glycerol for long-term strain storage. All *Pseudomonas* spp. strains were grown at 30°C, and all *E. coli* strains were grown at 37°C.

Species	Strain	Gv	Other Designations	Source <sup>1</sup>	Location	Date	Reference
<i>Ps. stutzeri</i>	ATCC 17684	1	Stainer 318	Clinical	Paris, France	Before 1966	Stainer <i>et al.</i> (1966)
	ATCC 17589	1	Stainer 222	Clinical	Copenhagen, Denmark	Before 1966	Rius <i>et al.</i> (2001)
	ATCC 17593	1	Stainer 226	Clinical	Copenhagen, Denmark	Before 1966	Stainer <i>et al.</i> (1966)
	ATCC 17595	2	Stainer 228	Clinical	Copenhagen, Denmark	Before 1967	Stainer <i>et al.</i> (1966)
	ATCC 17587	2	Stainer 220	Clinical	Copenhagen, Denmark	Before 1966	Stainer <i>et al.</i> (1966)
	ATCC 14405	2	Zobell	marine	Pacific Ocean, California	Before 1944	Zobell and Upham (1944)
	19smn4	4	DSM 6084	marine	Barcelona, Spain	1988	Rius <i>et al.</i> (2001)
	DNBP21	5	DSM 6082	WTP	Mallorca, Spain	1988	Rius <i>et al.</i> (2001)
	ATCC 17685	7	Stainer 319	Clinical	Paris, France	Before 1966	Stainer <i>et al.</i> (1966)
	DSM 50238	7	Stainer 419	Soil	Berkley, California	Before 1966	Rius <i>et al.</i> (2001)
	API-2-142	(7)		PPS	Ensenada, Argentina		J. Sikorski, Pers. Comm.
	YPF-41	(7)		PPS	La Plata, Argentina		J. Sikorski, Pers. Comm.
	RNAIII	(7)		Bioreactor	Jena, Germany		J. Sikorski, Pers. Comm.
	Q	(8)		Contaminated soil	Sydney, Australia	2003	Holmes <i>et al.</i> (2003)
	BAM17	(8)		Contaminated soil	Sydney, Australia	2003	Holmes <i>et al.</i> (2003)
	P1	(8)		Contaminated soil	Sydney, Australia	2003	Holmes <i>et al.</i> (2003)
	ATCC 17641	8	Stainer 275	clinical	Copenhagen, Denmark		Stainer <i>et al.</i> (1966)
	JM 300	8	DSM 10701	Soil	California	Before 1982	Rius <i>et al.</i> (2001)
<i>Ps. balearica</i>	SP1402	6*	DSM 6083	WTP	Mallorca, Spain		Rius <i>et al.</i> (2001)
	ATCC 17682	6*	Stainer 316	Clinical	Paris, France	Before 1966	Stainer <i>et al.</i> (1966)
<i>Ps. straminea</i>	KM91	N/A		Soil	Sydney, Australia	2002	McNichol (2002)
<i>Ps. fluorescens</i>	NCTC 7244	N/A		Unknown	Unknown	Unknown	Unknown
<i>Ps. aeruginosa</i>	NCTC 3756	N/A		Unknown	Unknown	Unknown	Unknown
<i>Ps. putida</i>	F1	N/A	DSM 6899				

Table 2.1 - Strains analysed for the presence of integrons in the present study. Genomovar designations shown in parentheses indicate strains in which genomovar classifications are known from surrogate measures, rather than DNA:DNA hybridisation. Reference and isolation details could not be obtained for all strains.

\* - Members of *Ps. balearica* were formerly classified as *Ps. stutzeri* Gv.6

<sup>1</sup> WTP - Wastewater treatment plant, PPS - Petrochemically polluted soil.



## **2.2 – DNA Extraction**

High molecular weight DNA was extracted from all strains using a CTAB/phenol/chloroform modified from Sambrook and Russell (2001). Cells from 5ml overnight cultures were pelleted using centrifugation, washed in several volumes of TE (10mM Tris-HCl pH 8.0, 1mM EDTA), and resuspended in 2 ml CTAB lysis buffer (50mM Tris-HCl pH 8.0, 10mM EDTA, 1M NaCl, 1% CTAB). Lysozyme (1mg/ml) and Ribonuclease A (100µg/ml) were added to the cell suspension, followed by incubation at 37°C for 1-2 hr, and extraction with one volume of chloroform:iso-amyl alcohol (25:1). After recovery of the aqueous phase, high molecular weight DNA was purified by a standard phenol/chloroform extraction and ethanol precipitation protocol as described previously (Sambrook and Russell, 2001). Recovered DNA was dissolved in 250 µl TE and stored at -20°C. Dilutions of the recovered DNA were electrophoresed in 0.8% agarose to assess both quantity and molecular weight.

## **2.3 – PCR**

All PCR amplification mixtures contained the following reagents unless otherwise indicated: 50pmol of each primer, each dNTP at a concentration of 200nM, 2mM MgCl<sub>2</sub>, 1U of *Taq* DNA polymerase (New England Biolabs) in the buffer supplied by the manufacturer, and 5-10 ng of template DNA. The PCR cycling conditions used varied depending on the primers used and are indicated where relevant. Amplification products were separated using agarose gel electrophoresis and visualised via staining with ethidium bromide, as described below.

## 2.4 – Agarose Gel Electrophoresis

Agarose gel electrophoresis was used to determine the size and concentration of all DNA recovered by direct extraction or amplified by PCR. An agarose (Promega) concentration of 0.8% - 2% in 0.5X TBE buffer was used, depending on the size DNA being analysed. All samples were run in a horizontal electrophoresis unit at 1 - 5 V/cm. Fractionated DNA was stained with ethidium bromide and visualised under UV-light.

## 2.5 – Southern Hybridisation

Southern hybridisation was performed as described by Southern (1975) and modified by Sambrook and Russell (2001). Approximately 5  $\mu$ g of genomic DNA of each strain was digested with 10U of either *Pst*I or *Pvu*II (New England Biolabs) in the buffer supplied with each enzyme. After incubation overnight at 37°C, chromosomal digests were separated by electrophoresis in 0.8% agarose gels, and visualised using ethidium bromide staining. DNA from the gel was then transferred to a nylon membrane (Hybond N+, Amersham) using capillary transfer under alkaline conditions (Sambrook and Russell, 2001).

Probes for southern hybridisation were labelled with digoxigenin-6-dUTP (DIG-6-dUTP), directly by PCR using the PCR direct labelling kit (Roche), according to the instructions and recommended PCR cycling parameters supplied with the kit. The efficiency and yield of the labelling reaction was assessed using standard gel electrophoresis of amplified probes. Band intensity after staining with ethidium bromide was used to gauge the yield of the labelling reaction, and as the addition of DIG-dUTP to a PCR product affects electrophoretic mobility, retardation of the

labelled PCR product relative to an unlabelled control was used to assess labelling efficiency. The approximate  $T_m$  of all probes was calculated using the Primer3 software package (Rozen and Skaletsky 1998) within 40bp windows in 10bp steps across the length of the sequence, with the average of these values being used as the  $T_m$  for that probe. The hybridisation temperature used was set at 25-28°C below the calculated  $T_m$  for each probe. All hybridisations were performed in *DIG Easy-Hyb* buffer (Roche) containing 25ng/ml DIG-labelled probe, and were incubated overnight at the hybridisation temperature calculated for each probe. Detection and visualisation of hybridisation products was performed using the DIG Nucleic Acid Detection Kit (Roche) according to the manufacturer's instructions. Washing conditions to remove unbound probe were varied depending on the expected level of sequence divergence between the probe and target DNA. When the level of sequence divergence between probe and target DNA was expected to be low (<10%), washes were performed at 68°C in 0.5 X SSC, 1% SDS. These washing conditions allowed hybridisation products to form in target DNA exhibiting up to 15% sequence divergence and will be referred to as 'high stringency' conditions. When the level of sequence divergence between probe and target DNA was expected to exceed 10%, washes were performed at 55°C in 1X SSC, 1% SDS, and allowed hybridisation signals to be obtained with target DNA exhibiting up to 40% sequence divergence. These washing conditions will be referred to as 'low stringency' conditions.

## **2.6 – Preparation of PCR Products and plasmid DNA for Sequencing**

Where possible, PCR amplification products were sequenced directly. At least 500ng of each PCR product was purified using the Wizard SV PCR and plasmid purification

kit (Promega) according to the manufacturer's instructions. The purified DNA was then sequenced directly at the Macquarie Sequencing Facility (Macquarie University, Australia) using an ABI Prism 377 (Perkin-Elmer Biosystems), and the primers which had been originally used to generate the PCR product.

When direct sequencing of PCR products was not possible, amplicons were directly ligated into a cloning vector. Approximately 50 ng of PCR product was ligated into the pGEM-T Easy Vector (Promega) following the manufacturer's instructions. The ligation mixture was transformed by electroporation into *E. coli* JM109 competent cells. Clones containing insert DNA were identified by blue/white screening.

Plasmids from clones containing inserts were isolated from 3 ml overnight cultures using the Wizard Plus Miniprep DNA Purification System (Promega) as per the manufacturer's instructions. DNA sequencing of cloned inserts was performed at the Macquarie Sequencing Facility (Macquarie University, Australia) using an ABI Prism 377 (Perkin-Elmer Biosystems), using primers flanking the insert region, pGEMF - (5'-CCGACGTCGCATGCTCC-3') and pGEMR - (5'-CTCCCATATGGTCGACCTG-3').

## **2.7 – Sequence Analysis**

All sequences were compiled and all sequence traces checked using the BioEdit software package (Hall, 2001). Multiple sequence alignments were generated using the CLUSTALW software package (Thompson *et al.*, 1994) and optimised using GeneDoc (Nicholas *et al.*, 1997), unless otherwise specified. Phylogenetic trees were constructed using the programs available in the PHYLIP software package (Felsenstein, 1989). All details of the models used to construct phylogenetic trees are

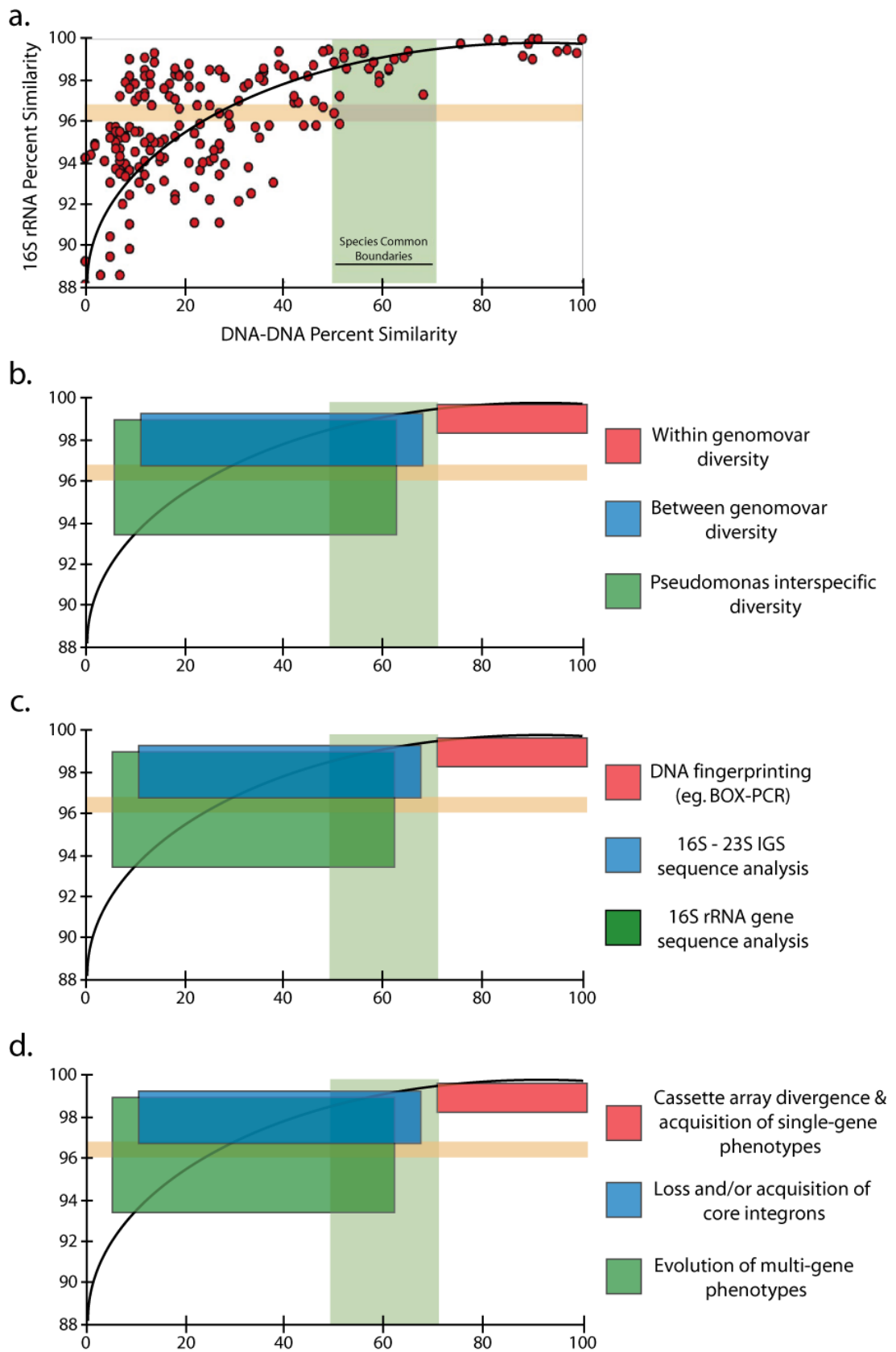
specified where relevant. Print-quality trees were produced using the TreeView software package (Page, 1996). All other details relating to sequence analysis techniques used are given in the relevant chapter.

## CHAPTER 3 – CHARACTERISATION OF STRAIN COLLECTION

### 3.1 – Introduction

Chromosomal integrons are a two-component system, consisting of the core integron (*intI* and *attI*) and cassette array. Each of these components evolve under different constraints. The core integron (or at least *intI* gene) of all integrons are homologous entities; consequently, evolutionary changes in these sequences primarily reflect point mutations of a common sequence, and diversity is observed at the species level of organisation. In contrast, the cassette arrays of different integrons may be considered as heterologous entities, in that evolutionary changes are primarily the result of the loss or gain of cassettes, but will also reflect point mutations, duplications and rearrangements. In any one lineage, cassette array composition is likely to be highly sensitive to changes in the environment, leading to exposure to new cassettes and/or exposure to new selective constraints. Consequently, cassette array diversity is expressed at the level of the clonal line. Changes in selective pressures are likely to rapidly alter cassette array composition, and array diversity is expressed at the level of the clonal line. Thus, to comprehensively examine the evolutionary history of a CI subfamily, a collection of independently isolated strains with a common ancestor and spanning a range of divergence times is required (for examples see (Gillings *et al.*, 2005; Rowe-Magnus *et al.*, 2001)). For example, analysis of cassette turnover requires strains which have recently diverged (eg. clonal lineages), while detection of cassette fixation events resulting in the acquisition of multi-gene phenotypes and the emergence of new ecotypes requires strains with deeper branching evolutionary relationships (Figure 3.1). Other evolutionary events such as movement between loci

Figure 3.1 (opposite) – Conceptual basis for selection of molecular methods to differentiate and characterise strains in the collection. A. Relationship between 16S rRNA percent similarity and total DNA-DNA percent similarity across bacterial strains representing a spectrum of evolutionary divergence (after Rosello-Mora and Amann 2001). The pale green bar and pale brown bar represent the cutoff values for species delineation by DNA-DNA similarity and 16S rRNA similarity, respectively. B. Approximate range of 16S rRNA and total DNA-DNA percent similarities encompassed by different groups within the Pseudomonadaceae. C. Different molecular methods detect genomic changes which occur over different evolutionary timescales. Particular methods are therefore suited to differentiation of strains across limited ranges of divergence. D. Integron events likely to be detected in strains encompassing different ranges of relatedness. It was hypothesised that changes that are likely to occur quickly (eg. acquisition and loss of cassettes) will be detected in closely related strains (red box) and changes that are likely to occur slowly (eg. integron translocation and evolution of multi-gene phenotypes) will be apparent in more distantly related strains (blue and green boxes).





by intra-genomic recombination, CI acquisition or loss by HGT, patterns in mutation (eg. amelioration; (Lawrence and Ochman, 1997) are likely to be evident in strains covering a specific range of divergences (Figure 3.1).

Several additional properties of bacterial genomes have the potential to complicate efforts to characterise the history of integrons within a bacterial group. Genomes are dynamic structures, and large segments may be subject to rearrangement (inversion), duplication, deletion or acquisition. As a consequence, recognising a particular core integron as homologous over a particular bacterial strain collection is not straightforward. For example, paralogous integron copies that may arise by duplication (eg. *Nitrosomonas europaea*; (Leon and Roy, 2003)) are not easily resolved if inversions mean they can be in more than one locus, and deletions mean detection efforts are hampered by inconsistent behavior of molecular probes due to the absence of expected regions in a homologous sequence (eg. *Xanthomonas* spp. integrons; Gillings *et al.*, 2005). All of these factors have the potential to limit the effectiveness of specific molecular probes and therefore complicate the process of detecting all integrons in a bacterial group. These properties of integrons reinforce the need for well characterised strain collections for examining CI subfamilies (Figure 3.1).

The collection used in the present study consisted of a total of 24 strains representing six *Pseudomonas* spp. (Table 2.1). The *Ps. stutzeri* species complex was represented by 18 strains from 6 different genomovars. Four genomovars were represented by multiple strains. Thus, in terms of conventional taxonomic classifications, the *Pseudomonas* spp. strains to be screened for integrons represent a broad spectrum of evolutionary divergence. However, empirical data on the specific evolutionary relationships between different strains is required to confirm this. Chromosomal framework gene phylogenies generally provide the most robust measure of evolutionary relatedness between strains, and members of the *Pseudomonas* genus, particularly *Ps. stutzeri*, have been extensively characterised in this way (Cladera *et al.*, 2004; Guasp *et al.*, 2000; Moore *et al.*, 1996; Sikorski *et al.*, 2005). A set of reference genes which provides robust estimates of the evolutionary relationships between members of the collection was therefore required to determine the levels of relatedness between strains in the collection.

The genomovar concept, while based on variation in a single measure of genetic variation, has proven a robust measure of the relatedness of members of the *Ps. stutzeri* complex, with several different molecular techniques lending strong support to the grouping of genomovars within the species complex (Cladera *et al.*, 2004; Guasp *et al.*, 2000; Sikorski *et al.*, 1999). Consequently, 'genomovar' will be the primary unit of classification used for *Ps. stutzeri* strains in this study. Total DNA:DNA hybridisation data must be used in order to formally determine the genomovar to which a given *Ps. stutzeri* strain belongs. However, screening multiple strains using this technique is labour intensive as each strain must be screened against a set of reference strains. As a result, much research has been dedicated to finding surrogate measures which will accurately determine the genomovar to which a given *Ps. stutzeri*

strain belongs. Exhaustive surveys of biochemical diversity have been carried out on several *Ps. stutzeri* strains in an attempt to find suites of biochemical properties which define particular genomovars, however few unambiguous diagnostic traits have been identified (Rainey *et al.*, 1994; Rossello-Mora *et al.*, 1994a; Rossello *et al.*, 1991; Vancanneyt *et al.*, 1996). More recently, several DNA-based methods have been tested as potential surrogates for DNA:DNA hybridisation, including RAPD PCR (Sikorski *et al.*, 1999), RFLP and sequence analysis of 16S-23S rDNA intergenic spacer (IGS1) regions (Guasp *et al.*, 2000), and Multi-Locus Sequence Typing (MLST) (Cladera *et al.*, 2004), and have collectively proved to be both rapid and powerful tools for differentiating *Ps. stutzeri* genomovars.

The specific aims of the work presented in this chapter were threefold: firstly, to confirm that all strains in the collection are independent isolates; secondly, to confirm the provenance of strains in the collection; and finally, to generate phylogenies with which to compare the evolutionary patterns of integron-associated genes (see Chapter 5). All *Pseudomonas* spp. strains in the collection were subjected to molecular tests to confirm their independence and identity. Species-level relationships between *Ps. stutzeri* strains and other *Pseudomonas* spp. were inferred from 16S rRNA gene phylogenies. Genomovar level relationships between members of the species complex were inferred from IGS1 phylogenies. Variation in the resolving power of these methods means that certain surrogate measures are superior differentiators of *Ps. stutzeri* genomovars with respect to others. Currently, the best known surrogate measure is sequence analysis of the 16S-23S rRNA intergenic spacer (IGS1). With the exception of members of Gv's 1 & 5 all genomovars represented in the present study were well differentiated on the basis of IGS1 divergence (Cladera *et al.*, 2004), and thus it represents a suitable marker for the

evolutionary divergence of the *Ps. stutzeri* strains in the collection. Finally, independence of all strains was confirmed using PCR-based DNA fingerprinting, as this technique has been shown to successfully differentiate closely related *Ps. stutzeri* strains (Sikorski *et al.*, 1999).

## 3.2 – Materials and methods

### 3.2.1 – BOX-PCR

Genomic fingerprints were generated for all strains in the collection using PCR.

Fingerprints were generated using BOX-PCR, and involves the use of a single primer, BOXA1R (Gillings and Holley, 1997; Louws *et al.*, 1994). Amplification mixtures were prepared as described in section 2.3 with the exception that MgCl<sub>2</sub> concentration was increased from 2mM to 4mM. As BOX-PCR uses a single primer, 100pmol of the BOX primer mix was added to amplification mixtures to provide the same total primer concentration used in all other PCRs. All BOX-PCRs were performed in triplicate to ensure reproducibility. The following cycling program was used: 94°C for 5 min for 1 cycle, 94°C for 30 sec, 53°C for 30 sec, 72°C for 8 min for 35 cycles, and 72°C for 5 min for 1 cycle. Amplification products were separated via electrophoresis in 1.5% agarose (Promega) and visualised as described in Section 2.4.

### 3.2.2 – Analysis of BOX-PCR profiles

The number and size of all bands generated using BOX-PCR was determined using the Quantity 1 software package (BioRad). Individual bands were detected using the automatic detect function within Quantity 1 (BioRad) and bands which were not detected were flagged manually. The presence or absence of band(s) from a set of size categories was determined within the software package for each strain. The resulting binary data matrix of strain by size category was exported and used for the construction of a similarity matrix according to the formula:

$$S = 2m_{x,y} / (m_x + m_y)$$

where  $m_{x,y}$  is the number of bands shared by both strain x and strain y, while  $m_x$  and  $m_y$  are the total number of bands present in strain x and strain y, respectively (Nei and Li, 1979). Similarity between the profiles of different strains (and groups of strains) was assessed by comparisons of values in the similarity matrix.

### 3.2.3 – 16S rDNA and 16S-23S IGS PCR

16S rRNA and the 16S-23S intergenic spacer (IGS1) were acquired from Genbank for most strains in the collection (Table 3.1). For those strains for which no sequence information was available, these regions were amplified using PCR. The primers f27 and r1492 (Table 3.2) were used to amplify 16S rRNA genes, and the primers IGS-F and IGS-R (Table 3.2) were used to amplify the IGS1 region. Amplification mixtures were prepared as described in section 2.3. The 16S rRNA PCR was performed using the following cycling program: 94°C for 5 min for 1 cycle, 94°C for 30 sec, 60°C for 30 sec, 72°C for 1 min 30 sec for 35 cycles, and 72°C for 5 min for 1 cycle. The IGS1 PCR was performed using identical cycling conditions, with the exception being that a 55°C annealing temperature and a 45 sec extension time was used. The nucleotide sequence of all 16S rRNA amplicons was determined by direct sequencing (as described in section 2.6) using the f27 and r1492 primers. Most IGS1 amplicons were directly sequenced using the primer IGS-F. The remaining IGS1 amplicons were cloned and sequenced, as described in section 2.6.

### 3.2.4 – RFLP analysis of 16S rDNA and IGS1 sequences

RFLP profiles of 16S rDNA PCR amplicons were generated for all strains in the collection, while RFLP analysis of IGS1 sequences was performed on *Ps. stutzeri* and *Ps. balearica* strains only. To generate 16S rDNA RFLP profiles, approximately 300ng

of PCR product was digested using 5U *Taq* I (New England Biolabs) in the buffer supplied by the manufacturer, and incubated for 3 hours at 65°C. IGS1 RFLP profiles were generated by digesting approximately 400ng of PCR product using the same enzyme and the same reaction conditions. RFLP profiles were resolved using electrophoresis in 1.5% agarose for 16S rDNA digests, and 2.5% agarose for IGS1 digests. Electrophoresis and visualisation of fractionated DNA was performed as described in Section 2.4. The likely identity of strains within the collection was inferred on the basis of comparisons between the RFLP profiles observed and those generated *in silico* from sequences available in public databases.

### **3.2.5 – Sequence Analysis**

Alignment of 16S rRNA sequences was performed using the alignment tool available on the Ribosomal Database Project II website (<http://rdp.cme.msu.edu/>). 16S-23S IGS sequences were aligned using the ClustalW software package (Thompson *et al.*, 1994) and the alignment optimised in the GeneDoc software package (Nicholas *et al.*, 1997). Phylogenetic trees were constructed from both alignments using the SEQBOOT, DNADIST, NEIGHBOUR and CONSENSE programs within the Phylip phylogenetic software package (Felsenstein, 1989). All bootstrap analyses were run for 1000 iterations. Print-quality trees were produced using the Adobe Illustrator CS2 software package.

Species	Strain	Abr.	16S rRNA Acc.	IGS1 Acc.
<i>Ps. stutzeri</i>	ATCC 17684	1A	U58660	Present Study
	ATCC 17589	1B	U25432	Present Study
	ATCC 17593	1C	AJ006104	Present Study
	ATCC 17595	2A	AJ006105	Present Study
	ATCC 17587	2B	U25431	AJ251902
	ATCC 14405	2C	U65012	AJ390590
	19smn4	4A	U22426	AJ251906
	DNSP21	5A	U26414	AJ251908
	ATCC 17685	7A	AJ006103	Present Study
	DSM 50238	7B	U26416	AJ251909
	API-2-142	7C	AJ410871	Present Study
	YPF-41	7D	AJ410872	Present Study
	RNAlll	7E	Present Study	Present Study
	Q	8A	Present Study ^	Present Study
	BAM17	8B	Present Study ^	Present Study
	P1	8C	Present Study ^	Present Study
	ATCC 17641	8D	AJ006106	Present Study
	JM 300	8E	X98607	AJ390581
<i>Ps. balearica</i>	SP1402	6A*	U26418	AJ279238
	ATCC 17682	6B*	AJ006107	Present study
<i>Ps. straminea</i>	KM91	Pstr	Present Study	Not Determined

Table 3.1 - Accession numbers for 16S rDNA and IGS1 sequences for strains analysed in the present study. All sequences determined in the present study are indicated.

^ - 16S rRNA gene sequences for these strains were generated by Holmes *et al.* (2003), but were not deposited in Genbank. The sequences were generated again in the present study.

\* - *Ps. balearica* was formerly classified as *Ps. stutzeri* Gv.6.

PCR assay	Primer name	Sequence (5' - 3')	Reference
BOX-PCR	BOXA1R	CTACGGCAAGGCGACGCTGACG	Louws <i>et al.</i> , (1994)
16S rDNA PCR	f27 r1492	AGAGTTTGATCMTGGCTCAG TACGGYTACCTTGTACGACTT	Weisburg <i>et al.</i> , (1991)
16S-23S IGS PCR	IGS_F IGS_R	TGCGGCTGGATCCCCCTCCTT CCGGGTTTCCCCATTTCGG	Normand <i>et al.</i> (1996)

Table 3.2 - Primers used for BOX-, 16S rDNA- and IGS1 PCRs. The reference column refers to the study from which each primer set was sourced. All primers listed are 'universal', being able to successfully generate amplicons from diverse bacteria. See sections 3.2.1 for PCR cycling conditions used for BOX-PCR and section 3.2.3 for 16S rDNA- and IGS1 PCRs.





### 3.3 – Results

#### 3.3.1 – BOX-PCR

The BOX-PCR uses primers which target the BOX element of *Streptococcus pneumoniae* (Martin *et al.*, 1992). While the BOX element is only known to occur in this species, the BOXA1R primer, when used under typical RAPD-PCR conditions produces characteristic, reproducible DNA fingerprints using a diverse range of DNA templates (Gillings and Holley, 1997). These fingerprints typically contain 5-20 bands of different sizes, and the differences between fingerprints can be used to differentiate between DNA samples of different origins. This technique is typically sensitive to small genomic changes and is thus well suited to differentiating between closely related bacterial strains (Lanoot *et al.*, 2004; Masco *et al.*, 2003; Sikorski *et al.*, 1999).

Triplicate DNA fingerprints were generated for all *Pseudomonas* spp. strains in the collection. Identical banding patterns were observed in all replicate PCRs within the parameters for band detection and classification used (data not shown). Unique fingerprints were observed for all strains tested, as shown by the representative BOX-PCR gel and similarity matrix in Figure 3.2. A total of 25 band positions were observed. With the exception of *Ps. stutzeri* 19SMN4 (4A) and *Ps. balearica* SP1402 (6A), distinct (<75% similarity to all other profiles) BOX-PCR profiles were observed for all strains screened. This confirmed the independence of these strains with respect to other strains in the collection. The BOX-PCR profiles of *Ps. stutzeri* 19SMN4 (4A) and *Ps. balearica* SP1402 (6A) shared 23 of 25 bands, with a calculated pair-wise identity of 92% (Figure 3.2). This result was unexpected as these strains were labelled as belonging to different species and suggested that one or both of these strains had been either mislabelled or cross-contaminated. The difference observed between

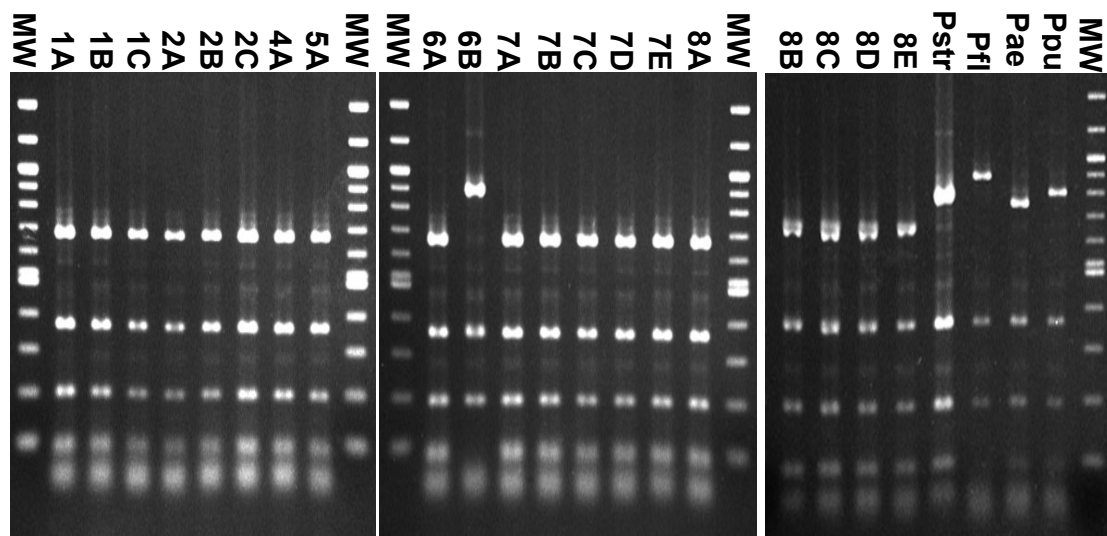
these two profiles was limited to the presence or absence of minor (faint) bands only, which is consistent with a mixed culture or variation in the purity of the genomic DNA used to generate the profiles. In order to further assess the provenance of each of these strains, RFLP and sequence analysis of the ITS1 region was employed, the details of which are given in Sections 3.3.2 and 3.3.3.

Correlations between BOX-PCR similarity scores and species and genomovar groupings were generally weak. The average level of similarity between members of the *Ps. stutzeri* species complex was  $34\% \pm 13.5\%$ . Relative to this average, members of genomovars 2 and 8 exhibited consistently higher levels of similarity of  $63\% \pm 7.2\%$  and  $55.3 \pm 13.1\%$ , respectively. Members of genomovars 1 and 7 in contrast, exhibited average similarities closer to the total similarity to all *Ps. stutzeri* strains of  $33.7\% \pm 6\%$  and  $35.6\% \pm 12\%$ , respectively. Other *Pseudomonas* spp. were not well defined with respect to members of the *Ps. stutzeri* species complex, exhibiting comparable levels of similarity to that observed for the *Ps. stutzeri* species complex as a whole (*Ps. straminea* KM91 [ $33\% \pm 12.5\%$ ], *Ps. fluorescens* NCTC 7244 [ $34\% \pm 11.8\%$ ], *Ps. putida* F1 [ $30\% \pm 8.4\%$ ], and *Ps. aeruginosa* 29.6%  $\pm$  8.9%).



### 3.3.2 – Analysis of 16S rRNA gene RFLP profiles

Partial 16S rRNA genes were PCR-amplified from all strains. In all cases, the amplicon was approximately 1450 bp in size. By performing *in silico* RFLP analyses on *Pseudomonas* spp. 16S rRNA gene sequences using several different restriction endonucleases, *Taq* I was found to be the most informative enzyme available, and was able to differentiate all strains in the collection at the species level of classification (Figure 3.3). 16S rDNA *Taq* I RFLP profiles of all *Ps. stutzeri* strains in the collection corresponded to the fragment sizes expected from *Ps. stutzeri* 16S rDNA sequences obtained from Genbank (Figure 3.3). RFLP profiles for *Ps. balearica* strains should have been the same as those of *Ps. stutzeri*; while this was the case for *Ps. balearica* SP1402 (6A), the profile observed for *Ps. balearica* ATCC 17682 (6B) did not correspond to the expected profile, and instead had a profile characteristic of *Ps. mendocina* or other *Pseudomonad* (Figure 3.3). To determine the correct identity of *Ps. balearica* ATCC 17682 (6B), the nucleotide sequence of the 16S rRNA gene was determined and subjected to phylogenetic analysis (see section 3.3.5). The RFLP profiles observed for all remaining strains were in accordance with that expected from database sequences (Figure 3.3).



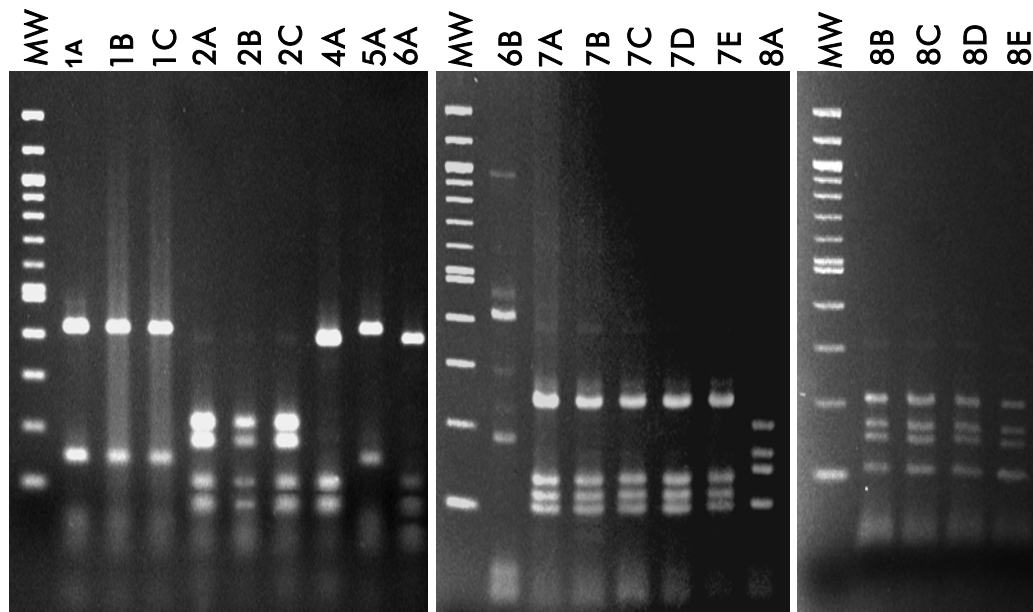
Species	Expected 16S rDNA <i>Taq</i> I restriction fragments (bp)				
<i>Ps. stutzeri</i>	<b>40-55</b>	<b>85-90</b>	<b>195</b>	<b>360</b>	<b>660</b>
<i>Ps. balearica</i>	<b>35-50</b>	<b>85-90</b>	<b>195</b>	<b>360</b>	<b>660</b>
<i>Ps. straminea</i>	<b>55</b>	<b>85</b>	<b>200</b>	<b>360</b>	<b>800</b>
<i>Ps. fluorescens</i>	<b>55</b>	<b>200</b>	<b>360</b>	<b>890</b>	
<i>Ps. aeruginosa</i>	<b>35-50</b>	<b>90</b>	<b>200</b>	<b>360</b>	<b>755</b>
<i>Ps. putida</i>	<b>35-50</b>	<b>85</b>	<b>200</b>	<b>360</b>	<b>805</b>
<i>Ps. mendocina</i>	<b>55</b>	<b>195</b>	<b>360</b>	<b>895</b>	

Figure 3.3 - *Taq* I RFLP profiles of 16 rDNA PCR amplicons from all *Pseudomonas* spp. strains used in the present study. Codes used to identify strains correspond to those given Table 3.1. All strains of *Ps. stutzeri* possess alphanumeric codes, where the numbers refer to the genomovar to which it belongs. Expected band sizes were calculated *in silico* from 16S rDNA sequences obtained from Genbank. Band sizes in bold red type indicate expected band sizes which were observed in the RFLP profiles above.

### 3.3.3 – Analysis of IGS1 RFLP profiles

IGS1 sequences were successfully amplified from all *Ps. stutzeri* and *Ps. balearica* strains in the collection. For all strains, IGS1 amplicons were approximately 600 bp in length. Figure 3.4 shows the *Taq*I RFLP profiles for 16S-23S ITS PCR amplicons for all strains. All bands greater than 50 bp in size were resolved for all strains screened. With the exception of bands < 50 bp in size, all *Ps. stutzeri* strains in the collection gave the expected RFLP profile (as indicated by the bold red type in the table accompanying Figure 3.4). All genomovars with the exceptions of Gv's 1 and 5 gave unique *Taq*I IGS1 RFLP profiles and all profiles were conserved within members of the same genomovar. This observation is consistent with results obtained previously using this technique (Guasp *et al.*, 2000). The IGS1 RFLP profiles for neither of the strains supplied as *Ps. balearica* (SP1402 [6A] nor ATCC 17682 [6B]) corresponded to the profile generated from *Ps. balearica* IGS1 sequences from Genbank. The *Taq*I IGS1 RFLP profile for *Ps. balearica* SP1402 (6A) was observed to be identical to that of *Ps. stutzeri* 19SMN4 (4A) and both corresponded to the profile expected for *Ps. stutzeri* Gv.4. Based on this observation it was concluded that *Ps. balearica* SP1402 (6A) had most likely been mislabelled or cross-contaminated, and was instead identical to, or very closely related to, *Ps. stutzeri* 19SMN4 (4A). This result was confirmed by sequence determination of the IGS1 region (see section 3.3.6). The *Taq*I IGS1 RFLP profile for *Ps. balearica* ATCC 17682 (6B) did not correspond to that expected for any member of the *Ps. stutzeri* species complex (including *Ps. balearica*) available in Genbank. This observation is not surprising considering the 16S rRNA RFLP profile of this strain was also different with respect to that expected for both of

these species. The nucleotide sequence of the 16S rRNA gene and IGS1 region for this strain was determined to ascertain its likely identity (see section 3.3.6).



Genomovar	Expected IGS1 TaqI restriction fragments (bp)			
Gv.1	<b>68</b>	<b>120</b>	<b>396</b>	
Gv.2	<b>50-75</b>	<b>106</b>	<b>170</b>	<b>200</b>
Gv.4	<b>74</b>	<b>107</b>	<b>371</b>	
Gv.5	<b>68</b>	<b>120</b>	<b>396</b>	
Gv.6*	105	121	260	
Gv.7	<b>90</b>	<b>105</b>	<b>135</b>	<b>240</b>
Gv.8	<b>107</b>	<b>125</b>	<b>164</b>	<b>198</b>

Figure 3.4 - *Taq*I IGS1 RFLP profiles for all *Ps. stutzeri* strains used in the present study. Codes used to identify strains correspond to those given Table 3.1. The number allocated to each strain indicates the genomovar to which it belongs. Expected band sizes were calculated *in silico* from IGS1 sequences obtained from Genbank. Band sizes in bold red type indicate bands which were observed in the RFLP profiles observed in the present study. \* - Gv.6 = *Ps. balearica*.



### **3.3.4 – Confirmation of strain independence and provenance**

Analysis of BOX-PCR profiles and 16S rDNA and IGS1 RFLP profiles allowed the independence and identity of all strains in the collection to be inferred and a summary of all PCR fingerprinting and RFLP results is provided in Table 3.3. The identity of a given strain was assumed to be correct if all observed restriction fragments were of the correct size based on the sequences accessed from Genbank. Under these criteria the identity of all but two strains in the collection was inferred to be correct (Table 3.3). The identity of strains at the species level of classification was inferred from 16S rDNA RFLP profiles, while *Ps. stutzeri* genomovar classifications were inferred from IGS1 RFLP profiles.

Species	Strain	Abr.	Correct 16S rDNA RFLP		IGS1 RFLP profile		BOX- PCR profile	Gv.	Strain proven- ance
			Obs.	Pred.	Obs.	Pred.			
<i>Ps. stutzeri</i>	ATCC 17684	1A	1	1	1	1	1	1	✓
	ATCC 17589	1B	1	1	1	1	2	1	✓
	ATCC 17593	1C	1	1	1	1	3	1	✓
	ATCC 17595	2A	1	1	2	2	4	2	✓
	ATCC 17587	2B	1	1	2	2	5	2	✓
	ATCC 14405	2C	1	1	2	2	6	2	✓
	19smn4	4A	1	1	3	3	7a	4	✓
	DNSP21	5A	1	1	4	4	8	5	✓
	ATCC 17685	7A	1	1	5	5	9	7	✓
	DSM 50238	7B	1	1	5	5	10	7	✓
	API-2-142	7C	1	1	5	5	11	7	✓
	YPF-41	7D	1	1	5	5	12	7	✓
	RNAIII	7E	1	1	5	5	13	7	✓
	Q	8A	1	1	6	6	14	(8)	✓
	BAM17	8B	1	1	6	6	15	(8)	✓
	P1	8C	1	1	6	6	16	(8)	✓
	ATCC 17641	8D	1	1	6	6	17	8	✓
	JM 300	8E	1	1	6	6	18	8	✓
	<i>Ps. balearica</i> SP1402	6A	1	1	3	7	7b	6*	✗
	ATCC 17682	6B	6	1	8	7	20	6*	✗
<i>Ps. straminea</i>	KM91	Pstr	2	2	N/A	N/A	21	N/A	✓
Control strains									
<i>Ps. fluorescens</i>	NCTC 7244		3	3	N/A	N/A	22	N/A	✓
<i>Ps. aeruginosa</i>	NCTC 3756		4	4	N/A	N/A	23	N/A	✓
<i>Ps. putida</i>	F1		5	5	N/A	N/A	24	N/A	✓

Table 3.3 - Summary of RFLP and BOX-PCR results to confirm strain independence and infer identity. The different numbers listed under each method (16S RFLP, IGS1 RFLP and BOX-PCR) refer to a unique profile, ie. strains with the same number gave identical profiles. Predicted and observed profiles are given 16S rDNA and IGS1 RFLP profiles. Strain provenance was assessed by comparison of observed and predicted RFLP profiles. The BOX-PCR profiles of strains 4A and 6A are labelled as 7a and 7b, respectively, as these profiles were near identical, differing in the presence of minor bands only. *Ps. balearica* SP1402 and ATCC17682 were the only strains found not to match their labelled identity. Genomovar numbers in parentheses indicate genomovar designations which are inferred from surrogate methods as opposed to DNA:DNA hybridisation data.

\* - *Ps. balearica* was formerly classified as *Ps. stutzeri* Gv.6.

### 3.3.5 – 16S rRNA gene phylogeny

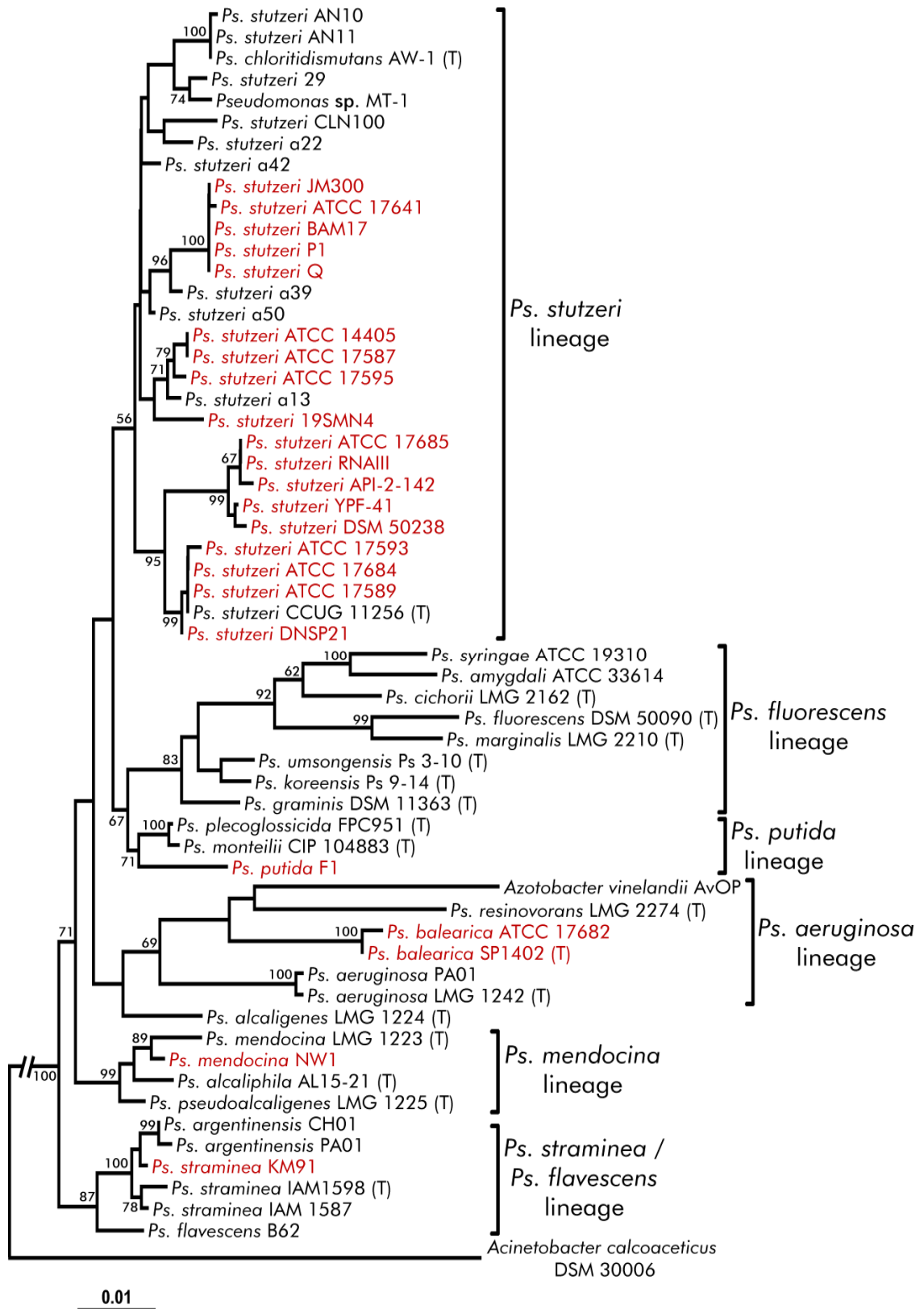
16S rRNA genes were amplified and directly sequenced for two strains in the collection (*Ps. straminea* KM91 [Pstr] and *Ps. balearica* ATCC 17682 [6B]), while 16S rRNA gene sequences for all other strains were obtained from Genbank (Table 3.1). The classification of strain KM91 (Pstr) as a strain of *Ps. straminea* was affirmed; the 16S rDNA sequence of this strain showed 99.3% similarity (1469/1484 pair-wise sequence identity) to *Ps. straminea* IAM 1587. This is consistent with observations made previously on this strain (McNichol, 2002). On the basis of 16S rRNA similarity to known strains, the 16S rRNA sequence of the strain labelled *Ps. balearica* ATCC17682 (6B) was found to not match the database sequence for this strain (94.1% [1168/1241] pairwise identity). Rather, this strain was putatively identified as a strain of *Ps. mendocina*, exhibiting a pair-wise similarity of 99.6% (1231/1236) with *Ps. mendocina* ATCC 25411. A partial IGS1 nucleotide sequence was also determined (data not shown) for this strain and was also found to match most closely to *Ps. mendocina* ATCC 25411, exhibiting 100% (141/141) pair-wise nucleotide identity. Since this strain was clearly not *Ps. balearica* ATCC 17682 its provenance is uncertain, and it was concluded that this strain was best classified within *Ps. mendocina* and will henceforth be referred to as *Ps. mendocina* NW1, with the abbreviation 'Pme'.

A reconstruction of the phylogeny of strains in the collection relative to other *Pseudomonas* spp. is provided in Figure 3.5. All members of the *Pseudomonas* genus included in the tree form a coherent group, separate from the closest non-*Pseudomonas* relative. The diversity of strains included in the tree is comparable to that apparent in other studies focussed on analysis of Pseudomonad diversity (Moore

*et al.*, 1996). At least six major clades are apparent in the tree in Figure 3.5, and all but two of these are represented by strains in the collection. The strains in the collection may therefore be considered to represent a broad spectrum of diversity within the *Pseudomonas sensu stricto*. The *Ps. flavescens* / *Ps. straminea* lineage occupies the deepest branch within the tree and all remaining *Pseudomonas* spp. strains in the tree occurred in a clade which received relatively strong support (71%) (Figure 3.5). *Ps. straminea* KM91 (Pstr) did not group with the other *Ps. straminea* strains included in the tree, but rather clustered with strains occurring in the nearest side branch to these strains. However, this clade did not receive high bootstrap support (39%) and *Ps. straminea* KM91 (Pstr) was found to cluster with other members of this species in other possible tree topologies. The *Ps. mendocina* lineage comprised a clade that received 99% bootstrap support and included *Ps. mendocina*, *Ps. pseudoalcaligenes* and *Ps. alcaliphila* strains. Within this lineage, *Ps. mendocina* NW1 (Pme) occurred in a well supported clade (89% bootstrap support) with *Ps. mendocina* LMG 1223, which was consistent with the inferred provenance of this strain. While all *Ps. stutzeri* strains do form a coherent group on the basis of 16S rRNA gene variation with respect to other *Pseudomonas* spp. ('*Ps. stutzeri*' lineage' in Figure 3.5), strong bootstrap support was not obtained (56%) for the monophyletic clustering of *Ps. stutzeri* strains included in the phylogeny reconstructed here (Figure 3.5). In contrast, strong bootstrap support for the monophyly of several *Ps. stutzeri* genomovars was observed. Of those genomovars represented in the collection used in the present study, members of Gv.7 and Gv.8 received strong bootstrap support as monophyletic clades (Figure 3.5).

Figure 3.5 (opposite) – Phylogenetic relationships between strains analysed in the present study and a representative set of *Pseudomonas* spp. The tree is based on maximum-likelihood and neighbour-joining analysis using nearly complete (approximately 1300bp) 16S rRNA gene sequences. Bootstrap values are given for all bifurcations which received greater than 60% support. Strains highlighted in red indicate strains included in the present study. Blue stars indicate strains from which 16S rRNA genes were sequenced in the present study. Orange diamonds indicate species (*Ps. fluorescens*, *Ps. aeruginosa* and *Ps. putida*) which are represented in the strain collection, but are not specifically represented in the tree. The scale bar given represents 1% estimated sequence divergence.

*Ps.*  
mendocina  
clade



### 3.3.6 – 16S-23S IGS phylogeny

IGS1 sequences were successfully amplified and sequenced from 11 strains, and sequences for all remaining strains were obtained from Genbank (see Table 3.1). About half of all IGS1 amplicons sequenced required cloning before a clean sequence could be obtained, suggesting that these strains contain multiple IGS1 regions which do not share the same sequence. Strains for which IGS1 sequences were cloned included all members of Gv.7 (five strains in total). For each cloned IGS1 PCR product, the *Taq*I RFLP profile of five random clones was analysed (data not shown). Indistinguishable *Taq*I RFLP profiles were observed for each cloned product. On the basis of this data, it was assumed that the variation between different IGS1 copies within these strains was the result of small differences in nucleotide sequence, such as the insertion or deletion of one or a few bases. Two independent clones were sequenced for each cloned IGS1 PCR product, giving a total of 10 independent clones from five strains. In all cases the sequence of both clones from each strain was identical. All remaining IGS1 PCR amplicons were directly sequenced. The IGS1 sequence of *Ps. balearica* SP1402 (6A) was identical to that of *Ps. stutzeri* 19SMN4 (4A) (Figure 3.6), confirming previous suspicions that the strain supplied as *Ps. balearica* SP1402 (6A) was most likely a member of *Ps. stutzeri* Gv.4, and consequently this strain was excluded from all further analyses.

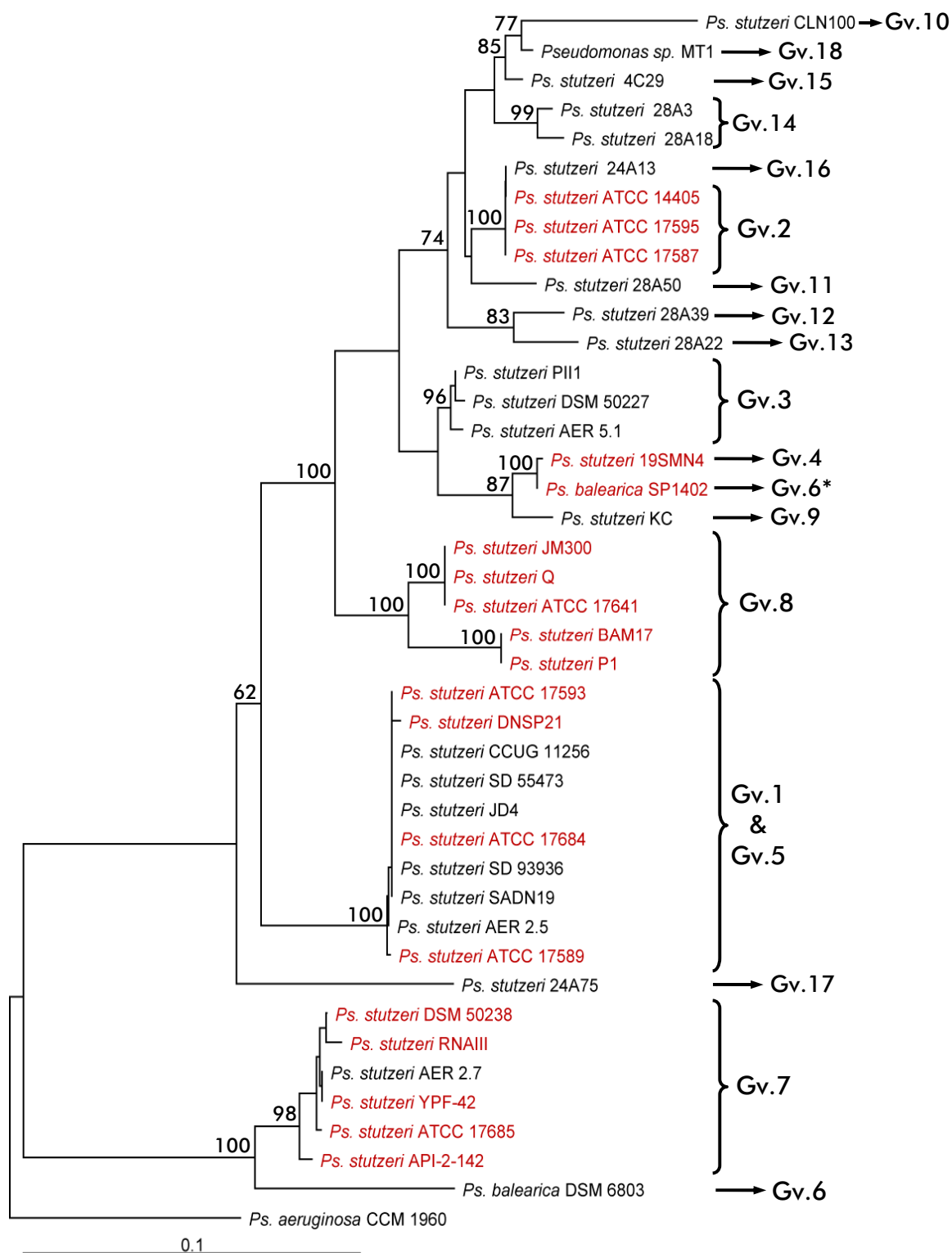
Figure 3.6 shows a phylogenetic tree constructed from an alignment of IGS1 sequences for all *Ps. stutzeri* strains in the collection in addition to a set of reference strains. With the exception of Gv's 1 and 5, all *Ps. stutzeri* genomovars represented in the collection (highlighted in red in Figure 3.6) formed well resolved clades which received >85% bootstrap support. While most genomovars were unambiguously

differentiated from their closest relatives, the deeper branches in the tree, which separate groups of genomovars, mostly receive poor bootstrap support. Interestingly, the amount of divergence observed among 16S rRNA genes in *Ps. stutzeri* Gv.7 and Gv.8 strains was not consistent with the pattern of diversity observed among IGS1 sequences. IGS1 sequences were more divergent than 16S rRNA genes in *Ps. stutzeri* Gv.8 strains (Figures 3.5 and 3.6), which is not surprising as IGS1 regions are predicted to experience less selective pressure for sequence conservation than 16S rRNA genes. In contrast, the opposite is true when examining these sequences in *Ps. stutzeri* Gv.7 strains. These strains exhibit significant 16S rRNA gene sequence divergence (Figure 3.5) relative to other *Ps. stutzeri* strains, and are relatively less diverse across IGS1 sequences (Figure 3.6). All IGS1 sequences from *Ps. stutzeri* Gv.7 strains generated here originated from cloned PCR products, rather than from direct sequencing of a PCR product. A possible explanation for the low level of divergence observed among IGS1 sequences in *Ps. stutzeri* Gv.7 strains is that variable copies of IGS1 exist in different rRNA operons, and the sequence of one of these variants for each strain is shown in Figure 3.6. The strains included in Figure 3.6 are representative of the total diversity of the *Ps. stutzeri* species complex; all 18 genomovars currently and formerly recognised are represented. All deep branching-groups within the tree contain members of the collection used here, despite the fact that only six *Ps. stutzeri* genomovars are represented. Thus, the *Ps. stutzeri* strains analysed in the present study may be considered representative of the total diversity apparent in the *Ps. stutzeri* species complex.



Figure 3.6 (opposite) - Phylogenetic reconstruction based on IGS1 sequences of *Ps. stutzeri*, *Ps. balearica* and *Ps. aeruginosa*. The tree is based on maximum-likelihood and neighbour-joining analysis using near complete IGS1 sequences; two small (5-30 bp) regions which could not be unambiguously aligned were excluded. Strains highlighted in red indicate those strains included in the present study. Percentage bootstrap values are given for all bifurcations which received >60% from 1000 iterations. The scale bar given represents 10% estimated sequence divergence.

\* - While this strain was labeled as *Ps. balearica* SP1402, in this tree it grouped with *Ps. stutzeri* 19SMN4 (Gv.4). The expected location for *Ps. balearica* strains is represented above by *Ps. balearica* DSM 6803.



### 3.4 – Conclusions

#### **All but two of the cultures in the collection showed typing data consistent with the provided identification**

BOX-PCR fingerprints indicated that all cultures analysed represented independent strains. Of the cultures that did not match the provided identity, one (*Ps. mendocina* NW1) was concluded to be an independent strain of *Ps. mendocina* and the other (*Ps. balearica* SP1402) was concluded to be an impure culture which appeared to be cross-contaminated with a strain related to *Ps. stutzeri* Gv.4.

#### ***Pseudomonas* spp. strains in the collection represent a spectrum of diversity within the genus *Pseudomonas* (sensu stricto)**

The strain collection used in the present study consisted of five different *Pseudomonas* species. Phylogenetic diversity of 16S rRNA gene sequences from these strains relative to a representative set of *Pseudomonas* spp. revealed deep branching evolutionary relationships. The divergence apparent between these strains is comparable to that observed in more extensive examinations of Pseudomonad diversity (Moore *et al.*, 1996); however, not all major Pseudomonad lineages were represented.

#### ***Ps. stutzeri* strains in the collection representative of total *Ps. stutzeri* diversity**

Of the 17 currently recognised *Ps. stutzeri* genomovars, six were represented in the present study. The *Ps. stutzeri* strains in the collection were scattered throughout the IGS1 tree and were present in most major clades. Thus, even with the recent expansion of *Ps. stutzeri* complex from 9 to 17 genomovars (Sikorski *et al.*, 2005), the

strains used in the present study may be considered representative of the breadth of diversity apparent in the entire *Ps. stutzeri* species complex.

**The strain collection represents a spectrum of evolutionary relationships suitable for investigating the impacts of integrons throughout a bacterial lineage**

The *Pseudomonas* spp. strains in the collection represent a spectrum of evolutionary divergence. Representation of several different *Pseudomonas* spp. will allow the long-term evolutionary stability of CIs in *Pseudomonas* to be assessed, as different *Pseudomonas* spp. are obviously related, but have genomes which diverged millions of years ago and have undergone significant rearrangement, mutation and gene loss and acquisition (Ginard *et al.*, 1997; Heuer *et al.*, 1998; Rainey *et al.*, 1994; Romling *et al.*, 1997; Sawada *et al.*, 2002). Members of *Ps. stutzeri* Gv's 1, 2, 7 and 8 each represent a group of strains which are very closely related, with members of each genomovar having identical or near identical 16S rRNA genes and IGS1 sequences. Each of these genomovars diverged relatively recently from their respective common ancestors and have undergone relatively little genomic change since, and are thus suited to investigating integron activity over short evolutionary timescales. Different *Ps. stutzeri* genomovars represent incipient species which have not diverged enough, at least in terms of phenotype, to be formally classified as different species. Collectively, the *Ps. stutzeri* species complex represents a level of divergence which is intermediate to that exhibited by different *Pseudomonas* spp. and individual *Ps. stutzeri* genomovars. More importantly, the *Ps. stutzeri* species complex represents several groups of bacteria which may be considered to be on the verge of

speciation, and allows the potential impact of integrons in speciation events to be assessed. The diversity of strains to be screened is therefore suitable for assessing evolutionary impacts of integrons which occur over different timescales (see Figure 3.1).

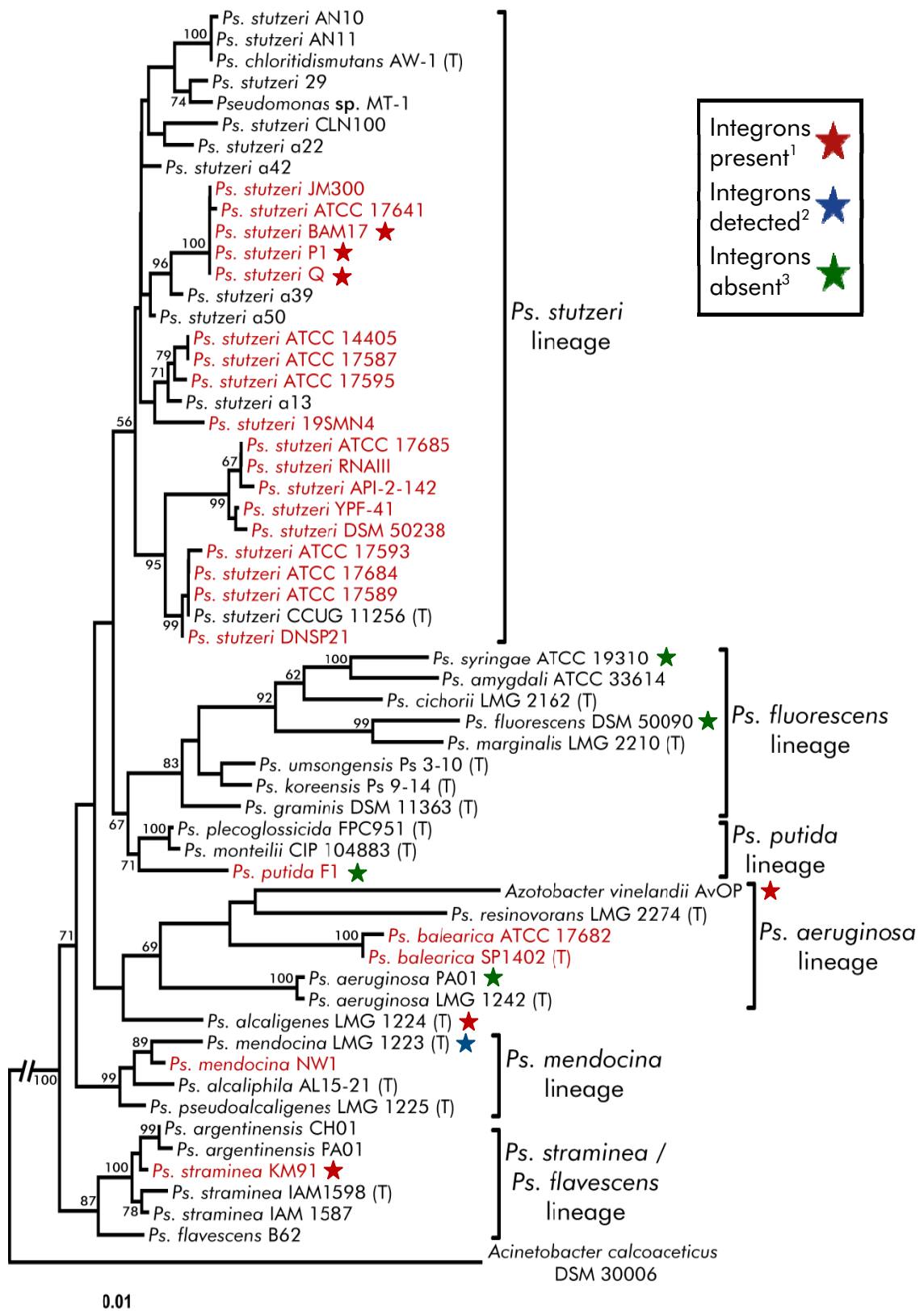
## CHAPTER 4 – DISTRIBUTION OF INTEGRONS AND GENE CASSETTES IN *PSEUDOMONAS*

### 4.1 - Introduction

Chromosomal integrons (CIs) are predicted to exhibit patterns of stable vertical inheritance in bacterial lineages and to exhibit properties, such as large cassette arrays with similar, lineage specific 59-be, indicative of a long residence time within a particular genome (the CI paradigm – see Section 1.7). In 2001, Vaisvila *et al.* described an integron and associated cassette array in *Ps. alcaligenes* ATCC 55044. A number of features of this integron were similar to those of CIs according to the model of Rowe-Magnus *et al.*, (2001). The authors used PCR targeting PAR elements (59-be of *Ps. alcaligenes*) to survey other Pseudomonads for the presence of related integrons, and detected their presence in strains of *Ps. stutzeri* and *Ps. mendocina*. On the basis of this limited data which was consistent with the CI paradigm, Vaisvila *et al.*, (2001) proposed that a CI family similar to that found in *Vibrio* spp. was present in *Pseudomonas*. Integrons exhibiting characteristics similar to that of the *Ps. alcaligenes* ATCC 55044 CI (InPal55044) have subsequently been detected and sequenced from strains of *Ps. stutzeri* (Holmes *et al.*, 2003b) and *Ps. straminea* KM91 (McNichol, 2002). *Pseudomonas* strains in which integrons have been either detected or sequenced are diverse (Figure 4.1) and *intI* gene phylogenies suggest that these integrons are lineage-specific (Holmes *et al.*, 2003b). Known integrons from *Ps. stutzeri*, *Ps. alcaligenes* and *Ps. straminea* share a conserved *intI* gene (minimum 71% [227/320] amino acid identity), a moderately conserved *attI* region, and 59-be sequences which contain a group-specific conserved sequence motif, and a highly

Figure 4.1 (opposite) – Known distribution of chromosomal integrons throughout the *Pseudomonas* evolutionary radiation. The tree was constructed from an alignment of 16S rDNA sequences spanning approximately 1300 bp, using maximum-likelihood and neighbour-joining analysis. All *Pseudomonas* spp. in which integrons are known to be either present or absent are indicated. Note – not all of the specific strains indicated above have been screened for the presence of integrons and are included here as representatives only.

- <sup>1</sup> 'Integrons present' refers to strains in which integron sequences are known through genome sequencing or specific recovery and sequencing.
- <sup>2</sup> 'Integrons detected' refers to strains in which integrons have been detected by DNA based molecular techniques but not sequenced.
- <sup>3</sup> 'Integrons absent' refers to genome sequences of strains in which integrons have not been detected.





diverse array of cassette encoded genes (Holmes *et al.*, 2003b; McNichol, 2002; Vaisvila *et al.*, 2001).

These observations suggest that an integron family is associated with the *Pseudomonas* evolutionary radiation, consistent with generalised CI concept (section 1.8) and similar to those associated with *Vibrio* and *Xanthomonas* spp. (Gillings *et al.*, 2005; Mazel *et al.*, 1998; Rowe-Magnus *et al.*, 2001; Rowe-Magnus *et al.*, 2002). However, the near absence of CIs in *Pseudomonas* genome sequences suggests that patterns in CI abundance are different in *Pseudomonas* spp.; only 1 of 17 *Pseudomonas* genome sequences (*Ps. mendocina* YMP) contains a close match (>60% pairwise identity) to core integron or 59-be sequences characteristic of known *Pseudomonas* integrons. In comparison, related CIs are found in 20/24 and 6/6 *Vibrio* and *Xanthomonas* genome sequences, respectively. Several *Ps. aeruginosa* strains have been found to contain mobile element-associated class 1 integrons (Aubert *et al.*, 2004; Bissonnette and Roy, 1992; Laraki *et al.*, 1999; Partridge *et al.*, 2001; Poirel *et al.*, 2000; Recchia *et al.*, 1994); however, no strains of this species are known to contain integrons closely related to those found in other *Pseudomonas* strains. Thus, while integrons which are clearly related are found in diverse *Pseudomonas* strains, their distribution within the *Pseudomonas* lineage appears to be patchy.

Integrons were originally functionally defined and were detected on the basis of phenotypes (multiple antibiotic resistance) associated with integron function. We now know that phenotypically cryptic elements are common and share two unambiguously recognisable features in common with integrons encoding known phenotypes, the *intI* gene and 59-be. The presence of these sequences signifies the presence of an

integron, or the evolutionary footprint thereof. Integron functionality is frequently inferred from sequence data alone (in the form of an *intI* gene predicted to encode a functional integrase). As all *IntI* genes predicted to encode functional integrases so far tested have been shown to be active in *in vivo* recombination assays (Biskri *et al.*, 2005; Drouin *et al.*, 2002; Holmes *et al.*, 2003b; Leon and Roy, 2003), and as the vast majority of integrons are known by sequence only, it is assumed that if all of the sequence features that constitute an integron (*intI*, *attI* site, and promoters) are present, then the sequence encodes a functional integron.

Several factors must be considered when adopting a sequence detection approach for the recovery of novel integrons. The level of sequence homology over known integrons is relatively low, and targets for molecular probes are limited to conserved sequence motifs within the core integron and 59-be. The divergence in these sequences is sufficiently great that specific molecular probes have limited target ranges. The use of low-stringency conditions in molecular detection assays is one solution but increases the probability of obtaining false-positive results. This difficulty is compounded by the fact that degenerate core integrons are common (Gillings *et al.*, 2005; Hansson *et al.*, 2002; Holmes *et al.*, 2003b) and detection of them by molecular probes or in databases can be prone to error. Multiple copies of the same core integron (Leon and Roy, 2003), and multiple divergent core integrons (Hochhut *et al.*, 2001); *Shewanella amazonensis* SB2B genome sequence - ZP\_00586789; *Marinobacter aquaeolei* VT8 genome sequence - ZP\_00818060; *Geobacter metallireducens* GS-15 genome sequence - NC\_007517.1 [see Table 1.3]) have been observed within single genomes. Precedents also exist for orphan gene cassettes or cassette arrays which are unlinked to a core integron (Drouin *et al.*, 2002; Mazel, 2006). Finally, while chromosomal integrons often contain cassettes which have 59-

be with conserved sequences (59-be subfamilies), these are not a universal feature, as illustrated by *Shewanella* spp. CIs which exhibit diverse 59-be sequences in their arrays. Thus, to comprehensively investigate the diversity and ubiquity of integrons within a bacterial lineage, a robust screening strategy should be employed which includes multiple molecular techniques and several molecular targets.

The aim of this thesis was to test the hypothesis that an integron lineage has shown vertical inheritance within *Pseudomonas*. The goal of the work presented in this chapter was to assess the distribution of integron-gene cassette system components showing homology to the integrons of *Ps. alcaligenes* ATCC 55044 (InPal55044) and *Ps. stutzeri* Q (InPstQ) within the Pseudomonadaceae. All strains in the collection were screened for the presence of core integron and gene cassette sequences using PCR and southern hybridisation. Several different molecular probes, designed to target a range of divergent integron sequences, were employed to screen for the presence of *Pseudomonas* subfamily integrons in addition to integrons belonging to other subfamilies.

## 4.1 - Materials and methods

### 4.2.1 - PCR screening for integrons

Three different PCR assays were employed to recover integrons from the strain collection, and will be referred to as *intI* PCR, integron PCR and cassette PCR (Figure 4.2). The primer sequences and PCR cycling conditions used for these assays are given in Table 4.1. The primers NW33 and NW2 were used for the *intI* PCR assay, and target the conserved *IntI* residues KTTMIYTH and QNQASLA, respectively, giving an amplicon 700 bp in size. The primers for this PCR assay were designed to target sequences with a maximal nucleotide divergence from *IntI* of approximately 30%.

The integron PCR assay was performed using the primers NW33 and ST1. The primer ST1 targets a conserved sequence motif in the (left) hand region of many *Pseudomonas*-type 59-be. This assay generates amplicons of variable sizes, depending on the size of the first cassette in the array, and may generate multiple amplicons, if more than one cassette is present in the array.

The cassette PCR assay was performed using the primers AJH17 and AJH27. These primers have been described previously (Holmes *et al.*, 2003b) and each target conserved sequence motifs within 59-be typically associated with the integrons of *Ps. alcaligenes* ATCC 55044 and *Ps. stutzeri* Q and BAM17. As both of the primers used in this PCR assay target repeat elements, multiple amplicons consisting of one or more gene cassettes are theoretically possible, as outlined in Figure 4.2.

The number and size of PCR products generated using each assay was determined using agarose gel electrophoresis (section 2.4). All *intI* PCR amplicons were sequenced directly, as described Section 2.6. As integron PCR and cassette PCR uses

one or more primers that target repeat elements and may produce multiple PCR amplicons, amplifications generated using this assay were cloned prior to sequencing (section 2.6).

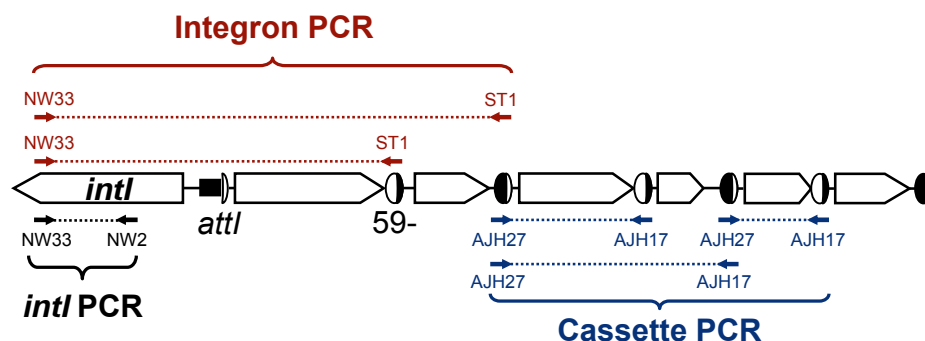


Figure 4.2 - Diagrammatic representation of PCR assays used for detection of integrons. PCR primers are represented by arrows and the dotted lines that join pairs of arrows (primers) represent potential PCR products formed using each assay. Using *intI* PCR, a single product of ~700 bp. Integron PCR and cassette PCR both give multiple products of an unpredictable size, as each uses at least one primer that targets a repeat element.

PCR Assay	Primer Name	Primer Sequence (5' - 3')	Annealing Temp (X)	Extension Time (Y)
<i>intI</i> PCR	NW33 NW2	GGGTRTAKATCWKSGTRGTYTTCAC GCARAACCAGGRYTKTCRGC	52°C	45 sec
Integron PCR	NW33 ST1	GGGTRTAKATCWKSGTRGTYTTCAC CCAGYGARCGARGYGARCG	58°C	2 min
Cassette PCR	AJH17 AJH27	CCCAGYGARCGARGYGAGCG GGCTGAAGCCVGCCCTTARC	65°C	2 min

Table 4.1 - Primer sequences and PCR conditions used for each integron detection assay. Cassette, integron, and *intI* PCRs were performed using the following cycling program: 94°C for 5 min for 1 cycle, 94°C for 30 sec, X°C for 30 sec, 72°C for Y min for 35 cycles, and 72°C for 5 min for 1 cycle.

#### 4.2.2 – Probe design for integron detection

A second approach to detecting integrons was by Southern hybridisation. Probes were chosen to maximize the chance of detecting all *Pseudomonas* IntI relatives that may be present (up to 35% amino acid divergence), and any additional *intI* genes present. All *intI* gene probes were constructed via direct PCR labeling with DIG-dUTP. The primer pair NW33 and AJH42 (Table 4.3) were used for amplification of an *intI* gene probe from *Ps. stutzeri* Q, and a second *intI* gene probe was constructed using *Ps. straminea* KM91 as template DNA and the primers NW33 and NW102 (Table 4.2). The sequence of the *intI* gene probes constructed from each of these strains exhibited 65% nucleotide identity (across the entire core integron) to InPal55044 and 69% nucleotide identity to each other. Using a 200 bp window, these sequences exhibit a maximum similarity of 76%. A third *intI* gene probe was constructed using plasmid R388 (*intI1*) as template DNA and the primer pair NW45 and NW46. Across the entire sequence, this probe exhibits 56% nucleotide identity to intIPstQ, and a maximum of 67% identity across any 200 bp region. All three of these probes span the entire core integron, with the exception of the last 15 amino acids of IntI (Figure 4.3). For all *intI* probes, PCR labeling reactions were prepared as described in Section 2.5, and were run using the following cycling program: 94°C for 5 min for 1 cycle, 94°C for 30 sec, X°C for 30 sec, 72°C for 1 min 15 sec for 35 cycles, and 72°C for 5 min for 1 cycle. Efficiency and yield of the labeling reaction was assessed as described in section 2.5.

Probes targeting 59-be (three in total) were selected to represent the diversity apparent across known *Pseudomonas* 59-be that contain conserved sequence motifs characteristic these integrons (Holmes *et al.*, 2003b; Vaisvila *et al.*, 2001). Other

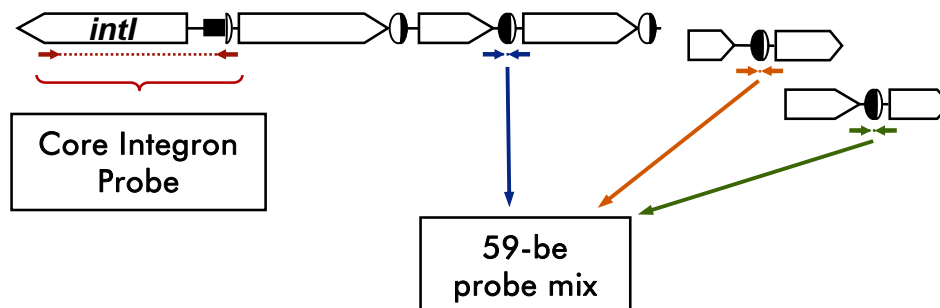


Figure 4.3 - Schematic representation DNA templates used for DIG-labelled probe synthesis. Arrows indicate forward and reverse primers used to synthesise each probe. See Table 4.2 for the names and sequences of all primers used for probe synthesis. All core integron probes spanned identical regions, encompassing all of *attI* and most of *intl*. Three different 59-be were synthesised and combined to form the 59-be probe mix.

Probe Target	Template DNA	Primer Name	Primer Sequence (5' - 3')	Anneal Temp. (X)
<b>Core Integron</b>	<i>Ps. stutzeri</i> Q	NW33 AJH43	GGGTRTAKATCWKSGTRGTYTTCAC CCGGAATTCGGATCCTTGAGGTTGGCGTGCG	60°C
	<i>Ps. straminea</i> KM91	NW33 NW102	GGGTRTAKATCWKSGTRGTYTTCAC CGCGAAATTTCCATGTGATAC	60°C
	R 388	NW45 NW46	TGTAAATCATCGTCGTAGAG ATGGCGACAAGCAAATCGAAC	56°C
<b>59be</b>	<i>Ps. stutzeri</i> DNSP 21	ST8 ST9	TGGCAGAAAACGTGATCTC GAGTTATCTGCGTAAATATC	52°C
	<i>Ps. mendocina</i> NW1	ST10 ST11	AAATGCACTCTAACAAGTCG AGGGTCATGAGCATTACCTC	60°C
	<i>Ps. straminea</i> KM91	ST12 ST13	TTCAGGTTGAGTCGCTTACC GATATTTTTCATAAAGGTTCCCTA	55°C

Table 4.2 - Primer sequences and PCR conditions used for the synthesis of labelled probes for Southern hybridisation. Annealing temperature (X) refers to the conditions used for each assay in the PCR described in section 4.2.2.



59-be sequences were not targeted due to the low level of sequence homology across diverse 59-be. All 59-be probes were also constructed via direct PCR labeling with DIG-dUTP. 59-be probes were synthesized using direct PCR labeling as described in Section 2.5, and were run using the following cycling program: 94°C for 5 min for 1 cycle, 94°C for 30 sec, X°C for 30 sec, 72°C for 1 min for 35 cycles, and 72°C for 5 min for 1 cycle. Primers for 59-be probe synthesis targeted sequences immediately upstream or downstream of the relevant 59-be, and resulted in 20-40 bp of the sequences flanking each element being included in the probe. While this may have increased the probability of false positive results, it was desirable to have probes spanning entire 59-be and due to their inverted repeat structure, 59-be are poor targets for PCR. Labeling efficiency and yield of probes was assessed as described in section 2.5. A final probe mix was prepared by combining approximately equal quantities of each 59-be probe.

#### 4.2.3 – Southern hybridization

*Pst*I and *Pvu*II chromosomal digests were prepared for all strains using 10U of enzyme and ~5µg genomic DNA (Section 2.5). *E. coli* JM109 was included as a negative control, as it is a distant relative of *Pseudomonas* and is known to not contain an integron. Plasmid R388 (*int1*) and *X. campestris* DAR 30538, both of which contain a core integron and gene cassettes, were included in order to test the limits of sensitivity of core integron probes. All hybridisation assays using 59-be probes were performed under 'high stringency' conditions, as described in section 2.5. Due to the relatively high level of core integron sequence variation observed among CIs, and the prevalence of core integron deletion mutants, all hybridisation assays using core integron probes were performed under low stringency conditions,

as defined in Section 2.5. Low stringency conditions were set to allow probe hybridisation to occur with target sequences exhibiting up to 30% sequence divergence. Hybridisation signals were detected using chemiluminescence (Section 2.5).

#### 4.2.4 – Sequence Analysis

All sequence compilations and alignments were performed using CLUSTALW initially and then optimised in GeneDoc (Section 2.7). Identification and localisation of integron-associated sequences was accomplished using various BLAST search techniques (BLASTN, BLASTP and BLASTX), in addition to manual sequence examination. Phylogenetic reconstructions were employed to determine the relationships between all *intI* genes recovered. Trees were constructed using *intI* amino acid alignments with the 'SEQBOOT', 'PROTDIST', and 'NEIGHBOUR' programs within the Phylip software package (Felsenstein, 1989). Print quality trees were produced using Adobe Illustrator CS2. Searches for putative  $P_c$  and  $P_{int}$  sequences were performed using the Neural Network Promoter Prediction program (available at [http://www.fruitfly.org/seq\\_tools/promoter.html](http://www.fruitfly.org/seq_tools/promoter.html)).

#### 4.2.5 – Public Database Searches

To further extend the dataset of integron sequences, database searches on all available *Pseudomonas* spp. genomes were performed using various *Pseudomonas* integron sequences as queries. Identical searches were also performed using the non-redundant Genbank database and the Genbank environmental DNA database. Standard nucleotide-nucleotide (BLASTN) and protein-protein (BLASTP) searches were performed using intIPstQ nucleotide and amino acid sequences, respectively. Four

59-be sequences, representative of the known diversity of 59-be in InPal55044, InPstQ and InPstBAM, were used as query sequences for BLASTN searches. Searches were also extended to the non-redundant Genbank and environmental DNA databases using the same query sequences as used for genome database searches.

## 4.3 – Results

### 4.3.1 – PCR detection of integrons

***intI* PCR** – PCR products of expected size (700 bp) were obtained from six strains (4 genomovar 8, one gv7, and KM91) using the *intI* PCR assay (data not shown). A PCR product smaller than expected was obtained for *Ps. stutzeri* RNAIII (Gv.7), and no PCR products were observed for any of the remaining strains. All *intI* PCR products were directly sequenced and all were found to be *intI* homologues, as determined by BLASTX searches of the generated sequence. The *intI* gene recovered from *Ps. stutzeri* RNAIII (Gv.7) contained a deletion of 77 amino acids, which accounted for the smaller than expected size of the *intI* PCR amplicon from this strain. The sequences recovered here are analysed in detail in section 4.3.4.

**Integron PCR** – Single products were observed for *Ps. stutzeri* Q, *Ps. stutzeri* BAM17 and *Ps. straminea* KM91. These amplicons were assumed to represent the expected product shown in figure 4.2 (near complete core integron and first gene cassette), as the size of all products corresponded to that expected from the known integron sequence of each of these strains (Holmes *et al.*, 2003b; McNichol, 2002). Single products were also observed for *Ps. stutzeri* DSM 50238, *Ps. stutzeri* API-2-142 and *Ps. mendocina* NW1 (data not shown). The nucleotide sequence of each of these PCR products was determined, and BLAST searches and manual sequence annotation confirmed that all six sequences included the expected region of the core integron and the first gene cassette. Integron sequences recovered here are analysed in detail in section 4.3.4.

**Cassette PCR** – Using the cassette PCR assay, signals were obtained from 13 of 23 strains screened, and in all but two of these, multiple bands were observed (Figure 4.4). Most strains which gave products in the *intI* or integron PCR assays also gave cassette PCR products. For all strains in which multiple bands were observed (except for *Ps. stutzeri* Q [8A] and BAM17 [8B], and *Ps. straminea* KM91 [Pstr]), two amplicons of different sizes were sequenced. In selecting clones for sequencing, larger clones were favoured as these were more likely to contain multiple gene cassettes. A total of 12 Cassette PCR amplicons were sequenced and of these, three were found to contain gene cassette arrays. A total of 8 cassettes were recovered, three from *Ps. stutzeri* ATCC 17589 (1B), two from *Ps. stutzeri* ATCC 17685 (7B), and three from *Ps. mendocina* NW1. All cassettes contained ORFs, and all 59-be identified belonged to the *Pseudomonas* 59-be subfamily (as defined by Holmes *et al.*, 2003). The remaining sequences were inferred to be the result non-specific amplification, as no sequences indicative of gene cassettes were identified within them (eg. partial or complete 59-be sequences, one or more complete ORFs). The gene cassette sequences recovered here are analysed in detail in Section 4.3.4.

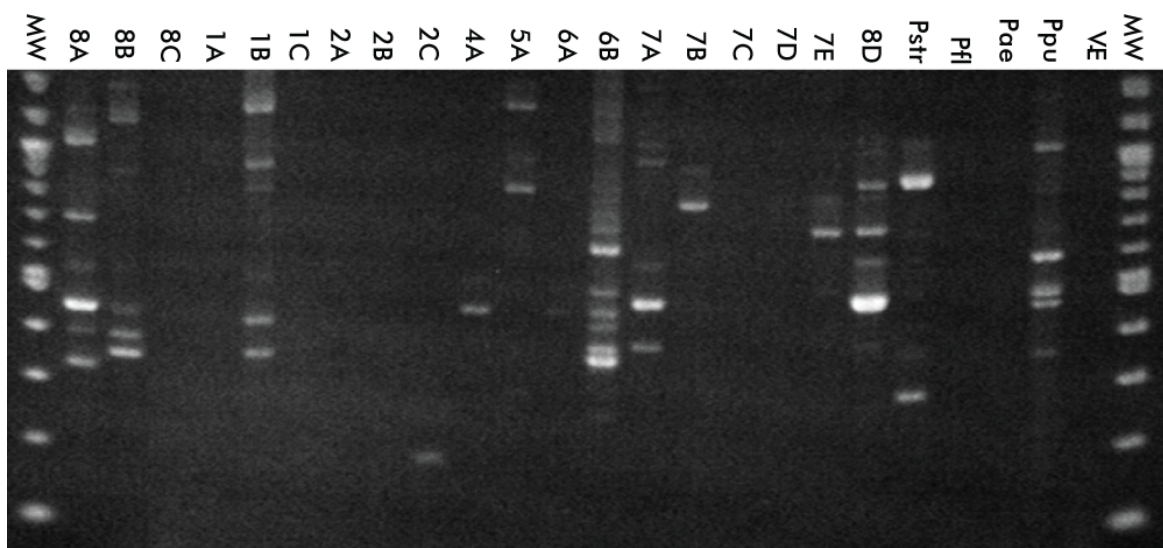


Figure 4.4 - Cassette PCR profiles for all strains in the collection. Relatively intense bands were observed in 13 of 23 strains. Each band in a particular lane is assumed to represent a unique gene cassette (or array of two or more cassettes). Labels used for each lane correspond to the strain codes given in Table 3.1. MW – molecular weight marker – band sizes from top to bottom (bp): 1500, 1200, 1000, 900, 800, 700, 600, 520, 500, 400, 300, 200, 100.

#### 4.3.2 – Southern hybridisation integron detection – Core Integron

Southern hybridisations using a probe targeting the core integron region of *Ps. stutzeri* Q on *Pst*I and *Pvu*II digests are shown in Figure 4.5a and 4.5b, respectively.

Hybridisation signals were observed for all *Ps. stutzeri* strains belonging to Gv's 2, 7 and 8, in addition to *Ps. mendocina* NW1 and *Ps. straminea* KM91. The same hybridisation signals were observed using *Ps. straminea* KM91 core integron probe as the equivalent *Ps. stutzeri* Q probe; however, signal intensity varied, *Ps. straminea* KM91 and *Ps. stutzeri* Gv.7 strains increased in relative intensity, and *Ps. stutzeri* Gv.8 and Gv.2 strains decreased in relative intensity (data not shown). No *intI* / *attI* sequences were detected in members of *Ps. stutzeri* Gv.1, Gv.4 or Gv.5. Neither Plasmid R388 nor *X. campestris* DAR 30538 gave a hybridisation signal when probed with either *intI*PstQ or *intI*PstrKM91 probes, which was consistent with the predicted behaviour of the probes under the conditions used (expected to allow up to 30% sequence divergence). Using *intI1* as a probe, the class 1 integron-containing plasmid R388 was the only sample which gave a positive signal (data not shown). No hybridisation signals were detected using any *intI* gene probe on *Ps. fluorescens* NCTC 7244, *Ps. aeruginosa* NCTC 3756 or *Ps. putida* F1. On the basis of hybridisation patterns observed using *intI*PstQ, *intI*PstrKM91 and *intI1* probes, the detection limits of *intI* gene probes under low stringency conditions was between 67% and 76% nucleotide sequence identity over any 200 bp region. It is therefore reasonable to assume that within the strain collection, all integrons exhibiting >76% nucleotide identity to any of the *intI* gene probes used have been detected.

Hybridisation data was consistent with the presence of variable numbers of *intI* genes (0, 1 or 2 when the whole dataset is considered). For several strains (*Ps. stutzeri* Gv.1

and Gv.4, *Ps. fluorescens* NCTC 7244, *Ps. aeruginosa* NCTC 3756 and *Ps. putida* F1, no hybridisation signals were detected using any *intI* probe. For 7 strains (*Ps. stutzeri* 2C, 7B, 7C, 7E & 8A, *Ps. mendocina* NW1 and *Ps. straminea* KM91), single hybridisation signals were observed in both *Pst*I and *Pvu*II digests in the same position using both *Pseudomonas* core integron probes. This observation strongly supports the existence of a single copy of a *Pseudomonas*-family core integron in each of these strains. All remaining strains in which signals were observed gave two or more core integron *Pvu*II hybridisation signals. Repeat hybridisations on *Pvu*II digests of strains 2A, 7A, 7D, 8B and 8C revealed identical banding patterns, in each case consisting of two bands of relatively strong intensity (data not shown). As single bands were observed in hybridisations on *Pst*I digests of these strains (Figure 4.5a), the presence of two *Pvu*II hybridisation signals in these strains is likely to be due to the presence of a *Pvu*II restriction site in the core integron of these strains. However, the presence of multiple core integrons cannot be ruled out before the core integron nucleotide sequence is determined. No *Pvu*II site is found within the core integron of *Ps. stutzeri* BAM17 (8B) (Holmes *et al.*, 2003). A hybridisation using the core integron probe on a *Pml*I digest of *Ps. stutzeri* BAM17 (8B) revealed a single band (data not shown), which is consistent with available sequence information for this strain. The source of the additional band in the *Pvu*II hybridisation for *Ps. stutzeri* BAM17 (8B) thus remains unknown, but may be due to an error in the original sequence deposition, or the presence of an additional *intI* gene in the genome. The banding pattern observed for strain 8D (*Ps. stutzeri* ATCC 17641) is characteristic of partial DNA digestion, and a repeat hybridisation on this strain confirmed the presence of a single *Pvu*II hybridisation signal, which corresponded to the smallest band (2.2 kb) observed for this strain in Figure 4.5b (data not shown).





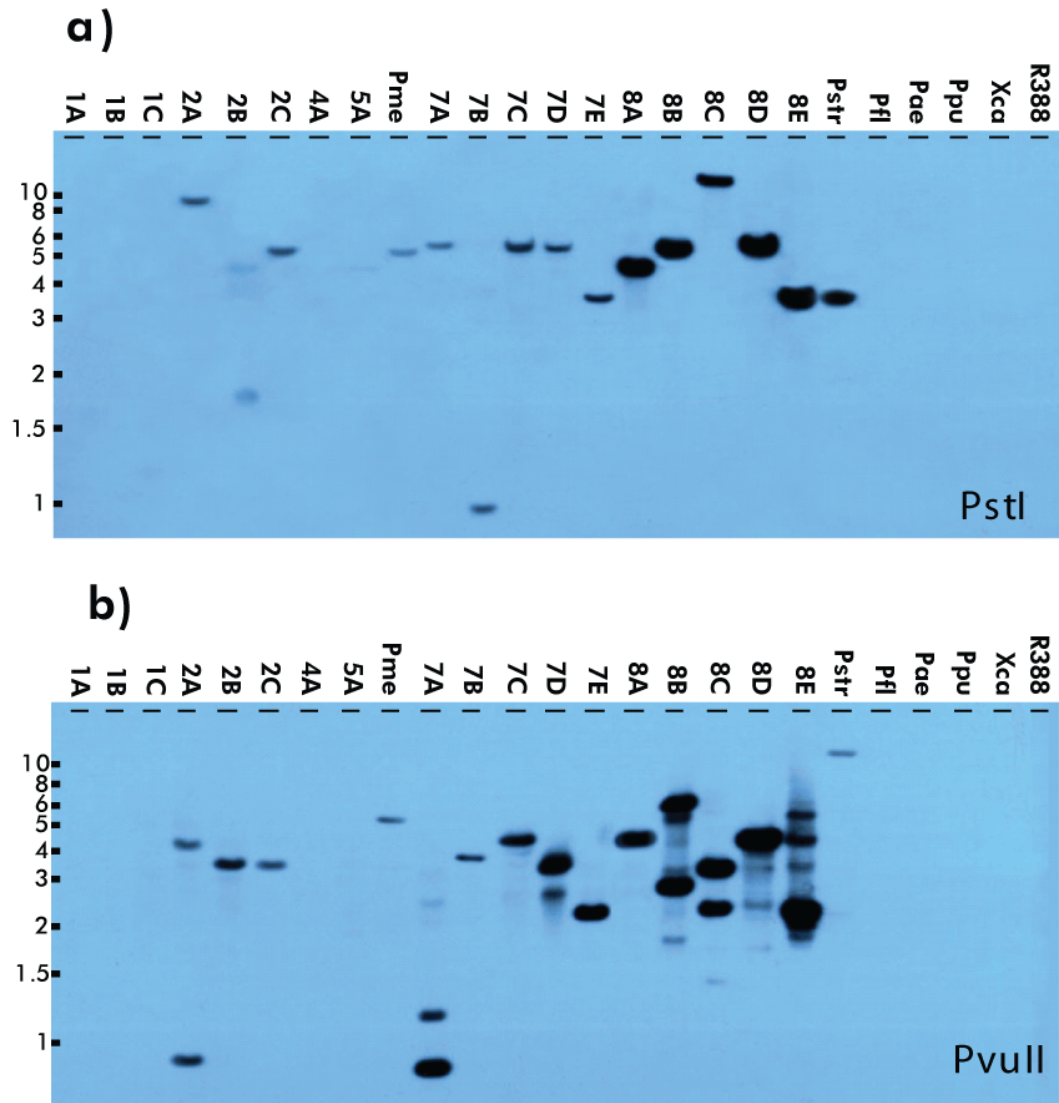


Figure 4.5 - Southern hybridisations on a) *Pst* I and b) *Pvu* II genomic digests using DIG-labelled probes targeting *Pseudomonas*-type core integron sequences. The probe used here was produced using *Ps. stutzeri* Q (Gv.8) gDNA as a template. Strain labels correspond to the abbreviations given in Table 3.1. Numbering on the left of each image indicates approximate band sizes in kilobases.

### 4.3.3 – Southern hybridisation integron detection – 59-be

Hybridisation signals were observed using probes targeting *Pseudomonas* subfamily 59-be for all *Ps. stutzeri* strains screened, with the exception of *Ps. stutzeri* API-2-142 (strain 7D) (Figure 4.6 a & b). Signals were also observed for *Ps. straminea* KM91 and *Ps. mendocina* NW1. For most strains in which hybridisation signals were obtained, multiple signals between 0.5 and 15 kb were observed, indicating the presence of multiple 59-be sequences, and by inference, multiple gene cassettes and/or cassette array(s) within each strain. By adding together the size of 59-be hybridisation signals observed for each strain, it was inferred that most contain cassette array(s) totalling 5 Kbp or more in size. *X. campestris* DAR 30538 and plasmid R388 both contain gene cassette arrays and multiple 59-be, although neither contains *Pseudomonas*-type 59-be. No hybridisation signals were observed for either sample. This observation is not surprising as the 59-be in these samples exhibit limited sequence identity to *Pseudomonas*-type 59-be. No hybridisation signals were observed for *Ps. fluorescens* NCTC 7244, *Ps. aeruginosa* NCTC 3756 and *Ps. putida* F1, which is again consistent with sequence information available from genome sequences of other strains belonging to these species (Section 4.3.5).

When probed with a 59-be probe, 16/19 strains which gave hybridisation signals included at least one very intense band. This contrasts with the core integron probe data where the majority of detected bands were faint. Several factors may be responsible for this: 1. multiple copies of 59-be targets in a single restriction fragment, 2. greater conservation of 59-be sequences relative to core integron sequences, or 3. the presence of degenerated core integron sequences. It is likely that a combination of all three factors is responsible for the differences between core

integron and 59-be hybridisation data. The fact that core integron hybridisations were performed at a low stringency and with large probes (relative to 59-be hybridisations) lends further support to this notion. Interestingly, strains belonging to *Ps. stutzeri* Gv's 1, 4 and 5, which did not give any hybridisation signals when screened with core integron probes, did give strong 59-be hybridisation signals. This may be due to the presence of cassette arrays that are either: 1. not linked to a core integron showing high identity to IntIPstQ, or 2. no longer linked to an intact core integron. A further interesting observation is that while core integron hybridisation signals were observed for *Ps. stutzeri* API-2-142 (7D), no 59-be hybridisation signals were observed. Two repeat hybridisations on this strain revealed identical results (data not shown). The presence of a core integron with no cassette array is a possible explanation for this. However, a core integron and adjacent gene cassette were recovered from this strain by integron PCR (section 4.3.1). This PCR assay uses a primer (ST1) which targets *Pseudomonas*-type 59-be and should therefore also have been a viable target for 59-be probes in hybridisation assays.



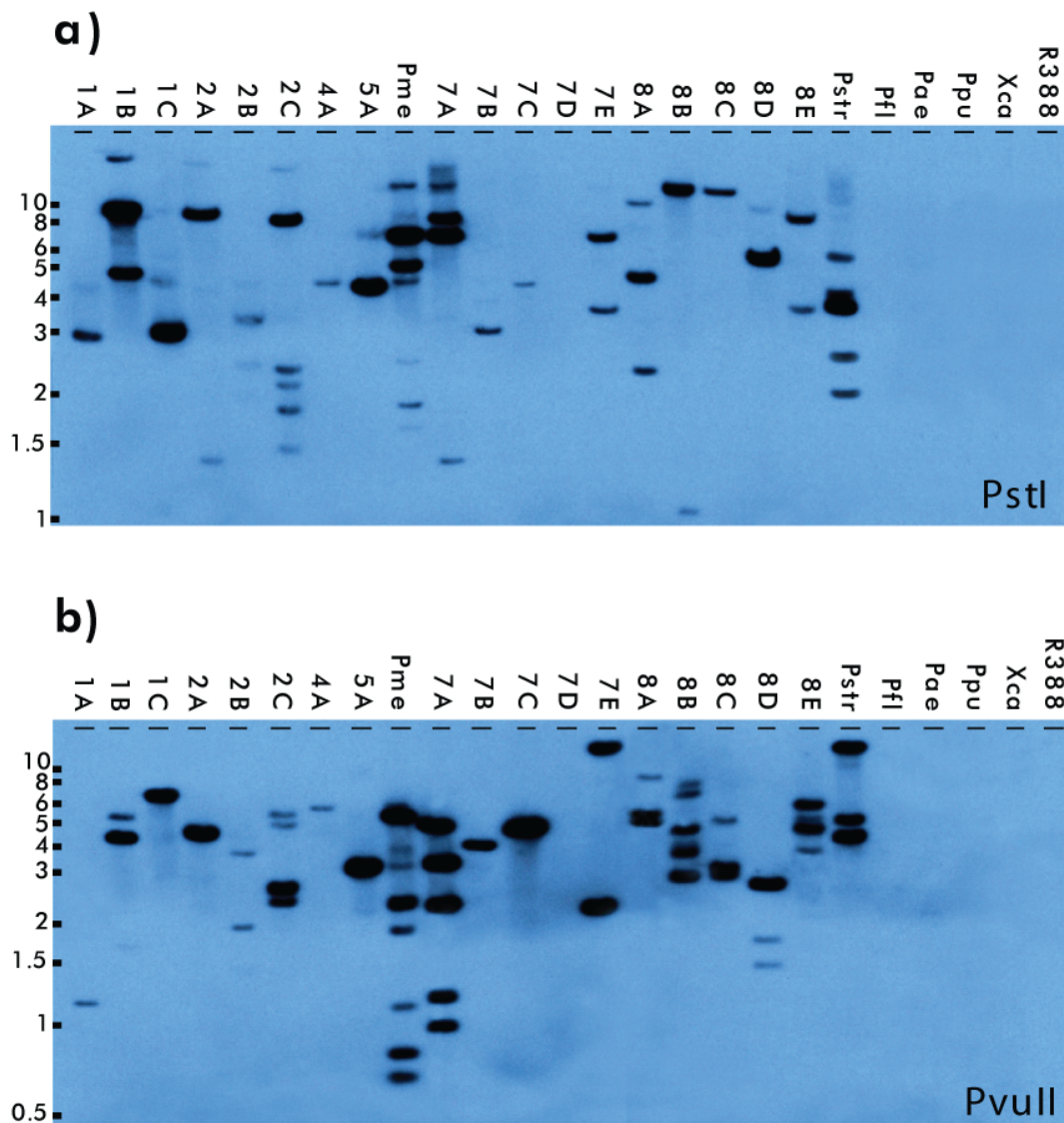


Figure 4.6 - Southern hybridisations on a) *Pst*I and b) *Pvu*II genomic digests using DIG-labelled probes targeting *Pseudomonas*-type 59-be. 59-be probes were prepared as described in section 2.5. Strain labels correspond to the abbreviations given in Table 3.1. Numbering on the left of each image indicates approximate band sizes in kilobases.

#### 4.3.4 – Analysis of Recovered integrons

Integron sequences recovered in the present study, in addition to those known previously from *Pseudomonas* spp. are represented diagrammatically in Figure 4.7, and the characteristics of all sequences recovered are described in the subsections below.

##### 4.3.4.1 – Analysis of recovered *intI* genes

Putative *intI* genes were identified in 10 strains, of which only five were predicted to encode functional proteins (Figure 4.7), as confirmed by alignment of predicted IntI amino acid sequences. *Ps. stutzeri* P1 (8C) was found to contain a partial *intI* gene (M. Gillings Unpublished data). A deletion of 125 amino acids between the residues equivalent to the K-165 and T-300 of intIPstQ, and a frameshift mutation at the residue equivalent the K-82 of intIPstQ were predicted to inactivate intIPstP1. The *intI* gene of *Ps. stutzeri* ATCC 17685 (7A) was found to be disrupted twice, containing a single nucleotide deletion at the residue corresponding to Pro-204 of intIPstQ, and a deletion of 2 nucleotides at the residue corresponding to Pro-230 of intIPstQ. Both deletions cause frame-shift mutations. As they collectively result in a deletion of 3 bp, the reading frame of the original protein is altered in between the two frameshift mutations only. The frameshifts occur within 87 bp of each other and thus, if transcribed from the original start, the resulting protein would contain a region of non-homologous sequence spanning 29 amino acids. Interestingly, the predicted shift in reading frame of this gene results in the exclusion of almost all of the *intI* specific patch, a conserved region characteristic of integron integrases (Nield *et al.*, 2001; Nunes-Duby *et al.*, 1998) which has been shown to contain residues essential for *intI* recombination activity (Messier and Roy, 2001). Thus, it is unlikely that the *intI* gene in

*Ps. stutzeri* ATCC 17685 encodes a functional integrase. For all subsequent analysis of this gene, the corrected amino acid sequence was used. All remaining inactivated *intl* genes were disrupted by deletions spanning several amino acids (all residue numbers given refer to corresponding region in intlPstQ): *Ps. straminea* KM91 – 8 amino acid deletion between Pro-225 and Arg-238, *Ps. stutzeri* RNAIII - 77 amino acid deletion between Met-136 and Ala-214, *Ps. stutzeri* DSM50238 - 176 amino acid deletion between Lys-99 and Thr-276, and *Ps. stutzeri* API-2-142 - 189 amino acid deletion between Leu-56 and Arg-247. All deletions also resulted in frameshift mutations and thus, the probability of any of the *intl* genes encoding functional integrases is very low. All deletions could be accounted for by a single event; the differing size and position of each deletion indicates that each occurred as an independent event.

The relationship between the *intl* genes recovered here and other integrase genes can be seen in the similarity matrix of amino acid identities presented in Table 4.3; integron integrases are clearly more similar to each other (minimum 45% amino acid similarity) than to XerD recombinases (maximum 29% amino acid identity), which is the most closely related gene family to the integron integrases. With the exception of the *Nitrosomonas europaea* Intl integrase (IntlNeu), *Pseudomonas* Intl integrases are more similar to each other (67-100% amino acid identity) than to other Intl (45-60% amino acid identity). Interestingly, *intl* genes from *Ps. stutzeri* Gv.7 strains are more similar to IntlPal (80-90% amino acid identity) than to the Intl of *Ps. stutzeri* Gv.8 strains (68-78% amino acid identity).



IntIvchA	100													
IntI1	0.45 (308)	100												
IntINeuA	0.47 (308)	0.57 (308)	100											
<i>E. coli</i> XerD	0.24 (308)	0.25 (308)	0.24 (308)	100										
IntIPal55044	0.47 (308)	0.55 (308)	0.64 (308)	0.21 (308)	100									
IntIPstQ	0.47 (308)	0.58 (308)	0.66 (308)	0.22 (308)	<b>0.71</b> <b>(308)</b>	100								
IntIPstBAM	0.47 (308)	0.57 (308)	0.66 (308)	0.22 (308)	<b>0.71</b> <b>(308)</b>	0.97 (308)	100							
IntIPstJM300	0.47 (308)	0.58 (308)	0.67 (308)	0.22 (308)	<b>0.71</b> <b>(308)</b>	0.98 (308)	0.98 (308)	100						
IntIPmeNW1	0.45 (308)	0.53 (308)	0.62 (308)	0.22 (308)	<b>0.73</b> <b>(308)</b>	0.70 (308)	0.69 (308)	0.70 (308)	100					
IntIPstrKM91	0.46 (301)	0.57 (301)	0.67 (301)	0.24 (301)	<b>0.72</b> <b>(301)</b>	0.76 (301)	0.76 (301)	0.77 (301)	0.68 (301)	100				
IntIPst17641	0.51 (227)	0.60 (227)	0.68 (227)	0.29 (227)	<b>0.71</b> <b>(227)</b>	0.98 (227)	0.98 (227)	0.98 (227)	0.67 (227)	0.74 (227)	100			
IntIPst17685	0.47 (227)	0.53 (227)	0.62 (227)	0.29 (227)	<b>0.90</b> <b>(227)</b>	0.71 (227)	0.71 (227)	0.71 (227)	0.70 (227)	0.65 (227)	0.71 (227)	100		
IntIPstRNAIII	0.48 (223)	0.57 (223)	0.65 (223)	0.24 (223)	<b>0.90</b> <b>(223)</b>	0.74 (223)	0.73 (223)	0.74 (223)	0.72 (223)	0.73 (223)	0.70 (129)	0.95 (129)	100	
IntIPst50238	0.51 (130)	0.59 (130)	0.70 (130)	0.25 (130)	<b>0.88</b> <b>(130)</b>	0.78 (130)	0.77 (130)	0.78 (130)	0.76 (130)	0.82 (130)	0.80 (37)	1.00 (37)	0.96 (130)	100
	IntIvchA	IntI1	IntINeuA	<i>E. coli</i> XerD	IntIPal55044	IntIPstQ	IntIPstBAM	IntIPstJM300	IntIPmeNW1	IntIPstrKM91	IntIPst17641	IntIPst17685	IntIPstRNAIII	IntIPst50238

Table 4.3 – Matrix of pair-wise identities for *Pseudomonas* spp. IntI and a reference set of integrases: IntIvchA (AKA intI4) - *Vibrio cholerae* O1 integron integrase, IntI1 – Class 1 integron integrase, IntINeuA – *Nitrosomonas europaea* integron integrase, *E. coli* XerD – XerD tyrosine recombinase. Pair-wise similarities are expressed as proportions and the number of paired residues used for each calculation is included in parentheses. Comparisons between IntIPal and the IntI sequences of other *Pseudomonas* spp. are highlighted in bold. All values to the right of this column represent comparisons between *Pseudomonas* spp.

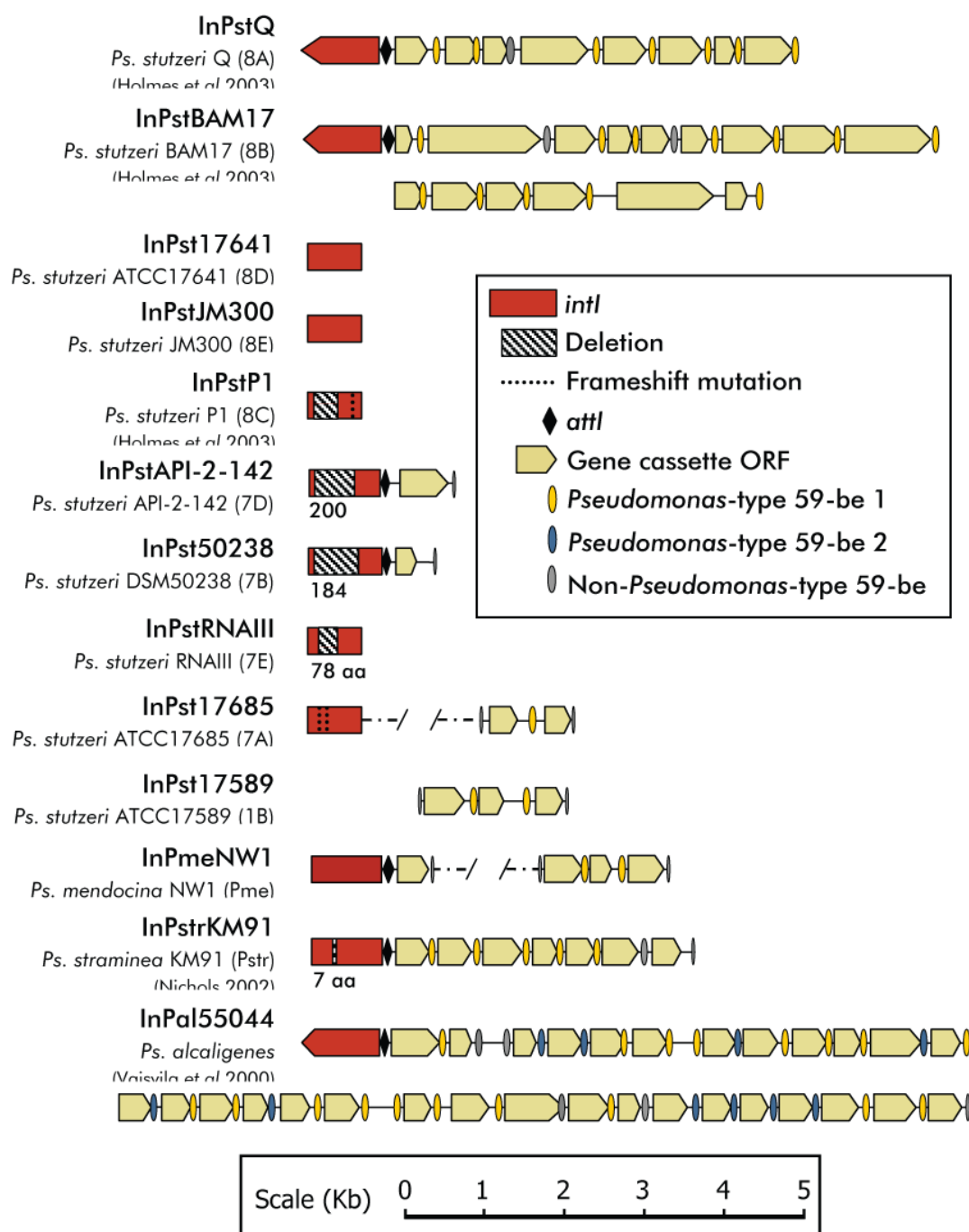


Figure 4.7 - Schematic representation of *Pseudomonas* spp. integrons recovered by PCR in the present study in addition to those known previously. The orientation of the point of arrow-shaped boxes indicates the direction of transcription. All features indicated are drawn to scale. Blank spaces represent non-coding regions.

#### 4.3.4.2 – Analysis of recovered *attI* regions

The position of the recombination crossover point of *attI* in *Ps. stutzeri* Q has been demonstrated experimentally (Coleman and Holmes, 2005; Holmes *et al.*, 2003b). All *Pseudomonas* spp. integrons recovered here, and previously, contain putative *attI* recombination crossovers approximately 200 bp 3' of the *intI* start codon (Figure 4.8). When these sequences are aligned, several regions of sequence conservation can be identified. The conserved region immediately 5' of the *attI* recombination crossover point encompasses the *attI* simple site as defined for class 1 integrons (Figure 4.8) (Partridge *et al.*, 2000). Sequence conservation in this region is suggestive of a role for these sequences, most likely in the recombination activity of *attI*. Two further regions of moderate conservation were observed which correspond to the strong and weak binding sites of *attI1* (Gravel *et al.* 1998; Partridge *et al.*, 2000), in terms of position relative to the *attI* recombination crossover (Figure 4.8). While these regions are conserved between the core integrons of the *Pseudomonas* strains included in the alignment, little if any conservation could be identified between these sequences and the sequence of *attI1* (data not shown). Experimental evidence is required to confirm any role for these sequences. The conserved regions boxed in red in Figure 4.8 do not correspond, in terms of position relative to the *attI* recombination crossover, to known functional sites in *attI1*. It is possible that these sequences are involved in regulation of *intI* expression (see section 4.3.4.4).

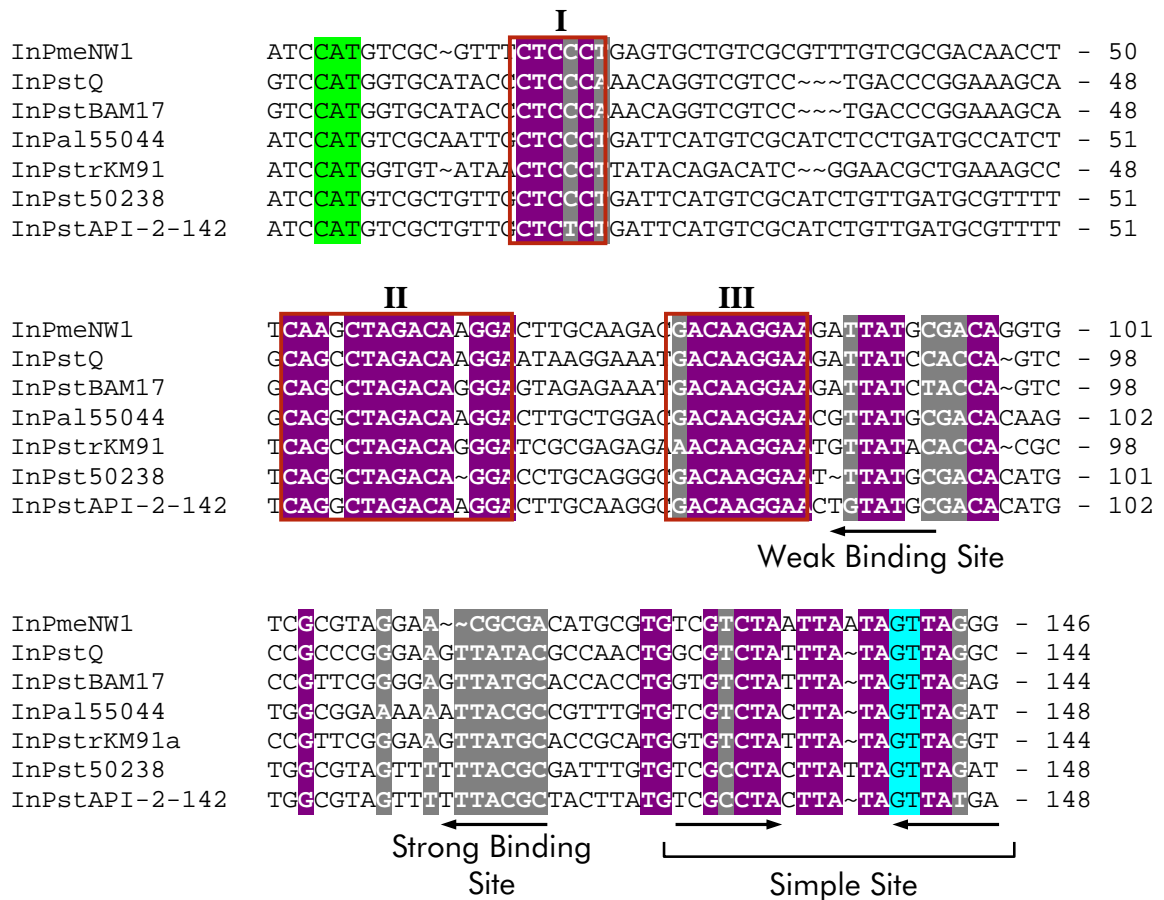


Figure 4.8 – Nucleotide alignment of the *attI* region of *Pseudomonas* spp. integrons. The alignment begins 3 bp before the start codon of *intI* (highlighted in green) and continues through to the *attI* recombination crossover point (highlighted in blue). Conserved sequence residues throughout the alignment are indicated in purple (universally conserved) and grey (>70% conserved). Regions corresponding to sequences involved in IntI1 recognition of *attI* (as defined by Partridge *et al.*, 2000) are indicated by arrows and brackets below the alignment. Conserved regions boxed in red do not correspond to any known *attI* sequences and consequently may be involved in regulation of *intI* expression. Sequence labels: InPmeNW1 – *Ps. mendocina* NW1, InPstQ – *Ps. stutzeri* Q, InPstBAM17 – *Ps. stutzeri* BAM17, InPal55044 – *Ps. alcaligenes* ATCC55044, InPstrKM91a – *Ps. straminea* KM91, InPst50238 – *Ps. stutzeri* DSM50238, InPstAPI-2-142 – *Ps. stutzeri* API-2-142.

#### 4.3.4.3 – Putative identification of P<sub>c</sub>

Previous studies have localized P<sub>c</sub> in class 1 and class 3 integrons to within the 5' end of the *intI* gene (-10 region corresponds to the Ser-30 codon of *intI1*) (Collis and Hall, 1995; Levesque *et al.*, 1994; Stokes and Hall, 1989; Swift *et al.*, 1981). A P<sub>c</sub> equivalent has been identified in *P. stutzeri* Q, but is not found at precisely the same site, with the -10 region corresponding to Asp-18 of *intI1* (Coleman and Holmes, 2005). Using comparisons to consensus sequences for 70-type promoters, a putative P<sub>c</sub> of TTGAGC-n16-TCTGAT was identified, and was also shown to account for almost all transcription of at least the first integrated cassette. The location and sequence of the putative *Ps. stutzeri* Q P<sub>c</sub> (as defined by Coleman and Holmes, 2005) is shown in Figure 4.9a; an equivalent sequence was identified in *Ps. stutzeri* BAM17 (Figure 4.9a), but not in any of the other *Pseudomonas* spp. analysed. An alternative site for P<sub>c</sub> was identified in the same position in all strains and corresponds to the P<sub>c</sub> identified for class 1 integrons, with the 1<sup>st</sup> base pair of the -10 region corresponding to the Ser-30 codon of *intI1* (Figure 4.9b).

The existence of potential promoters in all strains at a location equivalent to the P<sub>c</sub> of *intI1* strongly supports the existence of P<sub>c</sub> at this location; however experimental data is required to confirm this. In *Ps. stutzeri* Q and BAM17, the P<sub>c</sub> identified by Coleman and Holmes (2005) received a higher confidence score (as determined by the Neural Network Promoter Prediction program [see section 4.2.4]) and are thus favoured as promoter sequences. However, the test used by Coleman and Holmes (2005) to determine the correct P<sub>c</sub> for *Ps. stutzeri* Q was unable to discriminate between expression driven from each potential promoter due to their close proximity, and thus additional analyses are required to determine the correct P<sub>c</sub> in this strain.

## Bacterial Promoter Consensus Sequences

### Harley and Reynolds (1987)

T<sub>78</sub>T<sub>82</sub>G<sub>68</sub>A<sub>58</sub>C<sub>52</sub>A<sub>54</sub> -- 16<sub>21</sub>17<sub>52</sub>18<sub>19</sub> -- T<sub>82</sub>A<sub>89</sub>T<sub>52</sub>A<sub>59</sub>A<sub>49</sub>T<sub>89</sub>  
-35 region spacer -10 region

### Lisser and Margalit (1994)

T<sub>69</sub>T<sub>79</sub>G<sub>61</sub>A<sub>56</sub>C<sub>54</sub>A<sub>54</sub> -- 16<sub>17</sub>17<sub>43</sub>18<sub>17</sub> -- T<sub>77</sub>A<sub>76</sub>T<sub>60</sub>A<sub>61</sub>A<sub>56</sub>T<sub>82</sub>  
-35 region spacer -10 region

### a. P<sub>c</sub> defined by Coleman and Holmes (2005)

		Score
<i>Ps. stutzeri</i> Q	GT <b>TTGAGC</b> -----n16----- <b>TCTGAT</b> CAAGCAAGCG H K L R I R E R <b>M</b> Q D L L R	0.89
<i>Ps. stutzeri</i> BAM17	GT <b>TTGAGC</b> -----n16----- <b>TCTGAT</b> CAAGCAAGCG H K L R I R E R <b>M</b> Q D L L R	0.89

### b. P<sub>c</sub> equivalent to that of *IntI1*

<i>Ps. stutzeri</i> Q	AA <b>TAGACA</b> -----n16----- <b>AGTAGT</b> GTTTGAGCCG C Y V R E T R I <b>F</b> Y H K L R	0.82
<i>Ps. stutzeri</i> BAM17	AA <b>TAGACA</b> -----n16----- <b>AGTAGT</b> GTTTGAGCCG C Y V R E T R I <b>S</b> Y H K L R	0.81
<i>Ps. mendocina</i> NW1	CA <b>TAGACA</b> -----n16----- <b>AATAGT</b> TTCTGACTCG A Y V A E T R I <b>S</b> Y N R V R	0.84
<i>Ps. stutzeri</i> DSM50238	AA <b>TAGACA</b> -----n16----- <b>AATAGT</b> GTCGCAATCG L Y V A E T R I <b>S</b> Y H R L R	0.89
<i>Ps. stutzeri</i> RNAIII	AA <b>TAGACA</b> -----n16----- <b>AATAGT</b> GTCGCAATCG L Y V A E T R I <b>S</b> Y H R L R	0.89
<i>Ps. alcaligenes</i>	GA <b>TAGACA</b> -----n16----- <b>AATAGT</b> GTCGAAGCCG L Y V A E T R I <b>S</b> Y H R L R	0.85
<i>Ps. straminea</i> KM91	AA <b>TAGACA</b> -----n16----- <b>AGTAGT</b> GTCTGAGCCG C Y V R E T R I <b>S</b> Y H R L R	0.87

Figure 4.9 - Predicted cassette array promoter (P<sub>c</sub>) sequences for *Pseudomonas* spp. integrons. The consensus sequences for bacterial promoters shown indicate the degree of conservation (values in subscript) for each nucleotide in analyses performed on *E. coli* promoters. The predicted -35 regions and -10 regions are highlighted in yellow and red, respectively. Deviations from the consensus for bacterial promoters are highlighted in blue. The corresponding *intI* amino acid sequence for each strain is given beneath the nucleotide sequence. The *intI* amino acid used to refer to the location of the -10 region of P<sub>c</sub> is highlighted in green. The number given on the right of each predicted promoter provides a relative measure of confidence in the accuracy of the predicted promoter sequence. A value of 0.8 is expected to have a false positive prediction rate of 0.4%.

#### 4.3.4.4 - Putative identification of *intI* regulatory sequences

All integron sequences screened for putative  $P_c$  sequences were also screened for putative  $P_{int}$  sequences. The  $\sim 140$  bp region between the start codon of *IntI* and the recombination crossover of *attI* was analysed. No promoters with high confidence coefficients ( $>0.8$  as determined by the Neural Network Promoter Prediction program) were detected for any of the integron sequences screened. When a lower cutoff value (0.4) was used, however, promoters were detected at a conserved location in all *intI* genes screened (Figure 4.10). Putative Ribosome binding sites were also identified in the expected location for all *intI* genes analysed (highlighted purple in Figure 4.10). These sequences were identified on the basis of purine-rich regions (Shine-Dalgarno sequences) occurring 9-10 bp upstream of the *IntI* start codon. Putative  $P_{int}$  and RBS sequences were identified on the basis of conformation to consensus sequences and conserved location across divergent integrons only. Experimental data are required to confirm functionality of these sequences.



Figure 4.10 - Putative regulatory sequences associated with *Pseudomonas* spp. *intI* genes. Red lines beneath the alignment indicate the conserved sequence regions boxed in the *attI* alignment in Figure 4.8. The numbering beneath each line corresponds to the numbering of the conserved regions boxed in Figure 4.8. The locations of the putative *intI* promoter (-35 and -10 regions) and Ribosome Binding Site (RBS) are labeled above the alignment and included in bold, highlighted type. Deviations from the consensus for bacterial promoter sequences provided in Figure 4.9 are highlighted in red. Sequence labels: 1 – InPstQ, 2 – InPstBAM17, 3 – InPmeNW1, 4 – InPal55044, 5 – InPstrKM91α, 6 – InPst50238, 7 – InPstAPI-2-142.



#### 4.3.4.5 – Analysis of Gene Cassettes

Eleven new gene cassettes were recovered using the integron and cassette PCR assays and are represented diagrammatically in Figure 4.7. With the exception of a cassette in InPstBAM17 containing two ORFs in the forward orientation, all gene cassettes recovered contained a single ORF in the forward orientation (Figure 4.7). Most cassettes (88%) detected were associated with *Pseudomonas* subfamily 59-be (Figure 4.7). A more detailed analysis of the cassettes recovered here is provided in Section 5.3.1.5.

#### 4.3.5 – *in silico* analysis of *Pseudomonas* family integron distribution

To search for core integron sequences in published sequences available in public databases, *intI* amino acid sequences, in addition to the nucleotide sequence of the entire core integrons were used as query sequences for various BLAST searches. The core integrons of *Ps. stutzeri* Q (8A), *Ps. mendocina* NW1 (Pme) and *Ps. alcaligenes* ATCC 55044 were used as query sequences for all BLAST searches. To search for *Pseudomonas* subfamily 59-be and associated cassette ORFs, a set of three 59-be sequences (the same as those used for 59-be hybridisation probes – see Table 4.2) representative of the total variation apparent in this 59-be subfamily were used as query sequences for BLASTN searches against Genbank and environmental DNA databases.

The presence of sequences characteristic of *Pseudomonas* subfamily integrons outside members of the genus has been reported previously (Holmes *et al.*, 2003b). Three instances of gene cassettes containing *Pseudomonas* subfamily 59-be from non-*Pseudomonas* integrons were reported by Holmes *et al.*, (2003); orfN in the class 1

integron In31 of *Ps. aeruginosa* isolate 101/1477 (Laraki *et al.*, 1999)(accession number AJ223604 refers to this cassette as orf1 in In101), orfX in an unnamed class 1 integron in *Acinetobacter baumannii* BM4426 (Ploy *et al.*, 2000), and one site in an integron-like cassette array in the *Shewanella oneidensis* genome (Heidelberg *et al.*, 2002). The number sequences available to search in public databases has increased greatly since these observations were made and thus, additional sequences were expected to be detected in the searches performed here.

In addition to the 59-be sequences detected by Holmes *et al.*, (2003), two novel 59-be/gene cassettes were detected in published Genbank sequences. A *Pseudomonas* subfamily 59-be sequence was identified as part of a gene cassette (3<sup>rd</sup> position in array) in *Methylobacillus flagellatus* KT (CP000284.1), and as part of an orphan gene cassette in *Shewanella* sp. MR-7 (CP000444.1). A total of 17 *Pseudomonas* subfamily 59-be were detected in published environmental DNA sequences, all of which were found in the Sargasso Sea Marine Microbial Community database (Venter *et al.*, 2004). Details of all environmental DNA Blastn results are provided in the electronic appendix (Files within the folder – ‘environmental blast results\’). Only four pairs of cassettes were found within the same contig, and no sequence contig contained 3 or more cassettes. This observation is not surprising given the small size of most of the sequence fragments in this database (the largest contig which contained 59-be was 2.8 Kbp in size). All but one 59-be detected contained a potential ORF, which terminated near or within the 59-be sequence (data not shown). Thus, it is likely that 59-be detected in the Sargasso Sea Marine Microbial Community database are part of a gene cassette and at least 8 of 17 cassettes detected are part of multi-cassette arrays.

No additional *Pseudomonas* subfamily core integron sequences were detected in any Blast search using the Genbank database. Searching environmental DNA databases resulted in the detection of >20 putative *intI* integrase sequences, and again all representatives were from the Sargasso Sea Marine Microbial Community database (Venter *et al.*, 2004). However, the highest amino acid identity between one of these *intI* genes and a *Pseudomonas* subfamily *intI* gene was 59%, indicating that none of these *intI* are likely to be found in a *Pseudomonas* spp. strain.

## 4.4 – Conclusions

### **Members of the *Ps. stutzeri* species complex contain integron sequences with relatedness to InPstQ**

PCR and hybridisation were successfully used to detect integron sequences in several strains in the collection. While several integrons were successfully amplified using PCR, this method was prone to both false negative and false positive results. In contrast, hybridisation provided a more sensitive measure of the presence of integron-like sequences. Integron-like sequences were detected in all *Ps. stutzeri* strains screened, and their presence was confirmed in all Gv.8 strains, 4/5 Gv.7 strains, 2/3 Gv.2 strains and 1/3 Gv.1 strains. The presence of integrons was also confirmed in the *Ps. straminea* and *Ps. mendocina* strains screened. Several *Pseudomonas* strains screened showed no indication of the presence of integron sequences related to InPstQ, including *Ps. fluorescens*, *Ps. aeruginosa* and *Ps. putida*. Ambiguous cassette PCR data for *Ps. putida* F1 is likely to represent non-specific PCR amplification, as this assay was found to be prone to producing false positive results and all sequenced cassette PCR clones were non-specific products. In addition, no hybridisation signal was observed when screening this strain for 59-be sequences. *Ps. fluorescens*, *Ps.*

*aeruginosa* and *Ps. putida* strains gave clearly negative results in all hybridisation-based integron detection assays, which is consistent with information available from complete genome sequences. *X. campestris* DAR 30538 and Plasmid R388 gave clearly negative results in all *Pseudomonas*-type integron detection assays, which was consistent with the predicted behaviour of each probe under the hybridisation conditions used. Collectively, these observations indicate that the hybridisation screening method used was specific and provides strong evidence that all test strains in the collection contain integron sequences.

### **Many strains had integrons that are predicted to be non-functional**

Despite all Gv.2 strains giving moderate to strong *intI* and 59-be hybridisation signals, no specific PCR products could be generated with any of the assays used. It is likely that these strains do contain integrons; however, they are likely to be beyond the detection limits of the primers and PCR assays used here. Alternatively, integrons in Gv.2 strains may contain deletions which encompass one or more PCR primer target regions. The failure to detect core integron sequences in *Ps. stutzeri* Gv's 1, 4 and 5, despite the fact that these strains appear to contain gene cassettes on the basis of 59-be hybridisation data, indicates that the core integron is truncated, too divergent, or has been lost entirely in these strains. Taking into account the clear presence of *Pseudomonas* subfamily 59-be in these strains, it would be surprising if they contained a divergent core integron, and thus core integron loss or disruption is the favoured explanation. PCR independent methods employed in Chapter 5 will allow the nature of these integrons to be determined.

The large number of inactivated integrons observed here indicates that loss of core integron function is a frequent event in evolutionary terms. Half (6 of 12) of the

integrations currently known from *Pseudomonas* spp. contain inactivated core integrations. All observed inactivation mutations were unique, and consistent with each occurring as a single, independent event. All integrations in *Ps. stutzeri* Gv.7 strains contained deletions or frameshifts predicted to inactivate *intl*. An inactivated core integration is also present in *Ps. stutzeri* P1 (8C), despite the fact that putatively functional *intl* genes were detected in all remaining *Ps. stutzeri* Gv.8 strains. It is possible that core integration inactivation may be favoured in particular lineages and not in others as a means of stabilising cassette arrays, perhaps in response to changes in selective pressures.

### **An integration subfamily was acquired early in the evolution of *Pseudomonas***

Under a common origin hypothesis, integrations from members of the *Ps. stutzeri* species complex were expected to exhibit higher levels of identity to each other than to other *Pseudomonas* species. This was, however, not observed as the *intl* genes of *Ps. stutzeri* Gv.7 strains were more similar to intlPal55044 (*Ps. alcaligenes* ATCC 55044) than to the *intl* genes of *Ps. stutzeri* Gv.8 integrations. Nonetheless, the presence of similar integrations in *Pseudomonas* spp. which are separated by deep branching evolutionary relationships (see Figure 4.1), and the high level of similarity between most 59-bp from *Pseudomonas* spp. integrations, suggests that integrations were acquired relatively early in the evolution of Pseudomonads, and further supports the findings of Vaisvila *et al.*, (2001) and Holmes *et al.*, (2003) that an integration family is associated with the *Pseudomonas* lineage. The similarity between *Pseudomonas* spp. core integrations in the current dataset relative to other known core integrations is suggestive of a common origin for *Pseudomonas* spp. integrations in this genus; however, patterns in Intl identity observed among the current dataset indicate that a more detailed analysis of these sequences is required to confirm this. Several observations indicate that

further analyses are required to confirm this: 1. The distribution of integron sequences in Pseudomonads appears to be patchy, 2. Inconsistencies exist in Intl identity with respect to established evolutionary relationships, 3. The potential for multiple core integron copies in some strains suggests multiple integron acquisitions or core integron duplication. Further characterisation of the nature and extent of this integron subfamily is the subject of Chapters 5 and 6.



## CHAPTER 5 – GENOMIC CONTEXT OF *PSEUDOMONAS* INTEGRONS

### 5.1 - Introduction

It has become an informal convention to classify integrons on the basis genomic context as either mobile or chromosomal. The rationale for this classification is based on historical observations where the first chromosomal integrons analysed had distinctive features with regard to cassette array size and sequence relationships (see Section 1.6), and the assumption that these differences reflect variation in the properties of the integrons. Since mobile elements can insert into chromosomes, a means of detecting 'real' chromosomal integrons is needed. The most common method which has been used for this classification is characterisation of the genes immediately upstream and/or downstream of the integron. Mobile integrons are flanked by genes typical of mobile genetic elements such as transposons (Brown *et al.*, 1996; Hochhut *et al.*, 2001; Radstrom *et al.*, 1994; Sundstrom *et al.*, 1991). In contrast, a diverse array of genes have been found upstream of different CIs (Drouin *et al.*, 2002; Gillings *et al.*, 2005; Leon and Roy, 2003; Rowe-Magnus *et al.*, 2003), and usually encode enzymes of central metabolic pathways (eg. *ilvD* in *Xanthomonas* spp. [Gillings *et al.*, 2005] and *rpIT* in *Vibrio* spp. [Rowe-Magnus *et al.*, 2003]) or conserved hypothetical genes present throughout the lineage containing the CI (eg. PAL001 in *Pseudomonas* spp. [Vaisvila *et al.*, 2001]). These observations suggest that while CIs and the genes located immediately upstream are physically linked, they are not functionally linked.

It has been proposed that CIs are the ancestral form of mobilised integrons such as class1 integrons (Mazel, 2006; Rowe-Magnus *et al.*, 2001; Rowe-Magnus *et al.*,



2002). This hypothesis is consistent with the observation that CIs may exhibit stable vertical inheritance. However, it does not explain atypical codon usage of particular CI genes (Vaisvila *et al.*, 2001), the presence of divergent integrons in strains expected to related CIs (eg. divergent CI in *V. fischeri*), or the significant incongruence of the evolutionary relationships between *intI* and chromosomal framework genes. Thus, while CIs clearly form long-term associations with particular bacterial groups, these observations suggest that they are acquired initially by HGT before becoming fixed in the chromosome, and may also be re-mobilised by transfer of the CI from the chromosome to a mobile genetic element. How are integrons transferred between different genomic contexts? Presumably integrons arise in the parent cell as part of a larger genetic element, and are integrated into the chromosome occurring some time after acquisition (Figure 5.1). Integrons may be inserted into the chromosome as discrete units or as part of a larger genetic element, and additional, non-integron-associated genes may also be acquired subsequent to integron insertion. Determination of CI genomic context requires differentiation of the ancestral locus from genes of the parent element (as defined in Figure 5.1).

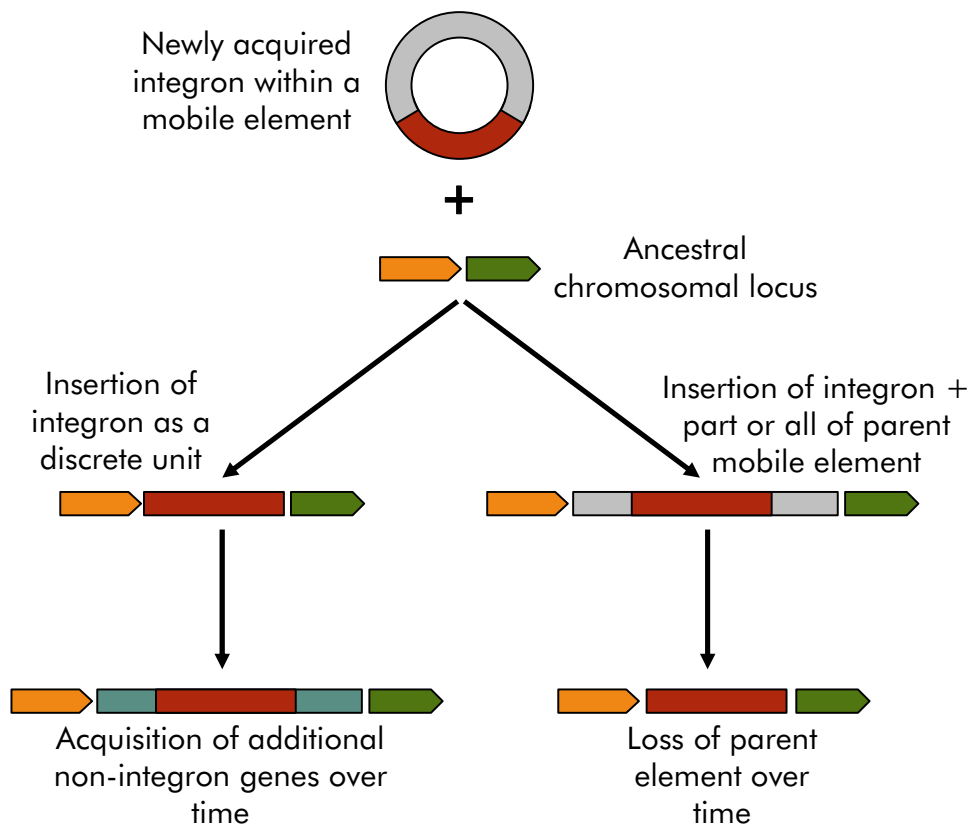


Figure 5.1 - hypothetical pathways for the integration of a CI into a host chromosome and its subsequent evolution. CIs may be integrated into the chromosome as discrete units, or may be integrated as part of a larger element. Additional elements or genes may also be acquired after integration of the CI. The mechanism(s) by which this may occur is unknown. Determination of genomic context requires differentiation of genes which belong to an ancestral parent element from those which constitute the ancestral chromosomal locus.

Analysis of genomic context in conjunction with phylogenetic diversity represents a powerful tool for reconstructing the evolutionary history of gene families. For example, several evolutionary events are apparent from examination of patterns in genomic context and phylogenetic diversity of CIs in *Vibrio* spp. strains. Phylogenetic data suggests that a CI was present in the common ancestor of many extant *Vibrio* spp., and that the CI has been stably vertically inherited over long evolutionary

timescales (Rowe-Magnus *et al.*, 2003). *Vibrio* spp. CIs have, however, undergone several intragenomic translocation events, some of which have resulted in the CI moving between different replicons (Boucher *et al.*, 2006; Mazel, 2006).

Independent acquisition of a divergent CI in *V. fischeri* is also evident from *intI* phylogenies and is further supported by genomic context data (Rowe-Magnus *et al.*, 2003). These observations indicate that while stable vertical inheritance clearly occurs for some CI sub-families, these elements may undergo intragenomic recombination or be transferred to a mobile genetic element, and divergent integrons may be acquired by HGT.

As part of the broader aim of reconstructing the evolutionary history of CIs in Pseudomonads, the primary aim of the work presented in this chapter was to assess the stability of inheritance and frequency of acquisition of CIs in Pseudomonads. The genomic context of integrons present in members of the strain collection was determined, and the phylogenetic diversity of integron-associated genes relative to genomic context explored. Based on extrapolation of the limited *Vibrio* spp. CI dataset by several researchers (Mazel, 2006; Rowe-Magnus *et al.*, 2001; Rowe-Magnus *et al.*, 2002), a number of specific hypotheses can be formulated regarding expected patterns of integron diversity in Pseudomonads using genomic context data:

1. An integron subfamily derived from a single ancestor exists in *Pseudomonas*.
2. Only one member of the subfamily will be found per *Pseudomonas* spp. strain.
3. The integron subfamily will be found at one (or few) chromosomal locus.
4. *Pseudomonas* subfamily integrons are not found within larger definable mobile genetic elements.

The above hypotheses can be addressed by determination of integron genomic context, provided that an appropriate sample set of strains is used. It was established in Chapter 3 that the sample set used in the present study comprises strains which exhibit a spectrum of evolutionary divergence, and in Chapter 4 that integrons are common features of *Pseudomonad* genomes. In the present chapter, whole-genome (fosmid) clone libraries were constructed from strains representative of this diversity, with the aim of recovering entire integron sequences in addition to both upstream and downstream flanking sequences. Genomic context was determined through examination of the genes adjacent to the integron.

## **5.2 - Materials and methods**

### **5.2.1 – Summary of cloning, detection and sequencing strategy**

To recover entire integron sequences in addition to genes upstream and downstream of the integron, large-insert libraries were constructed for 11 strains in the collection: *Ps. stutzeri* ATCC17589 (1B), *Ps. stutzeri* ATCC17595 (2B), *Ps. stutzeri* ATCC14405 (2C), *Ps. stutzeri* 19SMN4 (4A), *Ps. stutzeri* DNSP21 (5A), *Ps. stutzeri* DSM50238 (7B), *Ps. stutzeri* RNAIII (7E), *Ps. stutzeri* Q (8A), *Ps. stutzeri* JM300 (8E), *Ps. straminea* KM91 (Pstr), *Ps. mendocina* NW1 (Pme). Target sequences were detected by hybridisation, and pre-existing sequence information was used to generate additional upstream and downstream sequence information. Detailed methods and results of fosmid library construction, sequence detection and generation of additional sequence information are provided in Appendix 1. The present chapter was dedicated to analysing the data produced using these techniques.

### 5.2.2 - Analysis of gene cassettes

All gene cassettes were identified by manual annotation of nucleotide sequences. The probable function of cassette-associated ORFs was determined by performing Blastp searches (Altschul *et al.*, 1997) on the predicted amino acid sequence of the ORF using the Blast facility of the NCBI website (<http://www.ncbi.nlm.nih.gov/BLAST/>). A significant Blastp match was defined as a protein match that gave an e-value of less than 0.01. Cassette associated 59-be were also annotated manually, identified on the basis of their conserved core sequences and inverted repeat symmetry. All 59-be were classified as either belonging to the *Pseudomonas* subfamily of 59-be or not, based on the presence of a conserved (>90% nucleotide identity) 18 bp insert in the left hand region of the element, as defined by Holmes *et al.*, (2003).

### 5.2.3 - Sequence Analysis

All integron sequences and other open reading frames were identified using Blast searches and manual annotation of recovered sequences. All additional integron sequences recovered were analysed as described in Section 4.2.4. The strategy employed for the analysis of integron flanking sequences is outlined in Figure 5.2. The genomic context of each integron was determined by recovering at least one upstream and one downstream flanking gene which was characteristic of a particular genomic context (ie. chromosomal or mobile element), as inferred by similarity to known proteins. Testing for conserved integron-associated sequences outside the currently recognised integron boundaries was performed using multiple sequence alignments of the sequences immediately upstream and downstream of the integron and by local BLAST searches on sequence databases containing all known

*Pseudomonas* integrons, in addition to completely sequenced *Pseudomonas* genomes.

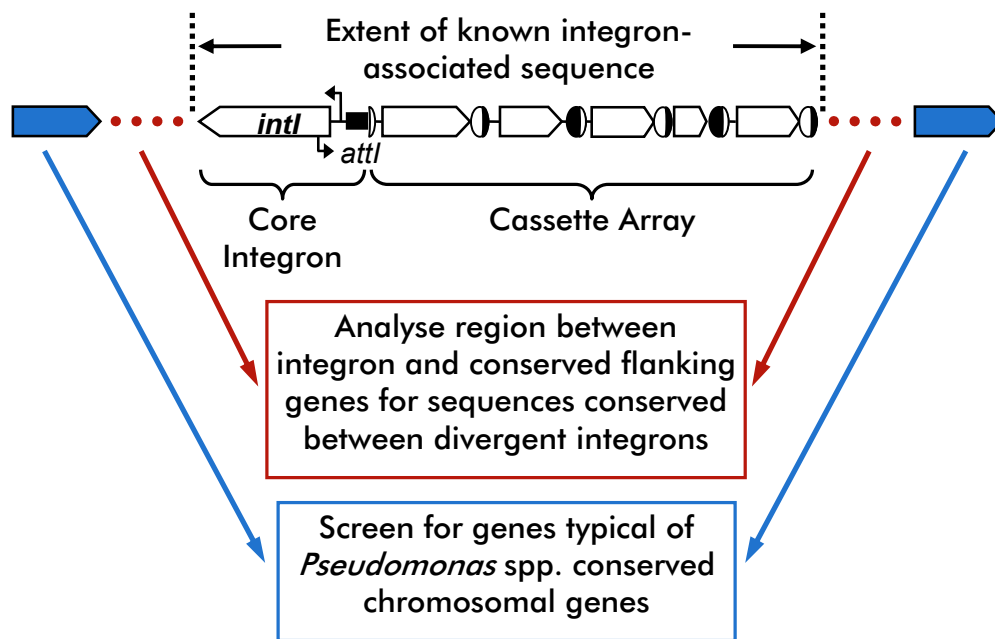


Figure 5.2 - Strategic approach to the analysis of integron flanking sequences. Non-coding regions between the last identifiable 5' and 3' integron sequence and the first flanking gene were analysed for the presence of conserved sequences present in other integrons, but not in sequenced *Pseudomonas* genomes. The term 'flanking gene' is used here to refer to the first gene upstream and downstream of the integron (highlighted in blue) which allows the genomic context of the integron to be determined (ie. the first genes which are characteristically chromosomal or typical of mobile elements).

#### **5.2.4 - Phylogenetic Analysis**

All alignments were constructed using CLUSTALX and optimised in GeneDoc as described in Section 2.7. Ambiguities in amino acid alignments were resolved by referring to the corresponding region in nucleotide alignments. Trees were constructed from optimised amino acids using the 'Seqboot', 'Protdist', 'Neighbour' and 'Consense' programs available within the Phylip software package, and print-quality trees were produced as described in Section 2.7.

## 5.3 – Results

Fosmid libraries which provided several-fold genome coverage were generated for 11 strains in the collection. Analysis of RFLP profiles of randomly selected clones indicated the presence of random-DNA inserts in each library (Figure A.2).

Hybridisation screening resulted in the detection of clones containing sequences homologous to *Pseudomonas* subfamily 59-be and/or *intI* sequences in all libraries (Figure A.3). Detailed results of fosmid library construction, coverage determination and hybridisation screening are provided in Appendix 1, and all sequences generated are analysed in detail below.

### 5.3.1 – Recovery of additional integron sequences

#### 5.3.1.1 – Extension of existing integron sequences

Existing sequence data was used as a template for primer design and extension of the nucleotide sequence of integrons in *Ps. stutzeri* RNAIII (7E), *Ps. stutzeri* DSM 50238 (7B), *Ps. stutzeri* ATCC 17589 (1B), *Ps. straminea* KM91 (Pstr), and *Ps. mendocina* NW1 (Pme) by primer walking. The entire integron was recovered for each strain, and a summary of the characteristics of these, and all additional integron sequences recovered is given in Table 5.1. The 3' end of *intI* was recovered for InPstRNAIII (7E), InPst50238 (7B) and InPstrKM91 (Pstr), and their cassette arrays were extended by three, two and three cassettes, respectively. Extension of the core integron and three-cassette array sequences amplified from *Ps. mendocina* NW1 (described in section 4.3.4) resulted in the recovery of a single integron, consisting of a cassette array of 17 cassettes and a complete core integron (Table 5.1). Extension of the three-cassette array of InPst17589a (1B) resulted in the generation of a sequence consisting of seven cassettes and a partial core integron (Table 5.1). A



putative *attI* site was identified which was characteristic of the *attI* of *Pseudomonas*-type integrons (Figure 5.3). The extent of this similarity extended approximately 100 bp upstream from the recombination crossover, and regions of sequence conservation corresponded to sequences thought to be involved in IntI recognition of *attI* (Partridge *et al.*, 2000). The sequence upstream of the putative *attI* site was subjected to manual examination and Blastn and Blastp database searches using all three possible translations of the nucleotide sequence, and an *intI* gene or remnant was not detected.

#### 5.3.1.2 – Characterisation of *Ps. stutzeri* 19SMN4 (4A) and DNSP21 (5A) integrons

Cassette PCR amplicons were generated from 59-be-containing fosmid clones from *Ps. stutzeri* 19SMN4 (4A) and DNSP21 (5A), and in all cases gene cassette sequences were recovered. Extension of Cassette PCR sequences from fosmid templates resulted in the recovery of a putative integron from both strains (InPst19SMN4 and InPstDNSP21; Table 5.1). Conserved sequences characteristic of the *attI* site of InPstQa were identified at the 5' end of each array, but an adjacent *intI* gene or remnant was not detected. The similarity between the putative *attI* sites of InPst19SMN4 and InPstDNSP21 and the corresponding regions of other *Pseudomonas* spp. integrons began at the *attI* recombination crossover and extended upstream approximately 100 bp (Figure 5.3). Sequences thought to be involved in IntI recognition of *attI* were conserved in both integrons, as was the putative -10 region of P<sub>int</sub> (as identified in section 4.3.4.2 - 4.3.4.4) (boxed in red in Figure 5.3). In both InPst19SMN4 and InPstDNSP21, the region exhibiting identity to *attI* was separated from the first upstream flanking gene by a relatively large non-coding sequence approximately 250 bp in length. No similarity between these sequences

and any *intI* gene was identified. InPst19SMN4 and InPstDNSP21 shared a nucleotide identity of 91% (244/269) across this region (data not shown), which was comparable to the identity of the putative *attI* regions (94% [99/105] pair-wise nucleotide identity) of these integron remnants. The high level of similarity between the region where *intI* would normally exist in InPst19SMN4 and InPstDNSP21 suggests that if a functional core integron once existed in these integrons, it was lost in a single event. As inactivation of the core integron is predicted to result in fixation of the associated cassette array, it was surprising to find that InPst19SMN4 and InPstDNSP21 contained cassette arrays composed of different cassettes.

Strain	Integron name	intI identified	intI complete/disrupted <sup>1</sup>	attI identified <sup>2</sup>	intI amino acids	% amino acid identity with:			Total cassettes <sup>3</sup>	Cassettes with BLASTp hits <sup>4</sup>		Pseudomonas subfamily cassettes
						intIPstQ	intIPal	intI1		Hyp.	Fun.	
<i>Ps. stutzeri</i> ATCC17589 (1B)	InPst17589a	x	-	✓ <sup>P</sup>	-	-	-	-	7	4	0	7/7
<i>Ps. stutzeri</i> ATCC17589 (1B)	InPst17589b	x	-	✓ <sup>P</sup>	-	-	-	-	5	1	0	4/5
<i>Ps. stutzeri</i> ATCC17595 (2B)	InPst17595	✓ <sup>P</sup>	T	x	164	77	71	59	?	-	-	-
<i>Ps. stutzeri</i> ATCC14405 (2C)	InPst14405	✓ <sup>P</sup>	T	x	164	77	71	59	?	-	-	-
<i>Ps. stutzeri</i> 19SMN4 (4A)	InPst19SMN4	x	-	✓ <sup>P</sup>	-	-	-	-	5	2	0	4/5
<i>Ps. stutzeri</i> DNSP21 (5A)	InPstDNSP21	x	-	✓ <sup>P</sup>	-	-	-	-	2	1	0	2/2
<i>Ps. stutzeri</i> ATCC17685 (7A)	InPst17685	✓ <sup>P</sup>	F	x	243	71	89	52	2	1	0	1
<i>Ps. stutzeri</i> DSM50238 (7B)	InPst50238	✓	D <sub>176</sub>	✓	139	79	89	60	3	0	0	3/3
<i>Ps. stutzeri</i> API-2-142 (7D)	InPstAPI-2-142	✓	D <sub>189</sub>	✓	122	71	89	52	1+	0	0	-
<i>Ps. stutzeri</i> RNAIII (7E)	InPstRNAIII	✓	D <sub>76</sub>	✓	244	73	90	56	4	0	0	3/4
<i>Ps. stutzeri</i> Q (8A)	InPstQa	✓	C	✓	320	100	72	58	10	5	1	9/10
<i>Ps. stutzeri</i> Q (8A)	InPstQb	x	-	x	-	-	-	-	3+	1	0	2/3
<i>Ps. stutzeri</i> BAM17 (8B)	InPstBAM17	✓	C	✓	320	98	72	57	14	8	4	12/14
<i>Ps. stutzeri</i> ATCC17641 (8D)	InPst17641	✓ <sup>P</sup>	C	x	235	98	71	59	?	-	-	-
<i>Ps. stutzeri</i> JM300 (8E)	InPstJM300	✓	C	✓ <sup>P</sup>	320	98	72	58	?	-	-	-
<i>Ps. straminea</i> KM91 (Pstr)	InPstrKM91a	✓	D <sub>8</sub>	✓	314	71	71	57	10	3	0	9/10
<i>Ps. straminea</i> KM91 (Pstr)	InPstrKM91b	x	-	✓ <sup>P</sup>	-	-	-	-	2+	1	0	2/2
<i>Ps. mendocina</i> NW1 (Pme)	InPmeNW1	✓	-	✓	321	70	74	54	17	4	0	16/17

Table 5.1 - Summary of all integron-containing sequence contigs recovered from strains in the collection. The data presented in the table constitutes a compilation of all integron sequence data generated from fosmid clones and PCR amplification presented in chapter 4 and the present chapter, in addition to pre-existing sequence data.

<sup>1</sup> F – frame-shift, T – Transposon, D – Deletion (number of deleted amino acids indicated in subscript).

<sup>2</sup> P – Partial sequence

<sup>3</sup> Cassette numbers followed by a '+' indicate incompletely recovered cassette arrays which may contain additional cassettes. '?' values indicate that no cassette array was recovered and the number of cassettes in the associated array is unknown.

<sup>4</sup> Hyp. – hypothetical gene, Fun. – functionally characterised gene. Details of selected cassettes containing ORFs which gave significant BLAST hits to characterised proteins are given in Table 5.2.

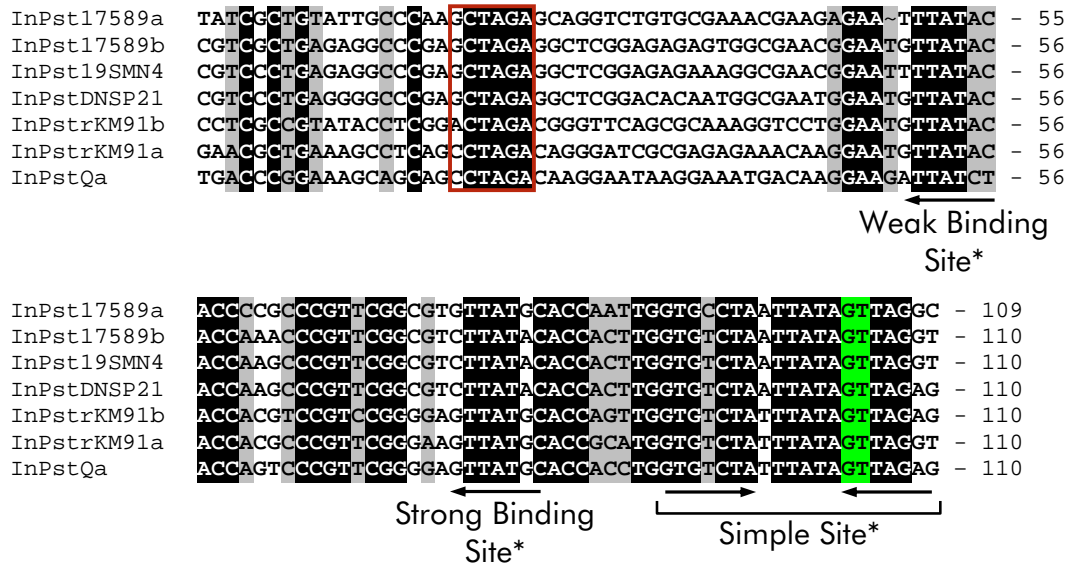


Figure 5.3 - nucleotide alignment of selected *Pseudomonas* spp. *attI* regions. Newly recovered *attI* regions (sequences 1-5) were aligned to the *attI* regions showing the highest level of similarity (InPstrKM91a and InPstQa). Universally conserved sequences are shaded black, while positions showing >70% conservation are shaded grey. The predicted *attI* recombination crossover is highlighted in green. The region boxed in red corresponds to the putative  $P_{int}$  -10 sequence identified in Figure 4.10.

\* - The arrows in the alignment above indicate the position of nucleotides shown to be involved in the recognition of *attI* by IntI1, relative to distance from the recombination crossover (Partridge *et al.*, 2000).

#### 5.3.1.3 – Characterisation of *Ps. stutzeri* Gv.2 integrons

As attempts to amplify integron sequences from *Ps. stutzeri* ATCC 17595 (2B) and ATCC 14405 (2C) fosmid templates were unsuccessful, fosmid clones were sub-cloned into pUC18 to generate a starting template for sequencing. Partial core integron sequences were recovered from both strains (Table 5.1). The predicted *intI* sequence of each strain exhibited 100% (163/163) pair-wise amino acid identity, and were found to be most similar to the *intI* gene of *Ps. stutzeri* ATCC 17641 (8D) (79% [129/163] amino acid identities). Both *intI* genes were interrupted by a transposase inserted at an identical location (equivalent to the Arg-168 residue of *intI1*), which belongs to the IS5 family of transposases found in a diverse range *Pseudomonas* spp. (eg. *Ps. resinovorans* CA10 IS5 family transposase - 186/187 [99%] amino acid identities). Attempts to recover additional sequence downstream of the transposon were unsuccessful; a sudden and precise loss of the sequencing signal observed in sequence traces suggests a strong sequence terminator prevented additional sequence generation using the primer walking approach.

#### 5.3.1.4 - Characterisation of additional integrons / cassette arrays

Several additional 59-be-containing fosmid clones from libraries in which 59-be sequences were overrepresented (*Ps. stutzeri* ATCC17589, *Ps. stutzeri* Q, *Ps. mendocina* NW1 and *Ps. straminea* KM91) were screened using cassette PCR to determine if these clones were likely to represent independent cassette arrays / integrons, or multiple representatives of a single cassette array. Cassette PCR profiles unique with respect to profiles obtained in Section 4.3.1 were observed in fosmid clones of *Ps. stutzeri* ATCC 17589, *Ps. stutzeri* Q, and *Ps. straminea* KM91 (data not

shown). Sequencing of the PCR amplicons resulted in each fosmid yielding novel gene cassettes, providing sequence information for the extension of each sequence from a fosmid DNA template.

A putative second integron consisting of 5 cassettes was recovered from *Ps. stutzeri* ATCC 17589 (InPst17589b; Table 5.1). An *intI* gene was not identified in the sequence upstream of the cassette array; however, a region clearly exhibiting homology to the *attI* site of *Ps. stutzeri* Q (8A) (Coleman and Holmes, 2005) was identified immediately adjacent to the first gene cassette. The region of InPst17589b exhibiting similarity to *Pseudomonas*-type *attI* sites extended approximately 100 bp upstream from the recombination crossover, and the regions of sequence conservation corresponded to (in terms of distance in bp from the recombination crossover) regions shown to be involved in *IntI* binding to *attI* in class 1 integrons (Partridge et al 2000) (Figure 5.3). The putative *attI* regions in InPst17589a and InPst17589b are clearly different (Figure 5.4a); this region InPst17589b was more similar to the *attI* sites of InPst19SMN4 and InPstDNSP21 (exhibiting 95% pair-wise nucleotide identity to both) than to that of InPst17589a (72% pair-wise nucleotide identity). Further, the region between *attI* and the first upstream flanking gene of InPst17589b exhibited significant identity to the corresponding region in InPst19SMN4 and InPstDNSP21 (Pair-wise nucleotide similarity, InPst19SMN4 – 91% [245/269], InPstDNSP21 – 91% [244/269]), while no regions of similarity could be identified between this sequence and the corresponding region in InPst17589a. These results suggest that InPst17589b, InPst19SMN4 and InPstDNSP21 derived from a common ancestor, while InPst17589a and InPst17589b have independent origins.

A putative second integron consisting of two gene cassettes was recovered from *Ps. straminea* KM91 (InPstrKM91b; Table 5.1). A region showing homology to the *attI* region of *Ps. stutzeri* Q (Coleman and Holmes, 2005) was identified in InPstrKM91b immediately adjacent to the first cassette in the array. Similar to the integron remnants identified in *Ps. stutzeri* ATCC 17589 (1B), 19SNM4 (4A) and DNSP21 (5A), this region extended approximately 100 bp upstream of the predicted *attI* recombination crossover, and contained regions of sequence conservation corresponding to sequences involved in IntI recognition of *attI* in class 1 integrons (Figure 5.3). The *attI* sites of InPstrKM91a and InPstrKM91b are clearly different (Figure 5.4b), however they shared a relatively high pair-wise identity (73% [76/104]). Interestingly, the putative *attI* region of InPstrKM91a exhibited a higher identity to the *attI* region of InPstQa (75% [78/104]) than to the corresponding region of InPstrKM91b. An *intI* gene could not be identified in InPstrKM91b, and the *attI*-like region was separated from the first upstream flanking gene by a non-coding sequence of approximately 260 bp. No regions of homology could be identified between this region and the corresponding region of InPstrKM91. BLASTN and BLASTX searches using the non-coding region between *attI* and the first upstream gene of InPstrKM91b as a query against both Genbank and the sequences generated in the present study also revealed no significant conservation to known sequences. The presence of conserved sequences in the *attI* region of InPstrKM91 beyond the region thought to be involved in IntI mediated recombination supports the notion that InPstrKM91b was once an independently functioning integron as opposed to acting as a secondary site for *in trans* IntI recombination. As outlined above, InPstrKM91a and InPstrKM91b are clearly different; however, their respective origins cannot be determined using the current dataset.

An incomplete cassette array/integron (InPstQb) consisting of 3 cassettes was recovered from *Ps. stutzeri* Q (InPstQb; Table 5.1). Efforts to generate additional sequence data upstream of this integron were unsuccessful, as a suspected termination sequence prevented further upstream sequencing using primer walking. As a result, evidence for a core integron-associated with this array was not recovered.

#### a. InPst17589a and InPst17589b

```
InPst17589a  TATCGCTGTATTGCCCAAGCTAGACAGGTCTCTCGAAAAGAAAGAA~TTTATA - 55
InPst17589b  CGTCGCTCAGAGGCCCGAGCTAGACGCTCGGAGACAGTGGCGAACGGAATCTTATA - 56

InPst17589a  CACCCCGCCCGTTCGGCGTGTATTGCACCAATTGGTGCTAATTATAGTTAGGC - 109
InPst17589b  CACCAAAACCGTTCGGCGTCTTATACACCACCTGGTGTCTAATTATAGTTAGGT - 110
```

#### b. InPstrKM91a and InPstrKM91b

```
InPstrKM91a  GAACGCTCAAAGCCTCAGCCTAGACAGGGATCGCGAGAGAAACAAGGAATGTTATA - 56
InPstrKM91b  CCTCGCTCTATACCTCGACTAGACGGGTTCAGCGCAAAGGTCTGGAATGTTATA - 56

InPstrKM91a  CACCACGCCCGTTCGGGAAGTTATGCACCGCATGGTGTCTATTATAGTTAGGT - 110
InPstrKM91b  CACCACGTCCGTCCGGGAGTTATGCACCAAGTGGTGTCTATTATAGTTAGAG - 110
```

Figure 5.4 - Alignments of putative *attI* sites from multiple integrons in *Ps. stutzeri* ATCC 17589 and *Ps. straminea* KM91. The predicted *attI* recombination crossover is highlighted in green. Conserved positions in each alignment are shaded grey.



#### 5.3.1.5 – Analysis of recovered gene cassettes

A total of 82 gene cassettes were recovered from 14 strains (Table 5.1). Of these, 76 cassettes contained a single ORF in the forward orientation (type A cassettes as defined by (Stokes *et al.*, 2001), and contained little non-coding sequence between the cassette ORF and its associated 59-be, all of which are typical features of gene cassettes (Stokes *et al.*, 2001; Holmes *et al.*, 2003). Four cassettes contained no identifiable ORFs and two cassettes contained two ORFs in the forward orientation (cassette types G and E, respectively, as defined by Stokes *et al.*, 2001). The vast majority (75/84) of cassettes recovered possessed *Pseudomonas* subfamily 59-be (Table 5.1). Two distinct types of 59-be were apparent within the dataset of *Pseudomonas* subfamily 59-be. Each type consisted of a characteristic number of nucleotides. The first (<sup>Ps76</sup>59-be) consisted of  $76 \pm 2$  bp and comprised 82% of all 59-be in the dataset, and the second (<sup>Ps89</sup>59-be) consisted of  $89 \pm 1$  bp, accounting for 13% of all 59-be in the dataset. These two 59-be types are not distinguished in Table 5.1, however a classification of all 59-be is provided in electronic appendix (Files within the folder – '59-be sequences\'). Across all possible pair-wise comparisons of <sup>Ps76</sup>59-be in the current dataset, identity ranged from 51-99%, and while the average similarity across all <sup>Ps76</sup>59-be was 81%, a pair-wise identity of >90% was observed in most comparisons (data not shown). Considering the larger size of the dataset analysed here, this result is comparable with the minimal similarity of 68% for *Pseudomonas* subfamily 59-be found by Holmes *et al.*, (2003), and lends further support to the hypothesis that *Pseudomonas* CIs share a specific (but not necessarily exclusive) association with one or more 'families' of 59-be. A more in depth analysis of the 59-be sequences recovered is beyond the scope of the present study, and will be presented elsewhere (S. Tetu unpublished data).

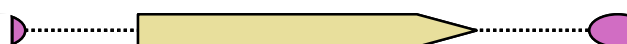
Consistent with patterns in CI cassette diversity observed previously (Gillings *et al.*, 2005; Rowe-Magnus *et al.*, 2001; Vaisvila *et al.*, 2001), vast majority of gene cassettes (79/84) recovered here contained ORFs with either no significant matches to proteins in public databases or gave matches to hypothetical genes only, as determined by Blastp searches (Table 5.2). Cassette ORFs which showed similarity to proteins of known or inferred function were limited to *Ps. stutzeri* BAM17 (Table 5.2). All of these genes were most closely related to genes found in *Ps. fluorescens* genome sequences. One gene is predicted to encode a ribonuclease inhibitor, and another a phosphatase. Two of the genes with a predicted function were found within a single gene cassette, and resemble a toxin/antitoxin module in the form of a pyocin gene and a pyocin immunity gene. Such toxin-antitoxin modules are common features of many prokaryotic genomes with roles in stability of extra-chromosomal elements and controlled cell death (Gerdes *et al.*, 2005). Several such modules have been identified in the cassettes of several *Vibrio* spp. integrons (Boucher *et al.*, 2006; Pandey and Gerdes, 2005; Rowe-Magnus *et al.*, 2003), and have been suggested to play a role in cassette array stabilisation (Rowe-Magnus *et al.*, 2003).

No two independent cassette arrays shared an identical gene cassette, however a pair of duplicated cassettes was observed in the cassette array of *Ps. mendocina* NW1 (data not shown). The duplicated cassettes were found at the 1<sup>st</sup> and 3<sup>rd</sup> positions in the cassette array and were 100% identical across their entire length. A total of five pairs of cassettes which contain related ORFs were recovered from different cassette arrays (Table 5.3). All related ORFs were identical in length and shared >80% pair-wise amino acid identity and >85% pair-wise nucleotide identity. For all but one pair of related cassette ORFs, >50% of nucleotide substitutions are synonymous, resulting in limited changes to the translated amino acid sequence

(Table 5.3). This observation provides indirect evidence that selection has operated on these ORFs to maintain the amino acid sequence and by inference the function of the protein. Most related pairs of cassettes were similar across their entire length, having similar 5'-be, non-coding regions and general organisation, in addition to similar ORFs (Table 5.3). Cassettes which exhibit similarity across their entire length are likely to represent homologous or paralogous entities, rather than the independent mobilisation of a homologous gene as a cassette on multiple occasions. However, two pairs of cassettes (InPst17589b [C2] & InPst19SMN4 [C5], and InPstQ [C10] & InPst19SMN4 [C3]) which were otherwise similar across their entire length (>85% nucleotide identity), contained 5'-be which were less similar than expected (<80% nucleotide identity, Table 5.3). The origin of the 5'-be and cassette ORF in these pairs of cassettes is questionable, and raises the possibility that the genes contained in these cassette pairs were mobilised as cassettes (ie. attached to a 5'-be) on multiple occasions. Alternatively, the 5'-be of these cassettes may be homologous entities in which the 5'-be have experienced a disproportionately high mutation rate.

Integron/ Strain	Position in array	ORF length	Species	Predicted function	ORF length	Identities	BLASTp e-value
InPstBAM17 8B	4th	95	<i>Ps. fluorescens</i> Pf-5	putative ribonuclease inhibitor	99	52/95 (55%)	$5 \times 10^{-25}$
InPstBAM17 8B	5th	114	<i>Ps. fluorescens</i> Pf-5	phosphatase	113	67/113 (59%)	$2 \times 10^{-34}$
InPstBAM17 8B	14th	86	<i>Ps. fluorescens</i> PfO-1	Colicin/pyocin immunity protein	84	51/84 (60%)	$6 \times 10^{-20}$
InPstBAM17 8B	14th	401	<i>Ps. fluorescens</i> PfO-1	S-type Pyocin	416	127/338 (37%)	$2 \times 10^{-41}$

Table 5.2 - Cassette associated genes which gave Blastp matches to genes of a known or predicted function (as indicated in Table 5.1).



Cassette 1	Cassette 2	5' spacer identity/size	ORF similarities			3' spacer identity/size	59be % identity/size
			Amino acid identities/size	Nucleotide identities/size	synonymous substitutions		
InPst17589b (C2)	InPst19SMN4 (C5)	90% (40)	92% (72)	94% (219)	45%	100% (4)	<b>76% (77)</b>
InPmeNW1 (C16)	InPstBAM17 (C1)	67% (6)	83% (67)	87% (204)	54%	N/A (-6)	95% (77)
InPstQa (C9)	InPstBAM17 (C8)	87% (15)	95% (226)	95% (681)	69%	N/A (-65)	97% (78)
InPstQ (C10)	InPst19SMN4 (C3)	85% (124)	90% (98)	90% (297)	61%	100% (3)	<b>77% (79)</b>
InPstrKM91a (C2)	InPstrKM91b (C1)	100% (6)	86% (136)	87% (411)	64%	100% (6)	95% (77)

Table 5.3 - Related cassettes associated with independent gene cassette arrays in *Pseudomonas* spp. integrons. Information in each column corresponds to the region of the gene cassette depicted above the table. 59-be identity values which were less than expected are indicated in bold type. Values in parentheses beneath each integron name indicate the position in the array of each gene cassette, i.e. – C5 refers the fifth cassette in an array.

### 5.3.2 - Recovery and analysis of integron flanking genes

The *intI* stop codon provides a sharp left-hand boundary of the integron, and thus any gene to the left of *intI* was considered a flanking gene. To the right of the integron, any gene not associated with a 59-be was considered a putative flanking gene, but further analysis is required as these genes may be associated with degraded 59-be. Each gene flanking an integron was characterised using Blastp searches of the predicted amino acid sequence, and was considered to be a typical chromosomal gene if homologues were present in >75% (11 of 14) of completely sequenced *Pseudomonas* spp. genomes, or if patterns in codon usage and G+C content were typical of *Pseudomonas* spp. chromosomal genes.

Including integron sequences recovered in the present study, evidence for 18 independent core integrons (Table 5.1) is apparent in *Pseudomonas* spp. on the basis of either partial or complete sequences of core integron components (*intI* or *attI*) that exhibit non-identical sequences. For all but one of the integrons recovered in the present study (InPstQb, see Figure 5.5), the locus was identified. The remaining 17 integrons or integron remnants occupied six different loci, based on the identity of upstream genes (Figure 5.5). Four of these loci were further defined by recovery conserved genes downstream of the integron. Genes characteristic of a chromosomal genomic context were recovered from the upstream or downstream region for all integrons except InPstQb. As described in Section 5.3.1.4, a suspected sequence terminator prevented recovery of additional upstream sequence data, and while several genes downstream of the last gene cassette were recovered, none was characteristic of a particular genomic context (Figure 5.5). Apparently non-integron-associated genes were frequently found between the last gene cassette and first

conserved flanking gene (Figure 5.5), and could not be unambiguously classified as being of integron or chromosomal origin. These genes are referred to in Figure 5.5 as 'unassigned' genes, and will hereafter be referred by this name.

For all but three of the flanking genes identified above, homologues were present in at least 11 of 14 completely sequenced *Pseudomonas* spp. genomes (Table 5.4). All of these genes, when present in a complete *Pseudomonas* spp. genome, existed as a single copy. Phylogenetic analyses performed on all genes present in >75% of sequenced *Pseudomonas* genomes revealed congruent phylogeny with respect to established relationships within the genus (see below for selected examples), suggesting that all of the integrons/integron remnants identified here are found at chromosomal loci. All *Ps. stutzeri* strains in the collection were screened for the presence of the flanking genes of integron locus 1 (PA0916 and PA3388) and 2 (PA3672 and PA3673), the details of which are presented in Appendix 1. All of these genes were detected in all *Ps. stutzeri* strains. These results further support the notion of stable chromosomal inheritance of the genes flanking integrons at locus 1 and 2 (Figure 5.5) in most Pseudomonads.

No evidence of an association with a larger mobile genetic element was observed for any of the integrons analysed. In all cases where genes upstream of the integron were recovered, no genes characteristic of an association with a mobile element were found between the integron and the first flanking gene. However, in the region downstream of the integron, or within the integron itself, several strains were found to contain IS or IS remnants (Figure 5.5), which raises the possibility of selective targeting of integron-associated sequences by IS elements. The discovery of several IS

elements found at specific sites within the 59-be of cassettes from *Ps. stutzeri* Gv.2 strains lends strong support to this possibility (S. Tetu, Unpublished data).

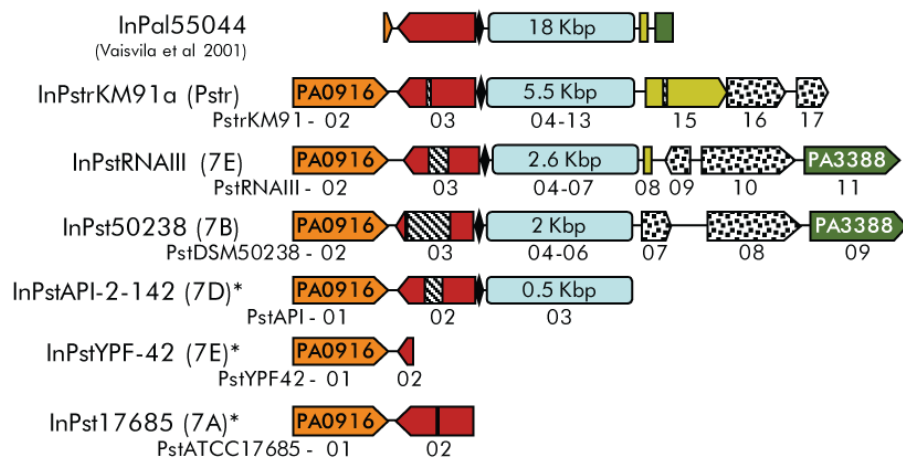
<i>Ps. aeruginosa</i> PAO1 gene	Locus	Flank	Amino acid identity to PAO1 <sup>1</sup>	Presence in other <i>Pseudomonas</i> genomes <sup>2</sup>	Conservation across all <i>Pseudomonas</i> <sup>3</sup>	Predicted function <sup>4</sup>
PA0916	1	5'	89%	14/14	77%	2-methylthioadenine synthetase
PA3388	1	3'	73%	14/14	73%	Conserved hypothetical protein
PA3672	2	5'	75%	6/14	75%	ATP-binding component of ABC transporter
PA3673	2	3'	75%	13/14	75%	Glycerol-3-phosphate acyltransferase
PA2414	3	5'	50%	13/14	49%	L-sorbose dehydrogenase
PA3994	4	5'	61%	5/14	53%	Hydrolase or acyltransferase
PA3733	4	3'	54%	11/14	53%	Acyl-CoA transferase/carnitine dehydratase
PA1859	5	5'	62%	11/14	56%	Transcriptional regulator
PA1561	6	5'	77%	13/14	77%	Aerotaxis receptor Aer
PA2496	6	3'	67%	4/14	67%	Conserved hypothetical protein

Table 5.4 – Homologues of genes flanking CIs recovered in the present study identified in the *Ps. aeruginosa* PAO1 genome. The coloured blocks shown next to each gene correspond to the colouring used for each of these genes in Figure 5.5.

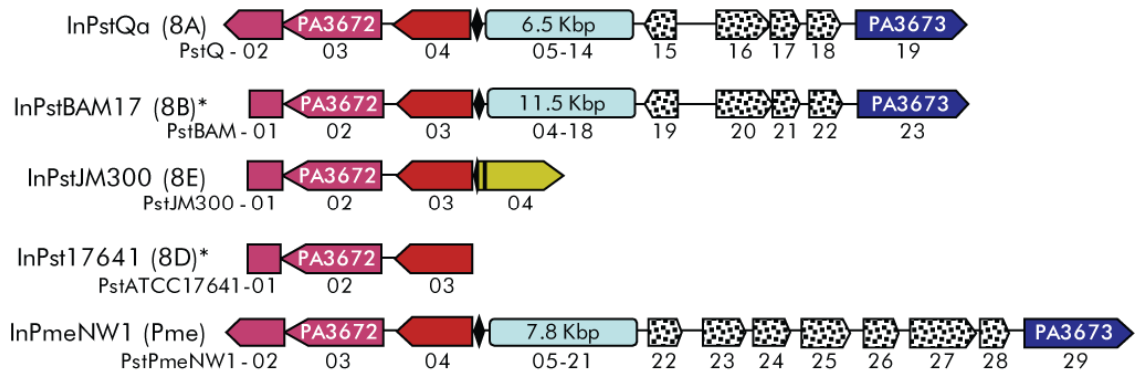
- <sup>1</sup> Values given refer to the range of conservation observed across all sequences recovered in the present study relative to *Ps. aeruginosa* PAO1.
- <sup>2</sup> Values given indicate the proportion of *Pseudomonas* spp. complete genomes (14 in total) that contain a gene homologous to the relevant flanking gene.
- <sup>3</sup> Values given indicate the minimum level of conservation for the relevant gene across all *Pseudomonas* genomes, in addition to sequences recovered in the present study.
- <sup>4</sup> Predicted functions listed correspond to those given in *Pseudomonas* spp. genome sequences.



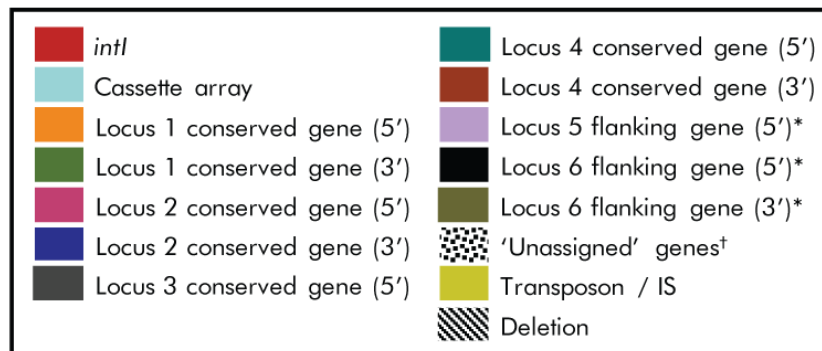
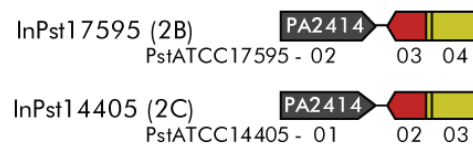
## a. Integron Locus 1



## b. Integron Locus 2



## c. Integron Locus 3



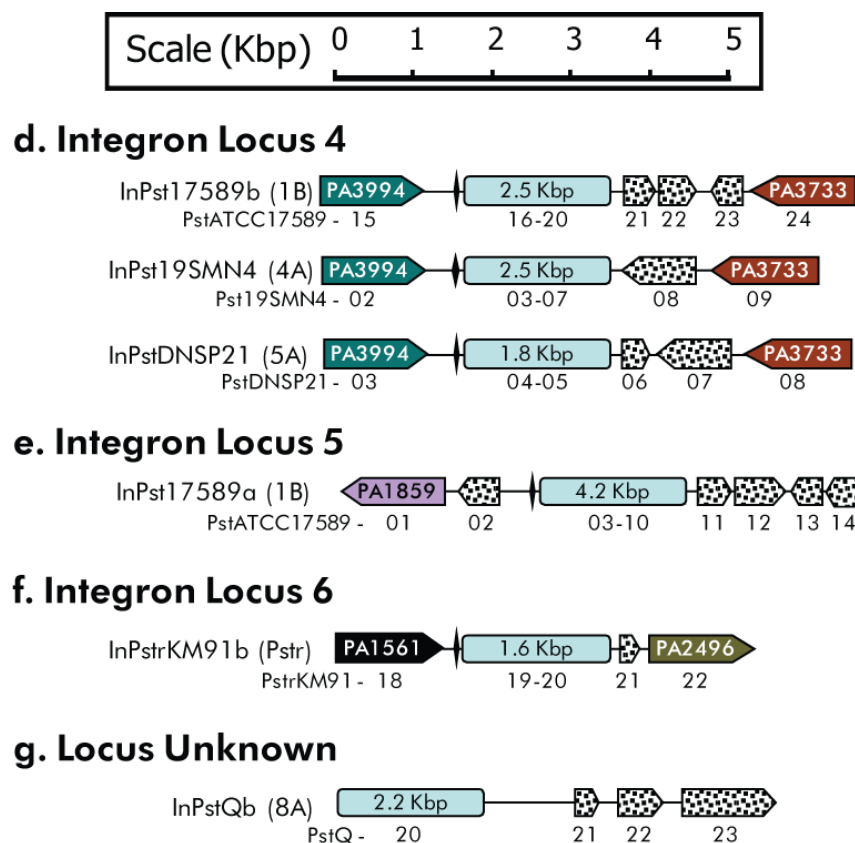


Figure 5.5 (above and opposite) - Schematic representation of integrator and flanking sequences recovered. Six integrator loci were identified (a-f). Each locus will hereafter be referred to according to the numbering system above. Homologous genes are colour-coded according to the legend opposite. The size of each cassette array is given inside the light blue boxes, and all other genes are drawn relative to the scale given above. The codes and numbers below each diagram indicate the reference used for the sequence of each gene provided in Appendix 2 (Filename – 'sequence maps\sequence diagrams.ppt'). All strain names labelled with an '\*' indicate sequences that were recovered or extended by PCR. All remaining sequences were generated from fosmid clone templates. Genes exhibiting homology to genes in *Ps. aeruginosa* PAO1 are indicated by numbers with a 'PA' prefix.

\* Flanking genes for integrator loci 5 and 6 are not known to be conserved, as each was represented by a single example. Rather, flanking genes were identified on the basis of similarity to chromosomal genes in *Pseudomonas* spp. genome sequences.

† 'Unassigned genes' refer to genes which could not be unambiguously classified as being either integrator-associated (lack of defining gene cassette features) or chromosomally associated (lack of homologues in *Pseudomonas* spp. complete genomes).

### 5.3.2.1 – Characterisation of integrons at locus 1

Integrans from *Ps. stutzeri* Gv.7 strains (InPstRNAIII and InPst50238) and *Ps. straminea* KM91 (InPstrKM91) were found to have the same upstream flanking gene as the integron described in *Ps. alcaligenes* ATCC 55044 (InPal55044) (Vaisvila *et al.*, 2001). The gene found immediately downstream of InPal55044 was also found downstream of InPstRNAIII and InPst50238, but not InPstrKM91. In InPal55044, the last gene cassette is separated from this gene by a non-coding sequence of 330 bp, and the region closest to the cassette array exhibited 92% (196/211) nucleotide identity to the 5' end of an ISPa21-like element in *Ps. aeruginosa* 257. This insertion sequence is inserted into an aadB gene cassette which is part of the cassette array of a class 1 integron in this strain. Several other class 1 integrons in *Ps. aeruginosa* have also been found to contain ISPa21-like elements (Poirel *et al.*, 2005). The corresponding region downstream of InPst50238 and InPstRNAIII, in contrast, was considerably larger, and contained several genes apparently not associated with the integron (Figure 5.5). A similar ISPa21-like remnant as that found after the last cassette of InPal55044 was also found immediately after the last cassette of InPstRNAIII (95% [187/197] pair-wise nucleotide identity), but was not detected in InPst50238. The extent of the IS remnant is the same in both *Ps. stutzeri* RNAIII (7E) and *Ps. alcaligenes* ATCC55044, suggesting that these remnants are homologous entities and were inactivated in a single event that occurred in a common ancestor of these strains. A putative tryptophan synthase gene was identified immediately upstream of the PA3388 homologue in both InPstRNAIII and InPst50238 (pair-wise amino acid identity 99.2% [390/393]). In both integrons approximately 1 Kbp of sequence separates the tryptophan synthase gene and the last gene cassette in the array; however, no regions of homology were detected. Several downstream, non-

integron genes were recovered from *Ps. straminea* KM91, however the homologue of PA3388 found downstream of other integrons at this locus was not observed (Figure 5.5). This gene was not detected in *Ps. straminea* KM91 by any molecular tests employed (see appendix 1 – A.3.5), and suggests that this gene has been lost from this strain.

#### 5.3.2.2 – Characterisation of integrons at locus 2

Integrons from *Ps. stutzeri* Q (InPstQa), *Ps. stutzeri* BAM17 (InPstBAM17) and *Ps. mendocina* NW1 (InPmeNW1) were found at a different locus sharing both upstream and downstream conserved flanking genes (Figure 5.5). Several apparently non-integron genes were found between the last cassette and the PA3673 homologue. *Ps. stutzeri* Q and BAM17 contained the same genes in the same arrangement in this region; however, none exhibited significant homology to genes in any published *Pseudomonas* spp. genome (see electronic appendix, filename – ‘sequence maps\sequence diagrams.ppt’ for details of Blastp results for these genes). Pair-wise alignments of the corresponding genes in this region revealed a lower similarity (average pair-wise similarity – 92.9%) relative to the identity of *intI*, PA3672 and PA3673 (average pair-wise similarity – 97.9%) between these strains (Figure 5.6). Interestingly, the identity between the genes in this region decreases as distance from the cassette array increases (Figure 5.6). The significance of this, if any, is unknown. The corresponding region between the last gene cassette and the PA3673 homologue in *Ps. mendocina* NW1 was considerably larger, spanning 4.7 Kbp, and contained a completely different suite of genes (Figure 5.5). A total of seven putative ORFs were identified in this region, none of which were homologous to genes found

in *Pseudomonas* spp. genome sequences. Some of these genes exhibited significant homology to genes found in non-*Pseudomonas* genomes, the details of which are given in the electronic appendix (Filename – 'sequence maps\sequence diagrams.ppt').

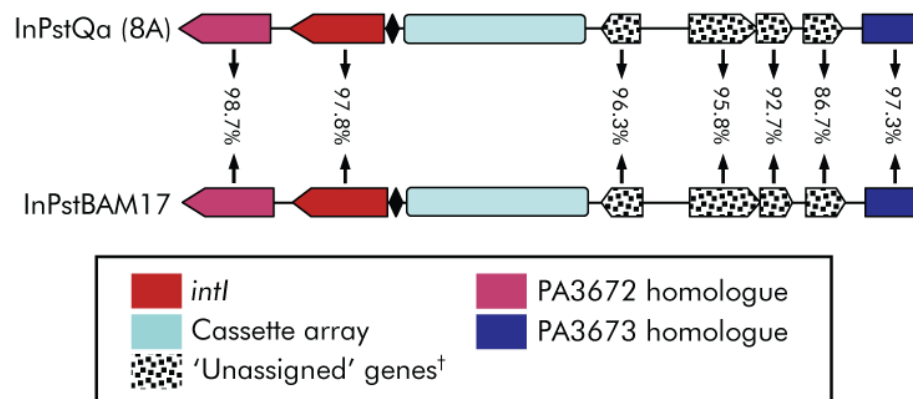


Figure 5.6 - Pair-wise amino acid identities for integron-associated genes and flanking genes of InPstQ and InPstBAM17. Cassette arrays were not included in the comparison as no genes are shared between the two integrons.

<sup>†</sup> 'Unassigned genes' refer to genes which could not be unambiguously classified as being either integron-associated (lack of defining gene cassette features) or chromosomally-associated (lack of homologues in *Pseudomonas* spp. complete genomes).

### 5.3.2.3 – Characterisation of integrons from remaining loci

Two *Ps. stutzeri* Gv.2 strains (2B and 2C) were found to contain an integron remnant consisting of a partial *intI* gene at a common locus (Locus 3 in Figure 5.5). The gene immediately downstream of *intI* exhibited homology to gene PA2414 of *Ps.*

*aeruginosa* PAO1. As described in Section 5.3.1.3, no additional integron sequences were recovered from these strains. Strains belonging to *Ps. stutzeri* Gv.1 (1B), Gv.4 (4A) and Gv.5 (5A) were also found to contain an integron remnant at a common locus (Locus 4 in Figure 5.5). All contained the same upstream flanking gene which exhibited homology to PA3994 from *Ps. aeruginosa* PAO1. This gene was separated from the first identifiable integron-associated sequence by a non-coding sequence of >200 bp in all strains. All of these strains also contained a common gene downstream of the integron remnant which was homologous to PA3733 of *Ps.*

*aeruginosa* PAO1. As observed in integrons at loci 1 and 2, integrons at this locus contained several apparently non-integron genes between the last gene cassette and the PA3733 homologue. None of these genes were shared between integrons of different strains and none exhibited significant similarity to genes in published *Pseudomonas* spp. genomes (Details of Blastp results for each of these genes is given in the electronic appendix, filename – ‘sequence maps\sequence diagrams.ppt’).

Integrons found at loci 5 and 6 (as defined in Figure 5.5) were each represented by a single example. Consequently, flanking genes which were conserved between different integrons could not be identified. Flanking genes in these strains were instead identified on the basis of homology to genes in *Pseudomonas* spp. genome sequences. Homologues of the gene immediately upstream of InPst17589a were not identified in any *Pseudomonas* spp. genomes, although the second upstream gene

was found to be homologous to PA1859 in the *Ps. aeruginosa* genome (Figure 5.5). Several genes after the last gene cassette of InPst17589a were recovered; however, none displayed homology to genes in *Ps. aeruginosa* PAO1, or any other *Pseudomonas* spp. genome (Figure 5.5). Homologues of genes upstream and downstream of InPstrKM91b were also detected in *Ps. aeruginosa* PAO1 (Figure 5.5). A single gene was identified between the last cassette and the PA2496 homologue, and did not exhibit homology to genes in complete *Pseudomonas* spp. genomes.



### 5.3.3 - Phylogenetic analysis of integron flanking genes

Phylogenetic trees constructed from PA0916 and PA3673 homologues are presented in Figures 5.7a and 5.7b, respectively. In both trees, *Pseudomonas* spp. strains form well-supported monophyletic clades with respect to other members of the gene family. *Ps. stutzeri* strains also form well supported clades with the *Pseudomonas* spp. lineage in both gene phylogenies. These observations are consistent with the phylogenetic diversity of 16S rRNA genes within the *Pseudomonas* lineage (Figure 5.7c). Several deeper branching relationships were also present in trees constructed from both PA0196 and PA3673 homologues; for example, *Ps. syringae*, *Ps. fluorescens*, *Ps. putida* and *Ps. entomophila* strains consistently group together. In the 16S rDNA tree (Figure 5.7c), *Ps. putida* and *Ps. entomophila* occur in a different clade to *Ps. syringae* and *Ps. fluorescens*; however, this branch received low bootstrap support and all these strains grouped together in other possible configurations of the tree. These data indicate that PA0916 and PA3673 homologues are typical of *Pseudomonas* chromosomal genes, and were likely to have been present in the last common ancestor of these strains. Collectively, the data presented in this section indicate that integrons in *Pseudomonas* spp. occur or have occurred at six or more independent loci, are not specifically associated with larger genetic elements predicted to confer mobility, and have conserved flanking genes which are characteristic of *Pseudomonas* spp. vertically transmitted genes.

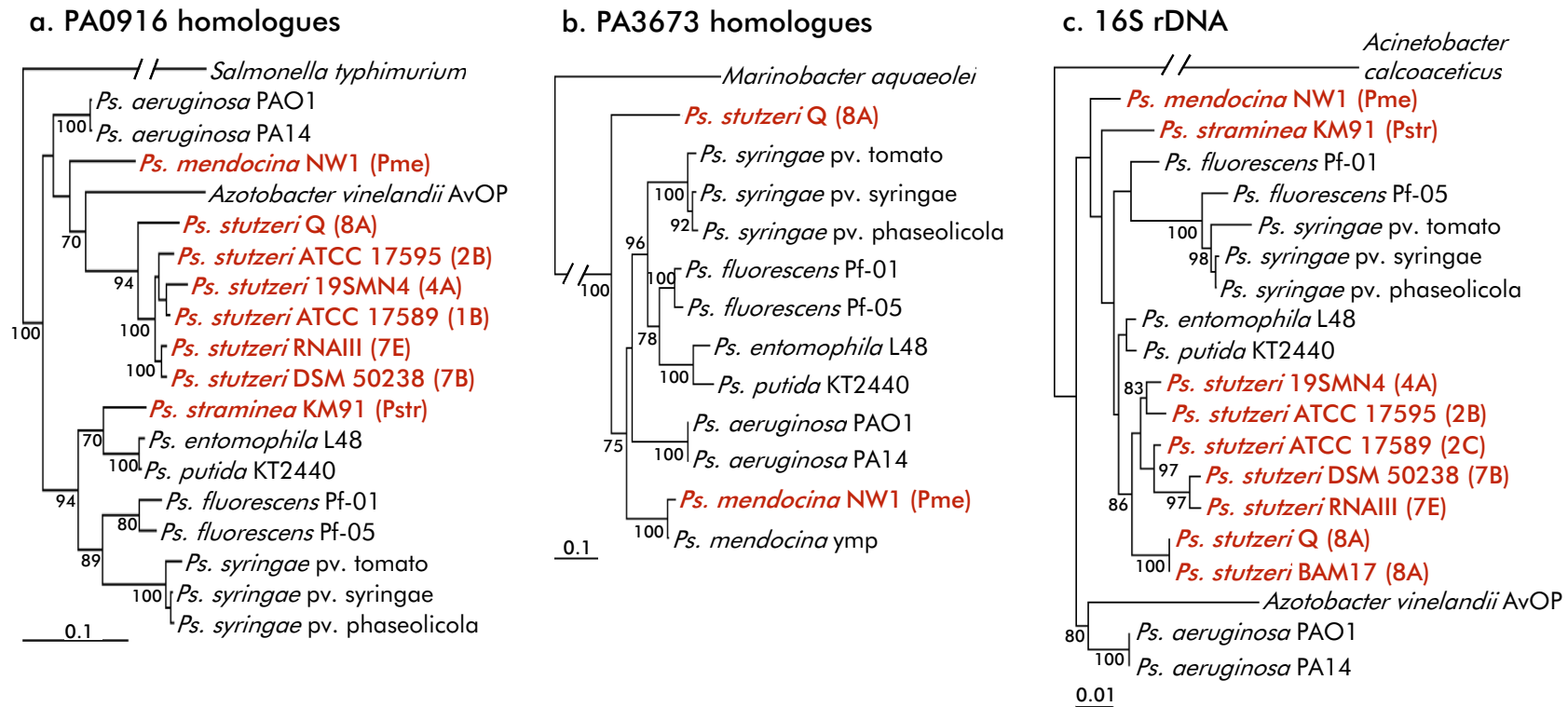


Figure 5.7 - Phylogenetic relationship between homologues of PA0916 (a.) and PA3673 (b.) relative to phylogenetic relationships determined by analysis of 16S rDNA (c.). The trees shown in a. and b. were constructed using neighbour joining from alignments spanning 443 and 324 amino acids, respectively. The tree shown in c. represents the best maximum likelihood tree constructed from alignment of 16S rDNA sequences spanning approximately 1300 nucleotides. Bootstrap values from 1000 iterations are given for all bifurcations which received >80% support. Strains highlighted in red indicate strain analysed in the present study. The scale bars for trees a. and b. represent 10% sequence divergence and the scale bar for tree c. represents 1% sequence divergence. The sequences and accession numbers of all sequences used to construct trees a. – c. are provided in the Appendix (Filenames: PA0916 – ‘PA0916 sequences.fas’; PA3673 – ‘PA3673 sequences.fas’; 16S rDNA – ‘16S sequences.fas’).

#### 5.3.4 - Phylogenetic analysis of *intI* relative to genomic context

Of all complete *Pseudomonas* subfamily *intI* genes, only four are uninterrupted and predicted to encode functional integrases (intIPstQ, intIPstBAM, intIPal55044, and intIPmeNW1). Several *intI* genes recovered were truncated, containing deletions of various sizes. Two additional partial *intI* genes were recovered using PCR and sequence information outside the primer target sites was not recovered. The limited amount of sequence information available for many *Pseudomonas* spp. *intI* genes creates problems when conducting phylogenetic analyses due to the lack of common regions of overlapping sequence (Figure 5.8). No single region of *intI* is universally present across the current dataset. To allow the evolutionary history of *intI* genes from all strains to be modelled separate trees from different sections of the alignment were constructed. The strains included in each tree were determined on the basis of available *intI* sequence data for each strain, as indicated in Figure 5.8. Of all the strains indicated in Figure 5.8, the *intI* gene of *Ps. stutzeri* API-2-142 (7D) was the only gene not included in phylogenetic analyses as the extent of the deletion in this gene resulted in insufficient remaining sequence for robust phylogenetic inferences to be made. However, this gene is clearly most similar to other *Ps. stutzeri* Gv.7 *intI* genes (see Table 4.3), and clustered with members of this genomovar in other phylogenetic analyses performed (data not shown).

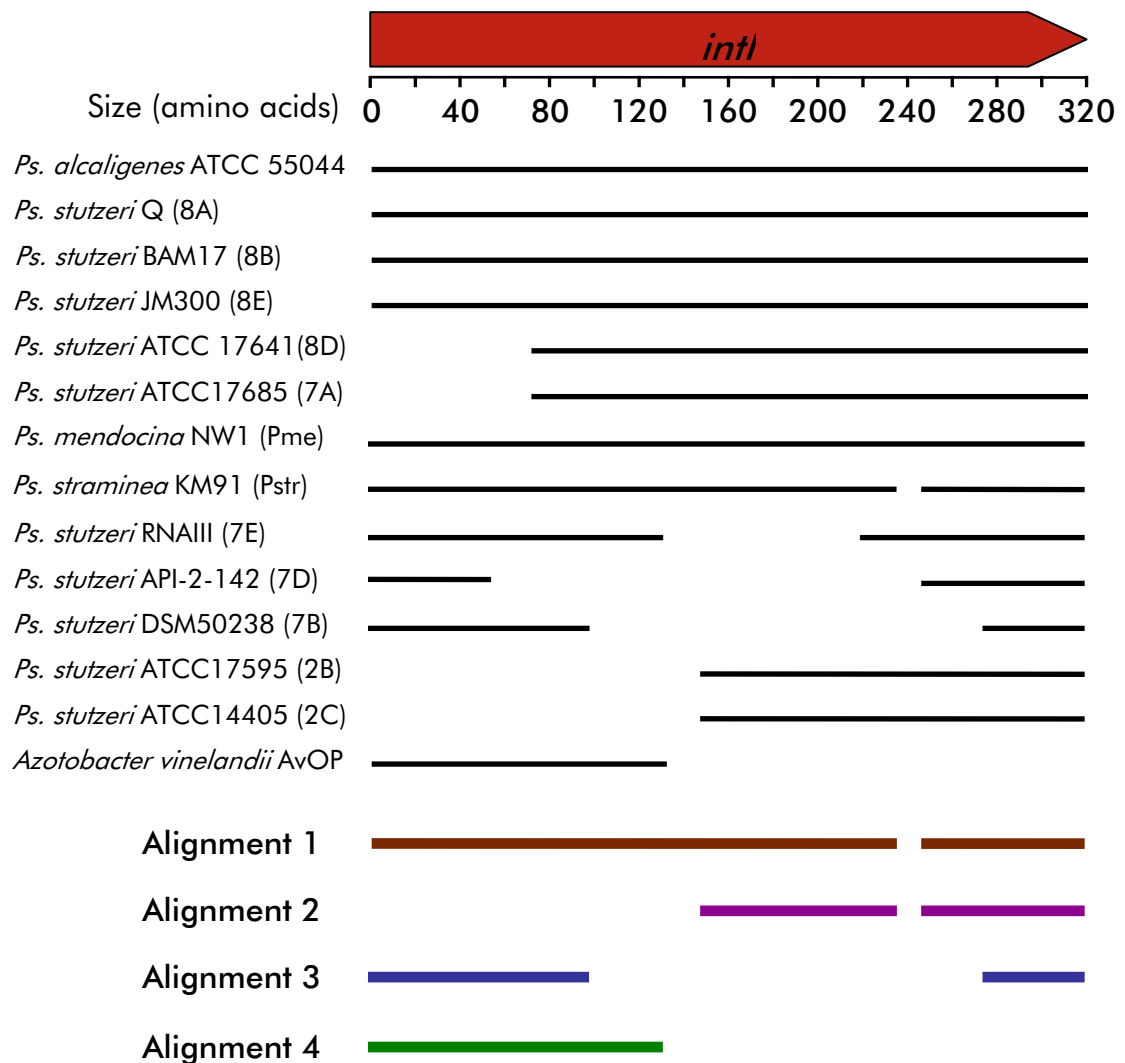


Figure 5.8 - Extent of *intI* gene sequence information available for *Pseudomonas* subfamily integrons. The line next to each strain represents the amount and the relative position of sequence information, according to the scale given above. The extent of different alignments used to investigate phylogenetic diversity in these genes is indicated in the lower section of the figure. For each alignment, only those strains which contain all the sequence indicated were included. The extent of each alignment set depending on aims and strains to be included: Alignment 1 – Longer alignment used to give greater confidence in tree topology, Alignment 2 – Phylogenetic characterisation of *Ps. stutzeri* Gv.2 *intI* genes, Alignment 3 – Phylogenetic characterisation of *Ps. stutzeri* Gv.7 *intI* genes, and Alignment 4 – Phylogenetic characterisation of *intI* gene of *A. vinelandii* AvOP. Partial *intI* sequence of *A. vinelandii* accessed from an incomplete genome sequence.

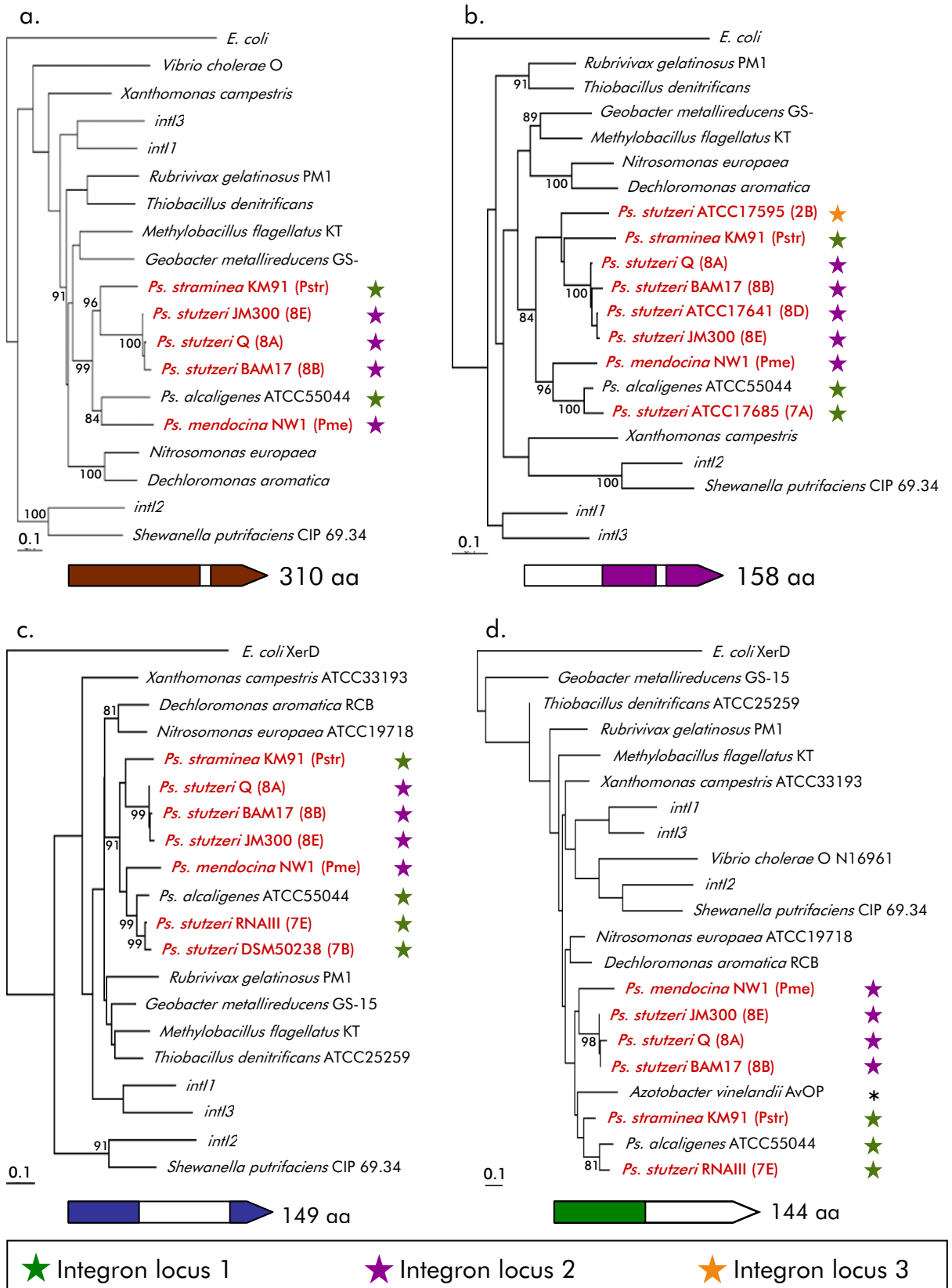
Phylogenetic trees constructed from alignments of the regions of *intl* indicated in Figure 5.8 are shown in Figure 5.9a-d. In all but one tree, *Pseudomonas* spp. *intl* genes form well supported monophyletic clades (receiving >85% bootstrap support) with respect to other known *intl* genes (Fig 5.9a-c). The *Pseudomonad* Intl clade received lower bootstrap support when only the 5' region of the gene was used to construct the tree (Figure 5.9d). This may be due to inclusion of the partial Intl sequence of *Azotobacter vinelandii* AvOP and/or loss of phylogenetic information resulting from the region of Intl used to construct the tree. Particular groups of strains clustered together in all trees. For example, *Ps. stutzeri* Gv.8 strains always clustered together (which is not surprising considering the high level of relatedness between these strains), and *Ps. stutzeri* Gv.7 and *Ps. alcaligenes* strains also occurred on the same branch when present in the same tree. The branching pattern within the *Pseudomonas* spp. clade did not mirror the patterns observed using other phylogenetic marker sequences (16S rRNA and IGS1) in that *Ps. stutzeri* Gv.7 and Gv.8 strains did not cluster together within the *Pseudomonas* Intl clade. In all of the trees shown in Figures 5.9a-d, *Ps. stutzeri* Gv.7 and Gv.8 strains occur in separate clades that receive >85% bootstrap support. This was consistent with earlier observations of Intl amino acid pair-wise identities in these strains (see Table 4.3).

Limited congruence of integron locus with Intl phylogeny was observed; clustering of *Ps. stutzeri* Gv.7 strains with *Ps. alcaligenes* ATCC 55044 was consistent with the locus of these integrons (Figures 5.9b-c). However, significant phylogenetic incongruence was evident in integrons at both loci 1 and 2; *Ps. straminea* KM91 (locus 1) formed a well supported clade (96% bootstrap support) with *Ps. stutzeri* Gv.8 strains (locus 2), while *Ps. mendocina* NW1 (locus 2) formed a well supported clade (84% bootstrap support) with *Ps. alcaligenes* ATCC 55044

(locus1). Thus, the current dataset does not support monophyly of the core integrons of CIs at particular loci. If an integron at a particular locus represents the single acquisition of an ancestral integron and subsequent stable vertical transmission, then integrons at a particular locus should form monophyletic clades in phylogenetic analyses. As this was not the case in the data presented in Figure 5.9, the observed phylogenetic diversity among *Pseudomonas* Intl in the current dataset can be parsimoniously explained only by independent acquisition of CIs at the same locus.

Figure 5.9 a.-d. (opposite) - IntI phylogeny relative to chromosomal locus. Each tree was constructed using neighbour-joining and a bootstrap analysis of 1000 iterations. Bootstrap values (as percentages) are given for all bifurcations which were present in >80% of all possible trees. The region of IntI included and the number amino acids in each alignment is indicated under each tree. Particular strains were included in a tree when *intI* sequence information was available across the region of *intI* being analysed. The locus of each integron is indicated according to the legend opposite.

\* - The locus of the integron in *Azotobacter vinelandii* AvOP is unknown as the *intI* gene occurs at the end of an incomplete shotgun genome sequence.

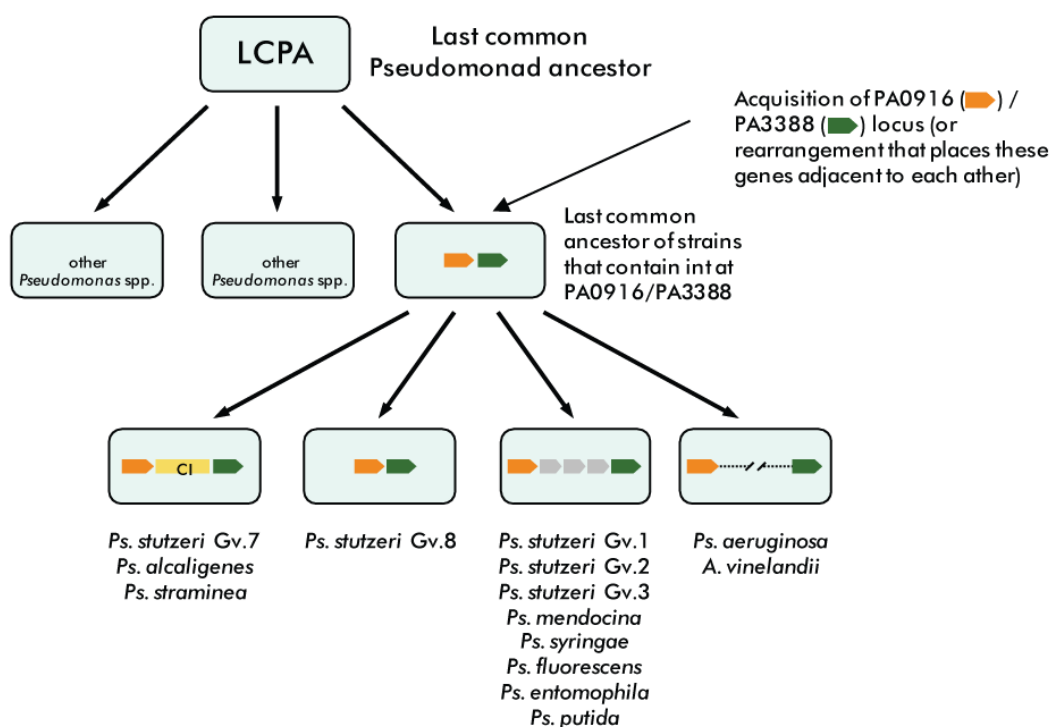




### 5.3.5 - Synteny and ancestry of integron loci across *Pseudomonas* spp.

On the basis of the arrangement and phylogenetic diversity of PA0916 and PA3388 homologues in extant *Pseudomonas* spp., a hypothetical evolutionary history of the association between these genes is outlined in Figure 5.10a. It was hypothesised that homologues of PA0916 and PA3388 were physically linked in the ancestral Pseudomonad that acquired an integron at locus 1 (Figure 5.10a). These genes were expected to remain in close proximity within the genomes of most extant Pseudomonads. However, it is also likely that synteny in these genes has been lost in some extant Pseudomonads, due the frequency of large scale genomic rearrangements in some members of this lineage (Ginard *et al.*, 1997; Heuer *et al.*, 1998; Rainey *et al.*, 1994; Romling *et al.*, 1997; Sawada *et al.*, 2002). Thus, it was hypothesised that a monophyletic group now exists within the *Pseudomonas* genus in which some lineages had the PA0916/PA3388 gene arrangement and retained it, some had it but have now diverged, being interrupted by genomic rearrangements or gene insertions, and some have acquired an integron in it (Figure 5.10a). Based on the tree in Figure 5.7a, the monophyletic group which acquired an integron consists of all *Pseudomonas* strains from which this gene is known, as strains which contain an integron at this locus are present in both major clades within the tree (*Ps. straminea* KM91 in one clade, and *Ps. stutzeri* Gv.7 strains in the other). One of two explanations for this observation is likely: 1. that an integron was present in the common ancestor of the strains in the PA0916 phylogenetic tree in Figure 5.7a, and was subsequently lost from most extant strains, or 2. that multiple independent acquisitions of CIs have occurred at this locus.

### a. Locus 1 – PA0916 and PA3388 homologues



### b. Locus 2 – PA3672 and PA3673 homologues

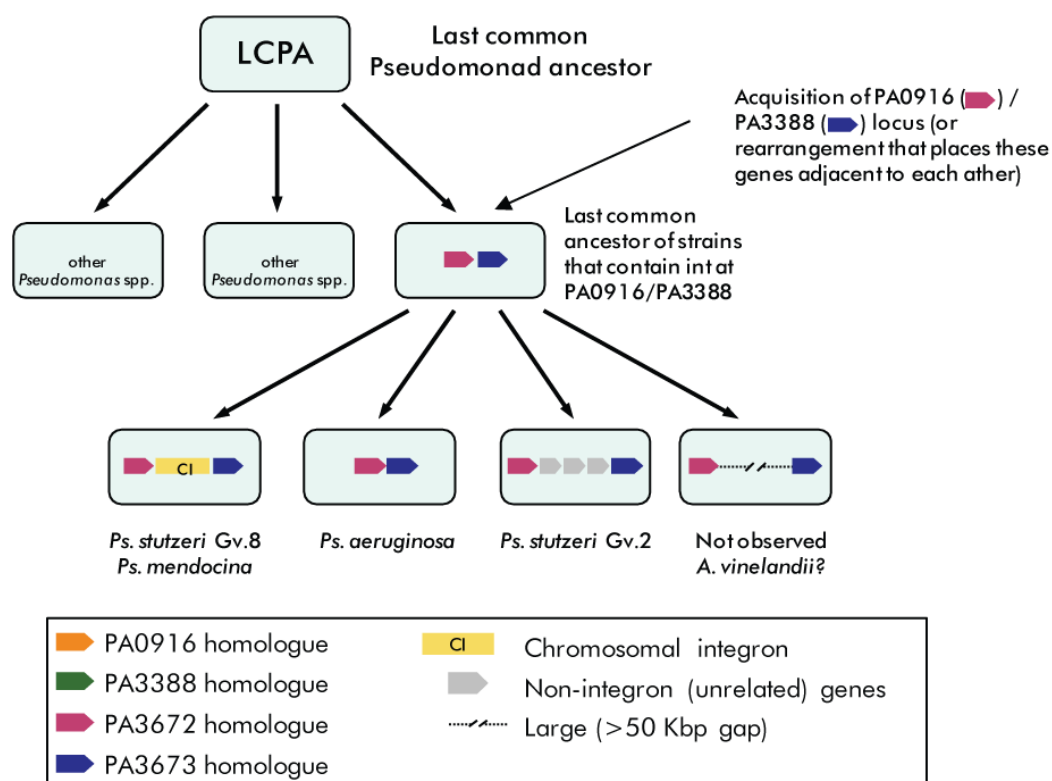


Figure 5.10 - Hypothetical association between the genes of integron locus 1 (a.) and 2 (b.) in the common ancestor of extant *Pseudomonas* spp. strains. The genes of each locus were hypothesised to have been linked in the common ancestor of extant strains which contain an integron at the locus. Diversification of each locus has led to multiple relative arrangements of these genes in different lineages.

Homologues of PA0916 and PA3388 were found to exist in close proximity in several *Pseudomonas* spp. strains (Figure 5.11). In the *Ps. aeruginosa* PAO1 and *Azotobacter vinelandii* AvOP genomes, PA0916 and PA3388 are located several megabases apart, suggesting that a large scale genomic rearrangement has placed these genes on opposite sides of the genome. In *Ps. putida*, *Ps. fluorescens*, *Ps. syringae* and *Ps. entomophila* strains PA0916 and PA3388 homologues are found within 5 Kbp of each other, which is consistent with the insertion of multiple genes at this locus. A similar pattern of gene arrangement was observed in several strains analysed in the present study. Homologues and PA0916 and PA3388 were found in close proximity (<30 Kbp apart) to *Ps. stutzeri* Gv.2 strains and *Ps. mendocina* NW1 (Table A.3), which is again consistent with the insertion of multiple genes at this locus. An example of the hypothetical ancestral locus described in Figure 5.10a was observed in *Ps. stutzeri* Q (8A). The presence of this gene arrangement in all other *Ps. stutzeri* Gv.8 strains was confirmed using a PCR assay which amplified the intergenic spacer between PA0916 and PA3388 (data not shown). The number and/or sequence of insertion events at this locus cannot be unambiguously determined with the current dataset. However, the variation in gene content which exists at the PA0916/PA3388 locus in *Pseudomonas* spp. strains strongly suggests that this locus is a hotspot for recombination in this genus. Data on the synteny of other Pseudomonad genes is required to support this assertion.

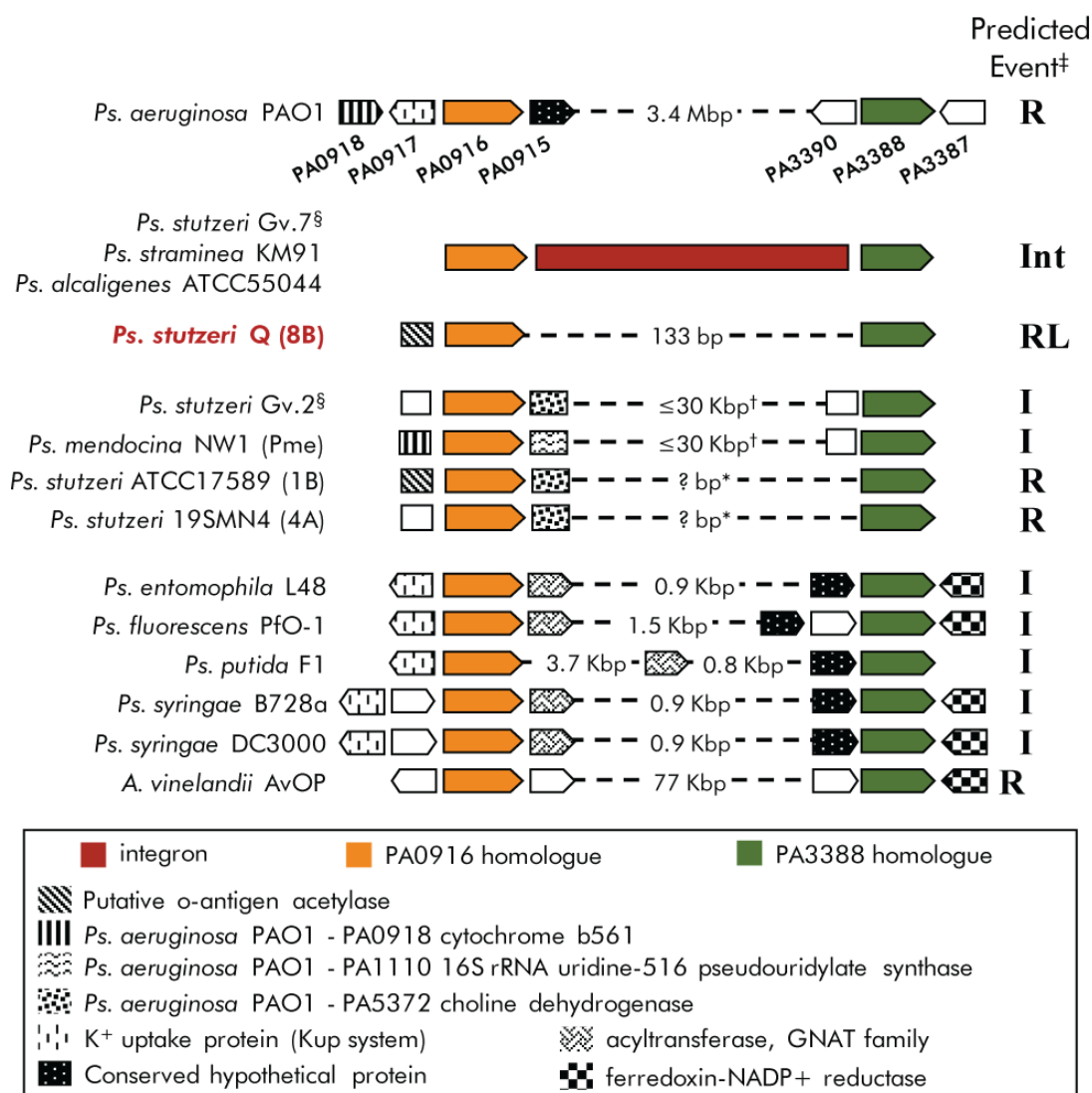


Figure 5.11 - Synteny at integrin locus 1 across members of the strain collection and other *Pseudomonad* genomes. Genes conserved between different strains are colour-coded according to the legend above, while genes which are not homologous to any other genes shown are coloured white. Genome reference codes are given beneath each *Ps. aeruginosa* PAO1 gene. Genes are not drawn to scale.

<sup>§</sup> *Ps. stutzeri* Gv.7 refers to strains DSM50238 and RNAIII, and *Ps. stutzeri* Gv.2 refers to strains ATCC17595 and ATCC14405.

\* As flanking genes were not detected in the same fosmid for these strains, the distance between them could not be determined.

<sup>†</sup> Maximum distance between genes known for these strains as both genes were detected within single fosmids (see Table A.3).

<sup>‡</sup> Letters to the right of each strain indicates the event(s) predicted to have given rise to the observed gene organisation: R – Large-scale recombination, Int – Integrin acquisition, RL – Retention of ancestral locus, I – Gene insertion.

A hypothetical evolutionary history of the conserved flanking genes of integrons locus 2 (PA3672 and PA3673) is represented diagrammatically in Figure 5.10b. Homologues of PA3672 were not present in any complete *Pseudomonas* genomes other than strains of *Ps. aeruginosa*. However, a PA3673 homologue was identified in 13/14 complete *Pseudomonas* spp. genome sequences (Table 5.4). In addition to strains which contain an integron at this locus (*Ps. stutzeri* Gv.8 strains and *Ps. mendocina* NW1), co-localised homologues of PA3672 and PA3673 were detected in several other *Ps. stutzeri* strains (Table A.3), and were recovered from *Ps. stutzeri* ATCC 17595 (2B) (Figure 5.12). Using a long-range PCR, the distance between the PA3672 and PA3673 homologues in *Ps. stutzeri* ATCC 17595 (2B) was estimated to be approximately 10 Kbp (data not shown). The hypothetical ancestral gene arrangement of PA3672/PA3673 was observed in *Ps. aeruginosa* PAO1 (Figure 5.12). The gene arrangement in *Ps. stutzeri* Gv.8 and *Ps. mendocina* strains is consistent with the insertion of an integron at this locus after divergence of these strains from *Ps. aeruginosa*. The arrangement of PA3672 and PA3673 homologues in *Ps. stutzeri* Gv.2 strains is consistent with the insertion of multiple genes at this locus in one or more insertion events. These results are consistent with the presence of the PA3672/PA3673 locus in common ancestor of *Ps. stutzeri*, *Ps. mendocina* and *Ps. aeruginosa*. The ancestral gene arrangement has remained in *Ps. aeruginosa*, while the acquisition of multiple genes has separated the PA3672 and PA3673 homologues in *Ps. stutzeri* and *Ps. mendocina*. However, the limited number of strains in the present dataset for which gene order of PA3672 and PA3673 homologues is known means the independence of these acquisitions cannot be unambiguously determined.

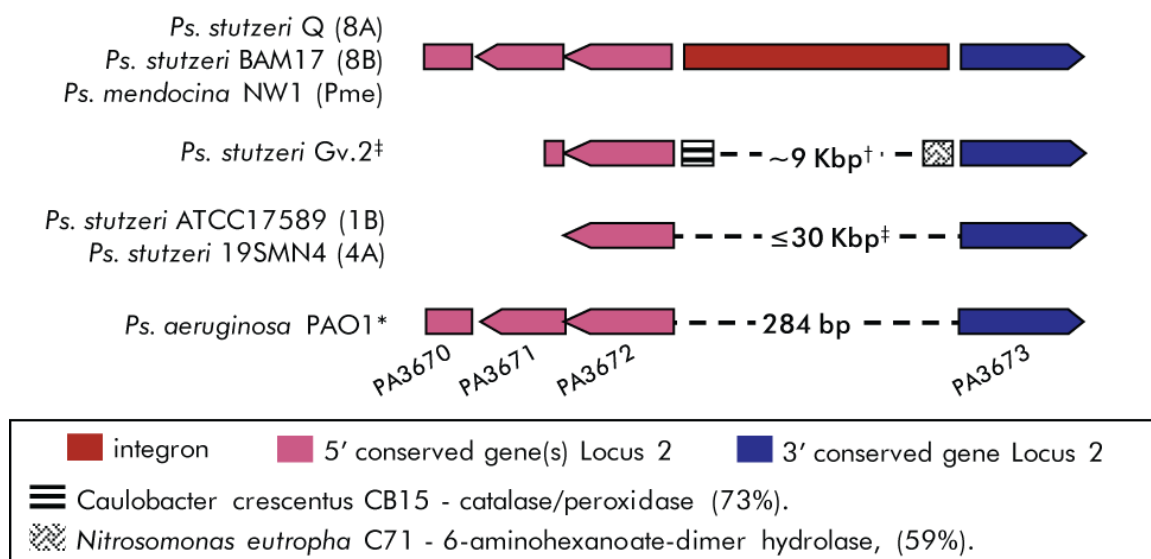


Figure 5.12 – Synteny at integrin locus 2 across members of the strain collection and sequenced *Pseudomonas* spp. genomes. Of the sequenced *Pseudomonas* genomes, homologues of PA3672 and PA3673 were present only in *Ps. aeruginosa*, and all sequenced *Ps. aeruginosa* genomes contained these genes in the same order as *Ps. aeruginosa* PAO1 (data not shown). Genes that have been characterised in previous sections are colour-coded according to the legend above. All remaining genes are coded in monochrome, and the top Blastp hit and predicted function of each is given in the legend above.

§ *Ps. stutzeri* Gv.7 refers to strains DSM50238 and RNAIII, and *Ps. stutzeri* Gv.2 refers to strains ATCC17595 and ATCC14405.

\* Gene organisation based on detection of both flanking genes within single fosmid clones. No sequence information for these genes was generated for these strains.

† Determined by PCR.

‡ Determined by PCR.

### 5.3.6 - Analysis of integron boundaries

The non-coding sequences immediately upstream (between *intI* and the first upstream flanking gene) of all integrons were analysed for the presence of repeat sequences indicative of an association with a larger mobile genetic element using Blastn searches. The non-coding regions immediately upstream of integrons from loci 1 and 2 were also examined for the presence of sequences conserved between different integrons using nucleotide alignments. The non-coding region immediately upstream of integrons at locus 3 were not analysed by alignment as both strains which contain integrons at this locus belonged to *Ps. stutzeri* Gv.2 and exhibited insufficient variation for the identification of conserved sequence motifs. As an *intI* gene was not identified in integrons from loci 4-6, a precise upstream non-coding sequence could not be defined, and these sequences were also not analysed using sequence alignments.

An alignment containing the intergenic spacer sequence between *intI* and the first upstream flanking gene was constructed for all strains with an integron at locus 1 (Figure 5.13). Not surprisingly, members of *Ps. stutzeri* (Gv.7) shared the highest level of similarity across this region (91%-100% nucleotide identity, Figure 5.13), as these strains form a related subgroup with respect to the other strains in the alignment (see Chapter 3). The non-coding region upstream of InPal55044 also exhibited a significant level of similarity to the corresponding region in *Ps. stutzeri* Gv.7 strains (68%-74% nucleotide identity, Figure 5.13). In contrast, this region is not well conserved in sequence upstream of InPstrKM91a, and conserved stretches of sequence are limited to the 5' and 3' ends of the alignment which are located within PA0916 and *intI*, respectively (Figure 5.13).

No regulatory sequences were expected to reside in this region as both adjacent genes terminate, rather than start, at the junction. In light of this, it is intriguing

that this spacer region is highly conserved between different *Pseudomonas* spp. strains (*Ps. stutzeri* and *Ps. alcaligenes*), as this region would be expected to experience neutral mutation pressures. Pair-wise alignments of the last 100 bp of nucleotide sequence of PA0916 and *intI* for integron from *Ps. stutzeri* Gv.7 strains and *Ps. alcaligenes* further highlights this, as the level of similarity between these genes (75-95% pair-wise nucleotide identity, data not shown) was comparable to that observed within the intergenic spacer. Within the *intI*/PA0916 intergenic spacer of *Ps. stutzeri* Gv.7 strains and *Ps. alcaligenes*, a conserved inverted repeat sequence characteristic of a REP element was located immediately adjacent to the *intI* stop codon. The first nucleotide of the 3' portion of the REP element is also the last nucleotide of the *intI* stop codon, suggesting that this sequence and the *intI* gene are linked. A divergent REP element was identified in *Ps. straminea* KM91, in the same region as the REP element in *Ps. stutzeri* Gv.7 and *Ps. alcaligenes* strains. Similar inverted repeat sequences were not observed downstream of the PA0916 homologue in *Ps. stutzeri* Q (8A) or *Ps. aeruginosa* PAO1 (data not shown). Collectively, these observations indicate that the REP element is associated with the integron rather than the PA0916 homologue in these strains.

It has been suggested that REP elements may serve as transcription attenuators (Espeli *et al.*, 2001) and the location of REP elements described would be consistent with a role in the attenuation of *intI* transcription. Database searches provided support for such a role; Blastn searches revealed that similar (>80% nucleotide identity) inverted repeat sequences are a common feature of *Pseudomonas* spp. and other proteobacterial genomes, and are frequently located within 50 bp of the 3' end of a diverse array of chromosomal genes (data not shown). More recently it has been hypothesised that REP elements do not serve as specific terminators (Aranda-Olmedo *et al.*, 2002), and several other



roles have been suggested (Espeli and Boccard, 1997; Gilson *et al.*, 1990; Tobes and Pareja, 2005), including serving as target sites for IS element transposition (Tobes and Pareja, 2006). In light of the observation that CIs in *Pseudomonads* have characteristics indicative of acquisition by horizontal transfer, it is tempting to speculate that the REP elements identified at the 3' end of *intI* served as a target for the acquisition of CIs at this locus. Several different IS elements have been shown to specifically insert into REP elements resulting in a portion of the original REP element being located on either side of the IS element (Tobes and Pareja, 2006). To test for the presence of a similar arrangement in CIs found at the PA0916/PA3388 locus, the sequences upstream of the REP element (indicated in blue in Figure 5.13) of InPstRNAIII, InPst50238 and InPal55044 were subjected to Blastn searches to determine if a portion of the REP element could be detected downstream of the integron. If the REP element continued downstream, then Blastn searches would result in the detection of inverted repeat matches to the query sequence in the region downstream of the integron. However, no such sequences were detected downstream of the last gene cassette in any of the integron sequences tested.

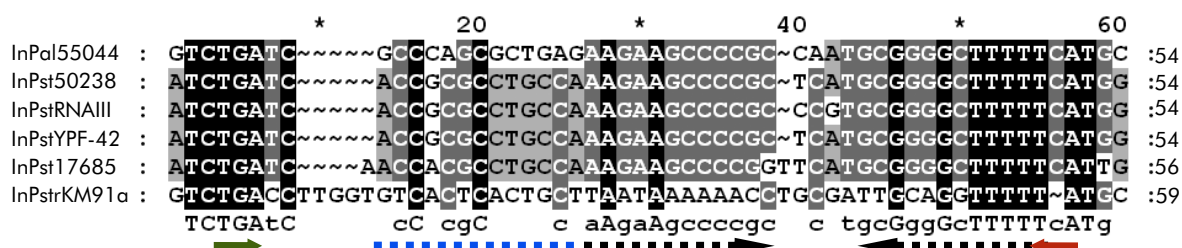


Figure 5.13 Alignment of the non-coding sequence between *intI* and the first upstream flanking gene of integrons at locus 1. The stop codon of PA0916 is indicated by a green arrow, and the stop codon of *intI* is indicated by a red arrow. Arrowheads indicate the direction of transcription of each gene. Conserved imperfect inverted repeat sequences were identified immediately adjacent to the *intI* stop codon, and are indicated by the dashed black arrows. The length of the arrow indicates the extent of the repeat sequence and the arrow heads indicate the orientation of the repeat. Searches were performed using the sequences indicated by the blue dashed line to determine if the REP sequences continued downstream of *intI*.

An alignment containing the intergenic spacer sequence between *intI* and the first upstream flanking gene was also constructed for all strains with an integron at locus 2 (Figure 5.14). Similar to the pattern observed for *Ps. stutzeri* Gv.7 integrons, the highest level of similarity for the upstream spacer of integrons at locus 2 was observed between members of *Ps. stutzeri* Gv.8 (95-99% nucleotide identity). The corresponding region of InPmeNW1, in contrast, exhibited limited identity to those in *Ps. stutzeri* Gv.8 integrons (46-48% nucleotide identity). Regions of similarity between the sequences included in Figure 5.14 were primarily limited to the 70-80 bp of the intergenic spacer closest to the start codon of PA3672 and are likely to be involved in the regulation of this gene. Blastn searches supported such a role for these sequences, as homologous sequences were identified preceding homologues of PA3672 in *Ps. aeruginosa* PAO1 and *Ps. stutzeri* ATCC 17595. Blastn searches using a small word size ( $n=7$ ) did not result in the detection of any additional conserved sequences in this region. Finally, no REP sequences analogous to those found flanking integrons at locus 1 were

observed. The intergenic spacer adjacent to *intI* in *Ps. mendocina* NW1 shared little homology with the corresponding intergenic spacer of *Ps. stutzeri* Gv.8 strains. This observation suggests either independent acquisition of an integron at the same locus or divergence of corresponding intergenic sequences after integron acquisition. Given the incongruence of *IntI* phylogeny with respect to integron locus in this strain, the former explanation is favoured.

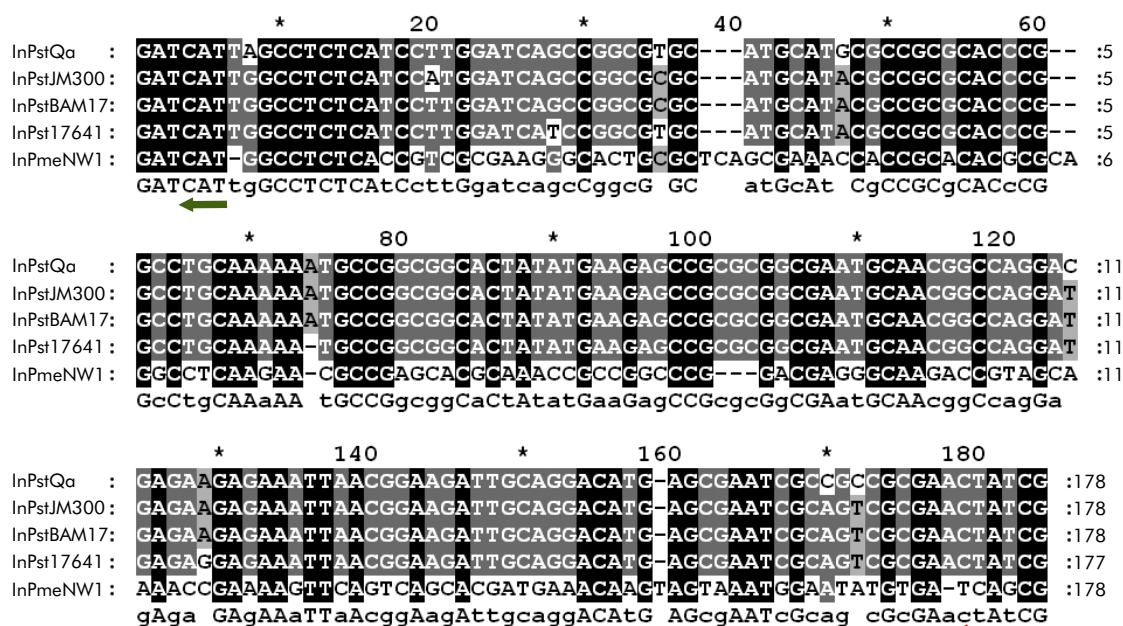


Figure 5.14 - Alignment of the non-coding sequence between *intI* and the first upstream flanking gene of integrons at locus 2. The stop codon of PA3672 is indicated by the green arrow, and the stop codon of *intI* is indicated by the red arrow. Arrowheads indicate the direction of transcription of each gene.

### 5.3.7 - Detection of a CI in the genome sequence of *Ps. mendocina* YMP

Shortly before the completion of this thesis, a shotgun sequence of the *Ps.*

*mendocina* YMP genome was published in Genbank, and was found to contain a *Pseudomonas* subfamily CI. The late timing of this publication precluded the inclusion of this strain in the analyses presented in this thesis. Instead, a separate analysis of this integron is included below.

The shotgun sequence contigs of *Ps. mendocina* YMP were subjected to several different BLAST searches to detect sequences characteristic of *Pseudomonas* subfamily CIs. A single core integron and at least eight gene cassettes were identified. A diagrammatic representation of *InPmeYMP* is provided in figure 5.15a. A putative *intI* gene and *attI* site was detected at the left hand end of

InPmeYMP, in the arrangement typical of other known integrons. The *intl* gene is annotated as a partial gene (missing the last 87 amino acids) in the genome annotation. However, upon closer inspection the 3' end of IntIPmeYMP was present, and a deletion of 12 amino acids between the residues corresponding to Gln-233 and Asp-244 of IntIPstQ which also caused a shift in the reading frame resulted in the failure to detect this gene fragment in the genome annotation. The sequence of the *attI* region of InPmeYMP contained regions of sequence conservation characteristic of other *Pseudomonas* spp. *attI* sites (see section 4.3.4.2) (data not shown). Unexpectedly, the sequence of intIPmeYMP was found to be most similar to IntIPstQ (8A) (pair-wise amino acid identity – 83.1% [266/320]), and exhibited only 70% (224/320) amino acid identity to IntIPmeNW1. This relationship was also apparent from examination of the *attI* regions of these integrons; the *attI* site of InPmeYMP exhibited 80% pair-wise nucleotide identity to the *attI* of InPstQ, and 49% pair-wise identity to the *attI* of InPmeNW1. Congruence between the divergence of *intl* and *attI* sequences is also consistent with co-evolution of the *intl/attI* complex. At least 5 cassettes were found immediately adjacent to the core integron, while three or more additional cassettes were detected in a different sequence contig (Figure 5.15a). All cassettes identified contained *Pseudomonas*-subfamily 59-be.

### a. CI in *Ps. mendocina* YMP genome annotation

*Ps. mendocina* YMP – whole genome shotgun Contig 80



*Ps. mendocina* YMP – whole genome shotgun Contig 59



*Ps. mendocina* YMP – whole genome shotgun Contig 81



### b. Comparison of PA0916/PA3388 gene order

*Ps. mendocina* NW1 (Pme)  $\leq 30 \text{ Kbp}^\dagger$

*Ps. mendocina* YMP 

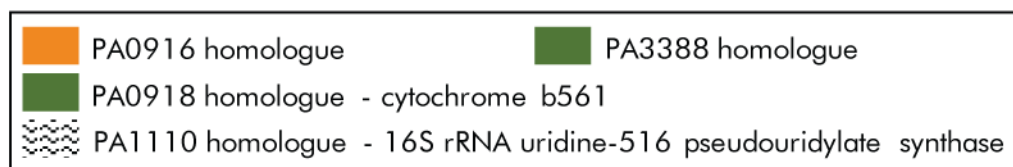


Figure 5.15 - Schematic representation of the InPmeYMP (a.) and the PA0916/PA3388 locus (b.) found in the *Ps. mendocina* YMP genome sequence. Where indicated, the beginning or end of the contigs shown was found within 1 Kbp of the last gene included. Genes in each section are colour coded according the legend shown below each diagram. Genes coloured white represent genes which are not homologous to any other gene shown. Accession numbers for each contig are as follows:  
Contig 59 - NZ\_AAUL01000028.1, Contig 80 - NZ\_AAUL01000010.1, and  
Contig 81 - NZ\_AAUL01000003.1.

<sup>†</sup> Distance between PA0916 and PA3388 homologues in *Ps. mendocina* NW1 known indirectly through the presence of these genes within a single fosmid clone.

The CI of *Ps. mendocina* YMP (InPmeYMP) appears to be found at the same locus as InPmeNW1 (Figure 5.15a). A homologue of PA3672 was identified immediately upstream of InPmeYMP. However, as the complete integron is not found within a single contig (Figure 5.15a), the genes found downstream of InPmeYMP could not be determined. Nonetheless, a homologue of PA3673 (3' conserved flanking gene for integron locus 2) was detected near the beginning of a different sequence contig (Figure 5.15a), and of the several genes found upstream of the PA3673 homologue, one is homologous to a gene upstream of PA3673 in *Ps. mendoncina* NW1 (gene PmeNW1-27 in Figure 5.5) (91.7% [242/264] pair-wise amino acid identity). The remaining genes upstream of PA3673 in *Ps. mendoncina* YMP were not typical of *Pseudomonas* spp. chromosomal genes, as no homologues were detected in any sequenced *Pseudomonas* spp. genomes (data not shown). Two IS remnants characteristic of those found in *Pseudomonas* spp. are also found upstream of the PA3673 homologue (Figure 5.15a).

The PA0916/PA3388 locus in *Ps. mendocina* YMP (Figure 5.15b) has a similar arrangement to that of *Ps. mendocina* NW1 (Figure 5.11). The same gene arrangement upstream of PA0916 was observed in both *Ps. mendocina* YMP and *Ps. mendocina* NW1. However, the gene adjacent to the PA0916 homologue in *Ps. mendocina* NW1 was adjacent to the PA3388 homologue in *Ps. mendocina* YMP. The gene arrangement in these strains is consistent with one or more independent insertions of multiple genes between PA0916 and PA388 in these strains (Figure 5.15b). These observations lend further support to the notion that homologues of PA0916 and PA3388 were ancestrally linked in Pseudomonads and that the PA0916/PA3388 locus is a hotspot for recombination in Pseudomonads.

With the inclusion of intIPmeYMP, *Pseudomonas* subfamily integrons remain a distinct monophyletic clade with respect to other known *intl* genes (Figure 5.16). The phylogenetic relatedness of intIPmeYMP to other *intl* genes was consistent with the *intl* pair-wise identity values given above; intIPmeYMP forms a well-supported monophyletic clade the *intl* genes of *Ps. stutzeri* Gv.8 strains, while intIPmeNW1 occurs in a different clade (Figure 5.16). This observation, in conjunction with the *intl* and *atl* identity values given above, is indicative of a common origin for the core integrons of *Ps. stutzeri* Gv.8 strains and *Ps. mendocina* YMP, and an independent origin for the core integron of *Ps. mendocina* NW1. It is noteworthy, however, that inclusion of *Ps. mendocina* YMP in phylogenetic analyses led to a significant decrease in the bootstrap support for the grouping of *intl* genes from *Ps. mendocina* NW1 and *Ps. stutzeri* Gv.7 strains (84-96% without *Ps. mendocina* YMP [Figure 5.9a&b] and 61% with the inclusion of *Ps. mendocina* YMP [Figure 5.16]). These observations indicate that the present dataset provides a somewhat incomplete picture of the evolutionary history of integrons in *Pseudomonas* spp., and sequences from additional strains are required to clarify the deeper branching relationships within this group.

The phylogenetic diversity observed among PA3672 homologues in these strains was (in contrast to Intl) congruent with established evolutionary relationships (as determined by 16S rRNA gene phylogenies [Figure 3.5]). *Ps. mendocina* NW1 and *Ps. mendocina* YMP exhibit 99% 16S rDNA pair-wise nucleotide identity and group with other members of this species in phylogenetic trees (data not shown). In trees constructed from alignments of PA3672, *Ps. mendocina* NW1 and *Ps. mendocina* YMP form a well-supported monophyletic clade (Figure 5.17), suggesting that a homologue of this gene was present in the common ancestor of *Ps. mendocina*. The incongruence of *intl* phylogenetic diversity with respect to



other chromosomal genes has interesting implications for the origins of CIs in *Pseudomonads*. On the basis of phylogenetic diversity in the current dataset, it appears that a divergent integron which is still a member of the *Pseudomonas* subfamily was acquired by *Ps. mendocina* NW1 and integrated at the same locus as the integron in *Ps. stutzeri* Gv.8 strains and *Ps. mendocina* YMP. Examination of alignments of the putative *attI* region (Figure 5.18) and intergenic spacer between *intI* and the PA3672 homologue (Figure 5.19) of *Ps. mendocina* YMP relative to other *Pseudomonas* spp. strains lends further support to this hypothesis. Both of these sequences in *Ps. mendocina* YMP were more similar to corresponding sequences in *Ps. stutzeri* Q (pair-wise nucleotide identity, *attI* – 75.7% [100/132], *intI* IGS – 58.6% [102/174]) than to *Ps. mendocina* NW1 (pair-wise nucleotide identity, *attI* – 47.7% [63/132], *intI* IGS – 34.8% [61/174]). In light of the apparent common origin of the CI in *Ps. stutzeri* Gv.8 strains and *Ps. mendocina* YMP, the conservation of the *intI*/PA3672 intergenic spacer in these strains and not in *Ps. mendocina* NW1 suggests that these sequences are associated with the CI rather than the chromosome (ie. were associated with the parent element which contained the CI and integrated at this locus).

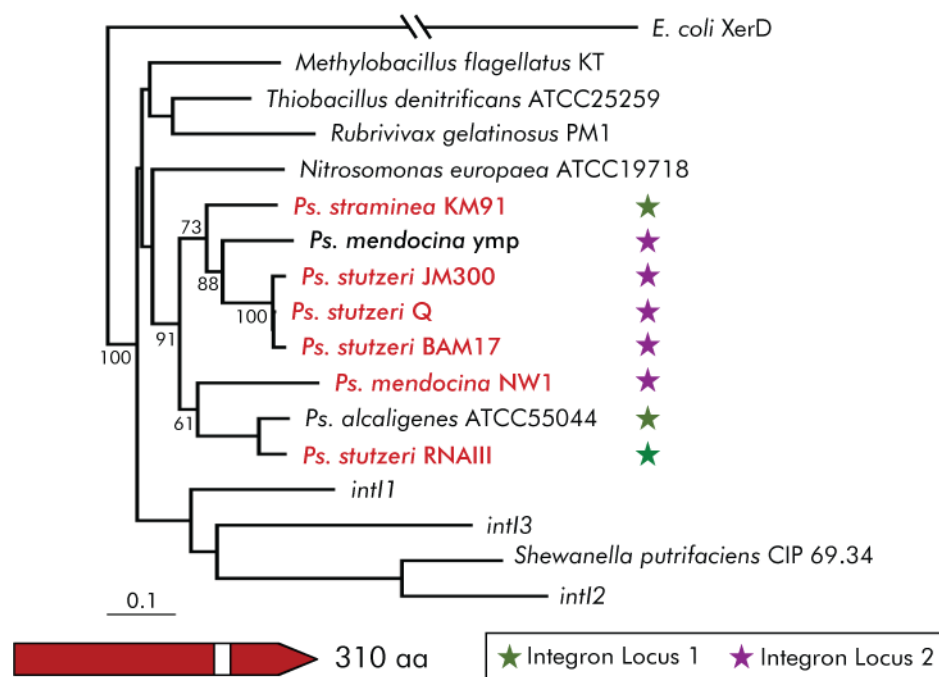


Figure 5.16 - Phylogenetic relationship between intlPmeYMP (*Ps. mendocina* YMP) and other *Pseudomonas* Intl. The tree shown was constructed using neighbour joining and a bootstrap analysis of 1000 iterations. Bootstrap values (as percentages) are given where possible. The region of Intl included in the alignment used to construct the tree is indicated by the red arrow-box. The locus of each integrin is indicated according to the legend above.

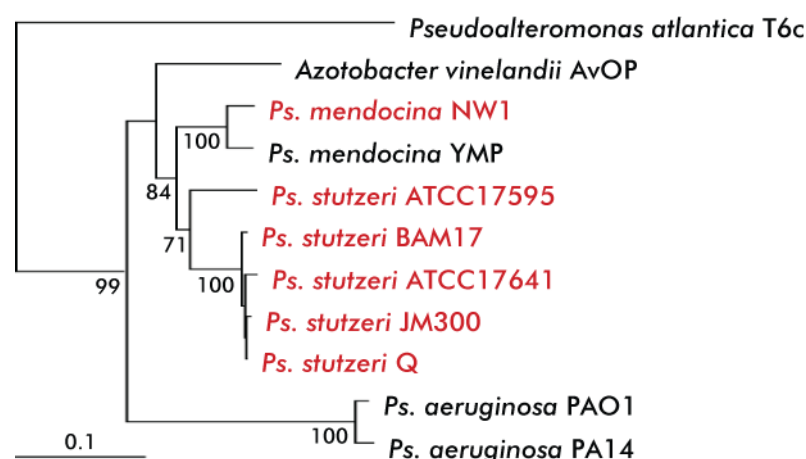


Figure 5.17 - Phylogenetic diversity of PA3672 homologues among *Ps. stutzeri*, *Ps. aeruginosa* and *Ps. mendocina* strains. The tree shown was constructed using neighbour joining and a bootstrap analysis of 1000 iterations. Bootstrap values (as percentages) are given where possible. An alignment of the entire gene (307 amino acids) was used to construct the tree.



```

      *      20      *      40
PmeYMP : CAT~~GCGGTATAGCCTCCCAAAATGATCATCCTGACCCGGAACCC : 46
PstQ (8A): CAT~~~~~ACCCTCCCAAAACAGGTCTCTCTGACCCGGAACGAG : 39
PmeNW1 : CATGTCGCGTTTCTCCTGAGTGCTGTGCGGTTTCTGCGGACAACTT : 48
          CAT   gcgt ta CctcccaaactG tc tccTGaCcCGaAA C

      *      60      *      80      *
PmeYMP : CAGCCTAGACAGGGAGTAGCGATATGACAAGGAAGCTTATCCCTCATT : 94
PstQ (8A): CAGCCTAGACAAGGAATAAGGAAATGACAAGGAAGATTATCCACAGT : 87
PmeNW1 : CAAGCTAGACAAGGACTTGCAGACGACAAGGAAGATTATGCGACAGG : 96
          CAgcCTAGACAaGGA TagcgA AtGACAAGGAAGaTTATcC cCAgt

      100      *      120      *      140
PmeYMP : GCCA~TGGG~AGGTTTATACACCATTTGGCGTCTATTTATAGTTAGGA : 141
PstQ (8A): CCGGCCCGGGAAGTTATACGCCAATGGCGTCTATTTATAGTTAGGC : 134
PmeNW1 : TGTCGCGTAGCAACGCGACATGCGTGTCTGCTAATTAATAGTTAGGG : 143
          cc gccgggaagttatACacca ttggcgctctATTtATAGTTAGG
                                     ↑

```

Figure 5.18 - Alignment of the putative *attI* region of InPmeYMP (*Ps. mendocina* YMP) relative to the corresponding regions of InPstQ (*Ps. stutzeri* Q) and InPmeNW1 (*Ps. mendocina* NW1). The start codon of *intI* is indicated by the red arrow, and the *attI* recombination crossover of InPstQ is indicated by the blue arrow.

```

      *      20      *      40      *      60
PmeYMP : GATCAT~GGCCTCTCAGCAGTGGGTCAAAGCCTGGCGCGCCGCTCGTGGGC~ACGAACC : 58
PstQ (8A): GATCATTAGCCTCTCATCTTGGATCAG~~~~~CCGGCGTGC~ATGCAATGCGCGCGCGACC : 55
PmeNW1 : GATCAT~GGCCTCTCAGCCTCGCGAAGGCGCATGCGCTGAGCGAAACCAACGCACACGCG : 59
          GATCAT gGCCTCTCAcC ttGcg g cc gGcgcgcg c tgcgcCaCgcgcC

      *      80      *      100      *      120
PmeYMP : AGGCTTGCAAAA~TGCCGGCGGAACTATAAGAAGTCGGGCGCGCTGATTGCAACCA~A : 116
PstQ (8A): CGGCCCTGCAAAAAATGCCGGCGGCACTATATGAAGAGCCCGCGCGCGAATGCAACGGCCA : 115
PmeNW1 : CAGGCTCAAGAA~~~~~CGCCGAGCAGCGCAACCGCGCGCGGACGAGGGCAAGACGT : 115
          cgGcCtgCAaAA tgcCGgCGg actataAagaaG cgggCgcG cGA tGCAAc gc a

      *      140      *      160      *      180
PmeYMP : TGGCGGCGCGCGAGATTTAGCGGAGGTTAGCCCTCCCCCGCG~TCACCC~~~~~~TCA : 168
PstQ (8A): GGACGAGAAGAGAAATTACGGAAGATTGCAAGACATGACGAATCGCGGCCGCGACTA : 175
PmeNW1 : ACCAAACCGAAGGTTTCAGTCAGCAGGATGAAACAAAGTAGTAATGGAATATGTGATCA : 175
          G cgagc GagAaaTT Ac gaagattac g aca g gcgaatcgcc atca
                                     ↓

```

Figure 5.19 - Alignment of the non-coding sequence between the PA3672 homologue and intIPmeYMP of *Ps. mendocina* YMP relative to the corresponding regions *Ps. stutzeri* Q (8A) and *Ps. mendocina* NW1 (Pme). The start codon of PA3672 homologue is indicated by the red arrow, and the stop codon of *intI* is indicated by the blue arrow.

## 5.4 - Conclusions

Integrans are a ubiquitous feature of *Ps. stutzeri* genomes; however, the majority contain inactivated core integrans. While integrans were recovered from members of all *Ps. stutzeri* genomovars tested, only members of *Ps. stutzeri* Gv.8 contained integrans which were predicted to be functional, and in several integrans, <150 bp of conserved core integran sequence was detectable. Some of these cassette arrays may have functional *attI* sites, while a complete integran, whose integrase operated *in trans*, has been deleted. Cassette arrays characteristic of *Pseudomonas* subfamily integrans were observed in all integrans for which cassette array sequence was generated, regardless of whether or not the integran contained a disrupted core integran. Core integran loss-of-function mutations are also a common feature *Xanthomonas* spp. integrans (Gillings *et al.*, 2005). These observations raise intriguing questions regarding the selective pressures operating on particular integran subfamilies. It appears core integran loss-of-function mutations are frequently favoured by selection, while selective pressures serve to maintain gene cassette arrays.

Gene cassettes represent a diverse genetic resource in Pseudomonads and in particular, *Ps. stutzeri*. All *Ps. stutzeri* strains screened in the present study contained gene cassette(s), and no two strains shared an identical cassette. Typical of other CIs, the vast majority of cassettes encoded proteins that showed limited homology to previously described proteins, and no phenotype could be attributed to any cassette encoded genes. Closely related strains (most notably *Ps. stutzeri* Gv.8) overwhelmingly contained unrelated cassettes, which indicates that cassette turnover occurs at a relatively rapid rate and that cassettes are accessed from a diverse pool. The presence of several pairs of related cassettes in the

arrays of different integrons indicates that particular cassettes have long life-spans, and may become relatively long-term residents of particular cassette arrays. The vast majority of cassettes (89%) were associated with *Pseudomonas* subfamily 59-be, which is consistent with patterns of 59-be diversity observed in other CIs (Boucher *et al.*, 2006; Gillings *et al.*, 2005; Rowe-Magnus *et al.*, 2003). These observations have interesting implications for the origin of cassettes associated with different CI subfamilies. They suggest that gene cassettes associated with particular CI subfamilies are acquired from common and restricted sources or that 59-be sequences are modified after cassette acquisition.

Examination of sequences flanking integrons in *Pseudomonas* spp. provided insights into the genomic context of integrons in this bacterial group. Several specific points (relating to the hypotheses stated in the section 5.1) can be made:

**CIs are present in many extant *Pseudomonas* spp. strains; however, the presence of this subfamily in the common ancestor of the *Pseudomonas* lineage cannot be confirmed**

Including the data generated in the present study, integrons are currently known to exist in four different species of *Pseudomonas*. This corresponds to 36% of species which have been screened for the presence of integrons (Vaisvila *et al.*, 2001; present study) or for which genome sequences exist; however, the frequency of integrons across all *Pseudomonas* spp. is likely to be considerably lower than this. Integrons have been detected in all *Xanthomonas* spp. strains screened (Gillings *et al.*, 2005), and are present in all (6/6) strains for which complete genome sequences exist. Integrons have also been detected in 85% (17/21) of *Vibrio* spp. complete genome sequences. The apparent frequency of integrons in *Pseudomonas* spp., in comparison, is significantly lower. This observation has interesting implications for the origin of integrons in

Pseudomonads. The current dataset indicates that integrons were either acquired after the divergence of the *Pseudomonas* genus, or if present in common ancestor of the *Pseudomonas* spp., integrons have been lost from most extant strains.

*Pseudomonas* spp. are clearly a monophyletic bacterial group that derived from a common ancestor on the basis of chromosomal framework gene phylogenies (Moore *et al.*, 1996). The same cannot be said for the integron subfamily within this bacterial group. Intl phylogenetic diversity was incongruent with respect to that of chromosomal framework genes (16S rDNA). Furthermore, integrons were observed at multiple loci, and conservation of locus between different strains did not correlate with established evolutionary relationships. This observation argues against movement of an ancestrally related CI between different loci. Rather, independent acquisition of integrons at locus 1 and 2 is the most parsimonious explanation for the observed patterns in genomic context and phylogenetic diversity in this CI subfamily. Collectively, these observations indicate that integrons in *Pseudomonas* spp. have a complex evolutionary history and argue against the presence of an integron in the common ancestor of the *Pseudomonas* lineage.

### **Multiple *Pseudomonas* subfamily integrons may be found in a single strain**

A single CI was identified in most strains. Two putative integrons/integron remnants were identified in three strains; however, no strains contained two active core integrons. No orphan cassettes or cassette arrays (gene cassettes which are not associated with a core integron) were detected in any strain. At least two strains contain multiple arrays (*Ps. straminea* KM91 and *Ps. stutzeri* ATCC 17589, Figure 5.5), and both arrays in each strain contain identifiable core integron remnants, consisting of a putative *attI* site and sequences which may be involved in *intI* regulation. This observation lends support to the notion that each of these

strains once had two functioning integrons. However, it is also possible that the putative *attI* sequences observed in integrons without an identifiable *intI* gene represent secondary sites for cassette insertion, and have never been associated with a functioning core integron. The origin of these multiple integrons cannot be resolved with the current dataset; each may have been acquired on independent occasions or have resulted from duplication of a parent integron.

In the majority of bacterial strains which contain CIs, the core integron occurs as a single copy; however, multiple core integrons within single genomes are not uncommon. Several *Vibrio* strains are known that contain a CI and class 1 integron (Ceccarelli *et al.*, 2006; Petroni *et al.*, 2002). Strains are also known which contain multiple copies of the same integron and multiple copies of divergent integrons (see Table 1.4). Such examples are relatively rare, occurring in 13% (14/108) of strains known to contain CIs (Table 1.4). In contrast, of the 16 *Pseudomonas* strains found to contain integrons in the present study, at least three contained multiple integrons or integron remnants. Given these observations, it appears that the presence of multiple integrons/integron remnants is higher in Pseudomonads than in other bacterial groups. This may, however, be an artefact of the currently available data; a limited number of strains with well characterised CIs are available, and to date no studies have examined the diversity of CIs in a bacterial group as extensively as the present study. Additional data on the abundance of CIs within different bacterial lineages are required to address this.

### ***Pseudomonas* subfamily integrons are found at multiple chromosomal loci and appear to have been acquired on multiple occasions**

At least six different integron loci were identified in the strains analysed, on the basis of variable flanking genes upstream and/or downstream of the integron. At least three strains contain multiple cassette arrays, but no strains contained two



active core integrons. The large number of loci observed is most likely the result of multiple independent integron acquisitions, intra-genomic recombination, or a combination of both. Evolutionary and genomic analyses will allow the relative contributions of these phenomena to be assessed (see Chapter 6).

Bacterial chromosomal genes typically differ from genes which frequently undergo HGT (the 'mobilome') in a number of ways. For example, chromosomal genes are present in most members of a lineage and exhibit phylogenetic congruence with established evolutionary relationships, while genes of the mobilome are typically present in few members of a lineage and exhibit patterns of diversity highly discordant with established relationships. The nature of the genes found flanking integrons in *Pseudomonas* spp. is indicative of a chromosomal locus for all integrons observed. Homologues were observed in >75% of all *Pseudomonas* spp. genome sequences for at least one flanking gene from each integron locus, and no genes characteristic of mobile elements were found between the integron and chromosomal flanking genes. Further, the conserved genes flanking integrons at both loci 1 and 2 were detected in all but one of the *Ps. stutzeri* strains screened in the present study (see section A.3.5). Conservation of gene content at particular loci (1 + 2) between different *Pseudomonas* spp. indicates stable inheritance of the integron platform, which is another key characteristic of chromosomal genes. While the phylogenetic diversity of *Pseudomonas* subfamily *intl* is incongruent with respect to established evolutionary relationships, some congruence between particular integron loci and *intl* phylogenetic diversity was observed, further supporting a history of stable vertical inheritance of CIs in these strains. Collectively, these observations provide a strong indication that all *Pseudomonas* subfamily integrons recovered here exist at chromosomal loci.

Phylogenetic data supports an independent origin of integrons at locus 1 and 2, as the *IntI* sequences of integrons from each locus clustered together (but not exclusively) in phylogenetic analyses (see sections 5.3.4 and 5.3.7). The origins of integrons occurring at the remaining loci (loci 3-6) cannot be unambiguously determined with current dataset due to a lack of multiple independent examples and/or phylogenetically relevant sequence information. These observations unambiguously indicate that integrons have been acquired and integrated at a chromosomal locus on at least two occasions. Interestingly, detailed internal comparisons of integrons at loci 1 and 2 revealed that integrons have been acquired at each locus on multiple independent occasions. Several lines of evidence supported this; incongruent *IntI* phylogeny with respect to established evolutionary relationships and locus, and a consistent relationship between the diversity of intergenic spacers downstream of *intI* and *IntI* phylogenetic diversity. On the basis of these data, a group of ancestrally related CIs at each locus were identified; integrons in *Ps. stutzeri* Gv.7 strains and *Ps. alcaligenes* ATCC55044 at locus 1, integrons in *Ps. stutzeri* Gv.8 strains and *Ps. mendocina* YMP at locus 2. Each of these groups formed well supported clades in phylogenetic analyses and contained conserved intergenic spacers downstream of *intI*. The same data supports the independent acquisition of a CI by *Ps. straminea* KM91 at locus 1 and *Ps. mendocina* NW1 at locus 2, as neither of these strains grouped with other strains with integrons at the same locus in phylogenetic analyses, and both contained divergent intergenic spacers downstream of *intI*. Collectively, these observations provide strong support for common ancestry of a subset of integrons at both locus 1 and 2, in addition to the independent acquisition of an integron at each of these loci. The significance of these results is discussed in detail in chapter 7.

***Pseudomonas* subfamily integrons are not found within larger definable mobile genetic elements.**

Several lines of evidence provide a strong case that none of the integrons recovered in the present study are contained within larger genetic elements. All completely sequenced integrons recovered here had relatively sharp boundaries, separated from the first upstream and downstream gene by a non-coding spacer of <200 bp. No genes were recovered either upstream or downstream of the integron which were indicative of it being part of a larger mobile genetic element. In some cases, IS elements or remnants were found downstream of the last cassette or within the integron itself. However, all IS elements detected appeared to be (or have once been) independently mobile and are thus unlikely to play direct role in integron mobilisation in their present form.

## CHAPTER 6 - EVOLUTIONARY ANALYSIS OF *PSEUDOMONAS* SPP. INTEGRONS

The use of phylogenetic data in conjunction with genomic context data is a powerful tool for reconstructing the evolutionary history of gene families. However, several other analytical techniques can provide additional information on the evolutionary history of particular genes. In particular, measures such as G+C% and codon usage analysis are powerful complements/alternatives to methods which are dependent on phylogenetic analyses, since related sets of genes are not required.

The G+C content of a genome and the codon usage of its genes are determined by selection and mutation pressures (Sueoka, 1988). As these evolutionary processes are characteristic for each species, the sequences of particular genomes share common patterns in nucleotide composition and codon usage. This property makes it possible to identify genes acquired by horizontal gene transfer (HGT) as those whose features are atypical with respect to the parent genome (Lawrence and Ochman, 1997). While the accuracy of such techniques has been drawn into question (Cortez *et al.*, 2005; Koski *et al.*, 2001), several studies have demonstrated that these approaches are relatively robust for the detection of alien genes (Daubin *et al.*, 2003; Grocock and Sharp, 2002; Medigue *et al.*, 1991).

The most powerful methods for the detection of HGT acquired genes are based on phylogenetic inference. Well-supported topological disagreement between inferred trees for particular gene families can often be parsimoniously explained only by invoking HGT (Brown and Doolittle, 1997; Nesbo *et al.*, 2001; Smith *et al.*, 1992). Indeed, as discussed in chapter 5, HGT is the most likely explanation for the phylogenetic incongruence between the *Pseudomonas* spp. *intI* and their 16S rRNA genes. However, it is difficult to extend this method to all genes (eg. most cassette

encoded genes) due to the limited phyletic distribution of many gene families. Consequently, several tree-independent methods for detecting genes acquired by HGT have been developed, including analysis of codon usage, nucleotide frequencies and codon-position specific G+C% (amelioration) (Hayes and Borodovsky, 1998; Koonin and Galperin, 1997; Koonin *et al.*, 1997; Lawrence and Ochman, 1997; Makarova *et al.*, 1999; Worning *et al.*, 2000). These and other methods which are based on analysis of nucleotide frequencies are collectively referred to as 'compositional' techniques, and since sets of homologous genes are not required, they are ideally suited to characterising the genes of CIs relative to their parent genome.

It was established in Chapter 5 that the phylogenetic diversity of *Pseudomonas* subfamily integrons is incongruent with respect to patterns observed among chromosomal framework genes such as 16S rRNA genes. This observation suggests that integrons were acquired after the divergence of the common ancestor of the Pseudomonadaceae lineage. Given the assumption that CIs are acquired by HGT, the genes that constitute the CI should exhibit patterns of codon usage and/or nucleotide composition which are atypical with respect to the parent genome (provided these parameters are different in the donor genome). Once acquired by HGT, CIs become fixed in the chromosome, and the core integron is assumed to become a static component of the CI. Consequently, the core integron will be exposed to similar mutational pressures to the parent genome over long evolutionary periods. Gene cassettes, in contrast, may continue to be lost and gained over time and consequently have a shorter residence time in the genome relative to the core integron. Consequently, they are likely to be exposed to the mutational pressures of the parent genome for shorter periods. Do patterns in codon usage/nucleotide composition and phylogenetic diversity reflect this?

Acquired genes with atypical codon usage are expected to progressively resemble the recipient genome (in terms of G+C%) over time, due to exposure to local mutational pressures. If the core integrons of CIs have a long residence time relative to gene cassettes, it may be expected that they have also mutated to match the G+C% of the host genome to a greater extent than cassettes in the array. This is yet to be specifically tested; however, some evidence exists that is consistent with this hypothesis. Codon usage of the cassettes of InPal55044 is highly atypical with respect to chromosomal genes, while the codon usage of *intI* is moderately atypical (Vaisvila *et al.*, 2001). Are similar patterns in codon usage and nucleotide composition observed across *Pseudomonas* spp. CIs? The generation of G+C content and codon usage data for the integrons recovered in Chapters 4 and 5 will allow this question, and several others, to be addressed.

The primary aim of the present chapter is to characterise nucleotide composition and codon usage of *Pseudomonas* spp. CIs to augment the phylogenetic and genomic context data generated in Chapter 5, and to provide a more comprehensive picture of the evolutionary history of CIs in these bacteria. Specifically, the codon usage of *Pseudomonas* spp. integron genes, in addition to the genes found adjacent to these integrons, were compared to that of typical *Pseudomonas* chromosomal genes. Using multivariate statistical analysis of codon usage values, the origin of different genes within and adjacent to the integron was inferred.

## 6.2 - Materials and methods

### 6.2.1 – Genomic Analyses and Codon Usage Analyses

Total G+C content, and codon usage statistics for all genes to be analysed were determined using the CodonW software package

(<http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html>; [(Peden, 1999)]).

Statistics were also calculated for CIs found in the genome sequences of *Vibrio cholerae* 01 N16961 and *Xanthomonas campestris* ATCC 33913 for comparative purposes. Total G+C% of 59-bp and other non-coding sequences was calculated using the Bioedit software package (Hall, 2001).

### 6.2.2 – Statistical analysis of codon usage data

Relative Synonymous Codon Usage (RSCU) was calculated for all genes analysed by dividing the observed codon usage by that expected when all codons for a particular amino acid are used equally (Sharp and Li, 1986). The use of relative measures (as opposed to absolute counts) removes the effect of amino acid composition, which is desirable when examining the use of a particular codon relative to its synonyms. RSCU values have been found to provide a statistically robust measure of variation in codon usage relative to raw codon usage data when detecting differences in codon usage in *Ps. aeruginosa* PAO1, particularly when aiming to detect genes acquired by HGT (Grocock and Sharp, 2002). To summarise differences in codon usage, Correspondence Analysis (CA) and Principal Component Analysis (PCA) were performed on matrices of RSCU values using the Minitab 14 statistical software package. Codon usage in *Ps. aeruginosa* PAO1 was selected as a surrogate for codon usage in *Ps. stutzeri*, as *Ps. aeruginosa* is the closest relative to *P. stutzeri* for which a complete genome exists. The suitability of *Ps. aeruginosa* PAO1 as a surrogate was tested by comparing

patterns in its codon usage against a reference set of *Ps. stutzeri* genes characteristic of both chromosomal and horizontally transferred genes.

## 6.3 – Results

### 6.3.1 – G + C content analysis

The G+C% of selected integrons and their respective flanking sequences is shown in Figure 6.1. All genes were analysed separately; however, average values for the cassette ORFs in each array are shown in Figure 6.1. Non-coding regions of *intI* (*attI*) were included in determination of G+C content of core integron sequences. The process by which 59-be become associated with an ORF to form a gene cassette is currently unknown; however, the high level of identity between the 59-be associated with particular CI subfamilies has led to the suggestion that these 59-be originate from the host genome, with 59-be becoming associated with recently acquired genes (Mazel, 2006). To account for this possibility, cassette-associated ORFs were treated separately in G+C content analyses to 59-be sequences. The G+C content of *Ps. stutzeri* strains has been reported to range from 60.9 to 65 mol%, however most strains have G+C contents between 63 and 65 mol% (Rossello *et al.*, 1991).

The G+C content of the conserved genes flanking integrons at different loci (as defined in Figure 5.5) was consistently high, ranging from 62.7 to 68.3 mol%, which is consistent with the total chromosomal G+C content of *Pseudomonas* spp. strains. In contrast, the G+C content of core integron and cassette ORF sequences was consistently lower than that of the chromosome in all integrons recovered (Figure 6.1). The G+C content of *intI* was relatively uniform across all *Pseudomonas* spp. CIs; both *Ps. stutzeri* Gv.2 strains contained *intI* genes with a high G+C content of 60.4 mol%, but in all remaining strains the G+C content of



*intI* ranged from 53.1 - 56.5 mol% (Figure 6.1). The G+C content of cassette ORFs ( $44.7 \text{ mol\%} \pm 3.1\%$ ) was extremely A+T rich with respect to both the core integron region and typical chromosomal genes, (Figure 6.1). In the unassigned genes (as defined in Section 5.3.2) of several integrons, a general trend of increasing G+C content with increasing distance from the last gene cassette was observed. This pattern was most obvious in the genes downstream of the cassette arrays of InPstQ, InPmeNW1 and InPst19589 (8A, Pme and 1B in Figure 6.1, respectively). Each gene downstream of the last cassette was found to be, on average,  $3.3 \text{ mol\%} \pm 2.7\%$  more G+C rich than the gene preceding it. The pattern was consistent across most genes in these regions, with the exception that the last gene before the PA3673 homologue in InPstQ and InPmeNW1 and the last gene recovered in InPst17589 were all more A+T rich than the preceding gene (Figure 6.1). Collectively, *Pseudomonas* subfamily 59-be were moderately A+T rich (average G+C content of  $56.4 \text{ mol\%} \pm 2.8\%$ ) with respect to typical chromosomal genes. While higher than the average G+C content of cassette ORFs, the G+C content of *Pseudomonas* subfamily 59-be is comparable with that of the core integron.

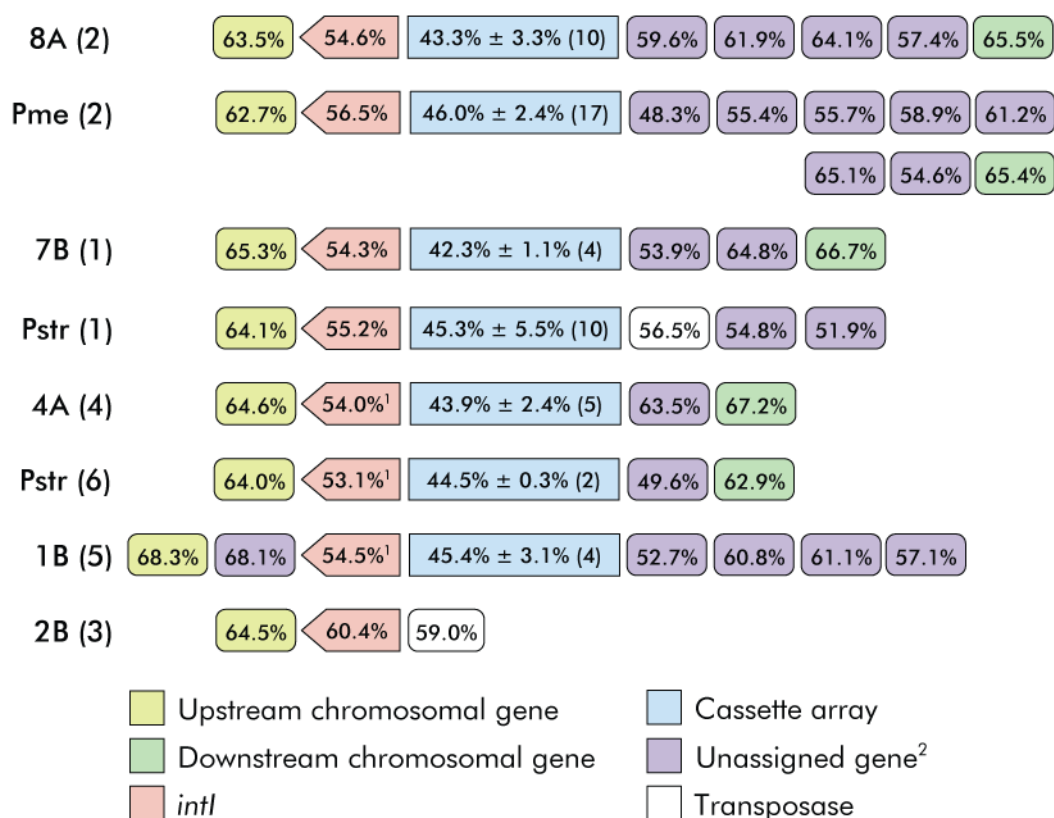


Figure 6.1 - Summary of patterns in total G+C% of integrons and flanking sequences. Regions of each sequence are represented as modules, colour coded according to the legend above. When measuring the G+C of gene cassettes, each was treated as a separate unit, and the values given indicate the average G+C% ± standard deviation. Numbers in parentheses after the G+C content of each cassette array refer to the number of cassettes in the array. Strain codes used for labelling correspond to the abbreviations given in Table 3.1. Numbers in parentheses after each strain label indicate the locus at which the integron is found, as described in Figure 5.5.

- <sup>1</sup> As complete *intI* genes were not detected in these integrons, total G+C% was calculated for the region immediately downstream of the 5' flanking gene up to, and including, the predicted *attI* site.
- <sup>2</sup> 'Unassigned gene' refers to genes which could not be unambiguously classified as being either integron-associated (lack of defining gene cassette features) or chromosomally associated (lack of homologues in *Pseudomonas* spp. complete genomes).

### 6.3.2 – Codon Usage Analysis

Variation in G+C content between different genes provides a useful but somewhat blunt measure of nucleotide compositional differences. A more extensive and quantitative analysis of the sources of variation among genes can be achieved using multivariate statistical analysis of codon usage data. Tabulating the synonymous codon usage for a set of genes results in a data matrix consisting of 59 dimensions (consisting of the 59 codons which have synonyms). Multivariate statistical methods may then be employed to reduce the dimensionality of the data, condensing the major trends in codon usage into a few dimensions. The resulting matrix of co-ordinates is then plotted in two or three dimensional space to compare patterns in codon usage between genes in the dataset. The closer two genes are in a plot, the more similar their codon usage. Trends in codon usage may also be correlated with biological properties (such as gene expression level, evolutionary history or position on the chromosome) to determine the causes of variation in codon usage.

Variation in codon usage has been widely used for the detection of genes which have been acquired by HGT (Garcia-Vallve *et al.*, 2000; Grocock and Sharp, 2002; Hooper and Berg, 2002; Lawrence and Ochman, 1997). However, the accuracy and sensitivity of this method has been shown to be limited when dealing with horizontally transferred genes of diverse origins (Koski *et al.*, 2001; Ragan, 2001; Wang, 2001). Nonetheless, variation in codon usage remains a powerful technique when detecting horizontally transferred genes which originate from genomes with a different G+C content to the recipient genome (Cortez *et al.*, 2005). For example, the predominant trend in codon usage (RSCU) in *Ps. aeruginosa* PAO1 (when controlling for amino acid composition) has been shown to be caused by variation in G+C%, and in particular GC3s (G+C content of the

third codon position) of different genes (Grocock and Sharp, 2002; see Figure 6.2a). Many genes in *Ps. aeruginosa* PAO1 that are thought to have been acquired by HGT (Stover *et al.*, 2000) have atypical G+C% contents and are significant contributors to this trend. RSCU is therefore a good candidate measure for the detection of acquired genes in other *Pseudomonas* spp.

Correspondence analysis (CA) is the multivariate statistical technique which has most commonly been used to characterise variation in codon usage between different genes (e.g. (Andersson and Sharp, 1996; Grantham *et al.*, 1981; Lafay *et al.*, 1999; Lafay *et al.*, 2000; Medigue *et al.*, 1991; Romero *et al.*, 2000)). However, CA has frequently been misused in the analysis of variation in codon usage (Perriere and Thioulouse, 2002). CA is designed for use with tables of counts (Hill, 1974); however, in studies involving analysis of codon usage, this technique has frequently been used on tables of relative measures such as RSCU (Fuglsang, 2003; Grocock and Sharp, 2002; Gupta *et al.*, 2004). The use of CA on datasets of relative measures has been shown to strongly affect the results obtained, increasing the probability of misinterpreting trends in the data (Perriere and Thioulouse, 2002). Principal components analysis is able to deal with tables of relative measures and is thus better suited to the analysis of RSCU values than CA. Consequently, PCA was used to analyse codon usage in the present study, and was tested against the CA approach used by Grocock and Sharp, (2002).

### 6.3.2.1 – Analysis of relative codon usage data using PCA and CA

PCA and CA were compared as methods for characterizing relative codon usage in *Ps. aeruginosa* PAO1. A data file containing the accession number and sequence of each *Ps. aeruginosa* PAO1 gene used is provided in the electronic appendix (Filename – 'PAO1 genome.fas'). The CA performed on RSCU values for *Ps. aeruginosa* PAO1 genes was identical to that performed by Grocock and Sharp, (2002); a similar spread of data-points was obtained (Figure 6.2a), and the same proportion of total variation explained by the first dimension (16.9%) was observed in both plots (data not shown). The spread of data-points obtained using PCA (Figure 6.2b) was also comparable to that obtained using CA. Visually, the spread of data-points from each category in the PCA plot (Figure 6.2b) was slightly greater relative to equivalent CA plots (Figure 6.2a), and the proportion of the total variation apparent in the dataset accounted for by the first PCA dimension increased from 16.9% to 17.5%. As observed in the CA plot (Figure 6.2a), the spread of chromosomal genes and genes thought to have arisen by HGT occurred primarily along the first PCA axis (x-axis in Figure 6.2b). These results indicate that both CA and PCA provide similar results in the analysis of relative codon usage in *Ps. aeruginosa* PAO1. As the use of relative measures does not violate the assumptions of PCA, this technique was used for all subsequent analyses of codon usage data.

#### 6.3.2.2 – Assessment of *Ps. aeruginosa* as a surrogate for codon usage in *Ps. stutzeri*

No genome sequences are currently available for members of the *Ps. stutzeri* species complex. *Ps. aeruginosa* is the most closely related species for which a genome sequence does exist. *Ps. stutzeri* and *Ps. aeruginosa* strains also exhibit similar total genomic G+C contents (64.1 mol% and 66.3 mol%, respectively). Codon usage in this species is therefore a good candidate to serve as a surrogate for codon usage in *Ps. stutzeri*; however, as empirical data was required to confirm this, a PCA was performed on the *Ps. aeruginosa* PAO1 genome and a representative set of genes from various *Ps. stutzeri* strains (Figure 6.3c). The *Ps. stutzeri* genes used in this analysis were acquired from Genbank, and a data file containing the accession number and sequence of each gene used is provided in the appendix (Filename – ‘Ps stutzeri Genbank genes.fas’). Genes were classified as chromosomal if they exhibited homology to genes involved in central metabolic pathways or genes present in most completely sequenced *Pseudomonas* spp. genomes. HGT-acquired genes were accessed from published *Ps. stutzeri* plasmid sequences. The spread of data-points obtained for *Ps. stutzeri* chromosomal and HGT acquired genes was comparable to that observed for *Ps. aeruginosa* PAO1 genes (Figure 6.3c). *Ps. stutzeri* chromosomal genes clearly clustered with the chromosomal genes of *Ps. aeruginosa* PAO1; 87.5% (42/48) of chromosomal genes occurred within the main cluster of *Ps. aeruginosa* PAO1 chromosomal genes. The sample of putatively acquired *Ps. stutzeri* genes was small (7 genes) which limits the robustness of any assertions which are drawn; however, all data-points representing *Ps. stutzeri* acquired genes clustered within the spread of *Ps. aeruginosa* PAO1 acquired genes, and >50% of *Ps. stutzeri* acquired genes occurred outside the main cluster of *Ps. aeruginosa* PAO1 chromosomal genes. These results indicate that *Ps. aeruginosa* PAO1 and *Ps. stutzeri* chromosomal

genes exhibit similar patterns of relative codon usage, and genes in both species thought to be of foreign origin are mostly outliers with respect to these chromosomal genes. Thus, *Ps. aeruginosa* PAO1 genes are a suitable surrogate for codon usage in *Ps. stutzeri* genes.

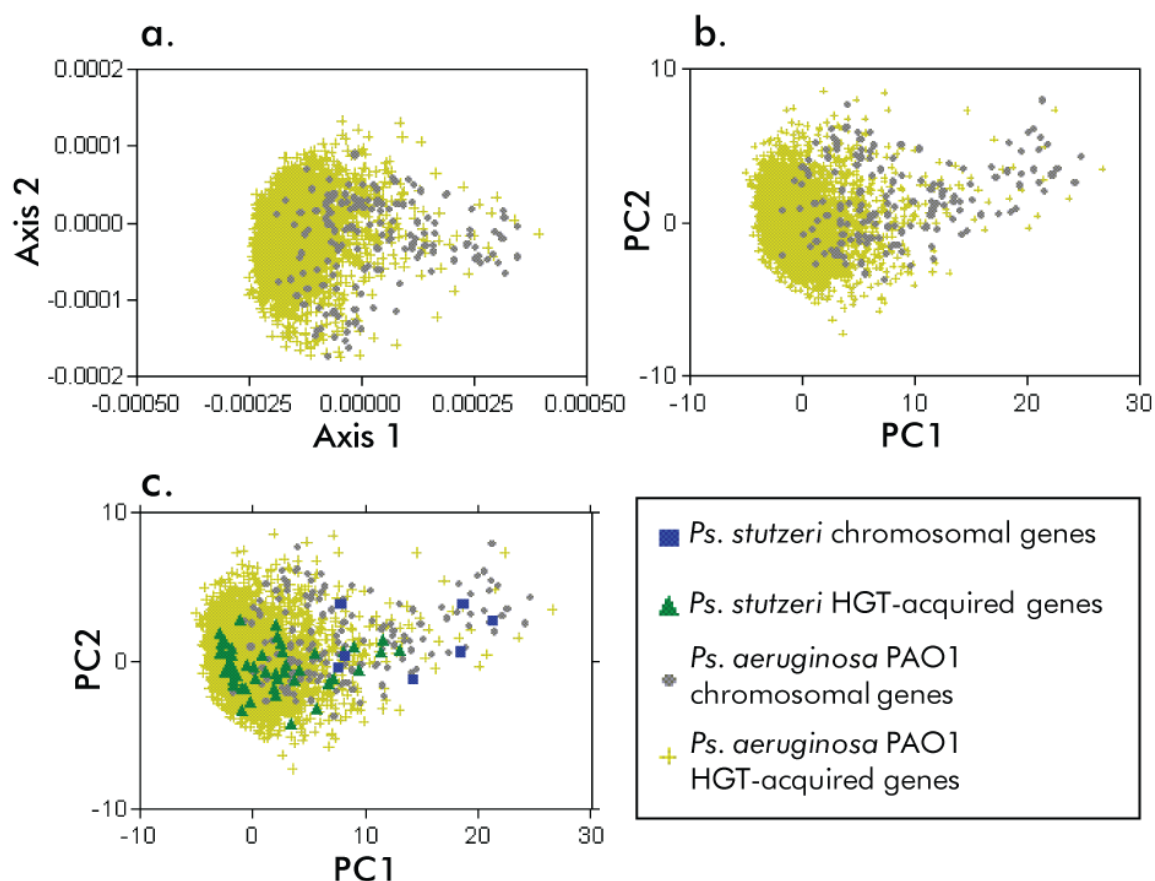


Figure 6.2 - Assessment of statistical method and surrogate genome choice for analysis of codon usage patterns. a. Correspondence analysis on *Ps. aeruginosa* PAO1 genome, b. Principal component analysis on *Ps. aeruginosa* PAO1 genome, and c. Comparison of codon usage in PAO1 and *Ps. stutzeri* genes. The first and second principal components are plotted on the x- and y-axis, respectively. The reference set of *Ps. stutzeri* genes was accessed from Genbank and the sequences and accession numbers are provided in the appendix (Filename – 'Ps stutzeri Genbank genes.fas'). Classification of *Ps. aeruginosa* PAO1 HGT-acquired genes were obtained from Stover *et al.*, (2001). The circles in graphs b. and c. indicate the regions in which >98% (104/5567) of all chromosomal (red) and horizontally transferred (black) genes occur.



### 6.3.2.3 – Codon usage of *Pseudomonas* spp. CIs relative to chromosomal genes

Genes from selected *Pseudomonas* spp. CIs were subjected the same analysis as that performed on *Ps. stutzeri* genes in Figure 6.2c, and the resulting principal components plots are shown in Figures 6.3a-c. A class 1 integron from *Ps. aeruginosa* plasmid R1033 was also analysed to compare patterns of codon usage between typical mobile integrons and CIs (Figure 6.3d). A similar spread of data-points was observed across all PCA plots presented in Figure 6.3; in all cases differences between typical chromosomal genes and putatively horizontally transferred genes were expressed in the 1<sup>st</sup> principal components axis. The proportion of the total variation apparent in the dataset that was explained by the 1<sup>st</sup> principal components axis was for the PCA plots in Figure 6.3 ranged from 17.5% to 18.2% (data not shown), which is high considering the large size of the dataset used to generate the plots.

Gene cassettes clustered separately to *Ps. aeruginosa* PAO1 chromosomal genes in all PCA plots (Figure 6.3). This observation indicates that synonymous codon usage of gene cassettes was consistently atypical with respect to chromosomal framework genes, and is consistent with the notion of gene cassettes as a pool of genes acquired by HGT. The *intI* genes analysed in Figure 6.3 also exhibited atypical RSCU with respect to the majority of *Ps. aeruginosa* PAO1 chromosomal genes, but were less extreme outliers in this regard than cassette encoded ORFs. Relative synonymous codon usage of the *intI* genes analysed may therefore be considered as intermediate with respect to chromosomal framework genes and cassette encoded ORFs. No consistent patterns were observed in the distribution of the 'unassigned' genes found between the last gene cassette and the first conserved flanking gene (see Figure 5.5) relative to *Ps. aeruginosa* PAO1 genes (Figure 6.3a-c). However, several of these genes did occur outside the region of

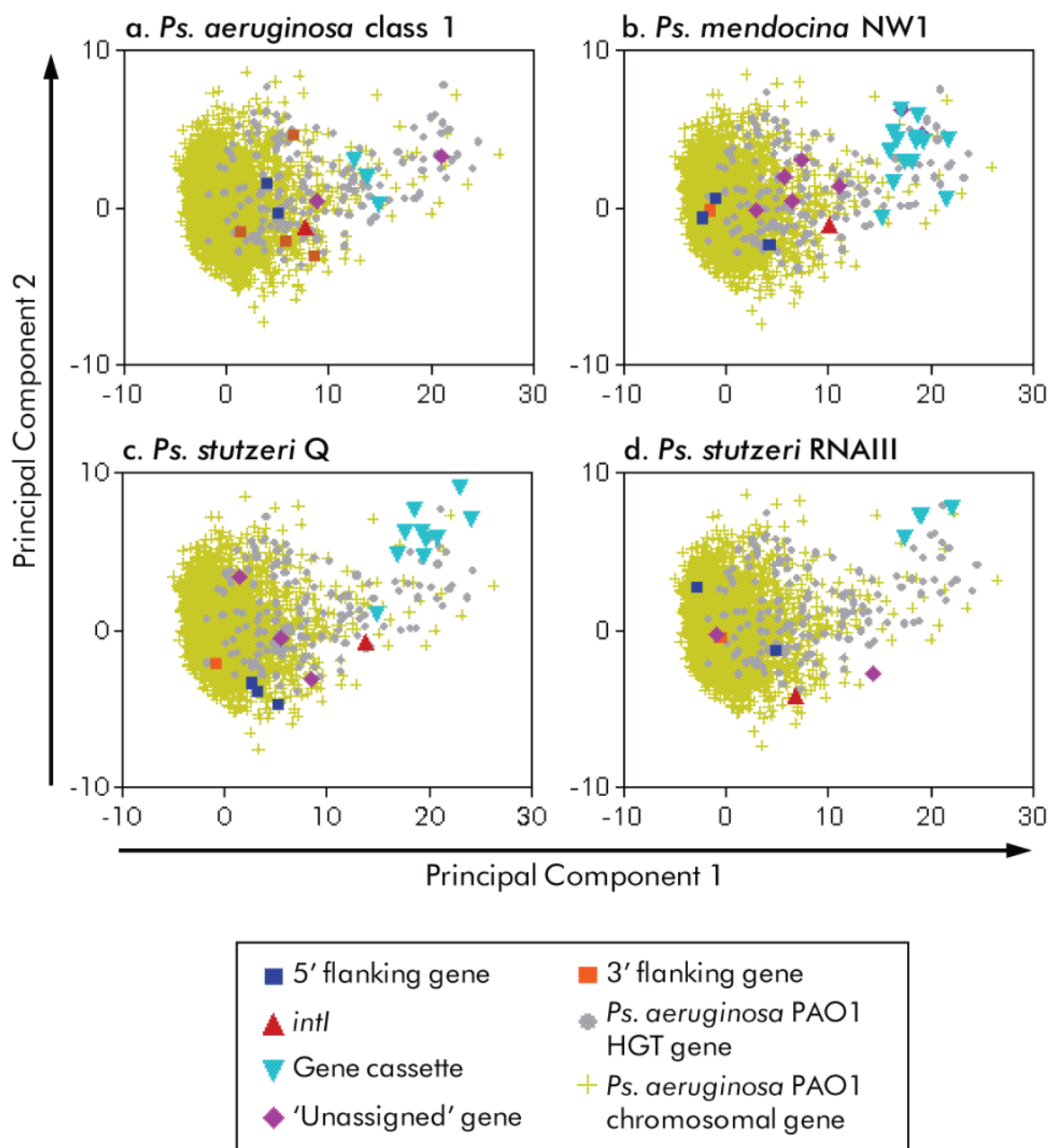


Figure 6.3a-d - Principal component analysis of relative synonymous codon usage values for selected *Pseudomonas* spp. integrons relative to genes in the *Ps. aeruginosa* PAO1 genome. The first and second principal components are potted on the x- and y-axis, respectively. The genes of the class 1 integron were extracted from the sequence of *Ps. aeruginosa* plasmid R1033 with accession number U12338. The circles in each graph indicate the regions in which >99% of all chromosomal (red) and horizontally transferred (black) genes occur.

<sup>†</sup> 'Unassigned genes' refer to genes which could not be unambiguously classified as being either integron-associated (lack of defining gene cassette features) or chromosomally associated (lack of homologues in *Pseudomonas* spp. complete genomes).

the graph containing >99% of all *Ps. aeruginosa* PAO1 chromosomal genes, which is suggestive of acquisition of by HGT. Overall variation in the RSCU of genes from the class 1 integron of *Ps. aeruginosa* plasmid R1033 (Figure 6.3d) were comparable to those observed in the other integrons analysed (Figure 6.3a-c). The entire integron and the plasmid genes flanking it occur within the region containing >99% of all transferred genes, which is consistent with the recent acquisition of this integron (and the parent element carrying it) by HGT. In contrast, At least one flanking gene from the CIs of Q, RNAlII and Pme occur outside this region, but inside the region occupied >99% of PAO1 chromosomal genes (Figure 6.3a-c). This observation is consistent with stable vertical inheritance of the conserved genes flanking these integrons.

## 6.4 – Conclusions

The results presented in this chapter reinforce several key points made in earlier chapters on the nature of integrons in *Pseudomonads*:

### **Integron-associated genes exhibit properties indicative of acquisition by HGT**

Multivariate analysis of RSCU values and analysis of G+C content has been found to be a relatively robust indicator of HGT in *Ps. aeruginosa* PAO1 (Grocock and Sharp, 2002), and similar patterns were observed when examining a representative set of *Ps. stutzeri* genes in the present study. The G+C content of integron-associated genes was consistently A+T rich with respect to most chromosomal genes, and variation in synonymous codon usage reflected these differences, with integron genes preferentially containing A+T rich codons over G+C rich codons. Genes acquired by HGT tend to be A+T rich with respect to the host genome (Daubin *et al.*, 2003; Lawrence and Ochman, 1997; Medigue *et al.*, 1991). These results were consistent with patterns in codon usage of *Ps. alcaligenes* ATCC 55044 (Vaisvila *et al.*, 2001). Genes acquired by HGT tend to be A+T rich with respect to the host genome (Daubin *et al.*, 2003; Lawrence and Ochman, 1997; Medigue *et al.*, 1991). In conjunction with the patchy distribution of integrons among *Pseudomonas* strains, the strain-specific nature of cassette encoded genes, and the incongruent phylogeny of *intI* with respect to chromosomal framework genes, these results strongly suggest that integrons in *Pseudomonas* spp. have been acquired by HGT on multiple occasions.

## **Integrans in Pseudomonads are not associated with larger mobile genetic elements**

Differences in G+C content and codon usage between integrin genes and the genes flanking each integrin support the notion that non-class 1 integrins in Pseudomonads are not contained within larger mobile genetic elements. A sharp change in both G+C content and codon usage was consistently observed between integrin genes (including 'unassigned' genes) and flanking chromosomal genes. The G+C content of integrin genes were consistently A+T rich relative to the chromosome, while the conserved flanking genes were consistently close to the G+C content of the chromosome. Where analysed, variation in codon usage reflected these differences in G+C content, and was consistent with patterns of codon usage bias in *Ps. alcaligenes* ATCC 55044 CI genes (Vaisvila *et al.*, 2001).

## **Unassigned genes were acquired after the acquisition of an integrin at the parent locus**

The origin of 'unassigned' genes (defined in Chapter 5) remains ambiguous; however, several lines of evidence suggest that these genes were also acquired by HGT either after or at the time of integrin acquisition. Codon usage and G+C content of the 'unassigned' genes found downstream of certain cassette arrays were mostly atypical with respect to chromosomal genes, and homologues of these genes were present in few, if any, *Pseudomonas* spp. genomes. Comparison of integrins at the same locus revealed variation in gene content of 'unassigned regions'. *Ps. stutzeri* Q (8A) and BAM17 (8B) contain homologous genes in this region, but a completely different suite of genes is present in *Ps. mendocina* NW1 and *Ps. mendocina* YMP. The G+C content of 'unassigned genes' tended to increase with distance from the integrin. A tempting explanation for this observation is the acquisition of these genes one at a time with insertion occurring

after the last cassette, if mutational pressure has acted to ameliorate the G+C content of the genes, to match the G+C content of the host. Collectively, these observations suggest that the genes in the unassigned regions were acquired after acquisition of the integron. Alternatively, these genes may have been acquired at the same time as the integron, being associated with a larger (parent) element that was acquired at the locus in a single event. Regardless, the mosaic nature of CIs makes them a good candidate for further investigation of changes in gene composition after acquisition by HGT. Amelioration is one such measure (Lawrence and Ochman, 1997). Using this technique, the timing of acquisition of integron and gene cassette associated genes may be assessed, shedding light on the patterns of acquisition, loss and evolution of integrons over evolutionarily relevant timescales. However, before amelioration analysis can be carried out on *Pseudomonas* spp. CIs, estimation of the intrinsic mutation rate of strains belonging to this genus is required.

## CHAPTER 7 – FINAL DISCUSSION

The significance of integrons lies in their interaction with gene cassettes. Site-specific integration and excision of cassettes leads to acquisition and rearrangement of genes at a single locus, and the integron promoter ( $P_c$ ) drives expression gene cassette ORFs. These properties mean that the integron-gene cassette system can engineer physical and transcriptional linkage of genes. Integrons are of enormous interest to genome evolution, as gene cassettes encompass a very high diversity and exceptional levels of novelty. Given these properties, integrons may be seen to impact genome evolution at different scales. Through the potential to assemble higher order phenotypes *de novo*, integrons may be seen as phenotype engineers. Over longer timescales, integrons have the potential to operate as species engineers. One of the barriers to HGT is pre-existence of the gene or function. An integron-assembled phenotype may allow occupation of a new ecological niche, resulting in less chance of gene exchange due to increased genetic isolation and, ultimately, speciation. Several factors must be considered in determining the influence of integrons on bacterial diversification. For example, what is the extent of integron exposure experienced by modern bacterial chromosomes? What outcomes have resulted from integron activity over evolutionary timescales? How are dynamic cassette arrays converted into stable phenotypes?

The data presented in this thesis has made a significant contribution to the body of knowledge on the evolutionary history of chromosomal integrons. Several new integron sequences from members of a single bacterial lineage were recovered, which may now serve as an additional and unique model group for investigation of the evolutionary significance of CIs. The present study also constitutes the most

comprehensive analysis to date of CI genomic context. Two key points of particular significance warrant further discussion:

1. CIs in *Pseudomonas* spp. exhibit some patterns of stable vertical inheritance, a substantial body of evidence indicates that multiple CIs in strains of this genus are common and that historically, loss by deletion and acquisition by HGT has been relatively common (see Section 7.1).
2. Integron loci in *Pseudomonas* spp. may act as hotspots for recombination (see Section 7.2).

## **7.1 – Integron acquisition and loss in *Pseudomonas* spp.**

CIs in *Pseudomonas* are atypical with respect to other characterised CI subfamilies (*Vibrio* and *Xanthomonas*). While CIs in *Pseudomonas* exhibit some patterns of stable vertical inheritance, the bulk of the evidence presented in Chapters 5 and 6 indicates that CIs in *Pseudomonas* are elements which undergo relatively frequent loss by deletion and acquisition by HGT, and also often exist as multiple copies. Interestingly, despite evidence for multiple acquisitions, all *Pseudomonas* Intl sequences clustered together in phylogenetic analyses, suggesting that integrons have been independently acquired from common or related sources in these bacteria. The incongruence between Intl and chromosomal marker gene phylogenies has been noted recently (Boucher *et al.*, 2007), and the data presented in this thesis was consistent with this (Figure 5.9). Collectively, these observations indicate that historically, horizontal transfer of integrons between phylogenetically distant organisms has been common, regardless of the genomic context in which they are found. In other words, on an evolutionary timescale, all integrons are mobile elements. In conjunction with the recognition that CIs from most if not all bacterial groups have been subjected to HGT (Boucher *et al.*, 2007),



evidence that integrons in *Pseudomonas* have been acquired on multiple occasions has interesting implications for chromosomal integron concept, and for the role of integrons in bacterial diversification. This is discussed in greater detail in Section 7.1.1.

Of all the Pseudomonad integrons analysed in this thesis, 73% (16/22) were predicted to be inactive integron remnants. In all cases, deletion mutations occurred within *intI*, while the *attI* region and cassette array remained intact. Inactivated core integrons are also common in other bacterial groups, constituting approximately 26% of all known integron sequences (see Tables 1.7 and 7.1; Gillings *et al.*, 2005). It is possible that the disproportionately large number of remnant integrons detected in Pseudomonads is the result of sampling bias. The present study constitutes the most extensive analysis to date of integron genomic context across a group of related bacterial genomes. It is possible that the frequency of degenerate integrons may be similarly high in other bacterial groups, but have thus far evaded detection due to the difficulties associated with identifying highly degraded genes either experimentally or using sequence database searches.

The frequency of inactivated integrons across all bacteria suggests that selection sometimes favours *IntI* loss-of-function mutations. Selective pressures which may drive *IntI* loss-of-function mutations may only be speculated upon. Toxicity of the *IntI* integrase is one possible explanation but is unlikely as functional *intI* genes appear to have been maintained over long periods in several *Pseudomonas* strains. Maintenance of *attI* and cassette sequences may occur as a result of *in trans* *IntI* activity from other integrons in cell. Prevention of rearrangement or loss of a complementary set of cassette encoded genes is the most attractive explanation for *IntI* loss-of-function mutations, in terms of its potential to impact

bacterial evolution. This process would represent the final stage in integron mediated operon engineering – integron activity resulting in the assembly of a complementary suite of genes, and loss of *IntI* activity preventing its disruption. However, in the absence of a demonstrable phenotype encoded by a suite of cassettes, this is pure speculation.

Integrans at locus 1 (PA0916/PA3388 locus as defined in Section 5.3.2) are flanked in at least one direction (immediately downstream of *intI*) by inverted repeat sequences. This may be seen as analogous to the situation observed in class 1 integrons, which are characteristically flanked by inverted repeat sequences both upstream and downstream of the integron. Such inverted repeat sequences are commonly associated with mobile genetic elements. Given this, it was not surprising that class 1 integrons were found to be flanked by inverted repeat sequences, as these elements are associated almost exclusively with mobile elements (Partridge *et al.*, 2001). It is therefore reasonable to speculate that the inverted repeat sequences associated with CIs at locus 1 are indicative of mobile history, perhaps constituting the remnants of a larger mobile genetic element. Given the weight of evidence indicating relatively common HGT events among *Pseudomonas* CIs, it may be expected that remnant mobile elements would be found associated with particular CIs. However, inverted repeat sequences also serve many other functions within cells and a role other than recombination sites for these sequences cannot be ruled out based on the present dataset.

At least three *Pseudomonas* strains in the collection were found to contain more than one integron or integron remnant. Across all integron-containing bacteria, multiple core integrons are also relatively common, occurring in approximately 10% of strains. The relatively high frequency with which multiple core integrons are observed in bacterial genomes lends further support to the notion that integrons

are essentially acquired elements, regardless of genomic context. Multiple core integrons arise by independent HGT events (eg. *Geobacter metallireducens* GS-15 and *Roseiflexus castenholzii* DSM 13941; Table 1.3), but may also arise by duplication (eg. *Nitrosomonas europaea* ATCC 19718; Leon and Roy, 2001). Most importantly, the presence of multiple, independently acquired integrons in single strains means at least some bacteria have the potential to acquire new integrons by HGT frequently, potentially restoring integron activity in strains which contain degenerated integrons. Such a property may be critical if integrons are to play a role in phenomena such as operon engineering that necessitates loss-of-function mutations.

## **7.2 – Integrons are associated with hotspots for recombination**

Examination of gene content of different integrons and the genes flanking them (locus 1 – PA0916 & PA3388 and locus 2 – PA3672 & PA3673) revealed several interesting patterns in gene loss and acquisition at conserved loci which contained integrons:

### **1. Cassette arrays are highly diverse in terms of gene content and are indicative of a high rate of cassette turnover.**

*Ps. stutzeri* Q (8A) and BAM17 (8B) are very closely related strains, exhibiting pair-wise identities of 100%, 97%, and 98% across 16S rRNA, IGS1, and IntI sequences, respectively. Furthermore, these strains were isolated from the same soil sample (Holmes *et al.*, 2003), raising the possibility that they have evolved sympatrically. Despite these similarities, *Ps. stutzeri* Q (8A) and BAM17 (8B) contain completely different gene cassette arrays (although one pair of cassettes from the arrays exhibit 95% pair-wise nucleotide identity, see Table 5.3). This

observation indicates that integron activity may lead to a rapid turnover of gene cassettes, and in turn has the potential to drive the diversification of clonal lines.

## **2. Non-integron gene content at loci which contained integrons was found to be highly variable.**

Analysis of gene content at the PA0916/PA3388 (locus 1) and PA3672/PA3673 (locus 2) loci across several strains in the collection, in addition to several genome sequences, revealed a high level of diversity in gene content at these loci (see Figures 5.11 and 5.12). The gene content of each locus was conserved within, but not (necessarily) between different *Ps. stutzeri* genomovars. In strains which contained an integron at one of these loci, gene content downstream of the integron ('Unassigned region' as defined in Section 5.3.2) was also found to vary between integrons at the same locus. Collectively, these observations support the notion that the loci at which integrons occur in Pseudomonads are frequently targets of gene insertions and/or rearrangements. Consequently, these loci may be considered hotspots of genomic diversity which evolve over relatively short timescales, comparable to the timescales over which different *Ps. stutzeri* genomovars diverge.

## **3. Phylogenetic relationships between Intl and flanking gene sequences can be parsimoniously explained only by independent acquisition of integrons at single loci on at least two occasions.**

The extensive body of evidence suggesting that integrons have been acquired at the same chromosomal locus on multiple occasions is discussed in detail in Section 5.4. The likelihood that integrons in Pseudomonads have been acquired at the same chromosomal locus on multiple occasions has intriguing implications for understanding the mechanisms by which integrons are acquired

at, or mobilised from, chromosomal loci. Independent acquisition of class 1 integrons by different transposon/plasmid backbones has been observed (Partridge *et al.*, 2001). This is not surprising given the mobile genomic context of class 1 integrons, the mosaic nature of mobile genetic elements, and the intense selective pressures experienced by class 1 integrons, which drive dissemination by HGT. In contrast, chromosomal integrons are assumed to be anchored in the chromosome, and are consequently rarely exposed to HGT. The data presented in this thesis has demonstrated not only that integrons have been acquired by *Pseudomonads* on multiple occasions, but at a single locus on multiple occasions. This is the most significant finding of the present study for several reasons: it is the first time independent acquisition of multiple chromosomal integrons at a single locus has been observed, it suggests that integrons (or the parent elements carrying them) may target specific loci upon integration into the chromosome, and it provides additional support for the model of frequent integron acquisition and loss in *Pseudomonas* spp. This is an intriguing finding worthy of further investigation through examination of additional integrons from related strains.

### **7.3 – Implications for the concept of the generalised CI**

The following characteristics have frequently been used to describe a chromosomal (or 'super') integron (Mazel 2006): (i) a chromosomal location; (ii) many (>20) associated gene cassettes; (iii) a high degree of sequence identity between the 59-bp sites of these cassettes; and (iv) a mostly vertical descent within a given lineage (i.e. little or no evidence of HGT of the integron core). These characteristics should be represented in the archetypical examples of this category: the *Xanthomonas*, *Pseudomonas* and *Vibrio* integrons. The primary defining feature of CIs is a fixed chromosomal locus (i). All other features (ii – iv) associated

with CIs are most likely a direct or indirect consequence this fundamental characteristic. Of these, the critical and most biologically significant consequence of a chromosomal locus is stable inheritance (iv), which is hypothesised to result in a reduced frequency of horizontal transfer and changes in the nature and/or stability of selective pressures experienced by the integron. The characteristics of typical CI cassette arrays (ii and iii) are in turn hypothesised to be a consequence of the long-persistence times and stable selective pressures afforded chromosomal genomic contexts, and may therefore be considered an indirect consequence of a chromosomal locus.

The degree to which integrons in each of these genera (in addition to several others) meet the above criteria was reviewed recently (Boucher *et al.*, 2007). It was argued that collectively, they failed to provide a clear distinction of a unique type of integron. With the recent expansion in the number of integron sequences known, distinctions between the cassette array size and 59-be sequence similarity of CIs and MIs were found to become far less distinct. The authors also noted that many integrons exhibit characteristics intermediate between typical mobile and chromosomal integrons, and that despite stable vertical inheritance in some bacterial groups, HGT appears to have impacted integron diversity in all bacterial groups in which they occur. Boucher *et al.*, 2007 concluded that due to the lack of consistency in defining features (other than a chromosomal locus), the chromosomal (or 'super') integron concept had limited if any usefulness as a generic model of all integrons occurring at chromosomal loci. These observations suggest that rather than a dichotomy of well defined groups of MIs and CIs, the characteristics of integrons across all bacteria represent a spectrum, with the archetypal MI and CI representing either extreme.

The characteristics of the expanded *Pseudomonas* integron dataset analysed in this thesis provide very strong support for a spectrum model of integron diversity. Indeed, it represents the most comprehensive and only unambiguous example of a group of related chromosomal integrons which clearly occupy the middle of the spectrum, exhibiting features of both CIs and MIs. The sparse distribution of CI across all *Pseudomonas* species (Table 7.1) suggests that integrons were acquired by HGT relatively late in the evolutionary history of this lineage. Core integron similarity and locus conservation indicate that integrons have been acquired by HGT on at least three and perhaps four or more occasions, and aberrant codon usage and G+C content of integron-associated genes further supported this. Despite the clear evidence of multiple integron acquisitions by HGT, *Pseudomonas* Intl sequences exhibited phylogenomic fidelity, clustering together in global analyses of Intl phylogeny.

This finding complicates interpretation of stable inheritance of CIs, the principal biological consequence of a chromosomal locus. It demonstrates that integrons vary in phylogenomic persistence and fidelity in a manner which is at least partially independent of genomic context. Distinct clades of Intl show fidelity to bacterial lineages (Table 7.1). This has only been observed for chromosomally located integrons. In some cases accompanied by locus persistence (*Vibrio* and *Xanthomonas*) and in some cases this is despite multiple integron acquisitions (*Pseudomonas*). Distinct clades of Intl can also be promiscuous, which is well illustrated by integrons of classes 1-3. A spectrum model of diversity implies that *Vibrio* and *Xanthomonas* represent relatively uncommon examples of lineages with stably inherited CIs. Indeed, based on the present dataset, *Vibrio* is the only bacterial genus that represents an unambiguous example of a long term association with a CI lineage (as defined for generalised CIs in Section 1.8). The

phylogenetic depth of *Xanthomonas* (maximum 16S rRNA divergence – 95.2%) strains known to contain CIs is considerably lower than CI-containing *Vibrio* strains (maximum 16S rRNA divergence – 90.2%), and most *Xanthomonas* CIs contain degenerate core integrons, which will ultimately lead to a loss of the CI (Table 7.1).





Bacterial Genus	Strains known to contain CIs	CI frequency <sup>1</sup>	Phylogenetic depth of sample set <sup>2</sup>	Integron persistence <sup>3</sup>	Phylogenomic fidelity exhibited <sup>4</sup>	Phylogenomic integrity of CI <sup>5</sup>	Array size range	Total cassette pool <sup>6</sup>	Associated with 59-be subfamily <sup>7</sup>	Most dominant 59-be family
<i>Vibrio</i>	31	83% (20/24)	90.2% (94.7%)	High (87%)	Y	Moderate (6)	52-211	> 681 cassettes	Y	133bp (~70%)
<i>Xanthomonas</i>	20	100% (6/6)	95.2% (98.8%)	Low (45%)	Y	High (1)	1-22	> 57 cassettes	Y	N.D.
<i>Pseudomonas</i>	18	6% (1/17)	92.9% (97.3%)	Low (19%)	Y	Moderate (6)	2-33	> 115 cassettes	Y	76bp (90%)
<i>Shewanella</i>	9	50% (8/16)	Not Determined	High (89%)	Y	V. Low (9+)	0-11	> 17 cassettes	N	N/A
Generalised CI subfamily	-	Most	-	High	Y	High	Large	-	Y	-

Table 7.1 - Characteristics of CIs associated with particular bacterial genera. The four genera listed represent the bacterial genera with the highest known CI abundance. Where relevant, the expected characteristics of a paradigmatic CI are given.

<sup>1</sup> CI frequency refers to the fraction of available genome sequences that contain CIs.

<sup>2</sup> The percentage given for each genus indicates the pair-wise sequence identity of rRNA sequences of the two most divergent strains within each genus which are known to contain integrons. The percentages given in parentheses indicate the mean 16S rRNA sequence identity for strains with each genus that are known to contain integrons.

<sup>3</sup> Integron persistence refers to the frequency with which functional integrons are maintained by particular bacterial groups. Percentages given in parentheses refer to the fraction of integrons in each bacterial group which are predicted to be functional.

<sup>4</sup> Phylogenomic fidelity refers to the presence of genus-specific clades, as determined by phylogenetic analysis of IntI sequences.

<sup>5</sup> Phylogenomic integrity refers to the observed level of locus/genomic context conservation within each genus. Numbers in parentheses indicate the total number of chromosomal loci within each genus known to contain integrons.

<sup>6</sup> Total cassette pool refers to the total number of cassettes known to occur in integrons of each bacterial group. The figure given in each case is likely to be greatly underestimated.

<sup>7</sup> Different bacterial groups were considered to be associated with 59-be subfamilies if more than half of all cassettes are associated with 59-be that collectively exhibit >80% nucleotide identity.

In light of these emerging patterns in diversity, CIs may best be viewed as genetic elements which are essentially mobile, are frequently integrated at and/or mobilised from chromosomal loci and are sometimes, given appropriate selective pressures, maintained in a lineage over long evolutionary periods by vertical transmission. Under this model, integrons which are stably inherited over long evolutionary timescales, such as *Vibrio* and *Xanthomonas*, constitute a somewhat unique type of integron which may be seen as occupying one extreme of a spectrum. Occupying the opposite end of this spectrum are integrons which do not form associations with particular lineages over long periods and are frequently subject to horizontal transfer (eg. class 1 integrons). The general concept of a 'chromosomal' integron nonetheless remains relevant, as *Vibrio* (and to a lesser extent *Xanthomonas*) integrons have characteristics which are clearly unique, and additional integron subfamilies with similar characteristics are likely to be identified in the future. Consequently, the classical 'chromosomal' integron concept may be seen as a generalised model for integrons occupying one extreme of a spectrum of diversity. However, in light of the emerging patterns in CI diversity in *Pseudomonas* and other bacterial groups, it is not appropriate to assume that all CIs exhibit genomic context-specific characteristics and careful reconstruction of the evolutionary history of integrons in different bacterial lineages is required to properly assess the impact of HGT and vertical inheritance.

## 7.4 – Future Work

There exists a considerable potential for future relevant research focussed on characterising integrons in *Pseudomonas* and other bacterial groups. Most work continuing on the themes of this thesis would involve expanding the dataset of integron sequences through extension of sequences generated in the present study, in addition to recovery of sequences from new strains. In particular, additional sequences of integrons at loci 1 and 2 will likely shed light on the timing and nature of multiple independent integron acquisition events and may provide clues on the mechanisms that lead to integron capture, integration and fixation at a chromosomal locus. Additional sequences of integrons at these loci will also allow the origin of genes located in ‘unassigned regions’ to be better assessed. The screening methods and sequencing strategies used in the present study provide a robust means of achieving these future goals, and numerous relatively well-characterised *Ps. stutzeri* strains (and other *Pseudomonas* spp. strains) are potentially available.

Several other questions outside the specific scope of this thesis also remain unanswered and may now be better addressed in light of the data presented in this thesis. Some such questions include: What is the significance and role (if any) of IS elements in integron function and mobility? Several *Pseudomonas* CIs contain, or are flanked by IS elements or IS remnants; Do integrons outside the clinical context have the ability to compile and express functionally complementary sets of gene cassettes? *Ps. stutzeri* Q is an ideal model organism for addressing this question as its integron has been shown to be active *in vivo*, and express genes inserted at *attI* (Coleman *et al.*, 2005); What are the causes and significance of integron loss-of-function mutations? The large number of independent core integron loss-of-function mutations

in *Pseudomonas* integrons suggests that the selective forces that drive these mutations are particularly active in this genus. In summary, the study of integrons in *Pseudomonas* offers many opportunities to make significant discoveries regarding the ways in which integrons can impact bacterial evolution through the generation of diversity in response to environmental change.

# **APPENDIX 1 – FOSMID LIBRARY CONSTRUCTION AND SCREENING**

## **A.1 - Introduction**

To enhance the clarity of the results presented in Chapter 5, details of large-insert clone library construction and screening are included below, while analysis of this data is presented in chapter 5.

## **A.2 - Materials and methods**

### **A.2.1 - Fosmid library construction**

In order to easily recover entire integrons and the genes flanking them, whole-genome, large-insert clone libraries were constructed for 11 strains in the collection (Table A.2) using the CopyControl™ Fosmid Library Production Kit (Epicentre) according to the instructions supplied with the kit, with the exception of the method of DNA shearing employed. High molecular weight genomic DNA prepared as described in section 2.2 was sheared by passing it rapidly through a 0.5 mm syringe needle 20-30 times. For each library constructed, between 384 and 768 clones were selected for further analysis. All libraries were stored at -80°C in 384-well plates in LB containing 12.5µg/ml chloramphenicol and 15% glycerol.

### **A.2.2 - Estimation of library coverage**

For every library, five clones were selected at random in order to estimate the average insert size and genomic coverage of each library. Fosmid was extracted from each of these clones and digested with both BamHI and EcoRI (both enzymes supplied by New

England Biolabs). The insert size of each clone was estimated by adding the sizes of all bands observed after electrophoresis in a 1.2% agarose gel. The genomic coverage of each library was estimated by calculating the total amount of insert DNA (in megabases) present in each library and dividing this total by the estimated total genome size of the relevant strain. Genome size in *Ps. stutzeri* has been shown to range from 3.75 Mbp – 4.64 Mbp (Ginard *et al.*, 1997; Rainey *et al.*, 1994), and an average genome size of 4Mb was assumed for the purposes of fold-coverage estimation.

### **A.2.3 - Colony hybridisation**

Detection of specific genes in fosmid clones was accomplished using colony hybridisation, as described previously (Sambrook and Russell, 2001). Clones from 384-well plates were transferred to replicate nylon membranes (Hybond N+, Amersham) using a 384 pin replicator (Nunc). After transfer, the nylon membranes were placed on LB agar containing 12.5µg/ml chloramphenicol, and incubated overnight at 37°C to allow sufficient growth of the clones for hybridisation. Colonies were then lysed *in situ* as previously described (Sambrook and Russell, 2001), and the DNA fixed to the membrane via UV irradiation (Stratalinker). Fosmid libraries were screened for the presence of both integron-associated sequences and integron flanking genes. Screening for core integron and 59-be sequences was performed as described in section 4.2.2, using identical probes and hybridisation conditions.

#### **A.2.4 - Fosmid purification**

When purified fosmid was required for DNA sequencing or other analysis, induction of the fosmid to high copy number was performed according to the instructions provided with the CopyControl™ Fosmid Library Production Kit. After the induction step, fosmid was extracted using an alkaline lysis plasmid purification protocol from Sambrook and Russell (2000). The yield of purified fosmid was estimated by comparing the fluorescence of dilutions of fosmid DNA against that of a DNA standard after electrophoresis and staining with ethidium bromide as described in section 2.4. Purified fosmid DNA was stored in TE buffer at -20°C.



### **A.2.5 - Sequencing of fosmid clones**

All integron sequences were extended upstream and downstream to recover the complete cassette array and flanking sequences using primer walking on fosmid clones. The strategy employed for sequence recovery is represented in the flowchart in Figure A.1. When sequence information for the integron of interest was available (from PCR screening), this sequence was used as the starting template for upstream and downstream extension of the sequence. When sequence information was not available for a particular strain, integron detection PCRs (as described in section 4.2.1) were repeated on fosmid clones to generate a starting template for sequencing. Finally, if a PCR approach was unsuccessful, fosmid clones were sub-cloned into pUC18 to create a small insert library and re-screened using hybridisation to detect sequences of interest.

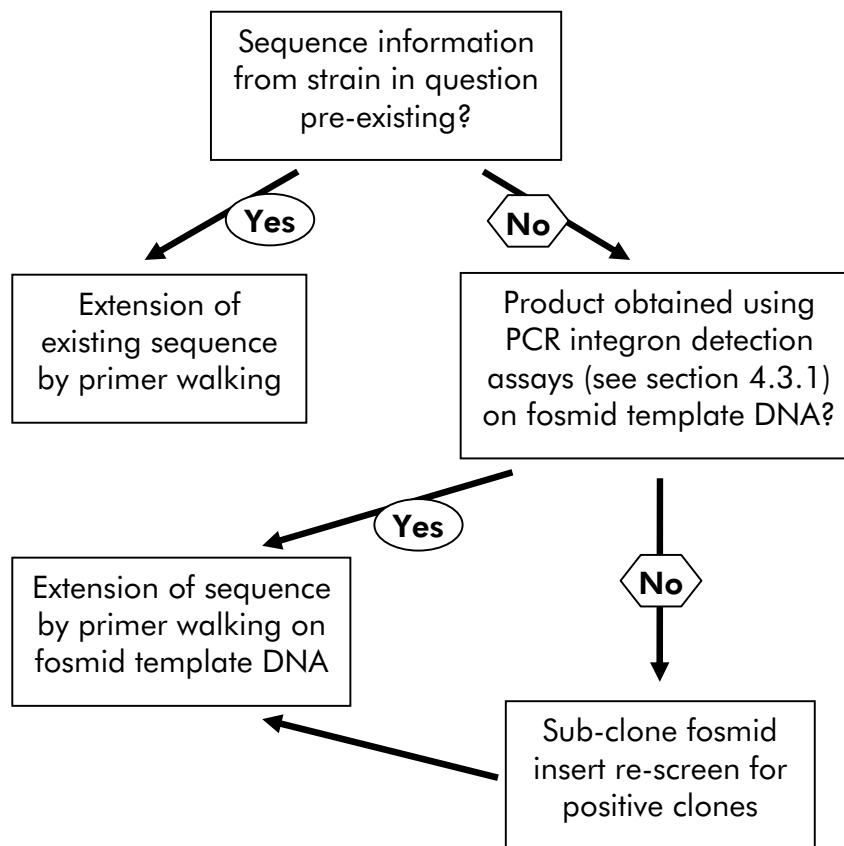


Figure A.1 - Flowchart representing the strategy employed to recover nucleotide sequence information from fosmid clones containing sequences of interest. Primer walking from pre-existing sequence was the favoured method. PCR to generate a starting template was the next favoured method, followed by sub-cloning of the fosmid clone to create a small-insert library.

### A.2.6 - PCR recovery of additional integron boundary sequences

Information obtained from fosmid clones on conserved sequences in the regions upstream and downstream of recovered integrons was used to design PCR assays aimed at recovering the upstream and downstream integron boundaries from selected strains in the collection. Two different PCR assays aimed at recovering upstream boundary sequences were used. For both assays, the PCR primer NW34 (5'-CCTTGGASAGVGTHTCGATG-3') was used to target *intI*. This primer binds to the same region of *intI* as NW33, but on the opposite strand. As several different upstream sequences were observed among the integrons sequenced from fosmid clones, two different primers each targeting a different conserved sequence motif in this region were used in conjunction with NW34 to amplify the upstream integron boundary from as many strains in the collection as possible. Primer NW59 (5'-GTGAARACYACSMWGATMTAYACCC-3') was designed on the basis of conserved upstream sequences of InPstQ and InPmeNW1 and binds within a gene that is homologous to gene PA3672 of *Ps. aeruginosa* PAO1. Primer NW42 (5'-GTGGTCAAYACCTGYGGYTT-3') was designed on the basis of conserved upstream sequences of InPstRNAIII, InPstDSM50238 and InPstrKM91 and binds within a gene that is homologous to gene PA0916 of *Ps. aeruginosa* PAO1. Identical conditions were used for both the NW34/NW59 and NW34/NW42 PCR assays. Amplification mixtures were prepared as described in Section 2.3, and were run using the following cycling program: 94°C for 5 min for 1 cycle, 94°C for 30 sec, 60°C for 30 sec, 72°C for 1 min 30 sec for 35 cycles, and 72°C for 5 min for 1 cycle. The nucleotide sequence of all amplicons generated from these assays was determined by direct sequencing as described in Section 2.6.

The downstream boundary region of *Ps. stutzeri* BAM17 was recovered via PCR using the primers NW93 (5'-TTGGAAATCTTCCTGCGTTG-3') and NW66 (5'-GAACACGCTSACCGGRAYGA-3'). The primer NW93 targets the ORF of last detectable gene cassette and primer NW66 targets the first gene downstream of the integron which is present in both InPstQ and InPmeNW1. Amplification mixtures were prepared as described in Section 2.3, and were run using the following cycling program: 94°C for 5 min for 1 cycle, 94°C for 30 sec, 60°C for 30 sec, 72°C for 3 min for 35 cycles, and 72°C for 5 min for 1 cycle. The nucleotide sequence of the amplicon generated from this assay was determined using a combination of cloning and direct sequencing as described in Section 2.6.

#### **A.2.7 - Hybridisation screening for integron flanking genes.**

Using sequence information obtained from integron sequences in fosmid clones, hybridisation probes were designed (as described in section 2.5) to target the upstream and downstream flanking genes of integrons at two different loci. The genes flanking integrons in *Ps. stutzeri* Gv.7 (homologous to PA0916 and PA3388 of *Ps. aeruginosa* PAO1) and *Ps. stutzeri* Gv.8 (homologous to PA3672 and PA3673 of *Ps. aeruginosa* PAO1) were targeted. The primer sequences used for the synthesis of probes targeting these genes are provided in Table A.1. Primers were designed to target conserved regions within the gene (determined from amino acid alignments of gene sequences from *Pseudomonas* spp.) and produce an amplicon 200-500 bp in size. *Ps. stutzeri* RNAIII and *Ps. stutzeri* Q cDNA were used as template for amplifying flanking genes from integron locus 1 and 2, respectively. Amplification mixtures for probe synthesis were prepared and quantified as described in Section 2.5, and were run using the following cycling program: 94°C for 5 min for 1 cycle, 94°C for 30 sec,

62°C for 30 sec, 72°C for 1 min for 35 cycles, and 72°C for 5 min for 1 cycle.

Hybridisations were performed on *Pst*I or *Pvu*II gDNA digests of all strains as described in Section 2.5. Using colony hybridisation, fosmid clone libraries were also screened for the presence of these genes using the same probes (Table A.1), and a representative set of clones were selected for sequencing. To generate a starting template for sequencing, PCRs were performed on fosmid templates using the same PCR assay as that employed to synthesise the probe to which it hybridised.

Integron Locus	Integron flank	Homologue in <i>Ps. aeruginosa</i> PAO1	Primers	Sequence (5' - 3')
1	5'	PA0916	NW42	GTGGTCAAYACCTGYGGYTT
			NW43	CVGGRAARCCSACGATGAA
1	3'	PA3388	NW87	AAGTTCRCCATYCKCGYCA
			NW88	RTCGATKCCGGAMAGCCAMA
2	5'	PA3672	NW58	ATCGGYTAYCTSCCSGARGG
			NW59	CCTTGGASAGVGTHTCGATG
2	3'	PA3673	NW65	CGCTCSGARACRATCARCCA
			NW66	GAACACGCTSACCGGRAYGA

Table A.1 - Primer sequences used for PCR synthesis of probes targeting integron flanking genes. The same primer pairs and PCR conditions were used both for screening of the entire strain collection and direct labelling of probes for hybridisation.

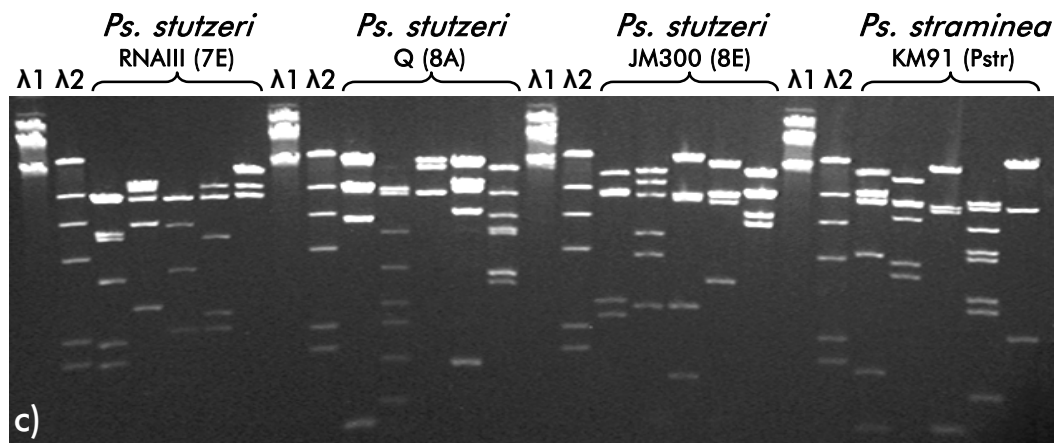
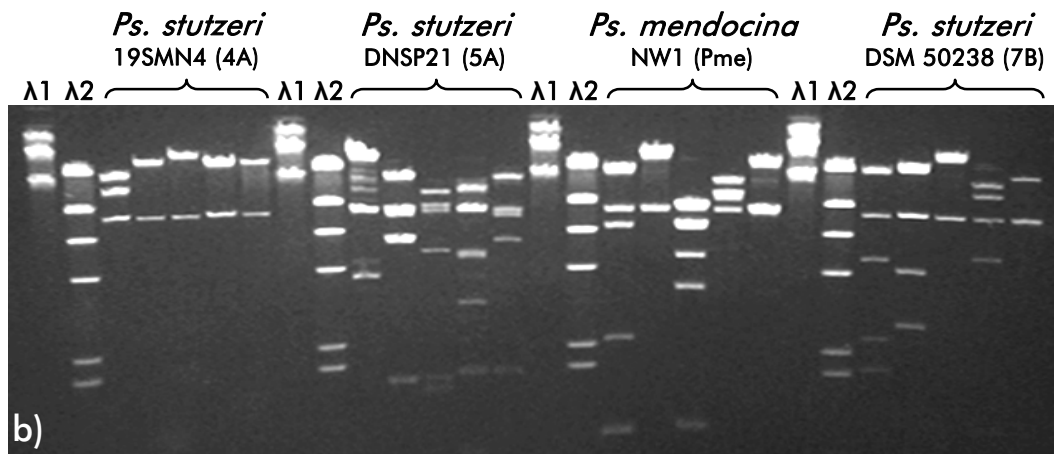
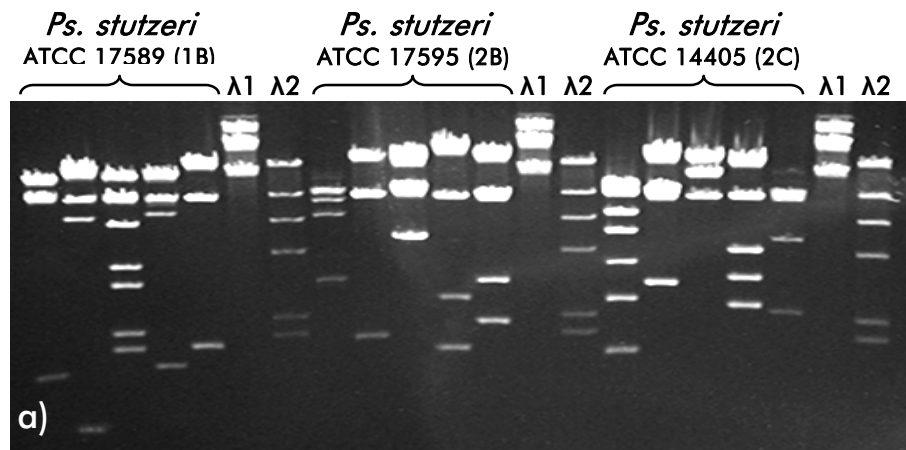
## A.3 - Results

### A.3.1 - Construction of fosmid libraries and estimation of genomic coverage

Fosmid clone libraries were successfully constructed for 11 strains in the collection (Table A.2). At least 384 clones from each library were selected for further screening. *Bam* HI RFLP profiles for five randomly selected clones are shown in Figure A.2. A unique RFLP profile was observed for all but five clones. These observations indicate that the majority of fosmid clones contain independent DNA-inserts. The absence of *Bam* HI sites in the insert DNA of several *Ps. stutzeri* 19SMN4 (4A) fosmid clones resulted in indistinguishable *Bam* HI profiles of 3 of 5 clones. Re-screening of these *Ps. stutzeri* 19SMN4 (4A) fosmid clones using a restriction enzyme which cuts more frequently (*Pvu* II) confirmed the presence of independent DNA inserts in all clones screened. Two *Ps. stutzeri* DNSP21 (5A) fosmid clones also produced identical *Bam* HI RFLP profiles (clones 2 and 5 in Figure A.2). The similarity of the RFLP profiles of these strains indicates that they are likely daughter clones. Unique RFLP profiles were observed for all remaining clones of this strain which were screened. The pCC1-FOS vector supplied with the fosmid kit (Epicentre) has *Bam* HI sites flanking the insert, resulting in a band of 8.1 kb in all *Bam* HI digests which represents the fosmid vector (Figure A.2). All remaining bands represent the cloned DNA insert, and the sum of these bands was used to determine the insert size of each clone. The average insert size across all fosmid libraries was determined to be approximately 30kb. Using this information, the genomic coverage of each library was estimated (Table A.2). The average genome coverage across all libraries was 5.3, with a maximum coverage of 5.8, and a minimum coverage of 2.8 (Table A.2).

Figure A.2 (opposite) - BamHI digests of five randomly selected fosmid clones for each strain for which libraries were constructed. A band of ~8.1 Kb is present in every fosmid digest and represents the entire fosmid vector. The remaining bands in each lane represent the cloned DNA insert. The sum of these bands gives the insert size for each clone. An average insert size of 30 Kbp was assumed for all fosmid clones.

$\lambda$ 1 – Undigested Lambda phage DNA and *Kpn*I Lambda phage DNA digest; band sizes top to bottom – 48.5 kb, 30 kb and 17 kb.  $\lambda$ 2 - HindIII Lambda phage DNA digest; band sizes top to bottom – 23.1 kb, 9.4 kb, 6.7 kb, 4.4 kb, 2.3 kb and 2 kb .





### A.3.2 - Screening fosmid clone libraries for integron sequences

A summary of hybridisation screening of fosmid libraries for integron sequences is provided in Table A.2. Multiple 59-be sequences were detected in all libraries, and signals occurred at frequencies approximately equal to or greater than the estimated coverage of the library. For strains 1B, 7B, 7E, 8A, Pstr and Pme, the number of 59-be-containing clones observed was at least double that expected on the basis of the estimated coverage of the libraries and/or the number of observed core integron-containing clones (Table A.2). As gene cassette arrays contain multiple 59-be at regular intervals and may greatly exceed 10 kb in size (Clark *et al.*, 1997; Vaisvila *et al.*, 2001), clones containing partial cassette arrays are predicted to be common. If a given cassette array was 15 Kbp in size in a clone library with an average insert size of 30 Kbp, the entire cassette array would be expected to be contained within a single clone approximately 50% of the time. As the cassette array of *Ps. alcaligenes* ATCC 55044 is 17 Kbp in length and the 59-be hybridisation data presented in Section 4.3.3 is indicative of the presence of several cassettes in strains 1B, 7B, 7E, 8A, Pstr and Pme, it is likely that the overrepresentation of 59-be clones in these strains is the result of clones which contain a fragmented portion of a large cassette array.

Alternatively, these clones may contain independent cassette arrays associated with divergent or disrupted core integrons, or gene cassette(s) inserted at secondary sites. To determine the cause of the overrepresentation of 59-be hybridisation signals, a selection of clones from these libraries were analysed further (see Section 5.3.1.4).

Core integron sequences were detected in all libraries except those of *Ps. stutzeri* Gv.1, 4 and 5 strains. Between 3 and 7 clones containing *intl*-like sequences were detected in each library (Table A.2). The frequency of core integron containing clones

for all libraries was approximately equal to or less than the estimated coverage library, which is consistent with the existence of a single-copy core integron. Core integron containing clones were underrepresented (occurring at less than half the expected frequency) in libraries from *Ps. stutzeri* DSM 50238 (7B) and RNAIII (7E) (Table A.2). This may have been due to random chance or to toxicity of genes located near the integron in these strains. In all strains except *Ps. stutzeri* ATCC 14405, •80% core integron signals co-occur with 59-be signals, indicating strong linkage between them. This is not surprising considering core integrons and gene cassettes are found in close proximity in virtually all characterised integrons. In *Ps. stutzeri* ATCC 14405 however, only 2 of 7 core integron signals co-occur with 59-be signals, suggesting that the core integron in this strain is not found in close proximity to gene cassettes with *Pseudomonas*-type 59-be.



Strain/abbreviation	Total fosmid clones	Est. DNA cloned (Mbp) <sup>1</sup>	Fold-genome coverage <sup>2</sup>	Total <i>intI</i> probe +ve	Total 59be probe +ve	Clones reacting to both probes <sup>3</sup>	<i>intI</i> <sup>+</sup> only Clones <sup>4</sup>	59be <sup>+</sup> only Clones <sup>5</sup>
<i>Ps. stutzeri</i> ATCC17589 <b>1B</b>	768	23.0	5.8	N/A	<b>17</b>	N/A	N/A	17
<i>Ps. stutzeri</i> ATCC17595 <b>2B</b>	576	17.3	4.3	5	4	4	1	0
<i>Ps. stutzeri</i> ATCC14405 <b>2C</b>	768	23.0	5.8	7	11	2	5	9
<i>Ps. stutzeri</i> 19SMN4 <b>4A</b>	768	23.0	5.8	N/A	6	N/A	N/A	6
<i>Ps. stutzeri</i> DNSP21 <b>5A</b>	768	23.0	5.8	N/A	5	N/A	N/A	5
<i>Ps. stutzeri</i> DSM50238 <b>7B</b>	768	23.0	5.8	2	7	2	0	5
<i>Ps. stutzeri</i> RNAIII <b>7E</b>	672	20.2	5.0	2	9	2	0	7
<i>Ps. stutzeri</i> Q <b>8A</b>	768	23.0	5.8	6	<b>12</b>	6	0	6
<i>Ps. stutzeri</i> JM300 <b>8E</b>	384	11.5	2.9	3	2	2	1	0
<i>Ps. straminea</i> KM91 <b>Pstr</b>	768	23.0	5.8	7	<b>23</b>	7	0	16
<i>Ps. mendocina</i> NW1 <b>Pme</b>	768	23.0	5.8	5	<b>25</b>	5	0	20

Table A.2 - Summary of genome coverage of fosmid clone libraries in addition to core integron and 59-be screening of fosmid clone libraries. Instances in which the number of positive clones exceeded the predicted coverage of the library by more than 2 are highlighted in bold type.

<sup>1</sup> Average insert size of 30 Kbp (as calculated from Figure A.2) used to estimate the total amount of DNA cloned for each strain.

<sup>2</sup> A genome size of 4 Mb was assumed for all strains to calculate genome coverage.

<sup>3</sup> Clones in which both a core integron and 59-be were detected.

<sup>4</sup> Clones in which a core integron was detected, but 59-be were not.

<sup>5</sup> Clones in which 59-be were detected, but a core integron was not.

### A.3.3 - Recovery of complete integrons and flanking sequences

Multiple positive clones were obtained in all fosmid libraries screened via hybridisation with 59-be and/or *intI* probes, as summarised in Table A.2. All positive clones were streaked out on LB agar containing 12.5 µg/ml chloramphenicol and single colonies re-screened using hybridisation to confirm the presence of target sequences. The strategy employed for sequence recovery from fosmid clones is outlined in Figure A.1. Existing integron sequence information, where available for a particular strain, was used as a template for the extension of the nucleotide sequence from fosmid clones by primer walking. As no sequence information was available for integrons in *Ps. stutzeri* ATCC17595 (2B), ATCC14405 (2C), 19SMN4 (4A) or DNSP21 (5A), integron detection PCR assays were employed (as described in section 4.2.1) to generate a starting template for sequencing from fosmid clones. This method was unsuccessful for *Ps. stutzeri* ATCC17595 (2B) and ATCC14405 (2C), and thus the final path outlined in Figure A.1 was employed to generate a starting template. Core integron-containing clones from these strains were first sub-cloned into pUC and then re-screened as described previously to detect sub-clones containing sequences of interest. A three-stage approach was used for sequence recovery and analysis, the details of which are presented in Chapter 5: 1. Extension of existing integron sequences and recovery of any new integrons from strains for which fosmid libraries were constructed to determine further the overall pattern of integron diversity in *Pseudomonas* spp. (Section 5.3.1.1 – 5.3.1.5), 2. Generation of sequence data upstream and downstream of all integrons recovered, and analysis of the regions adjacent to the integron for genes which are indicative of a conserved locus and particular genomic context (Section 5.3.2), and 3. Analysis of the sequences

between flanking genes and the integron to detect additional integron-associated sequences outside the currently defined boundaries (Section 5.3.6). The sequences of all primers used to generate additional nucleotide sequence data, and a detailed analysis of all sequences recovered are provided in the electronic appendix (Filenames – ‘primers.xls’ and ‘sequence maps\sequence diagrams.ppt’, respectively).

#### **A.3.4 - PCR recovery of additional integron boundary regions**

Sequence information generated in Section A.3.2 was used to design primers for PCR recovery of sequences upstream and downstream of integrons from strains for which fosmid libraries were not constructed. All sequences generated using this approach are analysed in detail in Section 5.3.2.

##### **A.3.4.1 - Recovery of 5' boundary regions**

The PCR assay targeting the 5' boundary of locus 1 integrons (as defined in Figure 5.5) resulted in the recovery of sequences from three *Ps. stutzeri* Gv.8 strains (BAM17 and ATCC 17641). PCR amplicons of identical size were obtained for all three strains. Determination of the nucleotide sequence of each of these revealed the expected sequences, the 3' end of *intI*, an intergenic spacer and the 5' end of the conserved upstream gene of this locus. These sequences are represented diagrammatically in Figure 5.5 and are analysed in detail in Section 5.3.2. No amplicons were obtained from the remaining strains in the collection using this PCR assay.

The PCR assay targeting the 5' boundary of locus 2 integrons (as defined in Figure 5.5) resulted in the recovery of sequences from three *Ps. stutzeri* Gv.7 strains (ATCC 17685, API-2-142 and YPF-42). All three strains gave PCR amplicons of identical size which contained the sequences expected on the basis of the corresponding sequences

in other strains with an integron at this locus. Each amplicon contained the 3' end of *intI*, an intergenic spacer and the 3' end of the conserved upstream gene of this locus. These sequences are represented diagrammatically in Figure 5.5 and are analysed in detail in Section 5.3.2. No PCR amplicons were obtained from the remaining strains in the collection using this PCR assay. This PCR-based approach was not employed for the recovery of upstream sequences from other integron loci, due to the lack of an *intI* gene at these loci (Figure 5.5).

#### **A.3.4.2 - Recovery of 3' boundary from *Ps. stutzeri* BAM17**

The 3' boundary of InPstBAM17 was successfully amplified using the NW93/NW66 PCR assay. A PCR amplicon of approximately 3.1 kb was obtained, which was the size expected if this region contained the same genes as the 3' integron boundary region of InPstQ. Determination of the nucleotide sequence of this PCR product revealed that all the identifiable ORFs did indeed correspond to the ORFs identified in the corresponding region of InPstQ (Figure 5.5). This sequence is analysed in detail in Section 5.3.2.2.

### A.3.5 - Screening strain collection for integron flanking genes

The conserved flanking genes of integron loci 1 (PA0916 and PA3388) and 2 (PA3672 and PA3673) were detected in the vast majority of test strains (Figure A.3). A moderate to strong, single hybridisation signal was observed in the vast majority of cases. Two signals were, however, observed for most strains in the assay targeting the 5' conserved gene of integron locus 1 (Figure A.3a). Two bands were expected in *Ps. stutzeri* Gv.7 strains, as all contain a conserved *Pst*I site within the region that the probe targets. This is a likely cause of multiple signals in other *Ps. stutzeri* strains. Thus, it is likely that when present, each flanking gene exists as a single copy, which is consistent with the observed copy-number of these genes in *Pseudomonas* spp. genome sequences. Particular flanking genes were not detected in some strains; the 3' flanking gene for locus 1 was not detected in *Ps. stutzeri* DNSP21 (5A) (Figure A.3b), and the 5' flanking gene of locus 2 was not detected in *Ps. straminea* KM91 (Figure A.3d). Re-screening of these strains confirmed this result, and these strains appear to not contain these genes. Collectively, these data indicate that all the flanking genes of integron loci 1 and 2 are present in 18 of 20 test strains. The only flanking gene detected in all control strains, *Ps. aeruginosa* NCTC 3756, *Ps. fluorescens* NCTC 7244 and *Ps. putida* F1 was PA0916 (Figure A.3a). A faint signal in *Ps. putida* F1 using the probe targeting PA3388, and another in *Ps. aeruginosa* NCTC 3756 using the probe targeting PA3673 were the only signals observed for the remaining hybridisation assays. All flanking genes except for PA3672 should be present in the control strains, on the basis of information available from completely sequenced *Pseudomonas* spp. genomes. Sequence divergence beyond the detection limits of the hybridisation assays used or loss of the target genes are equally likely reasons for the failure to detect these genes in the control strains. Nonetheless,



detection of homologues of PA0916, PA3388, and PA3673 in two or more distantly related *Pseudomonas* spp. indicates that these probes had a relatively broad specificity.

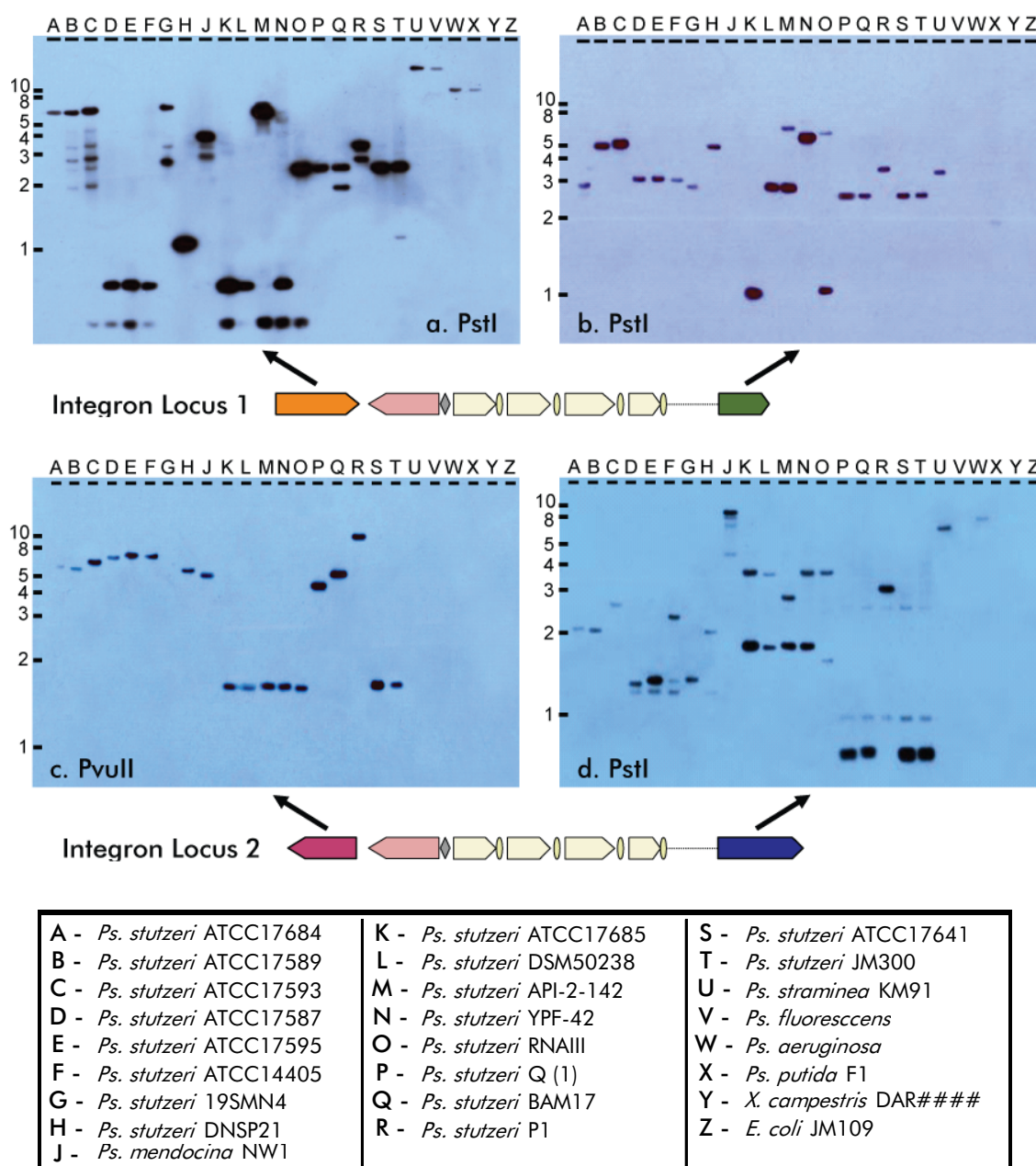


Figure A.3 - Screening for the presence of integrin flanking genes in members of the strain collection. Conserved flanking genes of integrin loci 1 and 2 were screened for using southern hybridisation. All auto-radiograms shown above represent transfers of *Pst*I or *Pvu*II gDNA digests. Lanes are labelled according to the table given above.

### A.3.6 – Screening fosmid libraries for integron flanking genes

Screening of fosmid libraries resulted in the detection of the conserved flanking genes of integron locus 1 and 2 in 9 of 11 fosmid libraries. Conserved flanking genes were not detected for integron locus 1 in *Ps. stutzeri* DNSP21, or for locus 2 in *Ps. straminea* KM91; however, fosmid clones containing each flanking gene were detected in all remaining strains (Table A.3). When detected, between 2 and 7 hybridisation signals were observed for all four genes screened (data not shown), which was consistent with each flanking gene existing as a single copy. These observations were consistent with hybridisation data obtained from gDNA digest screening (see Section A.3.6).

Results obtained regarding the co-localisation of genes from each fosmid within individual clones were consistent with available sequence data. For all strains which contain integrons at loci 1 and 2 with known 5' and 3' conserved flanking genes (*Ps. stutzeri* DSM50238 and RNAlII for locus 1 and *Ps. stutzeri* Q and *Ps. mendocina* NW1 for locus 2), fosmid clones were detected that contain the entire integron, in addition to both upstream and downstream flanking genes (Table A.3). The upstream and downstream flanking genes for integron loci 1 and 2 were also observed to co-occur in the same fosmid in strains which were not predicted to contain an integron at the relevant locus (5 of 8 strains for locus 1 and 4 of 8 strains for locus 2, Table A.3). The arrangement of these genes in strains predicted not to contain an integron is of interest in reconstructing the evolutionary history of integrons in Pseudomonads. It is possible that integrons in these strains have been lost. Alternatively, these strains may have never contained an integron at these loci, and contain integron flanking genes which have the gene order present prior to integron acquisition. Analysis of local gene

order was used to determine the nature of the genes flanking integron loci 1 and 2 across all *Pseudomonads* (See section 5.3.5).

	Integron Locus 1			Integron Locus 2		
	Both flanks detected <sup>1</sup>	Same fosmid	59be/ <i>intI</i> detected <sup>2</sup>	Both flanks detected <sup>1</sup>	Same fosmid	59be/ <i>intI</i> detected <sup>2</sup>
<i>Ps. stutzeri</i> ATCC17589 (1B)	Y	N	N	Y	Y	N
<i>Ps. stutzeri</i> ATCC17595 (2B)	Y	Y	N	Y	Y	N
<i>Ps. stutzeri</i> ATCC14405 (2C)	Y	Y	N	Y	Y	N
<i>Ps. stutzeri</i> 19SMN4 (4A)	Y	N	N	Y	Y	N
<i>Ps. stutzeri</i> DNSP21 (5A)	N	-	-	Y	N	N
<i>Ps. stutzeri</i> DSM50238 (7B)	Y	Y	Y	Y	N	N
<i>Ps. stutzeri</i> RNAIII (7E)	Y	Y	Y	Y	N	N
<i>Ps. stutzeri</i> Q (8A)	Y	Y	N	Y	Y	Y
<i>Ps. stutzeri</i> JM300 (8E)	Y	Y	N	Y	N	Y
<i>Ps. straminea</i> KM91 (Pstr)	Y	N	Y	N	-	-
<i>Ps. mendocina</i> NW1 (Pme)	Y	Y	N	Y	Y	Y

Table A.3 - Summary of fosmid library screening for flanking genes of integron loci 1 and 2 (as defined in Section 5.3.2).

- 1 – Data in this column refers to whether or not both the 5' and 3' flanking genes of each locus were detected.
- 2 - Data in this column refers to whether *intI* or 59-be hybridisation signals were co-localised with either or both flanking genes.

## A.4 - Conclusions

The implications of the data above presented above are discussed in detail in Chapter 5.



## REFERENCES

- 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.
- Andersson, G.E., and Sharp, P.M. (1996) Codon usage in the *Mycobacterium tuberculosis* complex. *Microbiology* **142** ( Pt 4): 915-925.
- Aranda-Olmedo, I., Tobes, R., Manzanera, M., Ramos, J.L., and Marques, S. (2002) Species-specific repetitive extragenic palindromic (REP) sequences in *Pseudomonas putida*. *Nucleic Acids Res* **30**: 1826-1833.
- Baggi, G., Barbieri, P., Galli, E., and Tollari, S. (1987) Isolation of a *Pseudomonas-stutzeri* Strain That Degrades O Xylene. *Applied & Environmental Microbiology* **53**: 2129-2132.
- Barlow, R.S., and Gobius, K.S. (2006) Diverse class 2 integrons in bacteria from beef cattle sources. *J Antimicrob Chemother* **58**: 1133-1138.
- Bennasar, A., Rossello-Mora, R., Lalucat, J., and Moore, E.R.B. (1996) 16S rRNA gene sequence analysis relative to genomovars of *Pseudomonas stutzeri* and proposal of *Pseudomonas balearica* sp. nov. *International Journal of Systematic Bacteriology* **46**: 200-205.
- Bennasar, A., Guasp, C., and Lalucat, J. (1998) Molecular methods for the detection and identification of *Pseudomonas stutzeri* in pure culture and environmental samples. *Microbial Ecology* **35**: 22-33.
- Biskri, L., Bouvier, M., Guerout, A.M., Boisnard, S., and Mazel, D. (2005) Comparative study of class 1 integron and *Vibrio cholerae* superintegron integrase activities. *J Bacteriol* **187**: 1740-1750.
- Boucher, Y., Nesbo, C.L., Joss, M.J., Robinson, A., Mabbutt, B.C., Gillings, M.R., Doolittle, W.F., and Stokes, H.W. (2006) Recovery and evolutionary analysis of complete integron gene cassette arrays from *Vibrio*. *BMC Evol Biol* **6**: 3.
- 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.
- Bouma, J.E., and Lenski, R.E. (1988) Evolution of a bacteria/plasmid association. *Nature* **335**: 351-352.

- Brown, H.J., Stokes, H.W., and Hall, R.M. (1996) The integrons In0, In2, and In5 are defective transposon derivatives. *J Bacteriol* **178**: 4429-4437.
- Brown, J.R., and Doolittle, W.F. (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev* **61**: 456-502.
- Burri, R., and Stutzer, A. (1895) Ueber Nitrat zerstörende Bakterien und den durch dieselben bedingten Stickstoffverlust. *Zentbl. Bakteriol. Parasitenkd. Abt. II* **1**: 257-265, 350-364, 392-398, 422-432.
- Cameron, F.H., Groot Obbink, D.J., Ackerman, V.P., and Hall, R.M. (1986) Nucleotide sequence of the AAD(2') aminoglycoside adenylyltransferase determinant aadB. Evolutionary relationship of this region with those surrounding aadA in R538-1 and dhfrII in R388. *Nucl. Acids Res.* **14**: 8625-8635.
- Carlson, C.A., Pierson, L.S., Rosen, J.J., and Ingraham, J.L. (1983) *Pseudomonas stutzeri* and related species undergo natural transformation. *J Bacteriol* **153**: 93-99.
- Ceccarelli, D., Salvia, A.M., Sami, J., Cappuccinelli, P., and Colombo, M.M. (2006) New cluster of plasmid-located class 1 integrons in *Vibrio cholerae* O1 and a dfrA15 cassette-containing integron in *Vibrio parahaemolyticus* isolated in Angola. *Antimicrob Agents Chemother* **50**: 2493-2499.
- Chauhan, S., Barbieri, P., and Wood, T.K. (1998) Oxidation of trichloroethylene, 1,1-dichloroethylene, and chloroform by toluene/o-xylene monooxygenase from *Pseudomonas stutzeri* OX1. *Appl Environ Microbiol* **64**: 3023-3024.
- Cladera, A.M., Bennasar, A., Barcel, M., Lalucat, J., and Garcfa-Valds, E. (2004) Comparative genetic diversity of *Pseudomonas stutzeri* genomovars, clonal structure, and phylogeny of the species. *Journal of Bacteriology* **186**: 5239-5248.
- Clark, C.A., Purins, L., Kaewrakon, P., and Manning, P.A. (1997) VCR repetitive sequence elements in the *Vibrio cholerae* chromosome constitute a mega-integron. *Mol Microbiol* **26**: 1137-1138.
- Coleman, N., Tetu, S., Wilson, N., and Holmes, A. (2004) An unusual integron in *Treponema denticola*. *Microbiology* **150**: 3524-3526.
- Coleman, N.V., and Holmes, A.J. (2005) The native *Pseudomonas stutzeri* strain Q chromosomal integron can capture and express cassette-associated genes. *Microbiology* **151**: 1853-1864.

- Collis, C.M., and Hall, R.M. (1992a) Site-specific deletion and rearrangement of integron insert genes catalyzed by the integron DNA integrase. *J Bacteriol* **174**: 1574-1585.
- Collis, C.M., and Hall, R.M. (1992b) Gene cassettes from the insert region of integrons are excised as covalently closed circles. *Mol Microbiol* **6**: 2875-2885.
- Collis, C.M., Grammaticopoulos, G., Briton, J., Stokes, H.W., and Hall, R.M. (1993) Site-specific insertion of gene cassettes into integrons. *Mol Microbiol* **9**: 41-52.
- Collis, C.M., and Hall, R.M. (1995) Expression of antibiotic resistance genes in the integrated cassettes of integrons. *Antimicrob Agents Chemother* **39**: 155-162.
- Collis, C.M., Kim, M.J., Stokes, H.W., and Hall, R.M. (1998) Binding of the purified integron DNA integrase IntI1 to integron- and cassette-associated recombination sites. *Mol Microbiol* **29**: 477-490.
- Collis, C.M., Kim, M.J., Partridge, S.R., Stokes, H.W., and Hall, R.M. (2002a) Characterization of the class 3 integron and the site-specific recombination system it determines. *J Bacteriol* **184**: 3017-3026.
- Collis, C.M., Kim, M.J., Stokes, H.W., and Hall, R.M. (2002b) Integron-encoded IntI integrases preferentially recognize the adjacent cognate attI site in recombination with a 59-be site. *Mol Microbiol* **46**: 1415-1427.
- Cortez, D.Q., Lazcano, A., and Becerra, A. (2005) Comparative analysis of methodologies for the detection of horizontally transferred genes: a reassessment of first-order Markov models. *In Silico Biol* **5**: 581-592.
- Dagan, T., and Martin, W. (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci U S A* **104**: 870-875.
- Daubin, V., Lerat, E., and Perriere, G. (2003) The source of laterally transferred genes in bacterial genomes. *Genome Biol* **4**: R57.
- Davies, R.L., Campbell, S., and Whittam, T.S. (2002) Mosaic structure and molecular evolution of the leukotoxin operon (lktCABD) in *Mannheimia* (*Pasteurella*) *haemolytica*, *Mannheimia glucosida*, and *Pasteurella trehalosi*. *J Bacteriol* **184**: 266-277.
- de Daruvar, A., Collado-Vides, J., and Valencia, A. (2002) Analysis of the cellular functions of *Escherichia coli* operons and their conservation in *Bacillus subtilis*. *J Mol Evol* **55**: 211-221.



- Drouin, F., Melancon, J., and Roy, P.H. (2002) The Intl-like tyrosine recombinase of *Shewanella oneidensis* is active as an integron integrase. *J Bacteriol* **184**: 1811-1815.
- Ermolaeva, M.D., White, O., and Salzberg, S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res* **29**: 1216-1221.
- Espeli, O., and Boccard, F. (1997) In vivo cleavage of *Escherichia coli* BIME-2 repeats by DNA gyrase: genetic characterization of the target and identification of the cut site. *Mol Microbiol* **26**: 767-777.
- Espeli, O., Moulin, L., and Boccard, F. (2001) Transcription attenuation associated with bacterial repetitive extragenic BIME elements. *J Mol Biol* **314**: 375-386.
- Felsenstein, J. (1989) PHYLIP -- Phylogeny Inference Package (Version 3.2). *Cladistics* **5**: 164-166.
- Fuglsang, A. (2003) The effective number of codons for individual amino acids: some codons are more optimal than others. *Gene* **320**: 185-190.
- Garcia-Valdes, E., Cozar, E., Lalucat, J., and Rotger, R. (1989) Molecular Cloning of Aromatic Degradative Genes from *Pseudomonas-Stutzeri*. *FEMS Microbiology Letters* **61**: 301-306.
- Garcia-Vallve, S., Romeu, A., and Palau, J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* **10**: 1719-1725.
- Gavini, F., Holmes, B., Izard, D., Beji, A., Bernigaud, A., and Jakubczak, E. (1989) Numerical Taxonomy of *Pseudomonas-Alcaligenes Pseudomonas-Pseudoalcaligenes Pseudomonas-Mendocina Pseudomonas-Stutzeri* and Related Bacteria. *International Journal of Systematic Bacteriology* **39**: 135-144.
- Ge, F., Wang, L.S., and Kim, J. (2005) The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol* **3**: e316.
- Gerdes, K., Christensen, S.K., and Lobner-Olesen, A. (2005) Prokaryotic toxin-antitoxin stress response loci. *Nat Rev Microbiol* **3**: 371-382.
- Gil, R., Silva, F.J., Pereto, J., and Moya, A. (2004) Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev* **68**: 518-537.
- Gillings, M., and Holley, M. (1997) Amplification of anonymous DNA fragments using pairs of long primers generates reproducible DNA fingerprints that are sensitive to genetic variation. *Electrophoresis* **18**: 1512-1518.

- Gillings, M.R., Holley, M.P., Stokes, H.W., and Holmes, A.J. (2005) Integrons in *Xanthomonas*: a source of species genome diversity. *Proc Natl Acad Sci U S A* **102**: 4419-4424.
- Gilson, E., Perrin, D., and Hofnung, M. (1990) DNA polymerase I and a protein complex bind specifically to *E. coli* palindromic unit highly repetitive DNA: implications for bacterial chromosome organization. *Nucleic Acids Res* **18**: 3941-3952.
- Ginard, M., Lalucat, J., Tuemmler, B., and Romling, U. (1997) Genome organization of *Pseudomonas stutzeri* and resulting taxonomic and evolutionary considerations. *International Journal of Systematic Bacteriology* **47**: 132-143.
- Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res* **9**: r43-74.
- Grocock, R.J., and Sharp, P.M. (2002) Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene* **289**: 131-139.
- Guasp, C., Moore, E.R.B., Lalucat, J., and Bennasar, A. (2000) Utility of internally transcribed 16S-23S rDNA spacer regions for the definition of *Pseudomonas stutzeri* genomovars and other *Pseudomonas* species. *International Journal of Systematic & Evolutionary Microbiology* **50**: 1629-1639.
- Gupta, S.K., Bhattacharyya, T.K., and Ghosh, T.C. (2004) Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection. *J Biomol Struct Dyn* **21**: 527-536.
- Hall, R.M., and Vockler, C. (1987) The region of the *incN* plasmid R46 coding for resistance to {beta}-lactam antibiotics, streptomycin/spectinomycin and sulphonamides is closely related to antibiotic resistance segments found in *IncW* plasmids and in Tn21-like transposons. *Nucl. Acids Res.* **15**: 7491-7501.
- Hall, R.M., Brookes, D.E., and Stokes, H.W. (1991) Site-specific insertion of genes into integrons: role of the 59-base element and determination of the recombination cross-over point. *Mol Microbiol* **5**: 1941-1959.
- Hall, T. (2001) BioEdit version 5.0.6. Carlsbad, California: Ibis Therapeutics.
- Hansson, K., Sundstrom, L., Pelletier, A., and Roy, P.H. (2002) *IntI2* integron integrase in Tn7. *J Bacteriol* **184**: 1712-1721.

- Hayes, W.S., and Borodovsky, M. (1998) How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res* **8**: 1154-1171.
- Heidelberg, J.F., Paulsen, I.T., Nelson, K.E., Gaidos, E.J., Nelson, W.C., Read, T.D., Eisen, J.A., Seshadri, R., Ward, N., Methe, B., Clayton, R.A., Meyer, T., Tsapin, A., Scott, J., Beanan, M., Brinkac, L., Daugherty, S., DeBoy, R.T., Dodson, R.J., Durkin, A.S., Haft, D.H., Kolonay, J.F., Madupu, R., Peterson, J.D., Umayam, L.A., White, O., Wolf, A.M., Vamathevan, J., Weidman, J., Impraim, M., Lee, K., Berry, K., Lee, C., Mueller, J., Khouri, H., Gill, J., Utterback, T.R., McDonald, L.A., Feldblyum, T.V., Smith, H.O., Venter, J.C., Neilson, K.H., and Fraser, C.M. (2002) Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. *Nat Biotechnol* **20**: 1118-1123.
- Hendrix, R.W., Lawrence, J.G., Hatfull, G.F., and Casjens, S. (2000) The origins and ongoing evolution of viruses. *Trends Microbiol* **8**: 504-508.
- Heuer, T., Burger, C., and Tummeler, B. (1998) Smith/Birnstiel mapping of genome rearrangements in *Pseudomonas aeruginosa*. *Electrophoresis* **19**: 495-499.
- Hill, M.O. (1974) Correspondence analysis: a neglected multivariate method. *Appl Stat* **23**: 340-354.
- Hilton, T., Rosche, T., Froelich, B., Smith, B., and Oliver, J. (2006) Capsular polysaccharide phase variation in *Vibrio vulnificus*. *Appl Environ Microbiol* **72**: 6986-6993.
- 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 constains. *Antimicrob Agents Chemother* **45**: 2991-3000.
- Holmes, A.J., Gillings, M.R., Nield, B.S., Mabbutt, B.C., Nevalainen, K.M., and Stokes, H.W. (2003a) The gene cassette metagenome is a basic resource for bacterial genome evolution. *Environ Microbiol* **5**: 383-394.
- Holmes, A.J., Holley, M.P., Mahon, A., Nield, B., Gillings, M., and Stokes, H.W. (2003b) Recombination activity of a distinctive integron-gene cassette system associated with *Pseudomonas stutzeri* populations in soil. *Journal of Bacteriology* **185**: 918-928.
- Holmes, B. (1986) Identification and Distribution of *Pseudomonas-Stutzeri* in Clinical Material. *Journal of Applied Bacteriology* **60**: 401-412.

- Hooper, S.D., and Berg, O.G. (2002) Detection of genes with atypical nucleotide sequence in microbial genomes. *J Mol Evol* **54**: 365-375.
- Horikoshi, K., and Grant, W.D. (1998) *Extremophiles. Microbial Life in Extreme Environments*. New York: Wiley-Liss.
- Hubalek, Z., Pacova, Z., Halouzka, J., Sedlacek, I., Dlouhy, M., and Honza, M. (1998) Selective isolation of *Pseudomonas stutzeri* from vertebrate faeces on Rambach agar. *Zentralblatt fuer Bakteriologie* **288**: 343-349.
- Itoh, T., Takemoto, K., Mori, H., and Gojobori, T. (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol* **16**: 332-346.
- Jacob, F., Sussman, R., and Monod, J. (1962) On the nature of the repressor ensuring the immunity of lysogenic bacteria. *C R Hebd Seances Acad Sci* **254**: 4214-4216.
- Komano, T. (1999) Shufflons: multiple inversion systems and integrons. *Annu Rev Genet* **33**: 171-191.
- Koonin, E.V., and Galperin, M.Y. (1997) Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr Opin Genet Dev* **7**: 757-763.
- Koonin, E.V., Mushegian, A.R., Galperin, M.Y., and Walker, D.R. (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol Microbiol* **25**: 619-637.
- Koski, L.B., Morton, R.A., and Golding, G.B. (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol* **18**: 404-412.
- Kutsukake, K., Nakashima, H., Tominaga, A., and Abo, T. (2006) Two DNA invertases contribute to flagellar phase variation in *Salmonella enterica* serovar Typhimurium strain LT2. *J Bacteriol* **188**: 950-957.
- Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M., and Wolfe, K.H. (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res* **27**: 1642-1649.
- Lafay, B., Atherton, J.C., and Sharp, P.M. (2000) Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology* **146** ( Pt 4): 851-860.

- Lanoot, B., Vancanneyt, M., Dawyndt, P., Cnockaert, M., Zhang, J., Huang, Y., Liu, Z., and Swings, J. (2004) BOX-pCR fingerprinting as a powerful tool to reveal synonymous names in the genus *Streptomyces*. Emended descriptions are proposed for the species *Streptomyces cinereorectus*, *S. fradiae*, *S. tricolor*, *S. colombiensis*, *S. filamentosus*, *S. vinaceus* and *S. phaeopurpureus*. *Syst Appl Microbiol* **27**: 84-92.
- Laraki, N., Galleni, M., Thamm, I., Riccio, M.L., Amicosante, G., Frere, J.M., and Rossolini, G.M. (1999) Structure of In31, a blaIMP-containing *Pseudomonas aeruginosa* integron phyletically related to In5, which carries an unusual array of gene cassettes. *Antimicrob Agents Chemother* **43**: 890-901.
- Lawrence, J. (1999) Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr Opin Genet Dev* **9**: 642-648.
- Lawrence, J.G., and Roth, J.R. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**: 1843-1860.
- Lawrence, J.G., and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**: 383-397.
- Lawrence, J.G. (2001) Catalyzing bacterial speciation: correlating lateral transfer with genetic headroom. *Syst Biol* **50**: 479-496.
- Leon, G., and Roy, P.H. (2003) Excision and integration of cassettes by an integron integrase of *Nitrosomonas europaea*. *J Bacteriol* **185**: 2036-2041.
- Leverstein-van Hall, M.A., Box, A.T., Blok, H.E., Paauw, A., Fluit, A.C., and Verhoef, J. (2002) Evidence of extensive interspecies transfer of integron-mediated antimicrobial resistance genes among multidrug-resistant Enterobacteriaceae in a clinical setting. *J Infect Dis* **186**: 49-56.
- Levesque, C., Brassard, S., Lapointe, J., and Roy, P.H. (1994) Diversity and relative strength of tandem promoters for the antibiotic-resistance genes of several integrons. *Gene* **142**: 49-54.
- Lorenz, M.G., and Sikorski, J. (2000) The potential for intraspecific horizontal gene exchange by natural genetic transformation: Sexual isolation among genomovars of *Pseudomonas stutzeri*. *Microbiology* **146**: 3081-3090.
- Louws, F.J., Fulbright, D.W., Stephens, C.T., and de Bruijn, F.J. (1994) Specific genomic fingerprints of phytopathogenic *Xanthomonas* and *Pseudomonas* pathovars and strains generated with repetitive sequences and PCR. *Appl Environ Microbiol* **60**: 2286-2295.

- Madigan, M.T., and Mairs, B.L. (1997) Extremophiles. *Sci Am* **276**: 82-87.
- Makarova, K.S., Aravind, L., Galperin, M.Y., Grishin, N.V., Tatusov, R.L., Wolf, Y.I., and Koonin, E.V. (1999) Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res* **9**: 608-628.
- Martin, B., Humbert, O., Camara, M., Guenzi, E., Walker, J., Mitchell, T., Andrew, P., Prudhomme, M., Alloing, G., Hakenbeck, R., and et al. (1992) A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. *Nucleic Acids Res* **20**: 3479-3483.
- Martinez, E., and De La Cruz, F. (1990) Genetic Elements Involved in Tn21 Site-Specific Integration a Novel Mechanism for the Dissemination of Antibiotic Resistance Genes. *EMBO (European Molecular Biology Organization) Journal* **9**: 1275-1282.
- Masco, L., Huys, G., Gevers, D., Verbruggen, L., and Swings, J. (2003) Identification of *Bifidobacterium* species using rep-PCR fingerprinting. *Syst Appl Microbiol* **26**: 557-563.
- 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.
- Mazel, D. (2006) Integrons: agents of bacterial evolution. *Nat Rev Microbiol* **4**: 608-620.
- McNichol, K. (2002) *Pseudomonas* integrons: the Role of Integrons in bacterial evolution. Honours Thesis, Macquarie University.
- Medigue, C., Rouxel, T., Vigier, P., Henaut, A., and Danchin, A. (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* **222**: 851-856.
- Messier, N., and Roy, P.H. (2001) Integron integrases possess a unique additional domain necessary for activity. *J Bacteriol* **183**: 6699-6706.
- Michael, C.A., Gillings, M.R., Holmes, A.J., Hughes, L., Andrew, N.R., Holley, M.P., and Stokes, H.W. (2004) Mobile gene cassettes: a fundamental resource for bacterial evolution. *Am Nat* **164**: 1-12.
- Mirkin, B.G., Fenner, T.I., Galperin, M.Y., and Koonin, E.V. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* **3**: 2.

- Moore, E.R.B., Mau, M., Arnscheidt, A., Bottger, E.C., Hutson, R.A., Collins, M.D., Van De Peer, Y., De Wachter, R., and Timmis, K.N. (1996) The determination and comparison of the 16S rRNA gene sequences of species of the genus *Pseudomonas* (sensu stricto) and estimation of the natural intrageneric relationships. *Systematic & Applied Microbiology* **19**: 478-492.
- Mosqueda, G., and Ramos, J.L. (2000) A set of genes encoding a second toluene efflux system in *Pseudomonas putida* DOT-T1E is linked to the *tod* genes for toluene metabolism. *J Bacteriol* **182**: 937-943.
- Mushegian, A.R., and Koonin, E.V. (1996) Gene order is not conserved in bacterial evolution. *Trends Genet* **12**: 289-290.
- Nei, M., and Li, W.H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A* **76**: 5269-5273.
- Nemergut, D.R., Martin, A.P., and Schmidt, S.K. (2004) Integron diversity in heavy-metal-contaminated mine tailings and inferences about integron evolution. *Appl Environ Microbiol* **70**: 1160-1168.
- Nesbo, C.L., Boucher, Y., and Doolittle, W.F. (2001) Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. *J Mol Evol* **53**: 340-350.
- Nicholas, K.B., Jr., N.H.B., and Deerfield, D.W.I., 4:14 (1997) GeneDoc: Analysis and Visualization of Genetic Variation. *EMBNEW.NEWS* **4**: 14.
- Nield, B.S., Holmes, A.J., Gillings, M.R., Recchia, G.D., Mabbutt, B.C., Nevalainen, K.M., and Stokes, H.W. (2001) Recovery of new integron classes from environmental DNA. *FEMS Microbiol Lett* **195**: 59-65.
- Noble, R.C., and Overman, S.B. (1994) *Pseudomonas stutzeri* infection. A review of hospital isolates and a review of the literature. *Diagn Microbiol Infect Dis* **19**: 51-56.
- Normand, P., Orso, S., Cournoyer, B., Jeannin, P., Chapelon, C., Dawson, J., Evtushenko, L., and Misra, A.K. (1996) Molecular phylogeny of the genus *Frankia* and related genera and emendation of the family Frankiaceae. *Int J Syst Bacteriol* **46**: 1-9.
- Nunes-Duby, S.E., Kwon, H.J., Tirumalai, R.S., Ellenberger, T., and Landy, A. (1998) Similarities and differences among 105 members of the *Int* family of site-specific recombinases. *Nucleic Acids Res* **26**: 391-406.

- Obradors, N., and Aguilar, J. (1991) Efficient Biodegradation of High-Molecular-Weight Polyethylene Glycols by Pure Cultures of *Pseudomonas-Stutzeri*. *Applied & Environmental Microbiology* **57**: 2383-2388.
- Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299-304.
- Osborn, A.M., and Boltner, D. (2002) When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. *Plasmid* **48**: 202-212.
- Ouellette, M., Bissonnette, L., and Roy, P.H. (1987) Precise Insertion of Antibiotic Resistance Determinants into Transposon 21-Like Transposons Nucleotide Sequence of the Oxa-1 Beta Lactamase Gene. *Proceedings of the National Academy of Sciences of the United States of America* **84**: 7378-7382.
- Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**: 2896-2901.
- Page, R.D.M. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* **12**: 357-358.
- Pal, C., and Hurst, L.D. (2004) Evidence against the selfish operon theory. *Trends Genet* **20**: 232-234.
- Palleroni, N.J., Doudoroff, M., Stanier, R.Y., Solnes, R.E., and Mandel, M. (1970) Taxonomy of the Aerobic *Pseudomonads* the Properties of the *Pseudomonas-Stutzeri* Group. *Journal of General Microbiology* **60**: 215-231.
- Palleroni, N.J., Kunisawa, R., Contopoulou, R., and Doudoroff, M. (1973) Nucleic-Acid Homologies in the Genus *Pseudomonas*. *International Journal of Systematic Bacteriology* **23**: 333-339.
- Palleroni, N.J. (1992) Present Situation of the Taxonomy of Aerobic *Pseudomonads*. In *Pseudomonas: Molecular Biology and Biotechnology; Third International Symposium on Pseudomonads: Biology and Biotechnology*. Galli, E., Silver, S. and B., W. (eds). Washington: American Society for Microbiology, pp. 105-115.
- Pandey, D.P., and Gerdes, K. (2005) Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Research* **33**: 966-976.



- Papapetropoulou, M., Iliopoulou, J., Rodopoulou, G., Detorakis, J., and Paniara, O. (1994) Occurrence and antibiotic-resistance of pseudomonas species isolated from drinking water in southern Greece. *Journal of Chemotherapy* **6**: 111-116.
- Partridge, S.R., Recchia, G.D., Scaramuzzi, C., Collis, C.M., Stokes, H.W., and Hall, R.M. (2000) Definition of the attI1 site of class 1 integrons. *Microbiology* **146** (Pt 11): 2855-2864.
- Partridge, S.R., Recchia, G.D., Stokes, H.W., and Hall, R.M. (2001) Family of class 1 integrons related to In4 from Tn1696. *Antimicrob Agents Chemother* **45**: 3014-3020.
- Peden, J.F. (1999) Analysis of codon usage. PhD Thesis. UK: University of Nottingham.
- Perriere, G., and Thioulouse, J. (2002) Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res* **30**: 4548-4555.
- Petroni, A., Corso, A., Melano, R., Cacace, M.L., Bru, A.M., Rossi, A., and Galas, M. (2002) Plasmidic extended-spectrum beta-lactamases in *Vibrio cholerae* O1 El Tor isolates in Argentina. *Antimicrob Agents Chemother* **46**: 1462-1468.
- Ploy, M.C., Denis, F., Courvalin, P., and Lambert, T. (2000) Molecular characterization of integrons in *Acinetobacter baumannii*: description of a hybrid class 2 integron. *Antimicrob Agents Chemother* **44**: 2684-2688.
- Poirel, L., Brinas, L., Fortineau, N., and Nordmann, P. (2005) Integron-encoded GES-type extended-spectrum beta-lactamase with increased activity toward aztreonam in *Pseudomonas aeruginosa*. *Antimicrob Agents Chemother* **49**: 3593-3597.
- Poole, A.M., Phillips, M.J., and Penny, D. (2003) Prokaryote and eukaryote evolvability. *Biosystems* **69**: 163-185.
- Price, M.N., Huang, K.H., Alm, E.J., and Arkin, A.P. (2005a) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res* **33**: 880-892.
- Price, M.N., Huang, K.H., Arkin, A.P., and Alm, E.J. (2005b) Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res* **15**: 809-819.
- Price, M.N., Alm, E.J., and Arkin, A.P. (2006) The histidine operon is ancient. *J Mol Evol* **62**: 807-808.

- Pupo, G.M., Lan, R., and Reeves, P.R. (2000) Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A* **97**: 10567-10572.
- Radstrom, P., Skold, O., Swedberg, G., Flensburg, J., Roy, P.H., and Sundstrom, L. (1994) Transposon Tn5090 of plasmid R751, which carries an integron, is related to Tn7, Mu, and the retroelements. *J Bacteriol* **176**: 3257-3268.
- Ragan, M.A. (2001) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol Lett* **201**: 187-191.
- Rainey, P.B., Thompson, I.P., and Palleroni, N.J. (1994) Genome and fatty acid analysis of *Pseudomonas stutzeri*. *Int J Syst Bacteriol* **44**: 54-61.
- Rainey, P.B., and Cooper, T.F. (2004) Evolution of bacterial diversity and the origins of modularity. *Res Microbiol* **155**: 370-375.
- Recchia, G.D., Stokes, H.W., and Hall, R.M. (1994) Characterisation of specific and secondary recombination sites recognised by the integron DNA integrase. *Nucleic Acids Res* **22**: 2071-2078.
- Rius, N., Fuste, M.C., Guasp, C., Lalucat, J., and Loren, J.G. (2001) Clonal population structure of *Pseudomonas stutzeri*, a species with exceptional genetic diversity. *Journal of Bacteriology* **183**: 736-744.
- Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., Tatusov, R.L., Szekely, L.A., and Koonin, E.V. (2002) Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res* **30**: 2212-2223.
- Romero, H., Zavala, A., and Musto, H. (2000) Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res* **28**: 2084-2090.
- Romling, U., Schmidt, K.D., and Tummeler, B. (1997) Large genome rearrangements discovered by the detailed analysis of 21 *Pseudomonas aeruginosa* clone C isolates found in environment and disease habitats. *J Mol Biol* **271**: 386-404.
- Rossello-Mora, R.A., Lalucat, J., Dott, W., and Kaempfer, P. (1994a) Biochemical and chemotaxonomic characterization of *Pseudomonas stutzeri* genomovars. *Journal of Applied Bacteriology* **76**: 226-233.
- Rossello-Mora, R.A., Lalucat, J., and Garcia-Valdes, E. (1994b) Comparative biochemical and genetic analysis of naphthalene degradation among *Pseudomonas stutzeri* strains. *Applied & Environmental Microbiology* **60**: 966-972.

- Rossello, R., Garcia-Valdes, E., Lulucat, J., and Ursing, J. (1991) Genotypic and Phenotypic Diversity of *Pseudomonas-Stutzeri*. *Systematic & Applied Microbiology* **14**: 150-157.
- Rothschild, L.J., and Mancinelli, R.L. (2001) Life in extreme environments. *Nature* **409**: 1092-1101.
- Rowe-Magnus, D.A., Guerout, A.M., Ploncard, P., Dychinco, B., Davies, J., and Mazel, D. (2001) The evolutionary history of chromosomal super-integrations provides an ancestry for multiresistant integrations. *Proc Natl Acad Sci U S A* **98**: 652-657.
- Rowe-Magnus, D.A., Guerout, A.M., and Mazel, D. (2002) Bacterial resistance evolution by recruitment of super-integron gene cassettes. *Mol Microbiol* **43**: 1657-1669.
- Rowe-Magnus, D.A., Guerout, A.M., Biskri, L., Bouige, P., and Mazel, D. (2003) Comparative analysis of superintegrations: engineering extensive genetic diversity in the Vibrionaceae. *Genome Res* **13**: 428-442.
- Sambrook, J., and Russell, D.W. (2001) *Molecular cloning : a laboratory manual* New York: Cold Spring Harbor.
- Sawada, H., Kanaya, S., Tsuda, M., Suzuki, F., Azegami, K., and Saitou, N. (2002) A phylogenomic study of the OCTase genes in *Pseudomonas syringae* pathovars: the horizontal transfer of the argK-tox cluster and the evolutionary history of OCTase genes on their genomes. *J Mol Evol* **54**: 437-457.
- Schmidt, K.D., Tummeler, B., and Romling, U. (1996) Comparative genome mapping of *Pseudomonas aeruginosa* PAO with *P. aeruginosa* C, which belongs to a major clone in cystic fibrosis patients and aquatic habitats. *Journal of Bacteriology* **178**: 85-93.
- Schrag, S.J., and Perrot, V. (1996) Reducing antibiotic resistance. *Nature* **381**: 120-121.
- Sentchilo, V.S., Perebituk, A.N., Zehnder, A.J., and van der Meer, J.R. (2000) Molecular diversity of plasmids bearing genes that encode toluene and xylene metabolism in *Pseudomonas* strains isolated from different contaminated sites in Belarus. *Appl Environ Microbiol* **66**: 2842-2852.
- Sharp, P.M., and Li, W.H. (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res* **14**: 7737-7749.

- Shi, L., Fujihara, K., Sato, T., Ito, H., Garg, P., Chakrabarty, R., Ramamurthy, T., Nair, G.B., Takeda, Y., and Yamasaki, S. (2006) Distribution and characterization of integrons in various serogroups of *Vibrio cholerae* strains isolated from diarrhoeal patients between 1992 and 2000 in Kolkata, India. *J Med Microbiol* **55**: 575-583.
- Sikorski, J., Graupner, S., Lorenz, M.G., and Wackernagel, W. (1998) Natural genetic transformation of *Pseudomonas stutzeri* in a non-sterile soil. *Microbiology* **144**: 569-576.
- Sikorski, J., Rossello-Mora, R., and Lorenz, M.G. (1999) Analysis of genotypic diversity and relationships among *Pseudomonas stutzeri* strains by PCR-based genomic fingerprinting and multilocus enzyme electrophoresis. *Syst Appl Microbiol* **22**: 393-402.
- Sikorski, J., Mohle, M., and Wackernagel, W. (2002) Identification of complex composition, strong strain diversity and directional selection in local *Pseudomonas stutzeri* populations from marine sediment and soils. *Environ Microbiol* **4**: 465-476.
- Sikorski, J., Lalucat, J., and Wackernagel, W. (2005) Genomovars 11 to 18 of *Pseudomonas stutzeri*, identified among isolates from soil and marine sediment. *International Journal of Systematic & Evolutionary Microbiology* **55**: 1767-1770.
- Smith, M.W., Feng, D.F., and Doolittle, R.F. (1992) Evolution by acquisition: the case for horizontal gene transfers. *Trends Biochem Sci* **17**: 489-493.
- Spiers, A.J., Buckling, A., and Rainey, P.B. (2000) The causes of *Pseudomonas* diversity. *Microbiology* **146** ( Pt 10): 2345-2350.
- Stainer, N.J., Palleroni, N.J., and Doudoroff, M. (1966) The aerobic pseudomonads: a taxonomic study. *Journal of General Microbiology* **43**: 159-271.
- Stokes, H.W., and Hall, R.M. (1989) A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. *Mol Microbiol* **3**: 1669-1683.
- Stokes, H.W., Holmes, A.J., Nield, B.S., Holley, M.P., Nevalainen, K.M., Mabbutt, B.C., and Gillings, M.R. (2001) Gene cassette PCR: sequence-independent recovery of entire genes from environmental DNA. *Appl Environ Microbiol* **67**: 5240-5246.
- Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrenner, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L.,

- Goltry, L., Tolentino, E., Westbrook-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.
- Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci U S A* **85**: 2653-2657.
- Summers, A.O. (2006) Genetic linkage and horizontal gene transfer, the roots of the antibiotic multi-resistance problem. *Anim Biotechnol* **17**: 125-135.
- Sundstrom, L., Roy, P.H., and Skold, O. (1991) Site-Specific Insertion of Three Structural Gene Cassettes in Transposon Tn-7. *Journal of Bacteriology* **173**: 3025-3028.
- Swift, G., McCarthy, B.J., and Heffron, F. (1981) DNA sequence of a plasmid-encoded dihydrofolate reductase. *Mol Gen Genet* **181**: 441-447.
- Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**: 4673-4680.
- Tobes, R., and Pareja, E. (2005) Repetitive extragenic palindromic sequences in the *Pseudomonas syringae* pv. tomato DC3000 genome: extragenic signals for genome reannotation. *Res Microbiol* **156**: 424-433.
- Tobes, R., and Pareja, E. (2006) Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements. *BMC Genomics* **7**: 62.
- Toussaint, A., and Merlin, C. (2002) Mobile elements as a combination of functional modules. *Plasmid* **47**: 26-35.
- Vaisanen, O.M., Mentu, J., and Salkinoja-Salonen, M.S. (1991) Bacteria in Food Packaging Paper and Board. *Journal of Applied Bacteriology* **71**: 130-133.
- Vaisvila, R., Morgan, R.D., Posfai, J., and Raleigh, E.A. (2001) Discovery and distribution of super-integrins among *Pseudomonads*. *Molecular Microbiology* **42**: 587-601.
- van der Woude, M.W. (2006) Re-examining the role and random nature of phase variation. *FEMS Microbiol Lett* **254**: 190-197.

- Van Niel, C.B., and Allen, M.B. (1952) A note on *Pseudomonas stutzeri*. *Journal of Bacteriology* **64**: 413-422.
- Vancanneyt, M., Torck, U., Dewettinck, D., Vaerewijck, M., and Kersters, K. (1996) Grouping of pseudomonads by SDS-PAGE of whole-cell proteins. *Systematic & Applied Microbiology* **19**: 556-568.
- Vandamme, P., Pot, B., Gillis, M., de Vos, P., Kersters, K., and Swings, J. (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev* **60**: 407-438.
- Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H., and Smith, H.O. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.
- Wang, B. (2001) Limitations of compositional approach to identifying horizontally transferred genes. *J Mol Evol* **53**: 244-250.
- Watanabe, H., Mori, H., Itoh, T., and Gojobori, T. (1997) Genome plasticity as a paradigm of eubacteria evolution. *J Mol Evol* **44 Suppl 1**: S57-64.
- Weisburg, W.G., Barns, S.M., Pelletier, D.A., and Lane, D.J. (1991) 16S ribosomal DNA amplification for phylogenetic study. *J Bacteriol* **173**: 697-703.
- Woese, C.R. (1987) Bacterial Evolution. *Microbiological Reviews* **51**: 221-271.
- Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S., and Koonin, E.V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res* **11**: 356-372.
- Worning, P., Jensen, L.J., Nelson, K.E., Brunak, S., and Ussery, D.W. (2000) Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*. *Nucleic Acids Res* **28**: 706-709.
- Wright, G.D. (2007) The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat Rev Microbiol* **5**: 175-186.
- Wright, S. (1977) *Evolution and the Genetics of Populations, Experimental Results and Evolutionary Deductions, vol. 3*. Chicago: Univ. Chicago Press.

- Xie, G., Keyhani, N.O., Bonner, C.A., and Jensen, R.A. (2003) Ancient origin of the tryptophan operon and the dynamics of evolutionary change. *Microbiol Mol Biol Rev* **67**: 303-342, table of contents.
- Yang, J., Nie, H., Chen, L., Zhang, X., Yang, F., Xu, X., Zhu, Y., Yu, J., and Jin, Q. (2007) Revisiting the molecular evolutionary history of *Shigella* spp. *J Mol Evol* **64**: 71-79.
- Zumft, W.G., and Korner, H. (1997) Enzyme diversity and mosaic gene organization in denitrification. *Antonie Van Leeuwenhoek* **71**: 43-58.