

GENOME ANALYSIS IN PLANT PATHOGENIC STREPTOMYCES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

José Carlos Huguet Tapia

February 2010

© 2010 José Carlos Huguet Tapia

GENOME ANALYSIS IN PLANT PATHOGENIC STREPTOMYCES

José Carlos Huguet Tapia, Ph. D.

Cornell University 2010

The soil bacteria *Streptomyces* have been extensively studied because of the variety of secondary metabolites that they produce, and their ability to degrade recalcitrant polymers in the environment. *Streptomyces* are mainly soil saprophytes, however, a relatively small number of species are pathogenic and cause lesions on potato and other root and tuber crops. All plant pathogenic *Streptomyces* species produce a phytotoxin, thaxtomin, which is required for pathogenicity. Additionally, these pathogens possess virulence genes that vary across species. Comparative genomic analysis of 11 *Streptomyces* spp. demonstrated that this genus possesses a surprisingly small core genome; this common set of orthologues encode biological processes that define a streptomycete. Among these are transcriptional regulators and proteins associated with morphological differentiation. Comparisons between the genomes of saprophytic and pathogenic species identified the portion of the accessory genome that is associated with pathogenesis, the patho-genome, including novel transcriptional regulators and secreted proteins with putative virulence functions.

Phylogenetic analysis of 11 species of *Streptomyces*, carried out using a subset of genes in the core genome resulted in only partial resolution of the evolutionary relationship of the genus. This analysis suggested that the plant pathogens *S. scabies* and *ipomoeae* have a common evolutionary history, while *S. turgidiscabies* is a newly evolved pathogen. The poor resolution of the phylogenetic relationships among these species resulted from extensive recombination within the core genome, affecting even informational genes. Our results suggest that homologous recombination is an important evolutionary force in the genus *Streptomyces*.

Analysis of the *Streptomyces turgidiscabies* genome revealed a large genomic island, PAISt, containing authentic and putative virulence genes and organization typical of an integrative conjugative element (ICE). PAISt contained genes expected to be involved in the integration and mobilization of the ICE. Based on bioinformatic and experimental analysis, the recombinase responsible for the integration of PAISt into the chromosome of *Streptomyces* spp. appears to represent a novel member of the tyrosine recombinase family.

BIOGRAPHICAL SKETCH

José Carlos Huguet Tapia was born on May 17, 1973, in Lima, Peru. He completed a B.A degree in Biology with mention in Microbiology at Mayor de San Marcos University, Lima, in 1998. Following graduation, José Carlos worked for four years at the National Institute of Health as a research assistant. During that time José Carlos studied the genetic variability of clinical isolates of epidemic strains of *Vibrio cholerae*. In 2003 José Carlos was accepted into the Ph.D. program in the Department of Microbiology, Cornell University.

ACKNOWLEDGMENTS

I thank Rosemary Loria, my advisor, who helped me all these years in my academic life. I would also like to thank my special committee members H el ene Marquis, Steve Zinder and Michael Stanhope for their advice and recommendations. I am grateful to the past and present members of the Loria lab for their friendship and support. I have to thank in particular Ryan Seipke, Dawn Bignell, Madhumita Joshi, Evan Johnson, and Simon Moll. I have to mention important people who helped in my research. In that sense, David Schneider, Tristan Lebefure and Genevieve DeClerck have been important people who supported my bioinformatics analysis. I have to mention my parents Rosa and Andres, my sister Carmen and brother Rodolfo. They are always in my mind and represent the most important motivation. Finally thanks to my friends here in Ithaca who collaborate to create a wonderful place to live.

TABLE OF CONTENTS

BIOGRAPHICAL SKETCH	iii
ACKNOWLEDGMENTS	iv
TABLE OF CONTENTS	v
LIST OF FIGURES	vii
LIST OF TABLES	x
CHAPTER 1: COMPARATIVE GENOMICS OF <i>Streptomyces</i> spp: PAN, CORE AND PATHO GENOMES.	
Abstract	1
Introduction	2
Methodology	4
Results and Discussion	7
Conclusions	23
References	24
CHAPTER 2: RECOMBINATION WITHIN THE CORE OF <i>Streptomyces</i> spp.	
Abstract	29
Introduction	30
Methodology	31
Results and Discussion	32
Conclusions	40
References	41

CHAPTER 3: COMPLETE GENOME SEQUENCE AND ANALYSIS OF THE
MOBILE PATHOGENICITY ISLAND (PAIS_t) IN *Streptomyces*

turgidiscabies Car8.

Abstract	44
Introduction	45
Methodology	46
Results and Discussion	50
Conclusion	77
References	78

CHAPTER 4: THE MOBILE PATHOGENICITY ISLAND PAIS_t IN

S. turgidiscabies Car8 POSSESSES A NOVEL TYROSINE RECOMBINASE

Abstract	84
Introduction	85
Materials and Methods	88
Results and Discussion	91
Conclusions	106
References	107

LIST OF FIGURES

Figure 1-1	Bioinformatic pipeline used to identify the pan, core and patho-genomes of the genus <i>Streptomyces</i> .	5
Figure 1-2	Orthologue accumulation curves for eleven <i>Streptomyces</i> spp.	10
Figure 1-3	Histogram showing the categories of orthologues that form the <i>Streptomyces</i> core genome.	11
Figure 1-4	Orthologues unique to pathogenic <i>Streptomyces</i> spp.	12
Figure 1-5	Categories of orthologues unique to pathogenic <i>Streptomyces</i> spp.	13
Figure 1-6	Proteins with atypical motifs in <i>Streptomyces</i> are found in the patho-genome.	17
Figure 1-7	SCP-like motif found in predicted proteins encoded in <i>S. scabies</i> and <i>S. ipomoeae</i> .	18
Figure 1-8	Chromosome plot of <i>S. scabies</i> and <i>S. ipomoeae</i> orthologues.	19
Figure 1-9	Chromosome plot of <i>S. scabies</i> and <i>S. turgidiscabies</i> orthologues.	20
Figure 1-10	Chromosome plot of <i>S. ipomoeae</i> and <i>S. turgidiscabies</i> orthologues.	21
Figure 1-11	16s rDNA tree of <i>Streptomyces</i> spp.	22
Figure 2-1	Recombinant regions detected by nucleotide substitution algorithms.	35
Figure 2-2	Consensus tree for 631 genes analyzed from <i>Streptomyces</i> core-genome.	36

Figure 3-1	PAISt structure and location in the chromosome of <i>S. turgidiscabies</i> Car8	56
Figure 3-2	Complete map of the PAISt in <i>S. turgidiscabies</i> Car8.	57
Figure 3-3	Predicted CDS for the PAISt classified by cell function categories.	59
Figure 3-4	Three clusters of genes within the PAISt containing putative integration/excision, conjugation, and replication functions.	63
Figure 3-5	Region of the PAISt syntenic with an integrated plasmid in <i>Corynebacterium glutamicum</i> .	64
Figure 3-6	Putative iron uptake operon encoded within the PAISt.	65
Figure 3-7	Conserved lantibiotic synthesis genes.	66
Figure 3-8	BLAST-fingerprinting plot of the PAISt.	68
Figure 3-9	PAISs1 and PAISs2 in <i>S. scabies</i> 87-22.	70
Figure 3-10	Comparison of regions containing the thaxtomin biosynthetic operon in PAISt and PAISs2.	71
Figure 3-11	Alignment and comparison of PAISt with <i>S. coelicolor</i> and <i>S. avermitilis</i> .	72
Figure 3-12	Nucleotide alignment and identity of two clusters containing the <i>tomA</i> in <i>S. turgidiscabies</i> Car8 with the syntenic region in <i>S. scabies</i> 82-22.	73
Figure 3-13	<i>S. turgidiscabies</i> Car8 possesses two copies of the gene cluster containing the <i>tomA</i> .	74
Figure 3-14	Phylogenetic tree showing the relation between the two copies of <i>tomA</i> found in <i>S. turgidiscabies</i> Car8 and <i>S. scabies</i> 87.22 PAISs1.	75

Figure 4-1	General model for tyrosine-type recombination.	87
Figure 4-2	The <i>S. turgidiscabies</i> pathogenicity island, PAISt.	94
Figure 4-3	Maps of the plasmids used for functional analysis of <i>intPAI</i> .	95
Figure 4-4	Alignment of IntPAI with characterized tyrosine recombinases	97
Figure 4-5	Structure analysis of IntPAI.	98
Figure 4-6	Functional analysis of <i>intPAI</i> .	99
Figure 4-7	Detection of multimeric integration of pIntPAI.	100
Figure 4-8	Conserved boxes of the proposed model for intPAI and recombinases obtained with intPAI as a query.	102
Figure 4-9	Alignment of recombinases in public databases obtained with intPAI as a query.	103

LIST OF TABLES

Table 1-1	<i>Streptomyces</i> spp. genomes used in this study.	6
Table 2-1	Actinobacteria used as out groups for reconstruction of phylogenetic trees.	34
Table 2-2	Genes in <i>S. scabies</i> and <i>S. ipomoeae</i> that have undergone recombination.	37
Table 3-1	Selected genomes used for comparison with PAISt.	48
Table 3-2	Predicted extra cellular protein encoded within the PAISt.	67

CHAPTER 1
COMPARATIVE GENOMICS OF *Streptomyces* spp: PAN, CORE AND
PATHO-GENOMES

ABSTRACT

Emergence of pathogenicity within a large and diverse saprophytic genus, such as *Streptomyces*, is an important evolutionary event. *S. scabies*, *S. ipomoeae* and *S. turgidiscabies* represent the best studied plant pathogenic species, and several virulence factors have been described in these bacteria. Despite this fact, the genome composition of these pathogens and the general description of categories of genes that define a pathogenic *Streptomyces* have not been studied in detail and remain unclear. We use comparative analysis of *Streptomyces* genomes to assess the gene content that defines the genus (core-genome) and analyze the accessory genome that is associated with plant pathogenic *Streptomyces* (patho-genome). Our results reveal that the *Streptomyces* core-genome is surprisingly small, estimated to consist of 1,151 orthologues. In the other hand, the pan-genome is extremely large, consisting of more than 26,540 orthologues. The pan-genome encompasses 4,259 orthologues that are present in at least one of the three pathogenic species. Among these we describe novel genes associated with transcriptional regulation, transport and metabolism. Our results suggest genes that encode plant cell degradation, such as cutinases and pectate lyases that are involved in virulence. In addition putative secreted proteins with atypical domains for *Streptomyces* are present exclusively in pathogenic species. We believe that this pool of genes represents the gene network associated with pathogenesis in *Streptomyces*. Finally, synteny analysis of conserved genes among pathogenic *Streptomyces* and phylogenetic analysis support the hypothesis that *S. scabies* and *S. ipomoeae* are closely related bacteria, while *S. turgidiscabies* has a different evolutionary history.

INTRODUCTION

Streptomyces species are a diverse group of aerobic, Gram positive, high G+C bacteria (Hopwood, 2006; Waksman and Henrici, 1943). This bacterial group is commonly found in soil and characterized by a complex life cycle (Buchanan, 1917). *Streptomyces* spores initially germinate and DNA replication takes place without cellular division, forming non-fragmenting filamentous vegetative hyphae (Flardh *et al.*, 1999). The hyphae branch and develop generating a filamentous network of mycelium. Nutrient deprivation stops the growth of vegetative hyphae and the mycelium switches to a reproductive phase of growth. In this phase the nutrients produced and stored in the vegetative hyphae are used to support the growth of new aerial hyphae (Flardh *et al.*, 1999). The aerial hyphae form septa and chains of unigenomic spores that germinate later to continue the reproductive cycle.

The complex life cycle is tightly associated with the diversity and number of compounds that *Streptomyces* can metabolize. These bacteria are able to metabolize alcohols, sugar, amino acids and aromatic compounds by producing a broad range of extra cellular hydrolytic enzymes (Crawford and McCoy, 1972; Crawford, 1978; Donnelly and Crawford, 1988). Additionally *Streptomyces* can synthesize diverse types of secondary metabolites that have antimicrobial, immunosuppressant, and anti-tumor characteristics (Challis and Hopwood, 2003). This exceptional metabolic diversity of *Streptomyces* is consistent with their large genome size of more than 10 Mb for some strains (Hopwood, 2006). The organisms have linear chromosomes and circular or linear plasmids containing a plethora of genes associated with diverse biological processes (Hopwood, 2006).

Streptomyces spp. are mainly saprophytic; however, some species are plant pathogens. The species *S. scabies*, *S. acidiscabies*, *S. ipomoeae* and *S. turgidiscabies* are the best-studied of the pathogenic streptomycetes (Loria *et al.*, 2006). These four

species produce a dipeptide phytotoxin, thaxtomin, which is assembled by two nonribosomal peptide synthetases (Loria *et al.*, 2008). In addition to thaxtomin, plant pathogenic *Streptomyces* spp. possess shared and unique virulence factors. For example, a secreted necrotic protein, Nec1, is produced by *S. scabies*, *S. acidiscabies* and *S. turgidiscabies* (Joshi *et al.*, 2007a; Kers *et al.*, 2005). In contrast, the cytokinin biosynthetic operon, *fas*, is found only in *S. turgidiscabies* (Joshi and Loria, 2007). These virulence factors are believed to have been acquired by lateral gene transfer and have been fixed in the respective genomes of these pathogens (Loria *et al.*, 2006).

Sequencing and genome comparative analysis have become a powerful tool for understanding the evolution of pathogenesis. Using comparative genome analysis it is possible to identify the core-genome, which is the common set of orthologues among a bacterial group (Lapierre and Gogarten, 2009; Lefebure and Stanhope, 2007). Comparative genomics also can be used to determine the pan-genome, which is the total pool of orthologues that are present in all the members of a bacterial group (Medini *et al.*, 2005). Consequently, characterization of the core and pan-genome can be used to determine the accessory or dispensable group of orthologues unique to specific bacterial groups.

While several virulence factors have been studied extensively in pathogenic *Streptomyces* (Loria *et al.*, 2006), we believe that these bacteria might possess a novel cluster of genes associated with pathogenesis that has not yet been described. In this study we used genome comparative analysis to discriminate and categorize a set of orthologues unique to pathogenic *Streptomyces*. We compared the genomes of eight saprophyte *Streptomyces* spp with three pathogenic *Streptomyces*: *S. scabies*, *S. ipomoeae* and *S. turgidiscabies*. The main objective of this study is to define the core-genome for the *Streptomyces* spp. and discriminate orthologues that are conserved in pathogenic *Streptomyces* spp.

METHODOLOGY

Determination of orthologue groups

Coding sequences of saprophytic *Streptomyces* and *S. scabies* genomes were obtained from Genbank (Table 1-1). Coding sequences of *S. ipomoeae* 91-03 and *S. turgidiscabies* Car8 were obtained from the Craig Venter Institute (Table 1-1). Orthologue groups were determined using OrthoMCL program (Li *et al.*, 2003) (Figure 1-1). The OrthoMCL program executes two main procedures. First it carries out reciprocal comparisons using BLAST (Altschul *et al.*, 1990) (Basic Local Alignment Search Tool) of each predicted protein. In a second step OrthoMCL uses the reciprocal e-values generated from the BLAST result and creates a matrix that is analyzed by a Markov Cluster Algorithm (MCL) (Li *et al.*, 2003). As a result of this analysis OrthoMCL yields sets of orthologues and recent paralogs. OrthoMCL was run with a BLAST e-value cut-off of 10^{-5} and inflation rate of 1.5. The output was used to construct a gene content table that contains common and unique orthologues for each genome (Figure 1-1). Genes coding for proteins smaller than 50 amino acids were filtered and not included in the analysis.

Determination of the core and pan-genomes

The gene content table with common and unique orthologues was used to generate cumulative orthologue curves for the pan and the core-genome. In the case of the pan-genome, the cumulative orthologue curve represents the number of unique and common orthologues resulting from the sequential addition of new genomes to the analysis. The cumulative curve for the core-genome describes only the orthologues in common with sequential addition of new genomes. For each point on the curves the procedure was repeated randomly 100 times. As a result of this procedure, means and standard errors were calculated and plotted (Figure 1-2). The statistical package

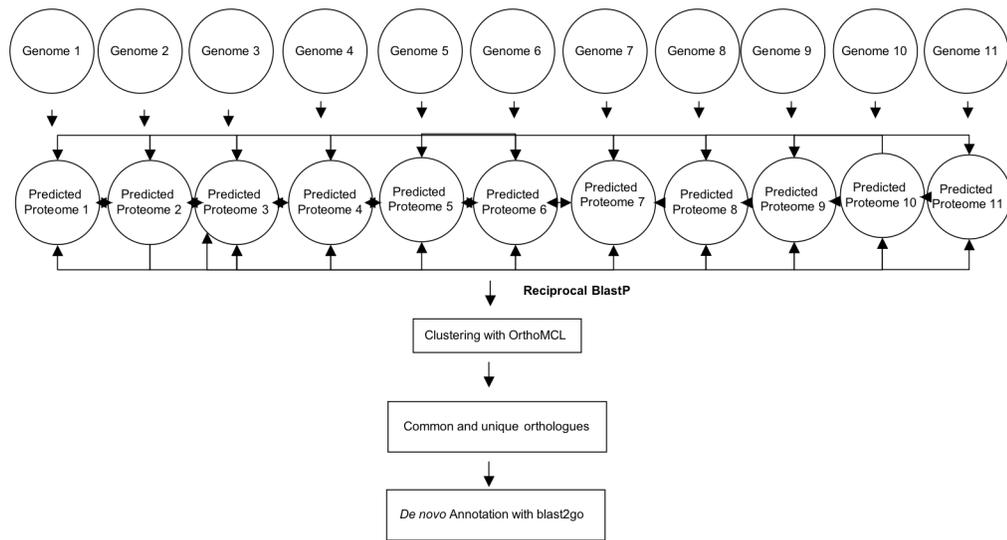


Figure 1-1. Bioinformatic pipeline used to identify the pan, core, and patho-genomes of the genus *Streptomyces*. Predicted proteins were obtained from Genbank files. Reciprocal BlastP of each protein was carried out using OrthoMCL. A Markov chain algorithm was used to discriminate between the common and unique orthologues among species. Annotation of core and patho-genomes was conducted using Blast2go.

Table 1-1 *Streptomyces* spp. genomes used in this study. Genome features and Genbank accession numbers are indicated.

	Genome status	Life style	Size	GC content	Reference or Genbank accessions
<i>S. coelicolor</i> A3(2) M145	Complete	Saprophytic	8.6 Mb (chromosome) 356 Kb (plasmid SCP1) 31 Kb (plasmid SCP2)	72.1 % (chromosome) 69.05% (plasmid SCP1) 72.11 % (plasmid SCP2)	AL645882 AL589148 AL645771
<i>S. avermitilis</i> MA-4680	Complete	Saprophytic	9.0 Mb (chromosome) 94 kb (plasmid SAP1)	70.7 % (chromosome) 69.2 % (plasmid SAP1)	BA000030 AP005645
<i>S. griseus</i> IFO 13350	Complete	Saprophytic	8.5 Mb	72.20%	AP009493
<i>S. scabies</i> 82-27	Complete	Plant pathogen	10.1 Mb	72%	FN554889
<i>S. ipomoeae</i>	Draft	Plant pathogen	10.4 Mb	70%	http://www.ttaxus.com/files/Streptomyces/gip.zip
<i>S. turgidiscabies</i> car8	Draft	Plant pathogen	10.8 Mb	70%	http://www.ttaxus.com/files/Streptomyces/gst.zip
<i>S. sviveus</i> ATCC 29083	Draft	Saprophytic	8.4 Mb	70%	ABJJ00000000
<i>S. clavuligerus</i> ATCC 27064	Draft	Saprophytic	6.7 Mb	71.10%	ABJH00000000
<i>S. pristinaespirales</i> ATCC 25486	Draft	Saprophytic	6.1 Mb	70%	ABJI00000000
<i>S. sp.</i> Mg1	Draft	Saprophytic	7.1 Mb	71%	ABJF00000000
<i>S. sp.</i> SPB74	Draft	Saprophytic	5.0 Mb	71%	ABJG00000000

R 2.2.1 (<http://www.r-project.org/>) was used to analyze the gene content table and calculate cumulative orthologue group curves for the pan, core-genomes as well as unique orthologue groups for each pathogenic *Streptomyces*.

Annotation and categorization of genes

Core and unique orthologues in pathogens were annotated “*de novo*” using Blast2Go annotation software (Conesa and Gotz, 2008). Databases used as a reference were the Gene Ontology database (AMIGO) (Carbon *et al.*, 2009), Pfam database (Bateman *et al.*, 2004) and the Clusters of Orthologous Groups database (COG) (Tatusov *et al.*, 2000).

RESULTS AND DISCUSSION

Characterization of the pan and core-genome.

The pan-genome of the eleven *Streptomyces* spp. analyzed in this study is estimated to consist of 26,540 orthologues. The genome accumulation curve shows an increasing tendency indicating that the actual size of the *Streptomyces* pan-genome is extremely large and cannot be estimated with the data set used in this study (Figure 1-2). In contrast to the pan-genome, the core-genome is surprisingly small and was determined to consist of 1,151 orthologues (Figure 1-2). This represents approximately 16 % of an average *Streptomyces* genome analyzed in this study (6,940 orthologues). The core-genome contains orthologue groups that can be classified in broad functional categories. One of the major categories is transcriptional process (10%) (Figure1-3). We believe that the pool of genes in this category might represent the complex regulatory network that commands the morphological development cycle of this bacterial group. Other categories that are important in the common biological processes in *Streptomyces* are cell morphogenesis (2.6%); aromatic compound

biosynthetic processes (1%); cellular macromolecule catabolic process such as degradation of complex polysaccharides processes (1.7%) Figure (1-3).

It is noticeable that the core-genome of the genus *Streptomyces* represents less than 4.3% of the pan-genome in this analysis. This proportion is smaller than percentages observed in other bacteria, such as *Streptococcus* species, in which the core is 10% of the pan-genome (Lefebure and Stanhope, 2007). Our results suggest that streptomycetes possess “open genomes” that readily accumulate novel genetic information that encodes the particular features of each species (Medini *et al.*, 2005). We discuss in the next paragraphs the accessory pool of orthologues that were found only in the pathogenic strains. It is clear to us that all of these orthologues that are not necessarily for pathogenicity. However, we believe that this group of genes is directly or indirectly involved in virulence, therefore, we denominate this cluster of orthologues the patho-genome.

Orthologue groups in pathogenic *Streptomyces* spp: Definition of the patho-genome

The number of orthologues that are unique to pathogenic *Streptomyces* is 4,259 (Figure 1-4). Inspection of the patho-genome reveals that previously reported virulence factors (Joshi and Loria, 2007; Kers *et al.*, 2005; Loria *et al.*, 2008) are distributed in the following fashion: The necrosis producing protein Nec1 and the tomatinase-like protein TomA are found in *S. scabies* and *S. turgidiscabies* but not in *S. ipomoeae*. Proteins that form the cytokinin biosynthetic pathway are only in *S. turgidiscabies*.

We were not able to discriminate the two nonribosomal peptide synthetases responsible for thaxtomin biosynthesis in the patho-genome, even though the pathway itself is conserved in pathogens. Instead, we found both enzymes clustered together in

the core-genome of the *Streptomyces* genus, along with other nonribosomal peptide synthetases found in non-pathogenic *Streptomyces*. This result is not surprising due to the fact that nonribosomal peptide synthetases are highly conserved, large proteins with multi-domain structure that are ubiquitous in streptomycetes (Amoutzias *et al.*, 2008; Kuo *et al.*, 2008). The high similarity among their conserved domains can interfere with the discrimination among nonribosomal peptide synthetases, even when they function to adenylate specific amino acids (Loria *et al.*, 2006).

Categories of genes in the patho-genome

Analysis of the patho-genome indicates that 66% of these orthologues do not have a match in the Genbank database and they are considered hypothetical proteins. 16% of the orthologues have homologues in the Genbank but do not have a putative function. These are considered hypothetical conserved proteins. Proteins associated with diverse functions comprise the rest of the patho-genome. Interestingly, a group of these proteins are expected to be involved in transcriptional regulation processes (3%) (Figure 1-5). Among these is the AraC type-regulator, TxtR, that has been shown to participate in the regulation of thaxtomin production (Joshi *et al.*, 2007b). Another interesting group is transport proteins (3% of the patho-genome). Proteins associated with transport could have many functions. One example is a putative operon that encodes for a glucitol-sorbitol permease unique to *S. ipomoeae* (*sip5766*, *sip5767* and *sip5768*).

The three pathogens possess genes coding for the biosynthesis of secondary metabolites. For example, *S. scabies* possesses genes for the biosynthesis of coronafacic acid. Among these genes *scab79631* and *scab79601* code for a coronafacic acid synthetase and coronafacic acid acyl carrier protein, respectively, that are unique to this pathogenic *Streptomyces*. Interestingly, coronafacic acid is a

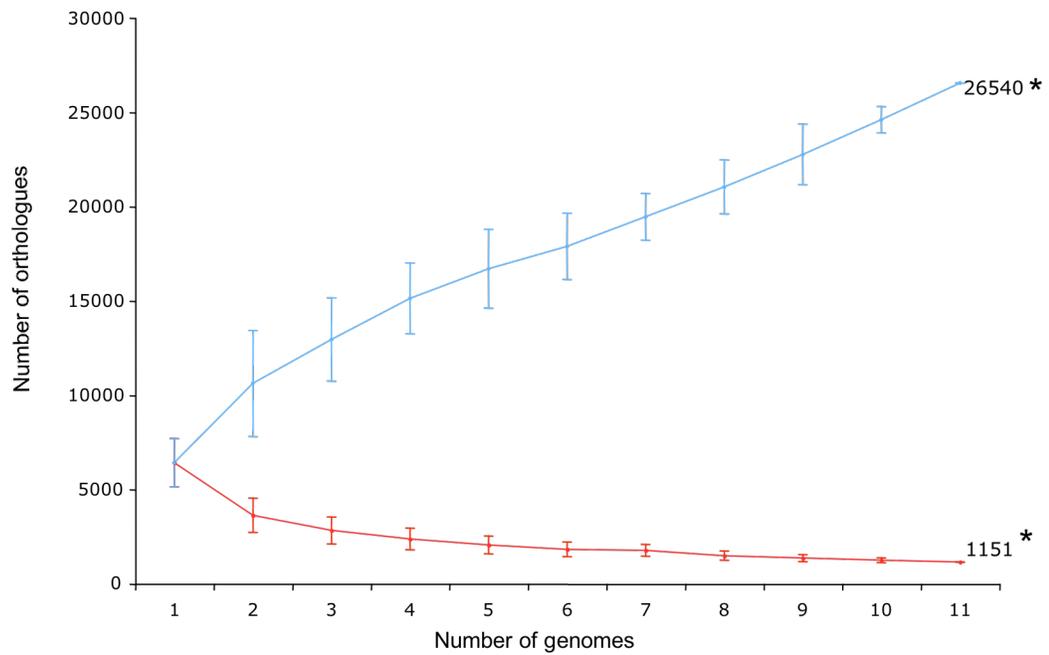


Figure 1-2. Orthologue accumulation curves for eleven *Streptomyces* spp. Pan-genome curve (blue line) and core-genome curve (red line). Standard deviations are indicated at each point in the curves and calculated base on 100 repetitions with random genome order. The number of orthologues across eleven genomes is indicated by *.

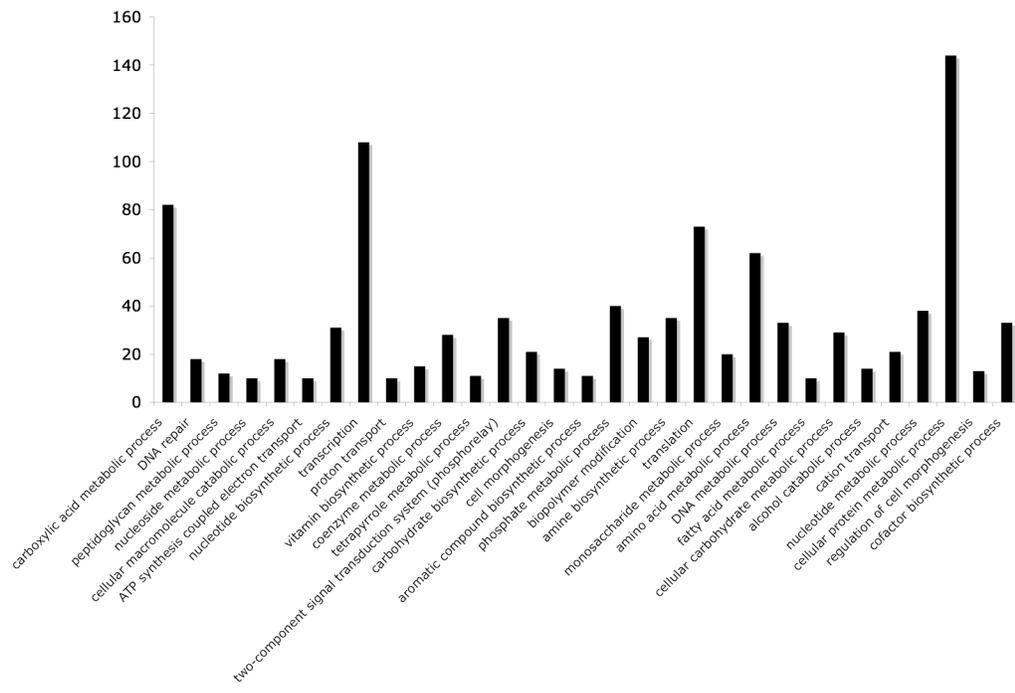


Figure 1-3. Histogram showing the categories of orthologues that form the *Streptomyces* core-genome. Orthologues were categorized by biological process (x-axis) and plotted base on frequency (y-axis). In some cases orthologues are included in more than one category.

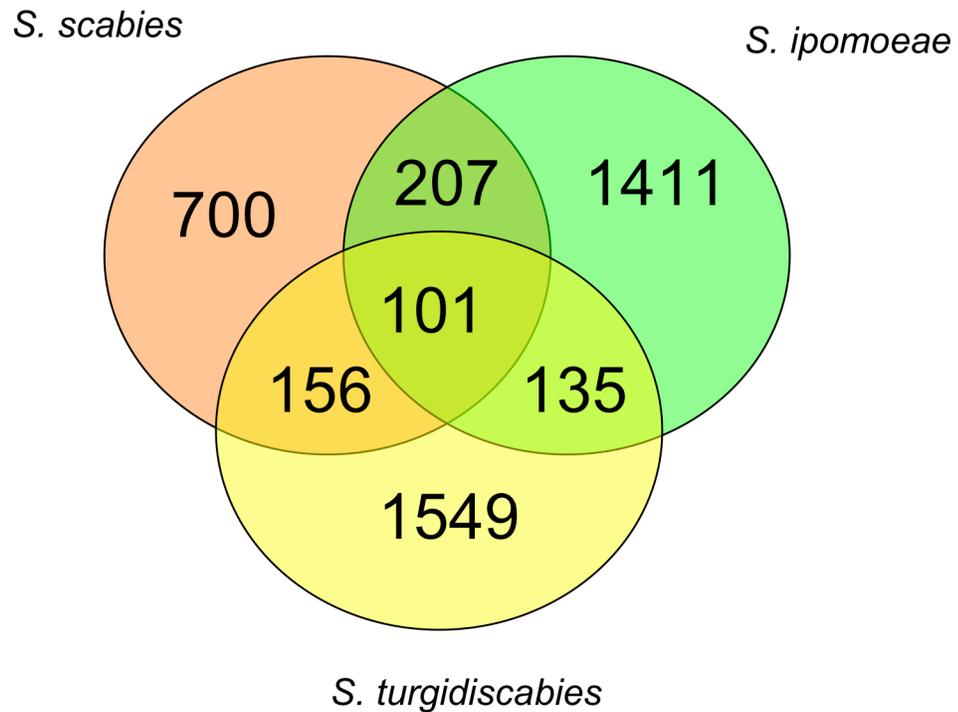


Figure 1-4. Orthologues unique to pathogenic *Streptomyces* spp. The Venn diagram shows the distribution of the 4,259 orthologues that are unique to the three pathogenic species. The central intersection contains the common group of orthologues in the three pathogenic species.

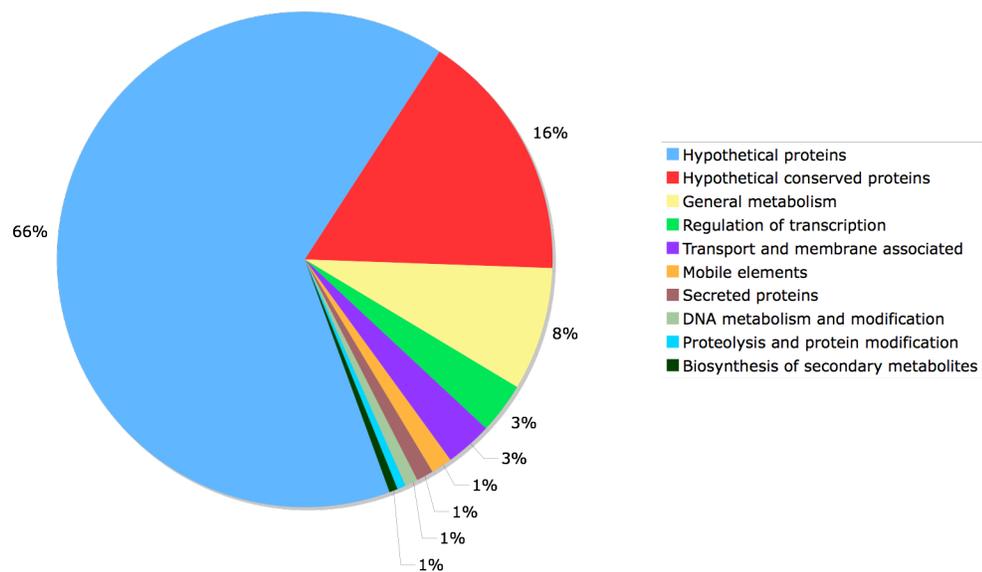


Figure 1-5. Categories of orthologues unique to pathogenic *Streptomyces* spp. The pie chart shows the categorization of the 4,259 orthologues that are conserved only in the pathogenic *Streptomyces* (refer to figure 1-4). Gene categories derived from annotation with Blast2Go software are indicated in the legend.

polyketide metabolite described in *Pseudomonas syringae* (Bender *et al.*, 1999) as an important virulence factor. Another example is a cluster of genes for the biosynthesis of a lantibiotic present only in *S. ipomoeae*. This streptomycete possesses the gene *sip4863* that codes for a predicted lanthionie synthetase and *sip3351* that codes for an ipomicin precursor. Production of a lantibiotic by *S. ipomoeae* may provide a competitive advantage in the colonization of novel niches.

Secreted proteins are part of the patho-genome and can be associated with pathogenesis. One case is SCAB78931 in *S. scabies*, STURG1136 in *S. turgidiscabies* and SIP1100 in *S. ipomoeae*. These proteins are 51% identical to the cutinase found in the actinobacterium *Thermonospora curvata* DSM 43183. Moreover, BLAST analysis against the non-redundant database indicates that these cutinases are similar to those found in *Phytophthora* spp., *Mycobacterium* spp. and some fungi. Cutinases have been studied extensively in plant pathogenic fungi where they play an important role in cuticle degradation and host colonization (Dantzig *et al.*, 1986; Schafer, 1993).

The predicted proteins SCAB82041, STURG9706c and SIP0228c represent another case of secreted proteins putatively associated with pathogenesis. The proteins are 52% identical to a putative pectate lyase found in the fungus *Aspergillus fumigatus* A1163. Pectate lyases participate in cell wall degradation and have been associated with pathogenesis (Roberts *et al.*, 1986). The role of these secreted proteins represents an important target for experimental research.

Genes coding for proteins with atypical *Streptomyces* motifs are found in the patho-genome.

The patho-genome contains genes that code for protein motifs atypical of *Streptomyces*. Among this group, are *scab0081*, *scab43241*, and *scab65341* in *S.*

scabies; and *sip8921* in *S. ipomoeae*, which code for putative patatin-like proteins (Figure 1-6). Patatins are plant glycoproteins that show lipid acyl-hydrolase activity (La Camera *et al.*, 2009). Their presence in plants has been associated with defense against bacterial and fungal infections (La Camera *et al.*, 2005). Interestingly, a secreted toxin ExoU in *Pseudomonas aeruginosa* with lipase and phospholipase activity contains defined regions of homology to potato patatins. The toxin ExoU is an important virulence factor in mammalian cells (Banerji and Flieger, 2004; La Camera *et al.*, 2005; Rabin and Hauser, 2005). The possibility that patatin-like proteins play a role in pathogenesis in *S. scabies* and *S. ipomoeae* is a subject of future studies.

Other example of atypical motifs in *Streptomyces* is the hemolysin calcium-binding repeated motifs found in proteins encoded in *S. scabies*. SCAB6751 and SCAB19481 are predicted to be secreted proteins that contain the repeat motifs L-X-G-G-X-G-N-D-X at their C-terminal regions (Figure 1-6). This motif has been described in pore forming proteins that group with cytolysins and cytotoxins (Welch, 1991). Interestingly, putative proteins SCAB6751 and SCAB19481 lack the RTX N-terminal domain that is characteristic of this type of toxins (Welch, 1991). Thus the function of these proteins in pathogenic *Streptomyces* is unclear. However, as they are predicted secreted proteins and the C-terminal region containing the calcium-binding motif binds lipid monolayer (Sanchez-Magraner *et al.*, 2007) it is tempting to speculate that they might be involved in a novel signaling process or cell membrane interaction with the host.

Another intriguing case is the presence of predicted secreted proteins encoded in *S. scabies* and *S. ipomoeae*. SCAB44951 and SIP5078 are proteins conserved only in these pathogenic *Streptomyces* and display a SCP-like domain that has been found in calcium chelating serine proteases and could be involved in various signaling processes (Fernandez *et al.*, 1997). SCP-like motifs have been described in some

Streptomyces before (Yeats *et al.*, 2003). However, the predicted proteins found in *S. scabies* and *S. ipomoeae* do not display significant similarity to those reported in non-pathogenic *Streptomyces*. More intriguing is the finding that Blast results indicate that the pathogenic SCR-like proteins are more similar to SCP-like proteins found in plants such as *Arabidopsis thaliana* (67% similarity), *Nicotiana tabacum* (65% similarity), *Ricinus communis* (64% similarity), and *Populus trichocarpa* (68% similarity) (Figure 1-7). In plants, SCP-like proteins have diverse functions and among them, the PR-1 family is synthesized during pathogen infection or environmental stress (Dixon *et al.*, 1991). The function of these proteins in pathogenic *Streptomyces* is unknown, however, their similarity to SCP-like proteins in plants suggests a possible signaling function during plant-pathogen interactions. SCR-like proteins in pathogenic *Streptomyces* are excellent targets for further studies.

Synteny of conserved genes in the pathogenic *Streptomyces* spp.

Analysis of chromosomal positions of shared orthologues between *S. scabies* and *S. ipomoeae* reveals a high level of synteny in these species, which is consistent with a common evolutionary history (Figure 1-8). Much less synteny exists between *S. scabies* or *S. ipomoeae* and *S. turgidiscabies* (Figure 1-9 and 1-10). The most noticeable difference in the synteny plots between *S. scabies* or *S. ipomoeae* with *S. turgidiscabies* is a large inversion in the central region of the chromosome (Figure 1-9 and 1-10). Large inversions also are observed when chromosomes of *S. scabies* or *S. ipomoeae* are compared with other *Streptomyces* species included in this study (data not shown). Highly syntenic regions among conserved genes of *Streptomyces* have been described before and levels of synteny is correlated with phylogenetic distances among *Streptomyces* spp. (Ventura *et al.*, 2007). The level of synteny between the

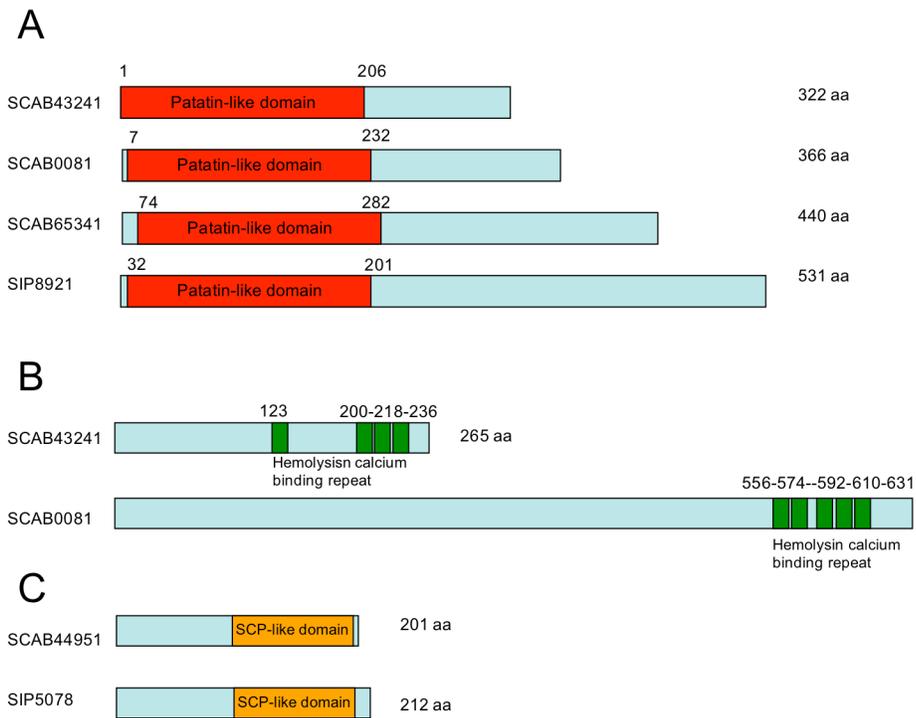


Figure 1-6. Proteins with atypical motifs in *Streptomyces* are found in the patho-genome. Predicted proteins displaying patatin-like domains (A); Hemolysin calcium binding repeats (green) (B); or, SCP-like domains (yellow) (C).

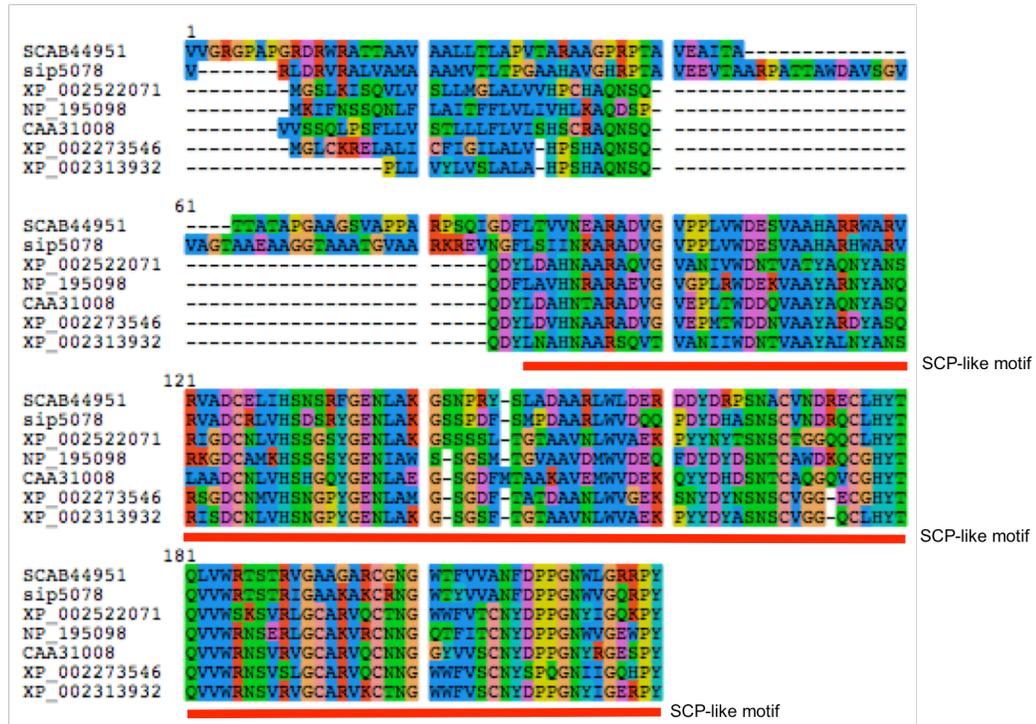


Figure 1-7. SCP-like motif found in predicted proteins encoded in *S. scabies* and *S. ipomoeae*. SCAB44951 (*S. scabies*) and SIP5078 (*S. ipomoeae*) are aligned with SCP-like proteins found in *Ricinus communis* (XP_002522071), *Arabidopsis thaliana* (NP_195098), *Nicotiana tabacum* (CAA31008), *Vitis vinifera* (XP_002273546), and *Populus trichocarpa* (XP_002313932). Red solid bar at the bottom of the alignment indicates the SCP motif.

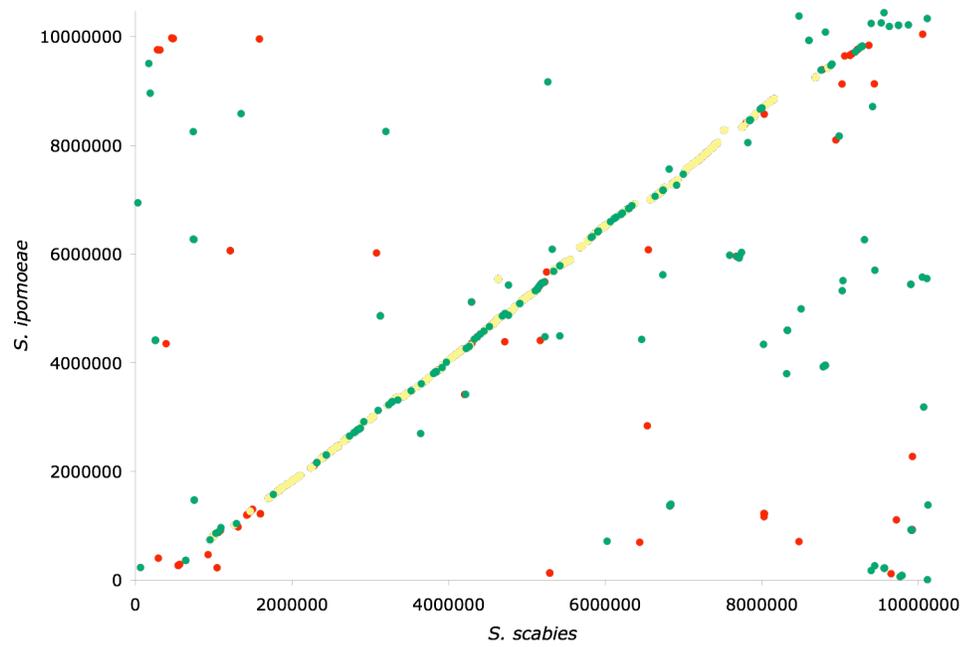


Figure 1-8. Chromosome plot of *S. scabies* and *S. ipomoeae* orthologues. Core-genome in yellow. Genes conserved in the three pathogens are red. Genes share by *S. scabies* and *S. ipomoeae* are green.

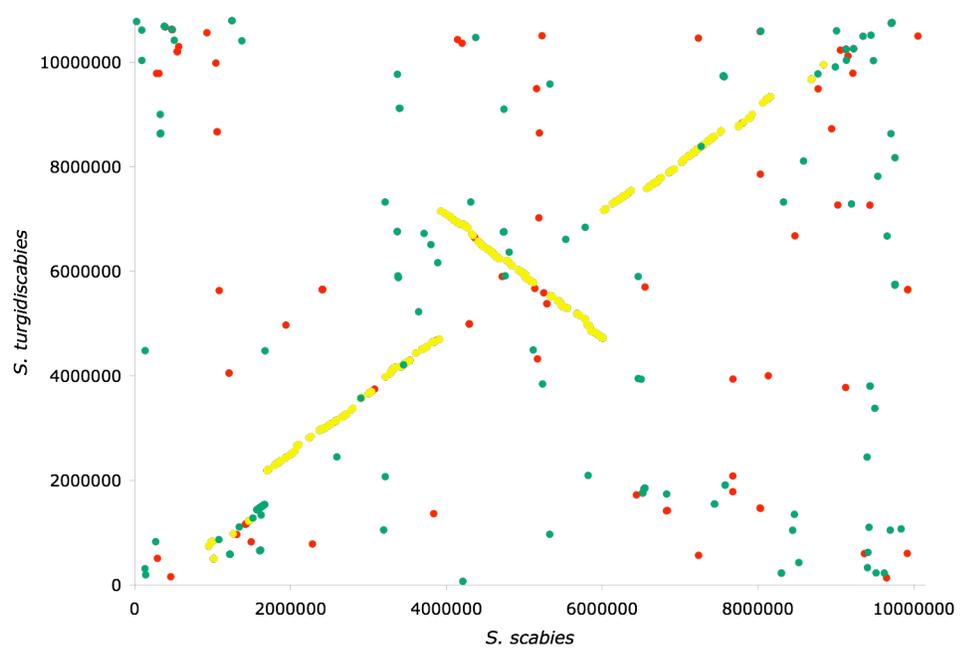


Figure 1-9 Chromosome plot of *S. scabies* and *S. turgidiscabies* orthologues. Core-genome in yellow. Genes conserved in the three pathogens in red. Genes share by *S. scabies* and *S. turgidiscabies* in green.

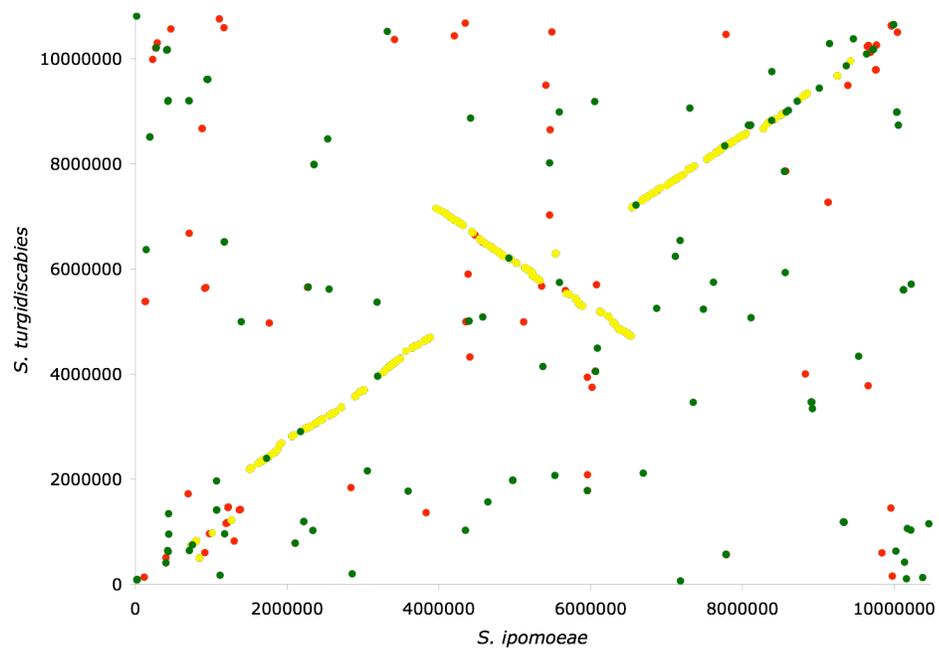


Figure 1-10: Chromosome plot of *S. ipomoeae* and *S. turgidiscabies* orthologues. Core-genome in yellow. Genes conserved in the three pathogens in red. Genes share by *S. ipomoeae* and *S. turgidiscabies* in green.

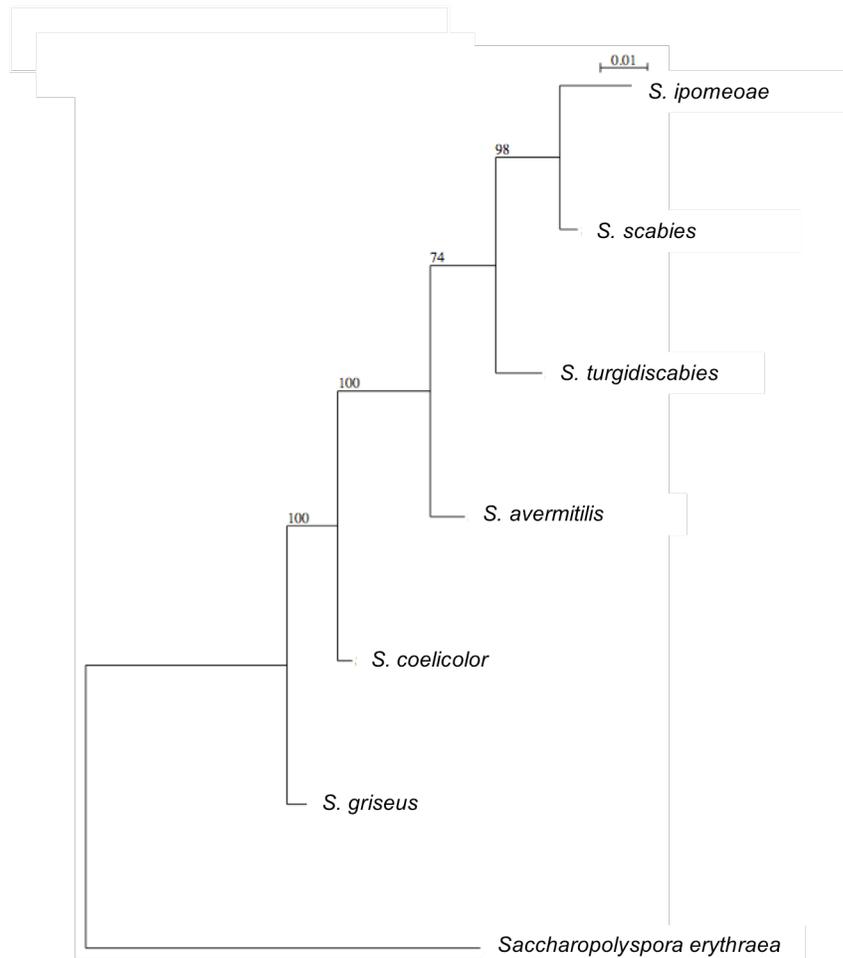


Figure 1-11. 16s rDNA tree of *Streptomyces* spp. Reconstruction of the tree was carried out with available sequences of 16S rDNA obtained from the sequence genome projects (Table 1-1). Maximum likelihood algorithm was used for tree reconstruction. The tree was rooted with the actinobacterium *S. erythraea*. Bootstrap values are showed in the branches. Scale indicates nucleotide change/position

pathogenic *Streptomyces* is congruent with the phylogenetic relationships among these species based on 16s ribosomal DNA (16s *rDNA*) (Figure 1-11). The level of synteny and the phylogenetic distances among the pathogens suggest that *S. scabies* and *S. ipomoeae* are an ancient lineage of plant pathogens sharing a common evolutionary history while *S. turgidiscabies* has a different evolutionary history.

CONCLUSIONS

Comparative analysis revealed the complex gene composition of *Streptomyces* genomes. We were able to calculate the size of the core-genome of eleven species of *Streptomyces* to be 1,151 orthologues. In contrast, the pan-genome for those species is extremely large at 2,640 orthologues. Our analysis indicated as well that the patho-genome of *Streptomyces* spp. contains novel transcriptional factors, secreted proteins and proteins associated with transport. Over 100 pathogenecity genes were conserved in the three pathogenic species. Some of them show putative functions as cutinases, pectate lyases and lipases that are consistent with a role in pathogenicity. The functional analysis of these genes will be tested in future studies.

REFERENCES

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403-410.
- Amoutzias, G.D., Van de Peer, Y., and Mossialos, D. (2008) Evolution and taxonomic distribution of nonribosomal peptide and polyketide synthases. *Future Microbiol* **3**: 361-370.
- Banerji, S., and Flieger, A. (2004) Patatin-like proteins: a new family of lipolytic enzymes present in bacteria? *Microbiology* **150**: 522-525.
- Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., and Eddy, S.R. (2004) The Pfam protein families database. *Nucleic Acids Res* **32**: D138-141.
- Bender, C.L., Alarcon-Chaidez, F., and Gross, D.C. (1999) Pseudomonas syringae phytotoxins: mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. *Microbiol Mol Biol Rev* **63**: 266-292.
- Buchanan, R.E. (1917) Studies in the Nomenclature and Classification of the Bacteria: II. The Primary Subdivisions of the Schizomycetes. *J Bacteriol* **2**: 155-164.
- Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., and Lewis, S. (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**: 288-289.
- Challis, G.L., and Hopwood, D.A. (2003) Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species. *Proc Natl Acad Sci U S A* **100 Suppl 2**: 14555-14561.
- Conesa, A., and Gotz, S. (2008) Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. *Int J Plant Genomics* **2008**: 619832.

- Crawford, D.L., and McCoy, E. (1972) Cellulases of *Thermomonospora fusca* and *Streptomyces thermodiastaticus*. *Appl Microbiol* **24**: 150-152.
- Crawford, D.L. (1978) Lignocellulose decomposition by selected *Streptomyces* strains. *Appl Environ Microbiol* **35**: 1041-1045.
- Dantzig, A.H., Zuckerman, S.H., and Andonov-Roland, M.M. (1986) Isolation of a *Fusarium solani* mutant reduced in cutinase activity and virulence. *J Bacteriol* **168**: 911-916.
- Dixon, D.C., Cutt, J.R., and Klessig, D.F. (1991) Differential targeting of the tobacco PR-1 pathogenesis-related proteins to the extracellular space and vacuoles of crystal idioblasts. *Embo J* **10**: 1317-1324.
- Donnelly, P.K., and Crawford, D.L. (1988) Production by *Streptomyces viridosporus* T7A of an Enzyme Which Cleaves Aromatic Acids from Lignocellulose. *Appl Environ Microbiol* **54**: 2237-2244.
- Fernandez, C., Szyperski, T., Bruyere, T., Ramage, P., Mosinger, E., and Wuthrich, K. (1997) NMR solution structure of the pathogenesis-related protein P14a. *J Mol Biol* **266**: 576-593.
- Flardh, K., Findlay, K.C., and Chater, K.F. (1999) Association of early sporulation genes with suggested developmental decision points in *Streptomyces coelicolor* A3(2). *Microbiology* **145 (Pt 9)**: 2229-2243.
- Hopwood, D.A. (2006) Soil to genomics: the *Streptomyces* chromosome. *Annu Rev Genet* **40**: 1-23.
- Joshi, M., Rong, X., Moll, S., Kers, J., Franco, C., and Loria, R. (2007a) *Streptomyces turgidiscabies* secretes a novel virulence protein, Nec1, which facilitates infection. *Mol Plant Microbe Interact* **20**: 599-608.

- Joshi, M.V., Bignell, D.R., Johnson, E.G., Sparks, J.P., Gibson, D.M., and Loria, R. (2007b) The AraC/XylS regulator TxtR modulates thaxtomin biosynthesis and virulence in *Streptomyces scabies*. *Mol Microbiol* **66**: 633-642.
- Joshi, M.V., and Loria, R. (2007) *Streptomyces turgidiscabies* possesses a functional cytokinin biosynthetic pathway and produces leafy galls. *Mol Plant Microbe Interact* **20**: 751-758.
- Kers, J.A., Cameron, K.D., Joshi, M.V., Bukhalid, R.A., Morello, J.E., Wach, M.J., Gibson, D.M., and Loria, R. (2005) A large, mobile pathogenicity island confers plant pathogenicity on *Streptomyces* species. *Mol Microbiol* **55**: 1025-1033.
- Kuo, M.C., Chou, L.F., and Chang, H.Y. (2008) Evolution of exceptionally large genes in prokaryotes. *J Mol Evol* **66**: 333-349.
- La Camera, S., Geoffroy, P., Samaha, H., Ndiaye, A., Rahim, G., Legrand, M., and Heitz, T. (2005) A pathogen-inducible patatin-like lipid acyl hydrolase facilitates fungal and bacterial host colonization in Arabidopsis. *Plant J* **44**: 810-825.
- La Camera, S., Balague, C., Gobel, C., Geoffroy, P., Legrand, M., Feussner, I., Roby, D., and Heitz, T. (2009) The Arabidopsis patatin-like protein 2 (PLP2) plays an essential role in cell death execution and differentially affects biosynthesis of oxylipins and resistance to pathogens. *Mol Plant Microbe Interact* **22**: 469-481.
- Lapierre, P., and Gogarten, J.P. (2009) Estimating the size of the bacterial pan-genome. *Trends Genet* **25**: 107-110.
- Lefebure, T., and Stanhope, M.J. (2007) Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol* **8**: R71.

- Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178-2189.
- Loria, R., Kers, J., and Joshi, M. (2006) Evolution of plant pathogenicity in *Streptomyces*. *Annu Rev Phytopathol* **44**: 469-487.
- Loria, R., Bignell, D.R., Moll, S., Huguet-Tapia, J.C., Joshi, M.V., Johnson, E.G., Seipke, R.F., and Gibson, D.M. (2008) Thaxtomin biosynthesis: the path to plant pathogenicity in the genus *Streptomyces*. *Antonie Van Leeuwenhoek* **94**: 3-10.
- Medini, D., Donati, C., Tettelin, H., Massignani, V., and Rappuoli, R. (2005) The microbial pan-genome. *Curr Opin Genet Dev* **15**: 589-594.
- Rabin, S.D., and Hauser, A.R. (2005) Functional regions of the *Pseudomonas aeruginosa* cytotoxin ExoU. *Infect Immun* **73**: 573-582.
- Roberts, D.P., Berman, P.M., Allen, C., Stromberg, V.K., Lacy, G.H., and Mount, M.S. (1986) Requirement for two or more *Erwinia carotovora* subsp. *carotovora* pectolytic gene products for maceration of potato tuber tissue by *Escherichia coli*. *J Bacteriol* **167**: 279-284.
- Sanchez-Magraner, L., Viguera, A.R., Garcia-Pacios, M., Garcillan, M.P., Arrondo, J.L., de la Cruz, F., Goni, F.M., and Ostolaza, H. (2007) The calcium-binding C-terminal domain of *Escherichia coli* alpha-hemolysin is a major determinant in the surface-active properties of the protein. *J Biol Chem* **282**: 11827-11835.
- Schafer, W. (1993) The role of cutinase in fungal pathogenicity. *Trends Microbiol* **1**: 69-71.
- Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**: 33-36.

- Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G.F., Chater, K.F., and van Sinderen, D. (2007) Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiol Mol Biol Rev* **71**: 495-548.
- Waksman, S.A., and Henrici, A.T. (1943) The Nomenclature and Classification of the Actinomycetes. *J Bacteriol* **46**: 337-341.
- Welch, R.A. (1991) Pore-forming cytolysins of gram-negative bacteria. *Mol Microbiol* **5**: 521-528.
- Yeats, C., Bentley, S., and Bateman, A. (2003) New knowledge from old: in silico discovery of novel protein domains in *Streptomyces coelicolor*. *BMC Microbiol* **3**: 3.

CHAPTER 2

RECOMBINATION WITHIN THE CORE-GENOME OF *Streptomyces* spp.

ABSTRACT

It is broadly accepted that recombination affects the core-genome of several bacterial groups, playing an important role in their evolution. The *Streptomyces* genus contains a small core-genome but relatively large pan-genome. The core-genome contains the essential genes that define the *Streptomyces* genus. In this study we use sequence analysis and phylogenetic approaches to investigate the impact of recombination on the core-genome of *Streptomyces*. Our results indicate that at least 38% of the core-genome has been affected by recombination. Surprisingly, recombination events within the core-genome affect even informational genes, such as those coding for elongation factors and ribosomal proteins. Our genomic analysis provides important information about the role of recombination in the evolution of the genus *Streptomyces*.

INTRODUCTION

Prokaryotes propagate by binary fission and, as a consequence of this feature, early models of evolution in bacteria were based on clonality and periodic selection. These models suggested that the stepwise accumulation of mutations was the uniquely important factor in bacterial evolution (Gogarten *et al.*, 2002). Several studies have shown, however, that bacterial genomes do recombine at rates that are high enough to affect the evolution in these organisms (Gogarten *et al.*, 2002; Gogarten and Townsend, 2005; Guttman and Dykhuizen, 1994). Recombination can occur between regions of genomes that share considerable regions of nucleotide identity. This type of DNA exchange is denominated homologous recombination and is one important force of evolution in genomes with high nucleotide identity (Majewski and Cohan, 1999; Majewski *et al.*, 2000). While homologous recombination is limited by the DNA identity of sequences, non-homologous recombination drives the interchange of DNA sequences by mechanisms that do not require any or only very short regions of identity. Transposition and site-specific recombination (Ochman *et al.*, 2000) are two types of recombination that can be classified as non-homologous recombination processes.

It has been shown that recombination can affect the core-genome of bacteria, which is the portion of the genome shared by members of the taxon, and that bacterial evolution is impacted (Lefebure and Stanhope, 2007; Wu *et al.*, 2009). For example, 18% of the core-genome of *Streptococcus* has a recombinant history (Lefebure and Stanhope, 2007). Other studies indicate that 28% of the *Rickettsia* core genes are recombinant (Wu *et al.*, 2009). In the previous chapter we described the core and pan-genomes of the genus *Streptomyces*. Comparative genome analysis demonstrated that *Streptomyces* possess a small core-genome but a very large pan-genome that represents the specific characteristics of each *Streptomyces* species. Having calculated

the core-genome composition of *Streptomyces* we want to determine if recombination occurs in this genus as it does in other bacteria.

METHODOLOGY

Recombination analysis

Analysis of recombination within the core-genome was based on three algorithms. Maximum X^2 (MaxChi) (Posada and Crandall, 2001a, b); Maximum mismatch X^2 (Chimerae) (Posada and Crandall, 2001a); and, phylogenetic incongruence (Lefebure and Stanhope, 2007). The first two are nucleotide substitution distribution algorithms and scan multiple alignments looking for significant differences in the proportions of variable and non-variable polymorphic alignment positions in adjacent regions of the sequence (Posada and Crandall, 2001a). MaxChi and Chimera are implemented in the Recombination Detection Program (RDP) (Martin *et al.*, 2005). The individual alignments of the core genes were concatenated in a single alignment and analyzed with the RDP software using the MaxChi and Chimera algorithms with a sliding window of 30, cutoff p-value of 0.05.

On the other hand, the phylogenetic incongruence algorithm reconstructs phylogenetic trees of a group of orthologues. The orthologue trees are collapsed into consensus. Each orthologue tree is then bootstrapped in order to obtain statistical support. Genes that produce trees that display different topologies than the consensus and have a significant bootstrap value are predicted to be genes with a recombinant history. For the phylogenetic approach, *Streptomyces* core genes (described in chapter 1) were grouped with an orthologue of a close actinobacteria. The actinobacterial orthologue was used as an out-group (root of the phylogenetic tree) (Table 2-1). *Streptomyces* orthologues without an out-group or those that have paralogues were excluded from the analysis. Nucleotide sequences of each orthologue were aligned

with the Probalign software (Roshan and Livesay, 2006). The program Probalign assigns a score based on the quality of the alignment. Regions in the alignments with less than 60% of quality were excluded. Remaining regions were translated to amino acid sequences conserving the reading frames of the gene. A final back-translation in each orthologue was conducted to obtain alignments at the nucleotide level. These final alignments were used to reconstruct the phylogenetic trees.

Phylogenetic trees were reconstructed using the maximum likelihood algorithm, which is implemented in the phyML program (Guindon and Gascuel, 2003). The Generalized Time Reversible (GTR) model (Lanave *et al.*, 1984) was used as a substitution model. Trees were recoded with the names of each *Streptomyces* taxon and a consensus tree was calculated using the program Consense-Phylip software version 3.6 (<http://cmgm.stanford.edu/phylip/>). Each tree of the consensus was analyzed for statistical support using non-parametric bootstrapping method. This methodology takes sub-samples of the sites in the alignment and creates trees based on those sub-samples. The process is iterated multiple times and the results are compiled to allow an estimated of the reliability of a particular tree. Bootstrapping analysis was performed with 100 repetitions. We considered a tree statistical reliable if it displayed 70 or more samples with congruent topology.

RESULTS AND DISCUSSION

Recombination is detected at high rates within the *Streptomyces* core-genome.

The nucleotide substitution algorithms, MaxChi and Chimera detected 701 and 631 points of recombination, respectively, within the concatenated alignment of the core. Both methods have been described before as good performance algorithms for detection of breakpoints of recombination (Posada and Crandall, 2001a). Analysis of the length of the recombinant regions detected by the nucleotide substitution methods

indicates that most of them are shorter than 500 bp (Figure 2-1). Considering the length of the recombinant regions, Chimera predicts that around 53% of the core is recombinant while MaxChi predicts that 49% of the core recombined. We believe that a more conservative estimate of recombination would be the intersection of both outputs, leading to the estimate of 38% of the core-genome is recombinant.

The phylogenetic algorithm indicates that the majority of branches of the consensus tree display a great number of genes with phylogenetic incongruence but with poor bootstrap support (Figure 2-2). It is believed that phylogenetic approaches can fail to detect small recombinant regions that other methods, such as nucleotide substitution, can identify (Lefebure and Stanhope, 2007; Posada and Crandall, 2001a). We believe that the short recombinant regions identified with the nucleotide substitution methods preclude the resolution of the tree derived from the phylogenetic approach.

In spite of the overall weakly supported tree, the phylogenetic approach indicates that two clusters have significant support (Figure 2-2). The first cluster contains *S. coelicolor*, *S. sviveus*, *S. avermitilis*, *S. turgidiscabies*, *S. ipomoeae* and *S. scabies*. Within this major cluster, 446 genes support a sub cluster that contains the species *S. scabies* and *S. ipomoeae* (Figure 2-2). The reconstructed trees of 44 genes display topological incongruence when they are compared with the consensus tree and have reliable bootstrap support values (>70). Based on the phylogenetic incongruence concept, these 44 genes are predicted to have a recombination history. 42 genes were confirmed with the Maxchi and Chimera algorithms.

Table 2-1. Actinobacteria used as out groups for construction of phylogenetic trees.

Species names	Genbank accession
<i>Arthrobacter aurescens</i> TC1	CP000475
<i>Acidothermus cellulolyticus</i> 11B	CP000481
<i>Corynebacterium diphtheriae</i> NCTC13129	BX248353
<i>Corynebacterium glutamicum</i> ATCC 13032	BA000036
<i>Clavibacter michiganensis</i> subsp. <i>sepedonicus</i> ATC 33113	AM711867
<i>Clavibacter michiganensis</i> subsp. <i>michiganensis</i> NCPPB 382	AM849034
<i>Frankia alni</i> ACN14a	CT573213
<i>Frankia</i> sp. CcI3	CP000249
<i>Frankia</i> sp. EAN1pec	CP000820
<i>Kineococcus radiotolerans</i> SRS30216	CP000750
<i>Leifsonia xyli</i> CTCB07	AE016822
<i>Mycobacterium leprae</i> TN	AL450380
<i>Mycobacterium</i> sp. MCS	CP000384
<i>Mycobacterium ulcerans</i> Agy99	CP000325
<i>Nocardia farcinica</i>	AP006618
<i>Nocardioides</i> sp. JS614	CP000509
<i>Propionibacterium acnes</i> KPA171202	AE017283
<i>Rhodococcus jostii</i> RHA1	CP000431
<i>Salinospora arenicola</i> ATCC BAA-917	CP000850
<i>Saccharopolyspora erythraea</i> NRRL 2338	AM420293
<i>Salinispora tropica</i> CNB-440	CP000667
<i>Thermobifida fusca</i> YX	CP000088

A

Method	Total recombinant region	Percentage of the core
MaxChi	287941	49.88%
Chimerae	305848	52.99%
MaxChi and Chimerae	221921	38.45%

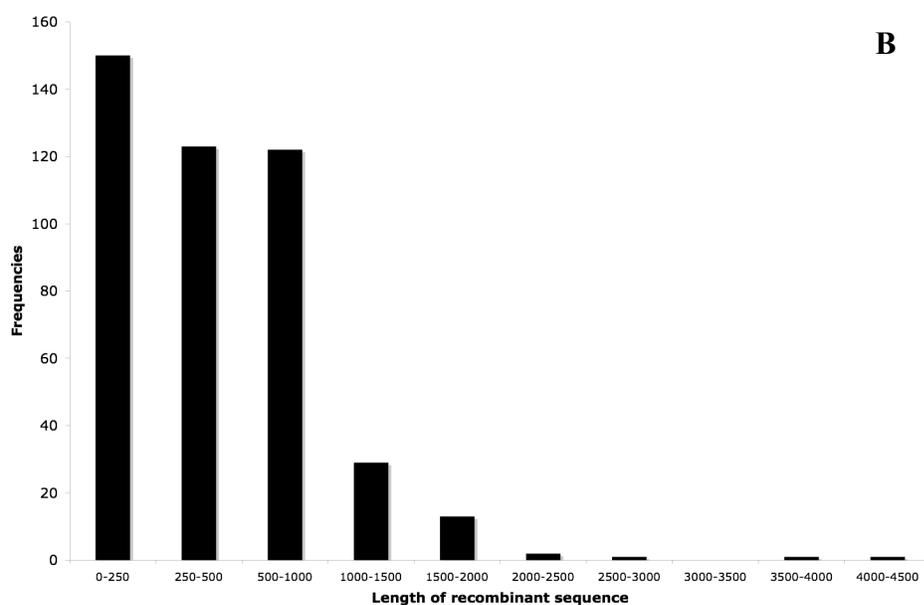


Figure 2-1. Recombinant regions detected by the nucleotide substitution algorithms. Table shows the results of MaxChi and Chimerae algorithms and the intersection of those results (A). The total recombinant region corresponds to the sum of the recombinant regions detected by the algorithm. The percentage was calculated using the length of the whole alignment produces with the core genes (57,7215 bp). Distribution of the length of recombinant sequences found by both algorithms (B).

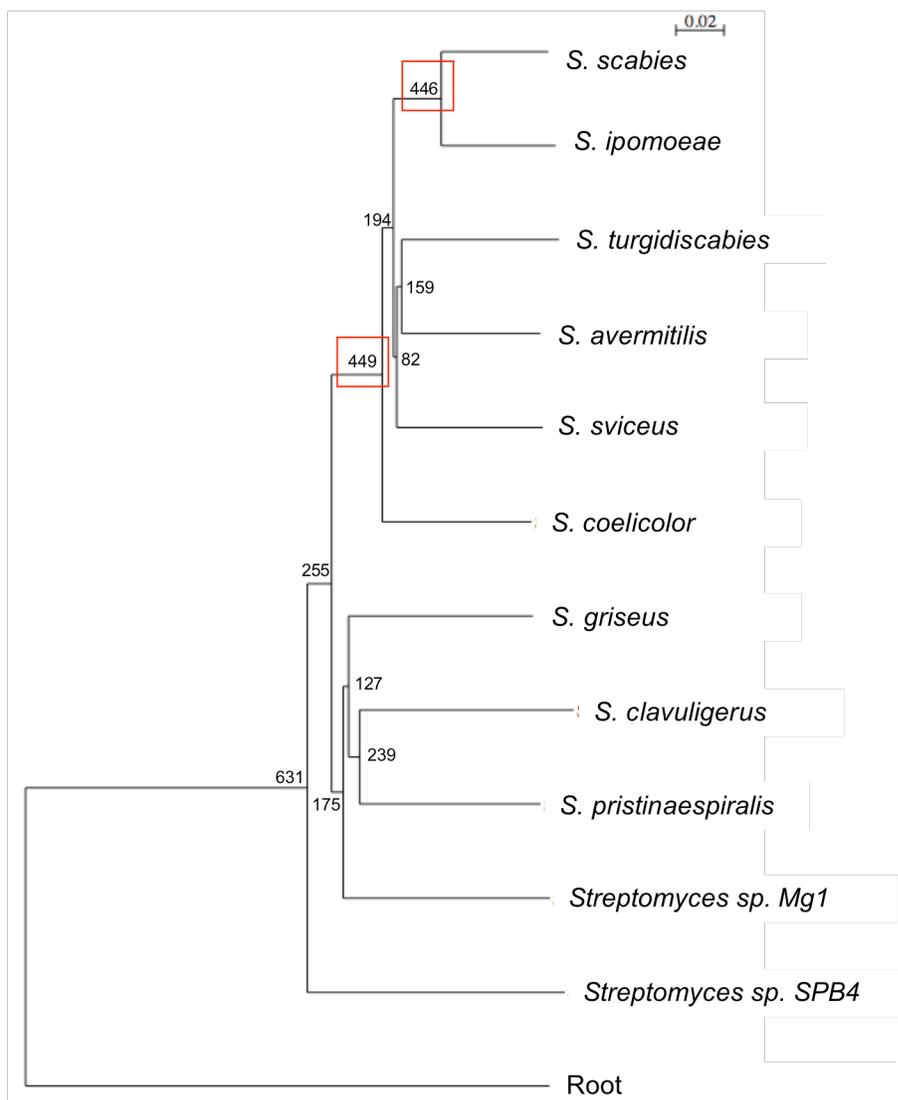


Figure 2-2. Consensus tree for 631 genes analyzed from the Streptomyces core-genome. The tree represents the best and most probable reconstruction for these Streptomyces spp. Numbers at the branches indicate the number of genes out of 631 that support the tree. Root indicates the related actinobacteria, which were used to root the tree. Significant clusters are indicated in red squares.

Table 2-2. Genes in *S. scabies* and *S. ipomoeae* that have undergone recombination. The first column describes the gene function. The second column describes the phylogenetic incongruence found in comparison with the consensus tree shown in Figure 2-2.

Gene	Phylogenetic incongruence
putative nucleotide sugar-1-phosphate transferase	<i>S. scabies</i> and <i>S. avermitilis</i> cluster together
putative extracellular solute-binding receptor	<i>S. scabies</i> and <i>S. sp S. Mg1</i> cluster together
guanosine pentaphosphate synthetase	<i>S. scabies</i> and <i>S. avermitilis</i> cluster together, <i>S. ipomoeae</i> and <i>S. sviceus</i> cluster together
putative polyamine transport protein	<i>S. scabies</i> and <i>S. sviceus</i> cluster together
putative polyamine ABC-transporter	<i>S. scabies</i> and <i>S. sviceus</i> cluster together
tRNA (guanine-N1)-methyltransferase	<i>S. ipomoeae</i> <i>S. avermitilis</i> and <i>S. turgidiscabies</i> cluster together
3-isopropylmalate dehydratase small subunit	<i>S. scabies</i> and <i>S. avermitilis</i> cluster together, <i>S. ipomoeae</i> and <i>S. sviceus</i> cluster together
putative DNA ligase	<i>S. scabies</i> and <i>S. coelicolor</i> cluster together
ATP synthase beta chain	<i>S. scabies</i> and <i>S. avermitilis</i> cluster together
ATP synthase gamma chain	<i>S. scabies</i> and <i>S. turgidiscabies</i> cluster together
ATP synthase alpha chain	<i>S. scabies</i> and <i>S. avermitilis</i> cluster together
ATP synthase B chain	<i>S. scabies</i> and <i>S. avermitilis</i> cluster together
ATP synthase A chain	<i>S. ipomoeae</i> and <i>S. sviceus</i> cluster together
conserved hypothetical protein	<i>S. scabies</i> and <i>S. avermitilis</i> cluster together
DNA-3-methyladenine glycosylase I	<i>S. ipomoeae</i> and <i>S. sviceus</i> cluster together
dihydropteroate synthase	<i>S. ipomoeae</i> and <i>S. sviceus</i> cluster together
putative elongation factor Tu	<i>S. ipomoeae</i> and <i>S. coelicolor</i> cluster together
50S ribosomal protein L13	<i>S. scabies</i> and <i>S. sviceus</i> cluster together; <i>S. ipomoeae</i> and <i>S. coelicolor</i> cluster together.
30S ribosomal protein S11	<i>S. ipomoeae</i> and <i>S. coelicolor</i> cluster together
50S ribosomal protein L30	<i>S. scabies</i> and <i>S. avermitilis</i> cluster together
30S ribosomal protein S5	<i>S. ipomoeae</i> and <i>S. coelicolor</i> cluster together
50S ribosomal protein L14	<i>S. scabies</i> and <i>S. turgidiscabies</i> cluster together
50S ribosomal protein L16	<i>S. scabies</i> , <i>S. avermitilis</i> , <i>S. turgidiscabies</i> cluster together
50S ribosomal protein L23	<i>S. ipomoeae</i> and <i>S. coelicolor</i> cluster together
30S ribosomal protein S10	<i>S. scabies</i> and <i>S. coelicolor</i> cluster together
elongation factor G	<i>S. ipomoeae</i> and <i>S. coelicolor</i> cluster together <i>S. scabies</i> and <i>S. turgidiscabies</i> cluster together
30S ribosomal protein S7	<i>S. ipomoeae</i> and <i>S. sviceus</i> cluster together
50S ribosomal protein L11	<i>S. scabies</i> and <i>S. turgidiscabies</i> cluster together
aspartate aminotransferase	<i>S. ipomoeae</i> and <i>S. coelicolor</i> cluster together
NuoF, NADH dehydrogenase subunit	<i>S. ipomoeae</i> cluster with <i>S. turgidiscabies</i>
putative GTP cyclohydrolase I	<i>S. scabies</i> and <i>S. sviceus</i> cluster together
putative phosphoribosyltransferase	<i>S. ipomoeae</i> cluster with <i>S. Mg1</i>
putative 30S ribosomal protein S6	<i>S. scabies</i> <i>S. turgidiscabies</i> and <i>S. sviceus</i> cluster together
50S ribosomal L25 protein	<i>S. ipomoeae</i> and <i>S. sviceus</i> cluster together
hydroxymethylglutaryl-CoA lyase	<i>S. scabies</i> and <i>S. sviceus</i> cluster together
GTP-binding protein	<i>S. ipomoeae</i> and <i>S. turgidiscabies</i> cluster together
putative GTP-binding elongation factor	<i>S. scabies</i> and <i>S. avermitilis</i> cluster together
30S ribosomal protein S1	<i>S. ipomoeae</i> and <i>S. sviceus</i> cluster together
conserved hypothetical protein	<i>S. scabies</i> and <i>S. avermitilis</i> cluster together
putative ABC transporter ATP-binding subunit	<i>S. scabies</i> and <i>S. sp Mg1</i> cluster together
putative LacI family regulator	<i>S. ipomoeae</i> and <i>S. avermitilis</i> cluster together
putative RNA pseudouridine synthase	<i>S. scabies</i> and <i>S. avermitilis</i> cluster together
glutamate N-acetyltransferase	<i>S. scabies</i> and <i>S. coelicolor</i> and <i>S. sviceus</i> cluster together
histidyl tRNA synthetase	<i>S. scabies</i> and <i>S. sviceus</i> cluster together

Analysis of the tree topologies indicates that *S. scabies* and *S. ipomoeae* have a history of recombination with other *Streptomyces* species Table (2-2). This occurs when one of these pathogens cluster with other *Streptomyces* and the branch displays high bootstrap support (>70). The majority of trees that display topological incongruence cluster *S. scabies* with *S. sviveus* and *S. ipomoeae* with *S. sviveus*. It is not possible to conclude that *S. sviveus* is the direct donor or receptor in the recombination history. However; our results suggest that *S. scabies* and *S. ipomoeae* have a recombination history with a group of *Streptomyces* related to *S. sviveus* (Table 2-2). Other *Streptomyces* species that have a recombination history with *S. ipomoeae* and *S. scabies* are *S. coelicolor*, *S. avermitilis* and *S. turgidiscabies*. For example, clusters of genes coding for an ATP synthetase appears to have been recombined between *S. scabies* and a group of *Streptomyces* close related to *S. avermitilis* (Table 2-2).

Interestingly, recombination is observed in a group of informational genes. The genes in this group code for ribosomal proteins and elongation factors (Table 2-2). Recombination in informational genes is considered rare but has been detected in other organisms. For example, genes encoding ribosomal proteins in *Mycoplasmas* (Brochier *et al.*, 2000), *Lactococcus lactis* (Makarova *et al.*, 2001), and the elongation factors Tu and EF-1 α in *Streptococcacea* (Ke *et al.*, 2000) and the archaea *Ethanopyrus kandleri* (Inagaki *et al.*, 2006), respectively, show phylogenetic evidence of recombination. Genetic recombination in *Streptomyces* has been documented under laboratory conditions. For example, *S. rimosus* and *S. aureofaciens* (Polsinelli and Beretta, 1966) genetically recombine when they are grown together, generating recombinant progeny.

CONCLUSIONS

We show that a large proportion (38%) of the core-genome of *Streptomyces* is affected by recombination. Furthermore, informational genes in *S. ipomoeae* and *S. scabies* have recombination signals. The high rates of recombination could be associated with the complex development process in *Streptomyces*. As mentioned in Chapter I, *Streptomyces* form filamentous structures with multiple copies of chromosomes in cellular compartments (Flardh *et al.*, 1999). In addition, *Streptomyces* are able to conjugate through fusion of cells with partial mobilization of chromosomal regions (Hopwood, 2006). Thus *Streptomyces* cells have a chimerical composition of hetero diploid chromosomes (Hopwood, 2006). Due to the high identity between core genes, homologous recombination could occur at high rates between the chromosomes.

Although phylogenetic reconstruction of the species tree with the core gene is not possible due extensive recombination events, partial resolution is observed among plant pathogenic *Streptomyces*. The resolution is observed in the branch that groups *S. ipomoeae* and *S. scabies*. This observation suggests that these two pathogens might have a common evolutionary history, as was suggested in Chapter I. Moreover, the level of synteny between their orthologues and the 16s rDNA tree supports this hypothesis. Unfortunately, at the time of this analysis we could not complete the 16s rDNA tree with the rest of the strains that were used in the core-genome phylogenetic tree. Further analysis and completion of the 16s rDNA sequences will produce more conclusive results about the relationships among of these species. However, we believe that *S. turgidiscabies* represents a newly evolved plant pathogen with features such as a virulence related genomic island that will be described in the next chapter.

REFERENCES

- Brochier, C., Philippe, H., and Moreira, D. (2000) The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends Genet* **16**: 529-533.
- Flardh, K., Findlay, K.C., and Chater, K.F. (1999) Association of early sporulation genes with suggested developmental decision points in *Streptomyces coelicolor* A3(2). *Microbiology* **145** (Pt 9): 2229-2243.
- Gogarten, J.P., Doolittle, W.F., and Lawrence, J.G. (2002) Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**: 2226-2238.
- Gogarten, J.P., and Townsend, J.P. (2005) Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* **3**: 679-687.
- Guindon, S., and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696-704.
- Guttman, D.S., and Dykhuizen, D.E. (1994) Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* **266**: 1380-1383.
- Hopwood, D.A. (2006) Soil to genomics: the *Streptomyces* chromosome. *Annu Rev Genet* **40**: 1-23.
- Inagaki, Y., Susko, E., and Roger, A.J. (2006) Recombination between elongation factor 1alpha genes from distantly related archaeal lineages. *Proc Natl Acad Sci U S A* **103**: 4528-4533.
- Ke, D., Boissinot, M., Huletsky, A., Picard, F.J., Frenette, J., Ouellette, M., Roy, P.H., and Bergeron, M.G. (2000) Evidence for horizontal gene transfer in evolution of elongation factor Tu in enterococci. *J Bacteriol* **182**: 6913-6920.
- Lanave, C., Preparata, G., Saccone, C., and Serio, G. (1984) A new method for calculating evolutionary substitution rates. *J Mol Evol* **20**: 86-93.

- Lefebure, T., and Stanhope, M.J. (2007) Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol* **8**: R71.
- Majewski, J., and Cohan, F.M. (1999) DNA sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics* **153**: 1525-1533.
- Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F.M., and Dowson, C.G. (2000) Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. *J Bacteriol* **182**: 1016-1023.
- Makarova, K.S., Ponomarev, V.A., and Koonin, E.V. (2001) Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biol* **2**: RESEARCH 0033.
- Martin, D.P., Williamson, C., and Posada, D. (2005) RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* **21**: 260-262.
- Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299-304.
- Polsinelli, M., and Beretta, M. (1966) Genetic Recombination in Crosses Between *Streptomyces aureofaciens* and *Streptomyces rimosus*. *J Bacteriol* **91**: 63-68.
- Posada, D., and Crandall, K.A. (2001a) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A* **98**: 13757-13762.
- Posada, D., and Crandall, K.A. (2001b) Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Mol Biol Evol* **18**: 897-906.
- Roshan, U., and Livesay, D.R. (2006) Probalalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics* **22**: 2715-2721.

Wu, J., Yu, T., Bao, Q., and Zhao, F. (2009) Evidence of extensive homologous recombination in the core-genome of rickettsia. *Comp Funct Genomics*: 510270.

CHAPTER 3
COMPLETE SEQUENCE AND ANALYSIS OF THE MOBILE
PATHOGENICITY ISLAND (PAISt) IN *Streptomyces turgidiscabies* Car8

ABSTRACT

Streptomyces turgidiscabies is a bacterial pathogen that causes potato scab disease. Pathogenesis is associated with a large mobile genomic island (PAISt) integrated within the chromosome. The island encodes the pathway for the synthesis of the dipeptide phytotoxin thaxtomin, the necrosis factor (Nec1), a cytokinin biosynthesis pathway, and a tomatine-degrading enzyme (TomA). The PAISt can be mobilized by conjugation from *S. turgidiscabies* to other non-pathogenic *Streptomyces* spp. and, in some cases, the recipient strain acquires a virulence phenotype. Analysis of the draft genome sequence of *S. turgidiscabies* Car8 allowed us to identify the integrated 675,092 bp PAISt sequence. The island has an overall G+C content of 68.46% and is predicted to encode 647 proteins. The organization of the PAISt resembles a complex integrative conjugative element (ICE) containing potential genes involved in integration, excision and mobilization. Here we describe novel putative virulence and fitness factors encoded within the PAISt. Comparative sequence analysis suggests how the PAISt has evolved in *S. turgidiscabies* Car8 and demonstrates that this mobile element plays an important role in lateral gene transfer.

INTRODUCTION

Integrative conjugative elements (ICEs) are mobile genetic elements found in prokaryotes and considered to play an important role in lateral gene transfer (Burrus and Waldor, 2004; Kers *et al.*, 2005; Mavrodi *et al.*, 2009; Pembroke and Piterina, 2006). These genetic elements evolve within the chromosome of the host in a symbiotic fashion, providing fitness and virulence genes in exchange for a “safe place” within the host genome that secures its maintenance. ICEs are variable in both size and gene content, and occur in many bacterial species. One of the smallest reported ICE is *Tn916*, which is 16.4 Kb and has been described from the *Enterococcus faecalis* genome; this element contains genes for tetracycline resistance (Clewell *et al.*, 1988). On the other side of the spectrum are large ICEs with complex gene organization, such as the 611 Kb element in *Mesorhizobium loti* that contains genes involved in nitrogen fixation and nodulation (Kaneko *et al.*, 2000).

ICEs integrate into host chromosomes at specific nucleotide sequences, denoted *att*-sites (Burrus and Waldor, 2004). The process of integration is carried out by a recombinase-integrase (Int) protein that is encoded within the element. The Int protein is responsible for the excision of the element, and is sometimes assisted by an excisionase (Xis) protein generally encoded upstream of the *int*. Once excised from the chromosome of its host the element exists as a transient circular structure that does not replicate autonomously, but transfers to other bacteria by conjugation, using a set of DNA mobilization proteins encoded within the ICE (Burrus and Waldor, 2004). Most of the ICEs transfer as a single DNA strand; however, conjugative elements in *Streptomyces* and related actinobacteria transfer as double stranded DNA (Grohmann *et al.*, 2003). The mechanism for double-stranded transfer resembles the process of chromosome partitioning during pre-spore formation in *B. subtilis* (Errington, 2001).

It is believed that Lateral Gene Transfer (LGT) plays an important role in the evolution of pathogenic streptomycetes (Loria *et al.*, 2006). The best-characterized example is the large genomic island (PAISt) that exists in *S. turgidiscabies*, a pathogen reported for the first time in the island of Hokkaido, Japan in the late 1900s (Kers *et al.*, 2005; Miyajima *et al.*, 1998). Previous studies demonstrated that the PAISt encodes at least four virulence factors: the biosynthetic pathway for the phytotoxin thaxtomin (Txt proteins), a tomatinase enzyme (TomA), the secreted necrogenic protein, Nec1, and the plant fasciation (Fas) biosynthetic pathway (Kers *et al.*, 2005). The PAISt can transfer during conjugation to other non-pathogenic *Streptomyces* as ~100 Kb or ~640 Kb modules. Both versions of the PAISt integrate specifically at the 3' end of the bacitracin resistance gene (*bacA*), recognizing the eight bp palindromic sequence, TTCATGAA (Kers *et al.*, 2005).

Completion of a draft genome sequence for *S. turgidiscabies* Car8 allowed us to identify, annotate, and analyze the entire sequence of the PAISt. A complete inventory of genes associated with integration and transfer, as well as putative and authentic virulence genes is presented here.

METHODOLOGY

Annotation of the PAISt

A pseudomolecule representing the chromosome of *S. turgidiscabies* Car8 was built from the scaffolds of the *S. turgidiscabies* Car8 genome-sequencing project (<http://www.ttaxus.com/files/Streptomyces/gst.zip>). Concatenation of scaffolds was conducted using Mummer software (Delcher *et al.*, 2003). The scaffolds were ordered using as a reference the finished sequence of the *S. scabies* 87-22 chromosome. Previously published sequences from the PAISt were used as genomic landmarks to delimit the boundaries of the PAISt in the pseudomolecule (Kers *et al.*, 2005). The

landmark sequences used were *necl*, the *txt* genes, and the 3' end of the bacitracin resistance gene (*bacA*). Genbank accession numbers: AY707080, AY707081 and AY707082, respectively.

Coding sequences in the PAISt were predicted using Glimmer 3.2 (Delcher *et al.*, 2007) trained with *S. scabies* 87-22 genes. Predicted genes were automatically annotated using Blast2GO (Conesa and Gotz, 2008). Manual annotation was performed using Artemis (Rutherford *et al.*, 2000). Predicted coding sequences in the PAISt were categorized using the Gene Ontology (GO) (GOC, 2006) and the Pfam (Bateman *et al.*, 2004) databases. Prediction of sub cellular localization of proteins was conducting using PSORTb version 2.0 (Gardy *et al.*, 2005).

Comparative sequence analysis

Predicted coding sequences were compared at the amino acid level with the genomes of Actinobacteria and plant pathogenic bacteria (Table 3-1). The Basic Local Alignment Sequence Tool (BLAST) (Altschul *et al.*, 1990) was used to identify homologues using e-values of $1e^{-5}$ as a threshold for significance. Identity scores were retrieved from the BLAST output and used to build a matrix of best hits. The matrix was used to generate a BLAST-fingerprinting plot with the R statistical package (<http://www.r-project.org/>). Additionally, sequence comparisons and analysis of gene synteny were carried out using the Artemis Comparison Tool (ACT) (Carver *et al.*, 2008).

Southern blot and hybridization

Genomic DNA was isolated from *S. turgidiscabies* Car8 cultures grown overnight in TSB using the Gram-positive MasterPure Kit (Epicenter) according to the manufactures instructions. Four micrograms of genomic DNA was digested with

Table 3-1. Genomes used for Blast finger printing of the PAISt. Brief description of each bacterium and accession numbers are provided. Accession numbers of plasmids (*). Genomes in draft status (**).

Bacterial species	Features	Source
<i>S. scabies</i> 87.22	Plant pathogen	FN554889
<i>S. avermitilis</i> MA-4680	Soil-dwelling producer of anti-parasitic agent avermectin	BA000030 AP005645
<i>S. coelicolor</i> A3(2)M145	Soil-dwelling producer of antibiotics	AL645882 AL589148 (*) AL645771 (*)
<i>S. griseus</i> IFO 13350	Antibiotic producer	AP009493
<i>S. pristinaespiralis</i> ATCC 25486	Antibiotic pristinamycin producer	ABJI00000000 (**)
<i>S. sviceps</i> ATCC 29083	Soil saprophyte actinomycete	ABJJ00000000 (**)
<i>S. clavuligerus</i> ATCC 27064	Producer of clavulanic acid and cephamycin C	ABJH00000000 (**)
<i>S. sp</i> SPB74	Saprophyte	ABJG000000 (**)
<i>S. sp</i> Mg1	Saprophyte	ABJF00000000 (**)
<i>S. erythraea</i> NRRL 2338	Erythromycin producer	AM420293
<i>Frankia sp.</i> EAN1pec	Symbiotic nitrogen-fixing plant commensal organism	CP000820
<i>Frankia alni</i> ACN14a	Symbiotic nitrogen-fixing plant commensal organism	CT573213
<i>Frankia sp.</i> Ccl3	Symbiotic nitrogen-fixing plant commensal organism	CP000249
<i>C. michiganensis</i> subsp. <i>michiganensis</i> NCPPB 382	Plant pathogen	AM711867 AM711865 (*) AM711866 (*)
<i>C. michiganensis</i> subsp. <i>sepedonicus</i>	Plant pathogen	AM849034 AM849036 (*) AM849035 (*)
<i>K. radiotolerans</i> SRS30216	Radiotolerant bacterium	CP000750
<i>S. arenicola</i> CNS-205	Marine actinomycete	CP000850
<i>S. tropica</i> CNB-44.0	Marine actinomycete	CP000667
<i>N. farcinica</i> IFM 10152	Pathogenic actinomycete	AP006618 AP006619 (*) AP006620 (*)
<i>Nocardiodex</i> sp. JS614	Vinyl chloride assimilating actinobacteria	CP000509 CP000508 (*)
<i>R. jostii</i> RHA1	Degradation of polychlorinated biphenyls.	CP000431 CP000432 (*) CP000433 (*) CP000434 (*)
<i>C. diphtheriae</i> NCTC 13129	Human pathogen	BX248353
<i>C. glutamicum</i> ATCC 13032 K (Kitasato)	Industrial producer of multiple amino acids	BA000036
<i>C. glutamicum</i> ATCC 13032 B (Bielefeld)	Industrial producer of multiple amino acids	BX927147
<i>Mycobacterium</i> MCS	Pyrene-degrading bacterium	CP000384 CP000385 (*)
<i>T. fusca</i> YX	Thermophilic bacteria	CP000088
<i>M. leprae</i> TN	Human pathogen	AL450380
<i>M. tuberculosis</i> CDC1551	Human pathogen	AE000516
<i>M. tuberculosis</i> F11	Human pathogen	CP000717
<i>M. ulcerans</i> Agy99	Human pathogen	CP000325 BX649209 (*)
<i>M. smegmatis</i> MC2 15.5	Non-pathogenic mycobacterium capable of causing soft tissue lesions	CP000480
<i>P. syringae</i> pv. <i>tomato</i> srt DC3000	Plant pathogen	AE016853 AE016854 (*) AE016855 (*)
<i>Ralstonia solanacearum</i> GM1000	Plant pathogen	AL646052 AL646053 (*)
<i>Ralstonia pickettii</i> 12J	Nosocomial opportunistic pathogen	CP001068 CP001069 (*) CP001070

NcoI and PstI. Fragments were separated by electrophoresis in a 1 % agarose gel. The DNA was depurinated with 0.1 M HCl, denatured with 0.5 NaOH and 1.5 M NaCl and neutralized with 1.5M NaCl and 0.5M Tris.HCl. DNA fragments were transferred overnight to a nylon membrane (Whatman) by capillary transfer using 20 x SCC. DNA was UV-crosslinked to the membrane by applying 120,000 $\mu\text{joules}/\text{cm}^2/\text{sec}$ for four minutes. The membrane was probed with a 400 bp. fragment of the *tomA* gene that was obtained by PCR using primers JH79 (5-CTGTTTCATCAACGAGTCGTTG-3) and JH80 (5-GTAGGCGTGCTTGTCGGTG-3). The PCR product was labeled with dioxigenin-11-UTP (Roche). Hybridization was performed in a rotary hybrid oven at 62° C. Stripping conditions were carried out with 0.5X sodium chloride/sodium citrate buffer (SCC) at 65° C for 20 minutes. Additional steps for hybridization and detection were performed according to the manufacturers instructions (Roche).

Phylogenetic trees reconstruction

Sequences of the *tomA* genes were aligned with homologues found in Genbank. *Kinetococcus radiotolerans* YP_001362287, *Fusarium oxysporum* BAB88658, *Aspergillus fumigatus* EAL89026, *Neosartoria fischeri* EAW16401 and *Clavibacter michiganensis* subsp. *michiganensis* AF39183. Alignment was conducted at the nucleotide level using ClustalX (Thompson *et al.*, 2002) and phylogenetic trees were reconstructed using the neighbor-joining algorithm (Saitou and Nei, 1987).

RESULTS AND DISCUSSION

Genome structure and general features

The PAISt is a complex genetic element of 674,225 bp. There are two sub elements delimited by three copies of the eight-palindrome sequence TTCATGAA

(Figures 3-1 and 3-2), which is the genomic integration site of the element (*att*-site) (Kers *et al.*, 2005). The two non-overlapping sub elements are 105,373 bp and 568,852 bp and have an average G+C content of 68.18%. The coding percentage of the PAISt is 84.6%, with 647 predicted genes having an average length of 883 bp.

Categories of genes in the PAISt

Annotation of the PAISt indicates that 197 (30%) of the predicted proteins do not have any match in the current Genbank database, 165 (25%) are conserved proteins without cellular process association. The remaining 287 (45%) can be classified in broad cellular process categories (Figure 4-3). The largest member of predicted proteins are related to signal transduction and transcription, with 51 (8%) predicted proteins. These proteins represent the regulatory network that might be expected to modulate the diverse biological processes encoded within the PAISt.

Forty-eight genes are predicted to encode transport, general secretion, and membrane-associated proteins, representing 7% of the predicted proteins. Several of these proteins are classified as putative ABC transporters and might function in nutrient uptake or secretion of secondary metabolites. Genes involved in DNA transposition and site-specific recombination represent 5% of the PAISt (47 genes). These elements are common in genomic islands and often play an important role in mobilization and acquisition of DNA (Burrus and Waldor, 2004). Other important categories found within the PAISt are secondary metabolism, 22 genes (3%); DNA metabolism, 16 genes (2%); and, carbohydrate metabolism, 13 genes (2%).

Analysis of categories of genes reveals integration and mobilization functions of the PAISt.

Previous studies suggested that the PAISt transfers from *S. turgidiscabies* Car8 by conjugation to other *Streptomyces* spp. and integrates by a site-specific recombination mechanism (Kers *et al.*, 2005). This evidence suggests that the PAISt behaves as an ICE. The recombinases-integrases of these elements are typically located at one end of the ICE, adjacent to the *att* site (Burrus and Waldor, 2004; Burrus *et al.*, 2006). Therefore, the ends of the ICE were scanned for the presence of putative recombinase genes.

As anticipated, the PAISt contains a coding sequence, stPAI0647, positioned just upstream of the palindromic eight bp sequence TTCATGAA (Figure 4-4). BLAST results indicated that the 466 amino acid sequence of stPAI0647 is 30% similar to a tyrosine-type recombinase found in *Alicyclobacillus acidocaldarius* (Genbank accession number: EED06447). This evidence suggests that stPAI0647 is the recombinase responsible for the integration and excision of the element in *S. turgidiscabies* Car8. Experimental data supporting this hypothesis is presented in Chapter IV.

The PAISt also possesses a cluster of genes similar to, and syntenic with, a plasmid integrated in the chromosome of *Coynebacterium glutamicum* ATCC 13032 (Figure 4-5). Among these genes, stPAI0112 encodes a protein similar to Cg2005 (Genbank accession CAF20165). The predicted protein encoded by stPAI0112 contains a P-loop NTPase domain, which is typically found in DNA motor-pump proteins (Gomis-Ruth *et al.*, 2001; Tato *et al.*, 2005; Tato *et al.*, 2007). Interestingly, conjugation of several actinobacterial ICEs occurs as double stranded DNA, with the participation of a DNA binding protein that contains domains similar to DNA motor-pump proteins (te Poele *et al.*, 2008a). It is tempting to speculate that StPAI0112 is involved in the mobilization of PAISt.

In addition to the integrative and mobilization genes found within the PAISt, this element is predicted to encode a set of proteins involved in DNA metabolism (Figure 3-4). The protein encoded by *stPAI0185*, which displays 46% similarity to the phage DNA primase cg1960 found in *C. glutamicum*, (Genbank accession number CAF20121). Another interesting protein is encoded by *stPAI0176*, which is 43% similar to a DNA helicase cg1963, (Genbank accession CAF20125). As observed with the putative mobilization gene *stPAI0112*, genes *stPAI0185* and *stPAI0176* are located in the syntenic region of PAISt with the integrated plasmid in *C. glutamicum* (Figure 3-5). Typically, ICEs do not possess autonomous replication functions (Burrus and Waldor, 2004). However, a unique group of ICEs in actinobacteria can autonomously replicate and possess genes implicated in DNA replication (te Poele *et al.*, 2008a; te Poele *et al.*, 2008b). While integration of PAISt has been demonstrated and conjugative properties are consistent with existing data (Kers *et al.*, 2005), autonomous replication has not previously been proposed.

The PAISt encodes proteins expected to be involved in virulence and fitness.

The PAISt encodes several virulence proteins that have been confirmed experimentally (Bukhalid *et al.*, 1998; Joshi and Loria, 2007; Kers *et al.*, 2005; Loria *et al.*, 2008; Seipke and Loria, 2008). In addition to these previously identified virulence genes, the PAISt possesses an operon predicted to be involved in iron acquisition (Figure 3-6). Sequence analysis identified proteins encoded by *stPAI0586*, *stPAI0587* and *stPAI058* as cytoplasmic membrane proteins. The putative molecular functions for *stPAI0586* and *stPAI0587* are an iron dependent peroxidase and an iron permease, respectively. The sequences are 86% and 70% similar to proteins in *Frankia sp.* CcI3 (Genbank accessions ABD12758 and ABD12757). On the other hand, *stPAI058* is 79% similar to an integral membrane protein with putative

peptidase functions found in *Streptomyces coelicolor* (Genbank accession CAA20090). These three ORFs are located downstream of *stPAI0589*, which codes a predicted transcriptional regulator that contains a motif typically found in penicillinase repressor type regulators (Figure 3-6).

Interestingly, the organization and the functions of the proteins encoded in this region of the PAISt resemble the ferrous iron transporter EfeUOB (YcdNOB) (Cao *et al.*, 2007). EfeUOB is a tripartite Fe²⁺ transporter that is cryptic in *Escherichia coli* K-12 because of a frame shift mutation. However; in the enterohaemorrhagic strain *E. coli* O157:H7 the EfeUOB is functional and has been identified as an important virulence factor (Cao *et al.*, 2007). The putative iron uptake operon within the PAISt is an excellent target for experimental research in *S. turgidiscabies*.

The PAISt possesses a novel lantibiotic biosynthesis operon

Besides the thaxtomin biosynthetic operon, the PAISt contains another cluster of ORFs predicted to be involved in nonribosomal peptide synthesis. The predicted genes *stPAI0322*, *stPAI0323*, *stPAI0324*, *stPAI0325*, *stPAI326*, *stPAI0327* are syntenic with a putative operon in *S. griseus* (Figure 3-7). The operon is predicted to be involved in lantibiotic synthesis. The predicted lantibiotic synthetase is encoded by *stPAI325*, and the other predicted proteins are a methyltransferase and genes associated with nonribosomal peptide modification. Lantibiotics are active against other Gram-positive bacteria (Nes and Tagg, 1996; Willey and van der Donk, 2007) and could be an important factor in *S. turgidiscabies* Car8 for rhizosphere competition and plant colonization.

Predicted extra cellular proteins

Fifteen genes were predicted to code for extra cellular proteins (Table 3-2). Among these, *stPAI0015* encodes a protein that contains a ricin-type beta-trefoil lectin domain. Although the function of the protein is unknown, the ricin-lectin domain binds to sugars and has been found in several toxins. A protein similar to *stPAI0015* is found in the fungus *Neosartoria fischeri*, an opportunistic human pathogen (Lonial *et al.*, 1997). Other predicted extra cellular protein is a beta-galactosidase encoded by *stPAI0009*. The gene is located downstream to a predicted ABC transport system (*stPAI0011*, *stPAI0012*, *stPAI0013*). BLAST results suggest that this cluster of genes may be involved in metabolism of complex carbohydrates.

NUDIX hydrolases within the PAISt and putative functions

NUDIX hydrolases have been proposed to be “house cleaning” proteins that hydrolyze toxic intermediate metabolites of biosynthetic pathways (Bessman *et al.*, 1996). Several NUDIX hydrolases are associated with defense against oxidative stress (Bessman *et al.*, 1996). MutT, for example, plays an important function in repair DNA oxidative damage (Bessman *et al.*, 1996). Other NUDIX hydrolases have been shown to regulate indirectly the activity of other proteins. For example a NUDIX hydrolase that stimulates the activity of a nicotinoprotein alcohol dehydrogenase in *B. methanolicus* (Kloosterman *et al.*, 2002). Also, the ICE pSAM2 in *S. ambifaciens* possesses the NUDIX hydrolase, Pif that regulates the activity of the repressor KorSA. KorSA controls the integration, mobilization and replication of the pSAM2. Pif prevents redundant transfers between cells that contain pSAM2 and is the first reported NUDIX hydrolase associated with the mobilization of ICEs (Possoz *et al.*, 2003).

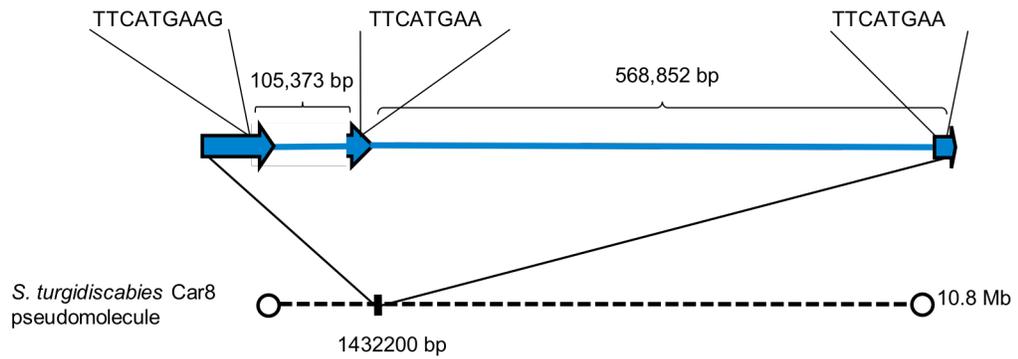
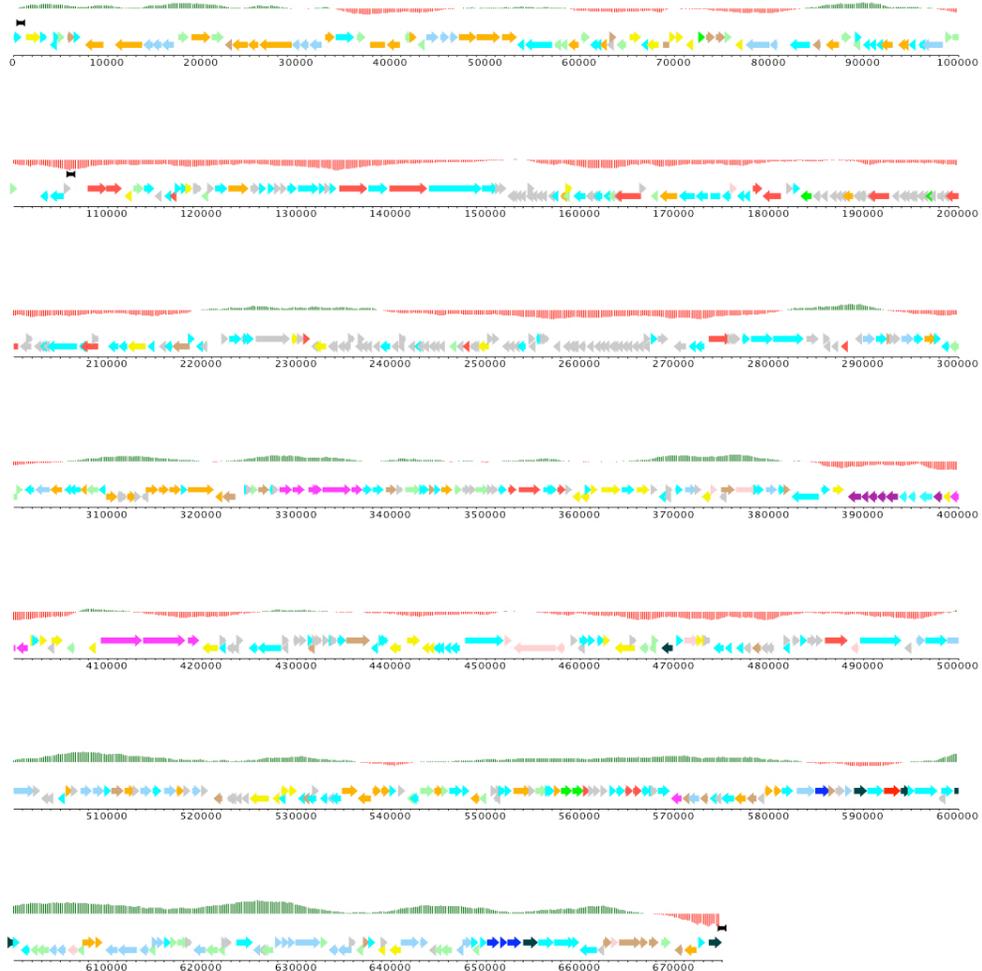


Figure 3-1. PAISt structure and location in the chromosome of *S. turgidiscabies* Car8. The cartoon shows the location of the PAISt in the pseudo molecule of *S. turgidiscabies* Car8. The eight bp recombination sites (*att* sites) are imbedded within the duplication of the 3' end of the bacitracin resistance gene.

Figure 3-2. Complete map of the PAISt in *S. turgidiscabies* Car8. Coding sequences are colored based on predicted biological functions. GC (%) content is indicated in red (below average), green (above average).



- | | |
|---|---|
| ■ Conserved hypothetical | ■ Transport/general secretion/membrane associated |
| ■ Hormone biosynthesis | ■ Signal transduction/regulation |
| ■ Insertion elements/transposons | ■ Secondary metabolism |
| ■ DNA metabolism | ■ Response to stress |
| ■ Proteolysis, Protein modification and fate | ■ Energy |
| ■ Site-specific Recombination - Integrases | ■ Miscellaneous |
| ■ Metabolism | ■ Hypothetical no homology in databases |

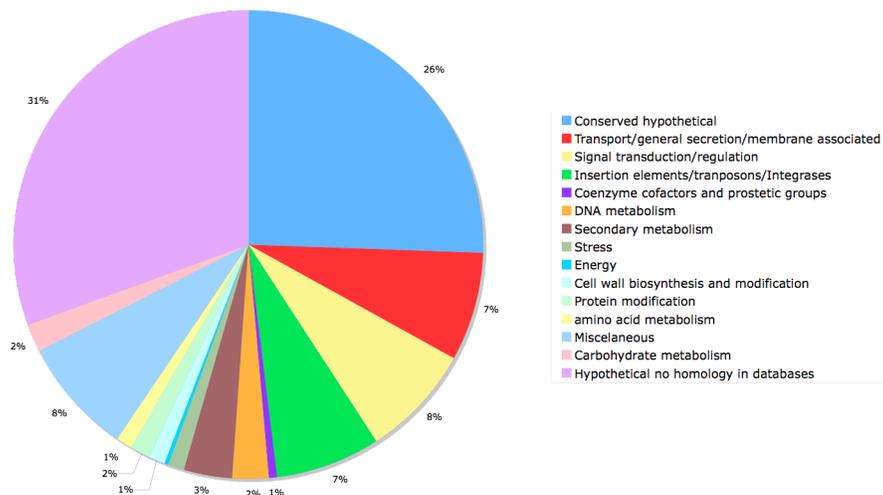


Figure 3-3. Predicted CDS for the PAISt classified by cell function categories.

Interestingly, PAISt possesses four predicted genes coding for putative NUDIX hydrolases, *stPAI0322*, *stPAI0631*, *stPAI0638* and *stPAI0642*. Two possible functions can be proposed for these clusters of NUDIX hydrolases. The first is that the NUDIX hydrolases play roles of detoxification of intermediary metabolites produced by the biological processes that are carried out within the PAISt. The second possibility is that some of the predicted NUDIX hydrolyses could play a role in the regulation of integration and/or conjugation of the PAISt.

Comparative sequence analysis of the PAISt suggests recombination events between *S. turgidiscabies* and *S. scabies*.

S. scabies and *S. turgidiscabies* are pathogenic *Streptomyces* that share common virulence factors. Previous studies demonstrated that *tomA*, *nec1* and the thaxtomin biosynthetic operon, are present in both species (Bukhalid *et al.*, 1998; Kers *et al.*, 2005; Seipke and Loria, 2008). Comparative analysis indicates that the first module of the PAISt (105,373 bp) is 95% identical to a genomic island located at 1,566576 bp in the chromosome of *S. scabies* 87-22 (PAISs1) (Figure 3-8 and 3-9). As observed in *S. turgidiscabies*, PAISs1 is integrated at the 3' end of the *bacA* gene. However, in PAISs1 the *att* site at the 3' end of the island is mutated from TTCATGAA to TGTATGAA. This mutation in the *att*-site of the PAISs1 might have resulted in the fixation of the island in *S. scabies* 87-22 (Figure 3-9).

A second conserved region of the PAISt is found in *S. scabies* 87-22 chromosome located at position 6, 531,990 bp (Figure 3-9). This region comprises other island (PAISs2) that contains the thaxtomin biosynthesis operon. The operon encodes for two nonribosomal peptide synthetases (TxtA and TxtB), an unusual P450 monooxygenase, a nitric oxide synthase (NOS) and an AraC-like regulator TxtR. The thaxtomin biosynthetic genes encoded in *S. scabies* 87-22 and *S. turgidiscabies* Car8

are syntenic and on average 89% identical at the amino acid level (Figure 3-10). The high identity between the islands found in *S. scabies* and *S. turgidiscabies* suggests recombination events between these *Streptomyces* species.

Possible models for the evolution of the PAISt

Reconstruction of recombination processes among the pathogenic islands in *S. turgidiscabies* and *S. scabies* is not possible. However, sequence comparisons suggest several possible scenarios. One possible scenario is that the PAISt was originally an ICE in *S. scabies* and progressively acquired novel DNA. The entire island might have been transferred from *S. scabies* or other intermediary donor to *S. turgidiscabies*. In *S. scabies* the continuous processes of gene erosion and recombination might have provoked the loss of the recombinase, conjugation genes, the *fas* operon and caused the degeneration of the integration sites. Separation of the island into two modules, PAISs1 that contains *necl* and *tomA*, and PAISs2 that contains the thaxtomin biosynthetic cluster. This recombination event could have fixed the PAI in the *S. scabies* 87-22 genome. In contrast, in *S. turgidiscabies* the island is still mobile (Kers *et al.*, 2005). Furthermore, the functional recombinase-integrase in the PAISt may serve in the acquisition of novel DNA into the ICE. It is tempting to speculate that this mechanism resembles super-integron elements, which “capture” foreign DNA and integrate it into the genome of their host by site-specific recombination (Biskri *et al.*, 2005; Fonseca *et al.*, 2008).

Another possible scenario is that *S. turgidiscabies* acquired a small version of the island from *S. scabies*. The thaxtomin biosynthetic cluster and *fas* operon could have been acquired subsequently during independent recombination events, forming genomic “islets” within the PAISt. Transposons and insertion elements in the vicinity of the thaxtomin genes and *fas* operon suggest that these regions might have been

acquired through transposition. We cannot discard the possibility that *S. scabies* and *S. turgidiscabies* might have a common ancestor that contained the island. It also is possible to hypothesize that the island in *S. scabies* has suffered deletions or intra chromosome recombination events that divided it in two independent islands. In contrast, in *S. turgidiscabies* the island has remained intact.

Comparative analysis of the PAIS_t with *S. avermitilis* and *S. coelicolor*.

In addition to the high identity with *S. scabies* genomic islands, discrete regions at the 3' end of the PAIS_t also display high identity with *S. avermitilis* and *S. coelicolor* (Figure 3-8). Alignment of the PAIS_t with *S. avermitilis* and *S. coelicolor* chromosomes reveals syntenic clusters of genes coding for putative two component systems and ABC transporters (Figure 3-11). The specific function of these clusters of genes within the island is unknown but it is possible that they play a role in regulation of conjugation through environmental sensing. These highly identical and syntenic regions in the PAIS_t might be the result of recombination events between *S. turgidiscabies*, *S. avermitilis* and *S. coelicolor*.

The region containing the *tomA* gene in the PAIS_t is amplified in *S. turgidiscabies* Car8.

Sequence analysis of the PAIS_t and the draft sequence of the *S. turgidiscabies* Car8 genome revealed an apparent duplication of a ~15 Kb region of the PAIS_t in the *S. turgidiscabies* Car8 chromosome (Figure 3-12). The duplicated regions are ~87% identical at the nucleotide level. Both regions contain the *tomA* gene, which encodes an enzyme that hydrolyzes the antimicrobial, plant defense compound tomatin to the nontoxic compounds tomatidine and beta-lycotetraose (Roldan-Arjona *et al.*, 1999).

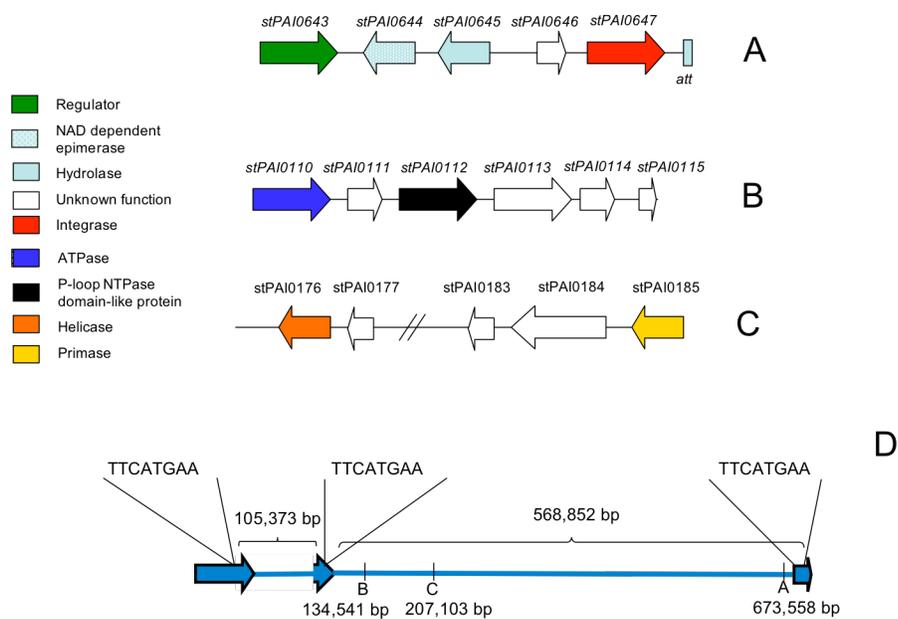


Figure 3-4. Three clusters of genes within the PAIS_t containing putative integration/excision (A), conjugation (B), and replication (C) functions. The positions of the three gene clusters within the PAIS_t; the *att* site (TTCATGAA) is the PAIS_t Integration site (D).

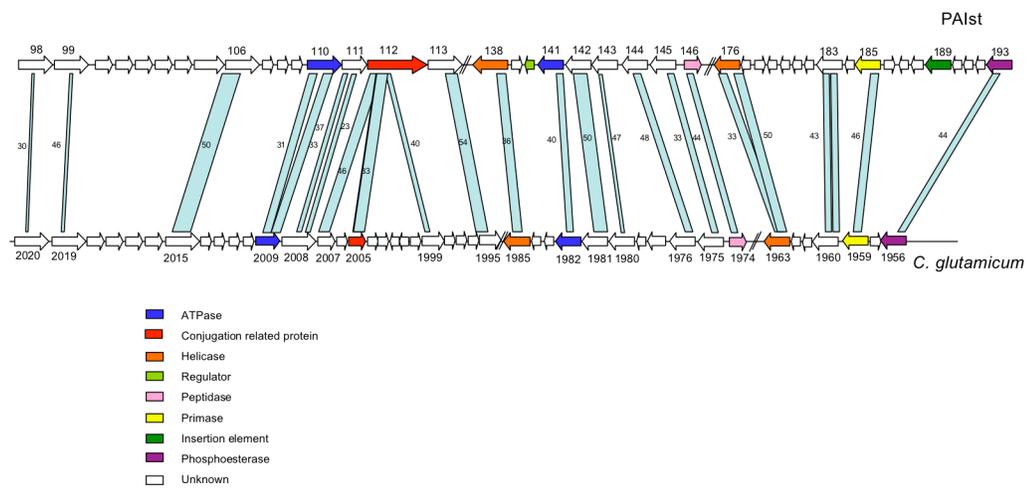


Figure 3-5. Region of the PAIst syntenic with the integrated plasmid in *Corynebacterium glutamicum*. Pale blue lines indicate homology. Percentage of identity at amino acid level is indicated. Numbers below or above the arrows indicate gene identifiers.

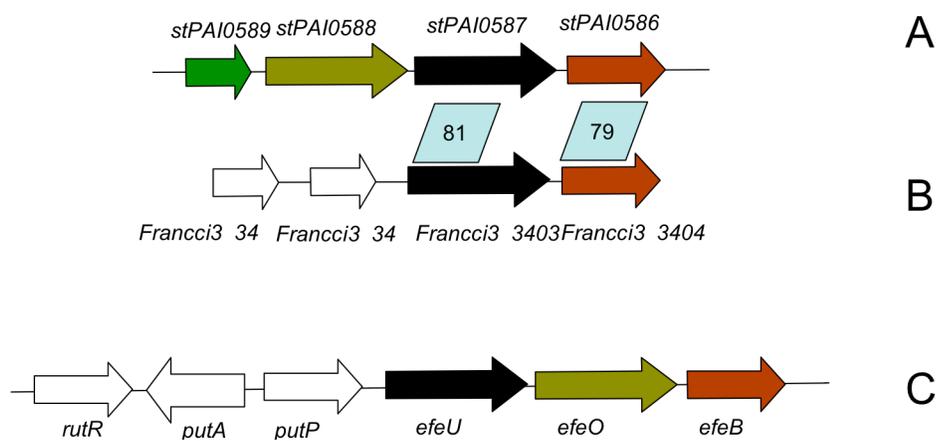


Figure 3-6. Putative iron uptake operon encoded within the PAISt. *StPAI0588*, *stPAI0587* and *stPAI0586*, encode a peptidase, an iron permease and a peroxidase, respectively, while *stPAI0589* encodes a transcriptional regulator (A). Homologues of *stPAI0587* and *stPAI0586* are found in *Frankia spp.*; ccl3 amino acid identity (%) with *stPAI* is indicated (B). Organization of the model iron uptake *efeUOB* operon found in *E. coli* in which *EfeU* is an inner membrane permease, *EfeO* is a periplasmic cupredoxin-peptidaseM75 protein, and *EfeB* is a periplasmic DyP-type haem peroxidase (C).

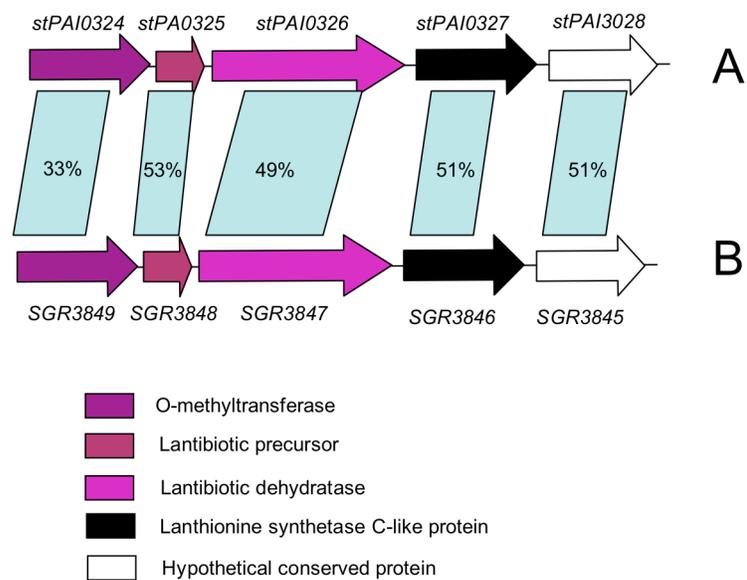
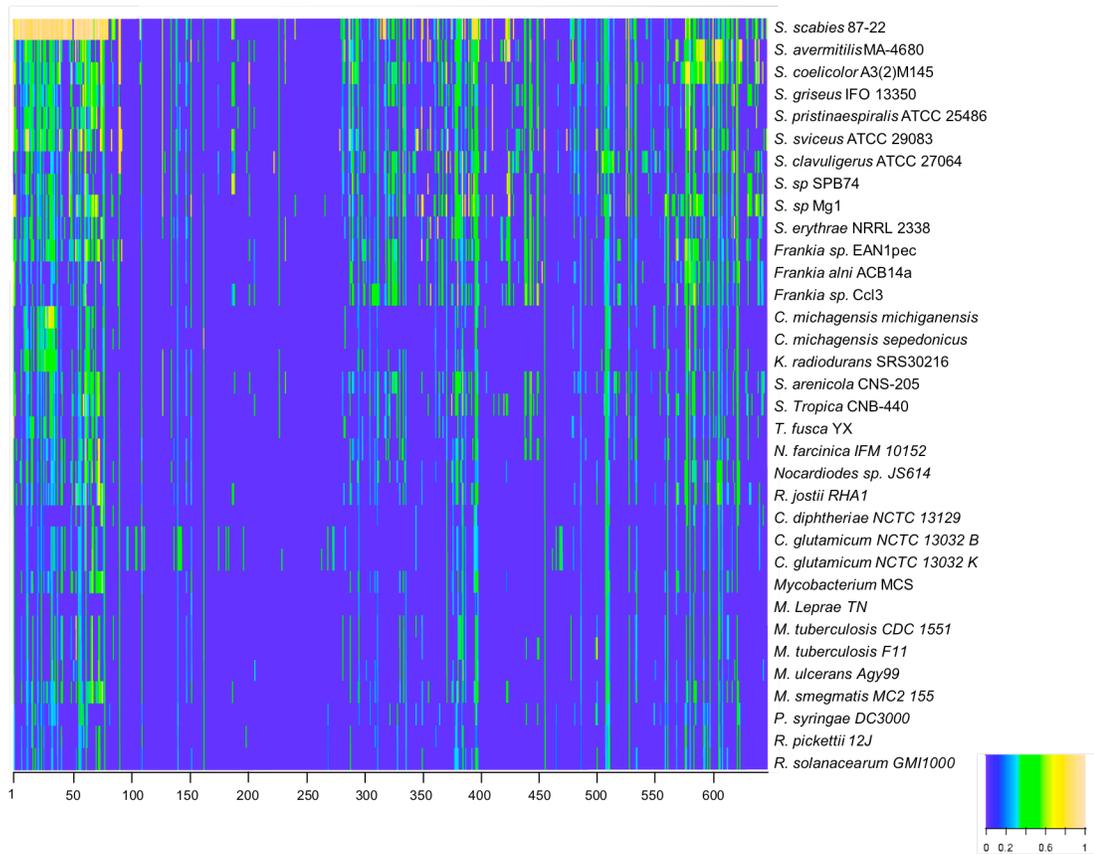


Figure 3-7. Conserved lantibiotic synthesis genes. Putative lantibiotic synthesis operon encoded in the PAISt (A). Aligned lantibiotic operon in *S. griseus* with amino acid identity (%) indicated (B)

Table 3-2. Predicted extra cellular proteins encoded within the PAISt. Prediction was performed using PSORTb. Scores are assigned from 0 to 10.

Sequence ID	Predicted product	Assigned score for PSORTb
stPAI0009c	glycosyl hydrolase	8.82
stPAI0015	Ricin motif containing protein	8.82
stPAI0027c	Tomatinsase	9.73
stPAI0163c	Unknown	8.82
stPAI0178	Unknown	8.82
stPAI0221	Unknown	8.82
stPAI0308c	Unknown	8.82
stPAI0361	Unknown	8.82
stPAI0402	Unknown	8.82
stPAI0437c	Unknown	8.82
stPAI0443	Unknown	8.82
stPAI0452	Unknown	8.82
stPAI0455c	Unknown	8.82
stPAI0558c	Unknown	8.82
stPAI0612	Unknown	8.82

Figure 3-8. Blast-fingerprinting plot of the PAISt. Coding sequences of the PAISt (horizontal line) were compared with diverse proteomes (vertical line). Percentage of identity is shown in the legend at the right bottom side.



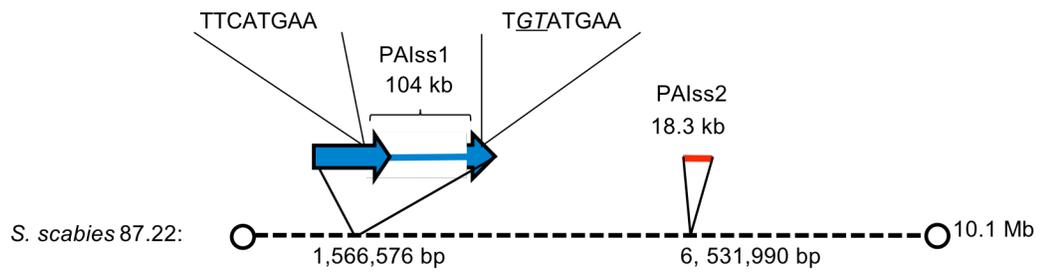


Figure 3-9. PAISs1 and PAISs2 in *S. scabies* 87-22. In *S. scabies* 87-22 the 3' end of the 104 Kb region shows a degenerate integration site. Red line indicates the location of the thaxtomin biosynthesis cluster.

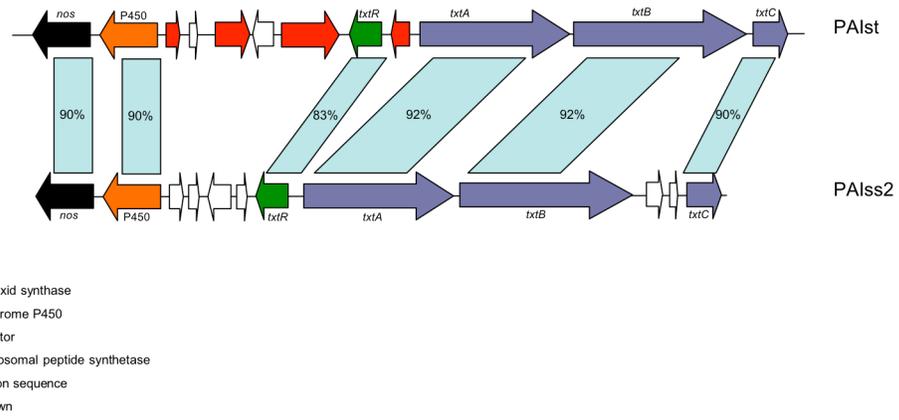


Figure 3-10. Comparison of the regions containing the thaxtomin biosynthesis operon in PAIst and PAIss2. Pale blue lines indicate homology and percentages of identity.

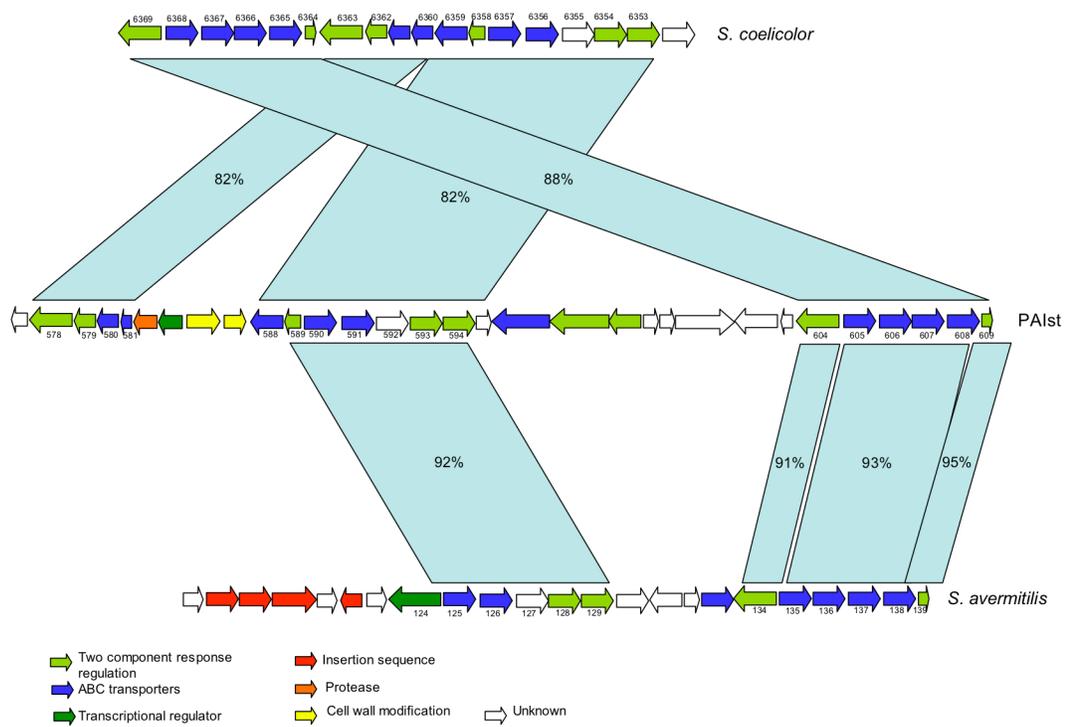


Figure 3-11. Alignment and comparison of PAIst with *S. coelicolor* and *S. avermitilis*. Pale blue lines indicate homology. Percentage of identity at amino acid level is indicated. Number below or above are gene identifiers.

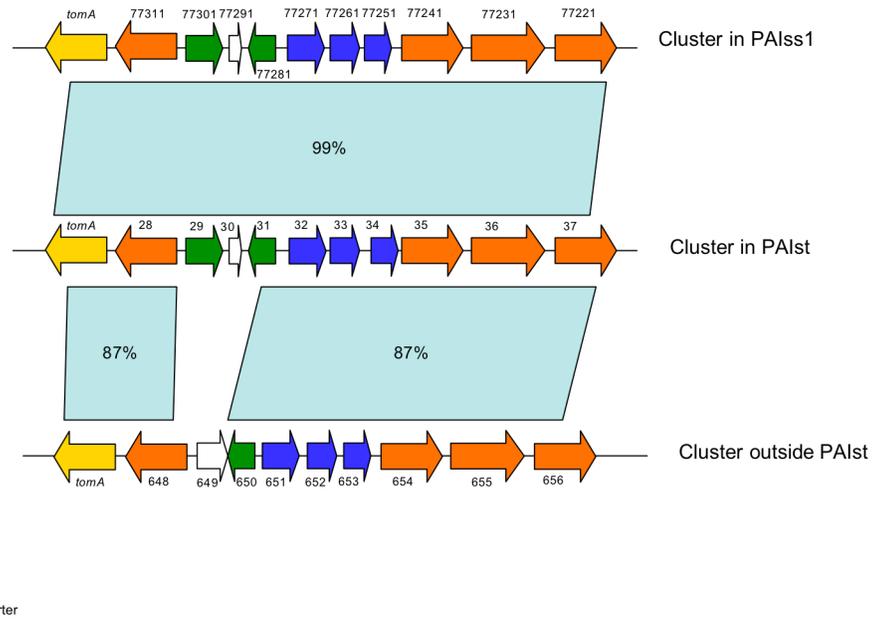


Figure 3-12. Nucleotide alignment and identity of two gene clusters containing *tomA* in *S. turgidiscabies* Car8 (Cluster in PAISt and cluster outside of PAISt) with the syntenic region in *S. scabies* 87-22 (Cluster in PAISs1).

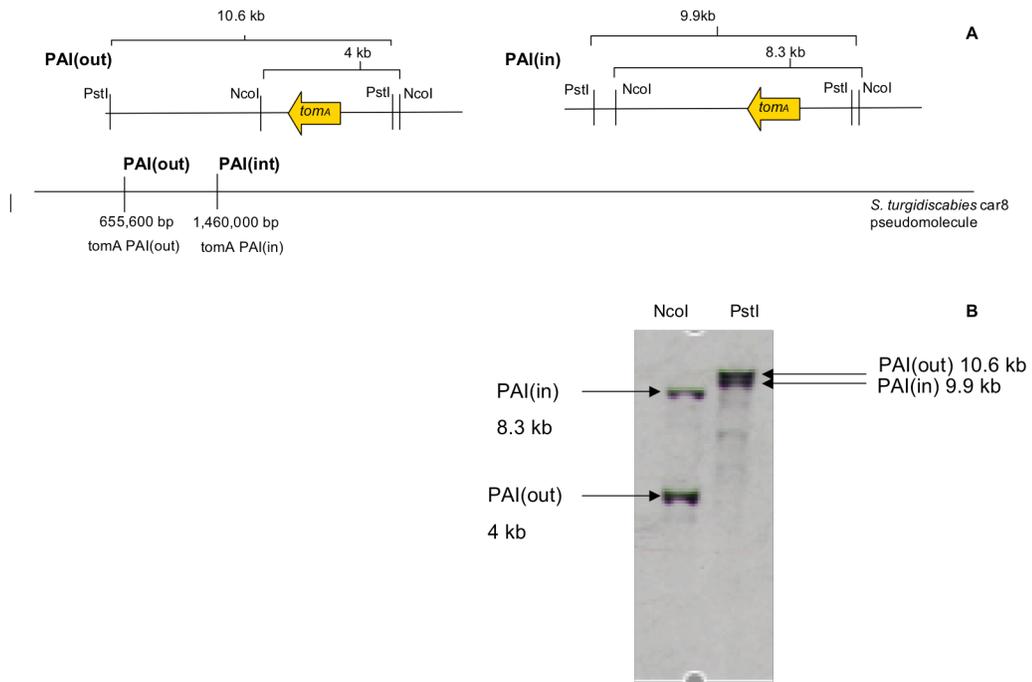


Figure 3-13. *S. turgidiscabies* Car8 possesses two copies of the gene cluster containing the *tomA* gene. Sequence map of the *tomA* gene and positions in *S. turgidiscabies* Car8 (A). Southern blot showing *S. turgidiscabies* genomic DNA digested with NcoI and PstI. PAI(out) indicates the *tomA* copy outside of the island PAISt, PAI(in) indicates the copy of *tomA* in the island PAISt (B).

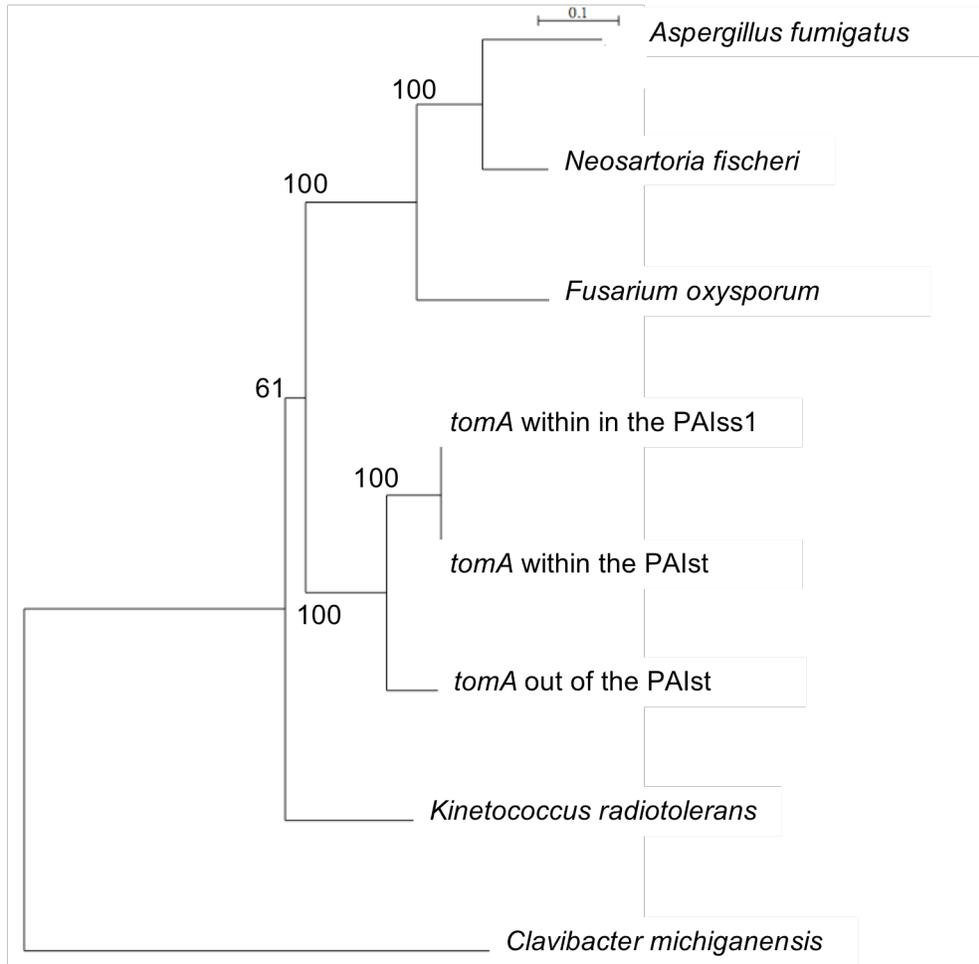


Figure 3-14. Phylogenetic tree showing the relation between the two copies of the *tomA* gene found in the *S. turgidiscabies* Car8 and *S. scabies* 87-22. The tree includes orthologs of *tomA* in other bacterial species (*Clavibacter michiganensis*, *Kinetococcus radiotolerans*) and fungi (*Fusarium oxysporum*, *Neosartoria fischeri* and *Aspergillus fumigatus*). Scale represents substitutions per site. Internal nodes show bootstrap values of 100 repetitions.

Tomatinases are virulence factors in fungi such as *Septoria lycopersici* and *Fusarium oxysporum* f. sp. *Lycopersici* (Bouarab *et al.*, 2002; Osbourn *et al.*, 1995; Roldan-Arjona *et al.*, 1999). The *tomA* gene is a conserved, secreted protein in plant pathogenic streptomycetes and is believed to be a virulence protein in *Streptomyces scabies*, though a virulence phenotype has not yet been demonstrated (Seipke and Loria, 2008). Other genes located within the duplicated region are *stPAI032*, *stPAI033*, and *stPAI034*. This genetic cluster codes for ABC transport proteins. In addition, a cluster of genes coding for putative glycosylases (*stPAI035*, *stPAI036* and *stPAI037*) is located in this duplicated region (Figure 3-12). The bioinformatics analysis was substantiated with Southern blot analysis. Two hybridization signals corresponding to the *tomA* gene were detected when total genomic DNA of *S. turgidiscabies* Car8 was digested with NcoI or PstI. The sizes of the hybridization products are 8.3 Kb and 4 Kb with NcoI and 10.6 Kb and 9.9 Kb with PstI. These results agree with the predicted digested products calculated from the *S. turgidiscabies* Car8 pseudomolecule (Figure 3-13).

Phylogenetic analysis was conducted using the nucleotide sequence of the two copies of the *tomA* gene located in *S. turgidiscabies* Car8 (one within the PAISt and the other out side of the PAISt) and the *tomA* copy in *S. scabies* 87-22. Additional homologues to *tomA* were used in the analysis as out-groups Figure (3-14). Results suggest that the *S. turgidiscabies* Car8 *tomA* copy within the PAISt is more closely related to *S. scabies* 87-22 than it is to second copy in *S. turgidiscabies* Car8, which lies outside of the PAISt (Figure 3-14). This evidence suggests that the duplicated PAISt regions in *S. turgidiscabies* Car8 have a different evolutionary histories and are not the product of a duplication event in the *S. turgidiscabies* Car8 genome.

CONCLUSION

The genomic island found in *S. turgidiscabies* Car8, PAISt, is an intriguing genetic element with a structure consistent with an ICE. PAISt is the largest ICE described to date and its complex structure and plethora of genes provide the plant pathogen *S. turgidiscabies* with virulence and fitness factors important for the process of plant infection.

Sequence comparisons reveal that the PAISt contains regions of homology with other *Streptomyces* spp. suggesting that the element has recombined and acquired novel genetic information. The specific mechanisms of recombination are still unclear. However, considering the organization of PAISt relative to other ICEs the recombinase-integrase located at the 3' end of the PAISt may function to integrate new DNA into the element. Functionality of the recombinase will be described in the next chapter.

REFERENCES

- The Gene Ontology Consortium. (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res* **34**: D322-326.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403-410.
- Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., and Eddy, S.R. (2004) The Pfam protein families database. *Nucleic Acids Res* **32**: D138-141.
- Bessman, M.J., Frick, D.N., and O'Handley, S.F. (1996) The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed, "housecleaning" enzymes. *J Biol Chem* **271**: 25059-25062.
- Biskri, L., Bouvier, M., Guerout, A.M., Boissard, S., and Mazel, D. (2005) Comparative study of class 1 integron and *Vibrio cholerae* superintegron integrase activities. *J Bacteriol* **187**: 1740-1750.
- Bouarab, K., Melton, R., Peart, J., Baulcombe, D., and Osbourn, A. (2002) A saponin-detoxifying enzyme mediates suppression of plant defences. *Nature* **418**: 889-892.
- Bukhalid, R.A., Chung, S.Y., and Loria, R. (1998) nec1, a gene conferring a necrogenic phenotype, is conserved in plant-pathogenic *Streptomyces* spp. and linked to a transposase pseudogene. *Mol Plant Microbe Interact* **11**: 960-967.
- Burrus, V., and Waldor, M.K. (2004) Shaping bacterial genomes with integrative and conjugative elements. *Res Microbiol* **155**: 376-386.

- Burrus, V., Quezada-Calvillo, R., Marrero, J., and Waldor, M.K. (2006) SXT-related integrating conjugative element in New World *Vibrio cholerae*. *Appl Environ Microbiol* **72**: 3054-3057.
- Cao, J., Woodhall, M.R., Alvarez, J., Cartron, M.L., and Andrews, S.C. (2007) EfeUOB (YcdNOB) is a tripartite, acid-induced and CpxAR-regulated, low-pH Fe²⁺ transporter that is cryptic in *Escherichia coli* K-12 but functional in *E. coli* O157:H7. *Mol Microbiol* **65**: 857-875.
- Carver, T., Berriman, M., Tivey, A., Patel, C., Bohme, U., Barrell, B.G., Parkhill, J., and Rajandream, M.A. (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**: 2672-2676.
- Clewell, D.B., Flanagan, S.E., Ike, Y., Jones, J.M., and Gawron-Burke, C. (1988) Sequence analysis of termini of conjugative transposon Tn916. *J Bacteriol* **170**: 3046-3052.
- Conesa, A., and Gotz, S. (2008) Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. *Int J Plant Genomics* **2008**: 619832.
- Delcher, A.L., Salzberg, S.L., and Phillippy, A.M. (2003) Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics* **Chapter 10**: Unit 10 13.
- Delcher, A.L., Bratke, K.A., Powers, E.C., and Salzberg, S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**: 673-679.
- Errington, J. (2001) Septation and chromosome segregation during sporulation in *Bacillus subtilis*. *Curr Opin Microbiol* **4**: 660-666.

- Fonseca, E.L., Dos Santos Freitas, F., Vieira, V.V., and Vicente, A.C. (2008) New qnr gene cassettes associated with superintegron repeats in *Vibrio cholerae* O1. *Emerg Infect Dis* **14**: 1129-1131.
- Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M., and Brinkman, F.S. (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **21**: 617-623.
- Gomis-Ruth, F.X., Moncalian, G., Perez-Luque, R., Gonzalez, A., Cabezon, E., de la Cruz, F., and Coll, M. (2001) The bacterial conjugation protein TrwB resembles ring helicases and F1-ATPase. *Nature* **409**: 637-641.
- Grohmann, E., Muth, G., and Espinosa, M. (2003) Conjugative plasmid transfer in gram-positive bacteria. *Microbiol Mol Biol Rev* **67**: 277-301, table of contents.
- Joshi, M.V., and Loria, R. (2007) *Streptomyces turgidiscabies* possesses a functional cytokinin biosynthetic pathway and produces leafy galls. *Mol Plant Microbe Interact* **20**: 751-758.
- Kaneko, T., Nakamura, Y., Sato, S., Asamizu, E., Kato, T., Sasamoto, S., Watanabe, A., Idesawa, K., Ishikawa, A., Kawashima, K., Kimura, T., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Mochizuki, Y., Nakayama, S., Nakazaki, N., Shimpo, S., Sugimoto, M., Takeuchi, C., Yamada, M., and Tabata, S. (2000) Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. *DNA Res* **7**: 331-338.
- Kers, J.A., Cameron, K.D., Joshi, M.V., Bukhalid, R.A., Morello, J.E., Wach, M.J., Gibson, D.M., and Loria, R. (2005) A large, mobile pathogenicity island confers plant pathogenicity on *Streptomyces* species. *Mol Microbiol* **55**: 1025-1033.

- Kloosterman, H., Vrijbloed, J.W., and Dijkhuizen, L. (2002) Molecular, biochemical, and functional characterization of a Nudix hydrolase protein that stimulates the activity of a nicotinoprotein alcohol dehydrogenase. *J Biol Chem* **277**: 34785-34792.
- Lonial, S., Williams, L., Carrum, G., Ostrowski, M., and McCarthy, P., Jr. (1997) *Neosartorya fischeri*: an invasive fungal pathogen in an allogeneic bone marrow transplant patient. *Bone Marrow Transplant* **19**: 753-755.
- Loria, R., Kers, J., and Joshi, M. (2006) Evolution of plant pathogenicity in *Streptomyces*. *Annu Rev Phytopathol* **44**: 469-487.
- Loria, R., Bignell, D.R., Moll, S., Huguet-Tapia, J.C., Joshi, M.V., Johnson, E.G., Seipke, R.F., and Gibson, D.M. (2008) Thaxtomin biosynthesis: the path to plant pathogenicity in the genus *Streptomyces*. *Antonie Van Leeuwenhoek* **94**: 3-10.
- Mavrodi, D.V., Loper, J.E., Paulsen, I.T., and Thomashow, L.S. (2009) Mobile genetic elements in the genome of the beneficial rhizobacterium *Pseudomonas fluorescens* Pf-5. *BMC Microbiol* **9**: 8.
- Miyajima, K., Tanaka, F., Takeuchi, T., and Kuninaga, S. (1998) *Streptomyces turgidiscabies* sp. nov. *Int J Syst Bacteriol* **48 Pt 2**: 495-502.
- Nes, I.F., and Tagg, J.R. (1996) Novel lantibiotics and their pre-peptides. *Antonie Van Leeuwenhoek* **69**: 89-97.
- Osborn, A., Bowyer, P., Lunness, P., Clarke, B., and Daniels, M. (1995) Fungal pathogens of oat roots and tomato leaves employ closely related enzymes to detoxify different host plant saponins. *Mol Plant Microbe Interact* **8**: 971-978.
- Pembroke, J.T., and Piterina, A.V. (2006) A novel ICE in the genome of *Shewanella putrefaciens* W3-18-1: comparison with the SXT/R391 ICE-like elements. *FEMS Microbiol Lett* **264**: 80-88.

- Possoz, C., Gagnat, J., Sezonov, G., Guerineau, M., and Pernodet, J.L. (2003) Conjugal immunity of *Streptomyces* strains carrying the integrative element pSAM2 is due to the pif gene (pSAM2 immunity factor). *Mol Microbiol* **47**: 1385-1393.
- Roldan-Arjona, T., Perez-Espinosa, A., and Ruiz-Rubio, M. (1999) Tomatinase from *Fusarium oxysporum f. sp. lycopersici* defines a new class of saponinases. *Mol Plant Microbe Interact* **12**: 852-861.
- Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16**: 944-945.
- Saitou, N., and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406-425.
- Seipke, R.F., and Loria, R. (2008) *Streptomyces scabies* 87-22 possesses a functional tomatinase. *J Bacteriol* **190**: 7684-7692.
- Tato, I., Zunzunegui, S., de la Cruz, F., and Cabezon, E. (2005) TrwB, the coupling protein involved in DNA transport during bacterial conjugation, is a DNA-dependent ATPase. *Proc Natl Acad Sci U S A* **102**: 8156-8161.
- Tato, I., Matilla, I., Arechaga, I., Zunzunegui, S., de la Cruz, F., and Cabezon, E. (2007) The ATPase activity of the DNA transporter TrwB is modulated by protein TrwA: implications for a common assembly mechanism of DNA translocating motors. *J Biol Chem* **282**: 25569-25576.
- te Poele, E.M., Bolhuis, H., and Dijkhuizen, L. (2008a) Actinomycete integrative and conjugative elements. *Antonie Van Leeuwenhoek* **94**: 127-143.
- te Poele, E.M., Samborsky, M., Oliynyk, M., Leadlay, P.F., Bolhuis, H., and Dijkhuizen, L. (2008b) Actinomycete integrative and conjugative pMEA-like

elements of Amycolatopsis and Saccharopolyspora decoded. *Plasmid* **59**: 202-216.

Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2002) Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* **Chapter 2**: Unit 2.3.

Willey, J.M., and van der Donk, W.A. (2007) Lantibiotics: peptides of diverse structure and function. *Annu Rev Microbiol* **61**: 477-501.

CHAPTER 4
THE MOBILE PATHOGENICITY ISLAND PAIS_t IN *S. turgidiscabies*
POSSESSES A NOVEL TYROSINE RECOMBINASE

ABSTRACT

PAIS_t is an Integrative Conjugative Element (ICE) located within the chromosome of the plant pathogen *Streptomyces turgidiscabies*. This mobile element carries clustered virulence genes and forms a pathogenicity island. This PAIS_t is mobilized to other *Streptomyces* spp., integrating by site-specific recombination at the 3' end of the bacitracin resistance gene (*bacA*). We have identified a gene located at the 3' end of the PAIS_t that encodes a tyrosine type recombinase (*IntPAI*). Experimental analysis demonstrates that IntPAI is able to integrate DNA vectors into the PAIS_t recombination site in the *S. coelicolor* and *S. lividans* chromosomes recognizing the integration site of the element at the 3' end of *bacA*. Our results indicate that IntPAI is the recombinase responsible for the integration of the PAIS_t in the chromosome of these *Streptomyces* species. Bioinformatics analysis suggests that the recombinase described here represents a novel tyrosine recombinase. Characterization of IntPAI allowed us to build and propose a profile hidden Markov model that describes intPAI and relate tyrosine recombinases.

INTRODUCTION

Tyrosine recombinases recognize specific sequences of DNA and catalyze rearrangements within them (Gopaul and Duyne, 1999; Grainge and Jayaram, 1999; Sauer, 1994). This site-specific recombination process is coordinated by four recombinase-monomers bound to the DNA substrates (Figure 4-1A). Strand exchange of DNA substrates is carried out at the catalytic core of the recombinase, which consists of two-conserved arginine residues and a tyrosine residue that executes a nucleophilic attack of the scissile phosphate in the substrate DNA, promoting the strand exchange.

Site-specific recombination is involved in a variety of biological functions such as resolution of dimer chromosomes during replication (Blakely *et al.*, 1991), resolution of concatenated plasmids (Cornet *et al.*, 1994), and the integration of phages and Integrative and Conjugative Elements (ICEs) into host chromosomes (Groth and Calos, 2004). Consistent with this diversity of functions, tyrosine recombinases constitute a large super family of proteins with conserved and variable regions in their amino acid sequences (Esposito and Scocca, 1997). Despite their complex architecture and diversity, tyrosine recombinases display three major non-continuous regions of similarity, designated Box A, B and C. Each box contains one of the highly conserved amino acid residues (RRY) that comprise the catalytic core of the enzyme. This amino acid triad is the molecular signature for the tyrosine recombinase super family (Esposito and Scocca, 1997).

ICEs have been described as mobile DNA elements that contain features of phages, while also possessing mechanisms of conjugation. Many ICEs carry tyrosine recombinases that allow them to integrate into the genome of host bacteria (Burrus *et al.*, 2006). The tyrosine recombinase recognizes specific sequences within the mobile element (*attP*) and within the genome of the host (*attB*). The recombinase carries out

the DNA strand exchange between these two sequences resulting in integration of the ICE into the host genome.

When integrated in the chromosome, the two *att*-sites flank the ICE. The flanking *att* sites are products of the original recombination event and function as substrates for a future recombination event leading to liberation of the ICE from the chromosome as a circular structure. Because ICEs lack the ability to replicate independently, the circular structure is transient. However; ICEs have DNA transfer machinery that can mobilize the element to another host by conjugation. Provided that the genome of the new host contains a copy of the *att* site, the recombinase can integrate the ICE into the new host genome (Figure 4-1B).

Streptomyces turgidiscabies is a pathogen that infects many plant species and causes the economically important disease potato scab (Joshi and Loria, 2007; Kers *et al.*, 2005; Miyajima *et al.*, 1998). The key virulence factors of *S. turgidiscabies* are the phytotoxin thaxtomin (Loria *et al.*, 2008), the secreted necrogenic protein Nec1 (Bukhalid *et al.*, 1998), the cytokinin-like hormone produced by the *fas* operon (Joshi and Loria, 2007) and the secreted protein tomatinase (TomA) (Kers *et al.*, 2005) (Figure 4-2). These virulence factors have been characterized experimentally and are located in a large pathogenicity island (PAISt) (Kers *et al.*, 2005). Previous sequence and experimental analysis of the PAISt demonstrated that the element is an ICE that can mobilize from *S. turgidiscabies* and integrate into the chromosome of *S. coelicolor*, *S. diastochromogenes* and *S. lividans* (Kers *et al.*, 2005). The integration of PAISt occurs at the eight-base palindromic sequence TTCATGAA located at the 3' end of the bacitracin resistance gene (*bacA*) (Kers *et al.*, 2005). Streptomycete transconjugants can contain the complete sequence of the PAISt, 674 Kb, or a truncated version of 106 Kb. This suggests that the mobile element can transfer and insert in a modular fashion, using an additional internal *att* site in the element. Here

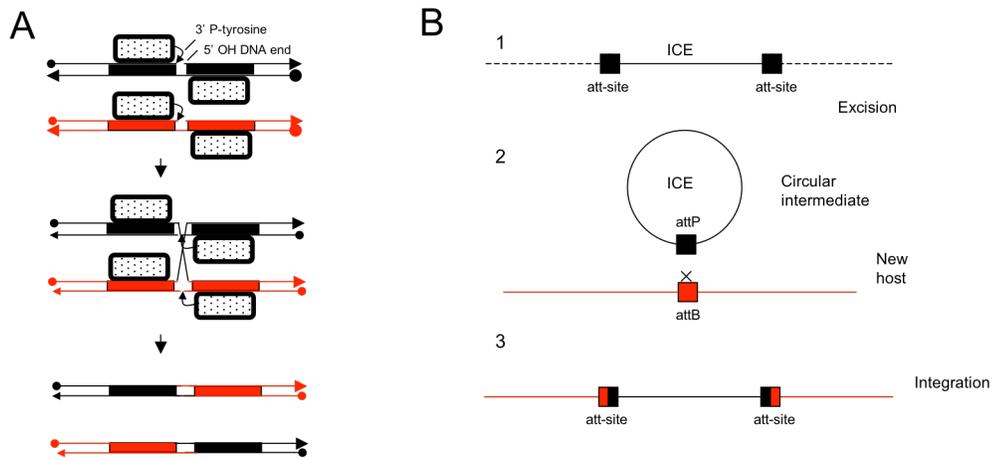


Figure 4-1. General model for tyrosine-type recombination (A). Four monomers of the tyrosine recombinase (rectangles), bind to the sites of recombination (in red and black). The tyrosine residue (arrows) acts and cleaves the phosphate of each DNA strand promoting the recombination of *att* sites. General mechanism of ICEs recombination (B). The element is integrated in the chromosome of the host (1) and using its recombinase could be released and make a circular structure (2) that could be mobilized by conjugation to other host that eventually will be a target of recombination if it contains the *attB* sites in its genome (3).

we analyze the sequence content of the PAISt in order to identify the putative recombinase that excises and integrates the element in the donor and recipient, respectively. We identified a gene coding for a novel tyrosine recombinase, IntPAI, which is located at the 3' end of the ICE (Figure 4-2). Experimental analysis confirmed that IntPAI functions as a recombinase that recognizes the *att* site described within the PAISt.

MATERIALS AND METHODS

Bacterial strains and clustering conditions

Escherichia coli strains were cultured in LB media at 37° C. *Streptomyces* strains were cultured at 28° C using International *Streptomyces* Project 4 (ISP4) agar media, manitol-soya flour (MS) and tryptic soy (TSB) broth. Antibiotic concentrations used in this study are as follows: for *E. coli* chloramphenicol at 25 µg/ml, kanamycin, at 50µg/ml, hygromycin at 100 µg/ml, apramycin at 100 µg/ml. For *Streptomyces* strains, chloramphenicol at 25 µg/ml, kanamycin and nalidixic acid at 25 µg/ml.

Plasmid construction

Plasmid pAmp001 is a derivative of pIJ10257 (Gregory *et al.*, 2003) (Figure 4-3), in which the phage integrase ϕ BT1 and its integration site were replaced by the ampicillin resistance gene (*amp*) using the NcoI and EcoRV restriction sites. Plasmid pIntPAI was derived from pAmp001. To construct pIntPAI, the coding sequence of *intPAI* and the thirty-seven bp downstream region were amplified from *S. turgidiscabies* Car8 total DNA by PCR using primer JH69 (5'-GGTCGACATATGCCCTACATCGAGTGGC-3') and JH70 (5'-TAGTAAAGCTTCGGCATGAACGACTTGGTTCG3-'). The downstream region of

the *intPAI* contains the sequence TTCATGAA, which is the site of recombination of the PAISt (Kers *et al.*, 2005). Primers JH69 and JH70 contain the NdeI and HindIII restriction sites. Consequently the PCR product was cloned in the NdeI-HindIII site in plasmid pAmp001 and positioned under control of the promoter ermEp* (Gregory *et al.*, 2003) (Figure 4-3).

Mating assays

E. coli ET1054 was transformed with plasmids by electroporation and selected on LB agar plates containing kanamycin, chloramphenicol and hygromycin. Transformed *E. coli* isolates were grown until OD=0.4 and 10 ml of the culture was mixed and incubated with 0.5 ml of *Streptomyces* spores, concentrated at 10^8 / ml. Incubation was carried out in MS agar supplement with 10 mM MgCl₂. After 24 hours of incubation 1 ml of dH₂O with nalidixic acid and hygromycin was added to select for transconjugants.

Detection of integrated plasmids in *S. coelicolor*

PCR and southern blot analysis were used to confirm the integration of the pIntPAI into the chromosome. Genomic DNA was isolated using the Gram-positive-MasterPure Kit (Epicenter) according to the manufacturer's instructions. Primers JH69 (5'-GGTCGACATATGCCCTACATCGAGTGGC-3') and JH125 (5'-AGTGAAGCTTCTATGACCACTTAGCCTTG-3') amplify the total CDS of *intPAI* (predicted product 1401 bp) and were used to confirm plasmid integration. For Southern blot hybridization, 4 µg of genomic DNA was digested with KpnI and fragments were separated by electrophoresis in a 1% agarose gel. The DNA was depurinated with 0.1 M HCl, denatured with 0.5 NaOH and 1.5 M NaCl and neutralized with 1.5 M NaCl and 0.5M Tris HCl. DNA samples were transferred

overnight to a nylon membrane (Whatman) by capillary action, using 20X of Sodium Chloride-Sodium Phosphate-EDTA Buffer (SCC). DNA was UV-cross linked to the membrane by applying 120,000 $\mu\text{joules}/\text{cm}^2/\text{sec}$ for 4 minutes. The membrane was probed with the PCR product obtained with primers JH69 and JH125. The PCR product was labeled with dioxigenin-11-UTP (Roche). Hybridization was performed in a rotary hybrid oven at 58° C. Stripping conditions were carried out with 0.5x SCC solution at 62°C for 20 minutes. Additional steps for hybridization and detection were performed according to the manufacturer's instructions (Roche).

Sequencing of recombination sites in *S. coelicolor*

Two pairs of primers JH139 (5'-GACCGACGCTCTTCGCCAC-3') with JH140 (5'-CGGAGGCCAAACGGCATTGAG-3') and JH134 (5'-CACAGGAGCGAGCGGGCAG -3') with JH106 (5'-CGGCCGTGGAGCGCTGGAAG-3') were used to amplify by PCR the DNA structures of recombination formed by the integration process in *S. coelicolor*. Primers JH139 and JH148 amplify the region that contains the 5' end of the integrated pIntPAI plasmid. Primers JH134 and JH106 amplify the 3' end of the integrated pIntPAI. Multiple integration events were detected and confirmed using primers JH107 (5'GGGTGTCGAGGGGTTGTACTC3-') and JH126 (5'CCGTGCCAATCGGATCAGC3').

Sequence analysis for the intPAI

Six functionally characterized integrase sequences (Figure 4-4) were used to compare and align IntPAI in order to identify conserved domains. Additionally, the IntPAI amino acid sequence was used as a query to retrieve one hundred annotated recombinases from the Non-redundant database in the Genbank, using Position

Specific Iterated BLAST (PSI-BLAST) (Altschul *et al.*, 1997). Sequences were aligned using the Multiple sequence comparison by log-expectation (MUSCLE) algorithm (Edgar, 2004). A Hidden Markov model (HMM) (Krogh *et al.*, 1994) was constructed from the alignment of recombinases using Hmmer software 2.3.2. Visualization of the HMM was carried out using logoMat-M (Schuster-Bockler and Bateman, 2005). Structural modeling was carried out using the EsyPred3D modeling server (Lambert *et al.*, 2002). Visualization of structures was conducted using Swiss-PdbViewer protein modeling software (Guex and Peitsch, 1997; Guex *et al.*, 2009).

RESULTS AND DISCUSSION

The PAIS_t contains a gene that encodes a tyrosine recombinase at its 3' end.

Sequence analysis of the PAIS_t revealed the presence of a predicted gene, *intPAI*, located at the 3' end of the element. The *att* site, an eight-bp-sequence (TTCATGAA) that allows the recombination of the element is located at the 3' end of *intPAI* (Figure 4-2). The predicted gene *intPAI* encodes a 466 amino acid protein with a predicted molecular weight of 54 kDa. Alignment of the predicted amino acid sequence from the IntPAI and experimentally characterized tyrosine recombinase suggests that IntPAI is a tyrosine recombinase (Figure 4-4). Amino acid alignment demonstrates that discrete regions (boxes), typical of tyrosine recombinases are shared by IntPAI. In IntPAI the predicted conserved boxes are located at amino acid positions 197 to 236 (Box A), 372 to 386 (Box B) and 392 to 409 (Box C). In addition, the amino acid catalytic triad RRY is predicted to be located in the respective box regions at positions 209(R), 375(R) and 407(Y). Structural predictions and comparisons of IntPAI were carried out using a *Vibrio cholerae* integrase (IntI4) (Genbank accession ZP_01682564, protein structure code 3a2Vb). The 3D model of intPAI displays the RRY triad in a spatially associated cluster (Figure 4-5). The results obtained with the

alignment and the 3D analyses suggest that the amino acid triad RRY found in intPAI forms the catalytic core of the recombinase.

Functionality of IntPAI in *Streptomyces* spp.

Previous experiments have shown that several *Streptomyces*, including *S. lividans* and *S. coelicolor* are suitable hosts for the mobile PAISt (Kers *et al.*, 2005). Based on this rationale, the suicide plasmid pIntPAI was conjugated into both *S. lividans* and *S. coelicolor*. Transconjugants were obtained at frequencies of 3.6×10^{-6} for *S. coelicolor* and 2×10^{-6} for *S. lividans*. Integration events were confirmed by PCR using primers JH69 and JH125 that amplify the entire *intPAI* gene (1401 bp), which is integrated in the transconjugants chromosome (Figure 4-6). In all the cases, the transconjugants remained resistant to hygromycin (the selection marker of the plasmid) after five continuous passages in ISP4 media without antibiotic selection. This suggests that the integration of pIntPAI in the host chromosomes is stable.

Vector pIntPAI integrates at the 3' end of the *bacA* gene in *S. coelicolor* and forms multimeric structures

Previous report indicates that the PAISt integrates in *S. coelicolor* chromosome at the bacitracin resistance gene (*bacA*) recognizing the specific palindrome sequence TTCATGAA (*att* site). We continue the analysis of integration in *S. coelicolor* taking advantage of the complete genome availability of this streptomycete (Bentley *et al.*, 2002). The vector pIntPAI is a mini-PAISt with the recombination functions, consequently IntPAI should recognize the *att* site at the *bacA* gene if this recombinase is responsible for the integration of the PAISt. Moreover, our rationale indicates that single integration of pIntPAI at the *att*-site in the *bacA* gene should produce a structure

that can be recognized by DNA digestion with KpnI and Southern blot analysis as a hybridization signal of 1.5 kb (Figure 4-7A and 4-7B).

Our prediction was confirmed when we detected the hybridization of 1.5 Kb fragment (Figure 4-7A). An additional signal of 6.2 Kb was observed in the Southern blot. The second signal represents a second integration event of pIntPAI at one of the new *att*-sites that flank the previously integrated pIntPAI. (Figure 4-7A and 4-7B). It is important to mention that besides the *att*-site located in the *bacA*, *S. coelicolor* chromosome contains 58 additional palindrome sequences identical to the *att*-site. However, the observed pattern of hybridization represents only the integration at the *bacA* gene and it was observed in 25 independent transconjugants that we analyzed.

Moreover, amplification by PCR and sequencing confirmed the products of integration observed in the Southern blot. Three DNA structures containing the *att* sites are illustrated in figures 4-7B, 4-7C and 4-7D. Two DNA structures represent the “scar” of the first recombination event between the *att*-sites located in pIntPAI and the *S. coelicolor* chromosome. As a consequence of this event, a linear *att*-flanked version of pIntPAI is integrated in the chromosome. The third DNA structure is the result of a second recombination event in which a copy of pIntPAI integrates at one of the flanking *att* sites. Our data suggest intPAI recognizes the specific *att* site located within the *bacA* gene and possible integration event in pseudo-*att* sites in the chromosome could cause deleterious effect in the cell.

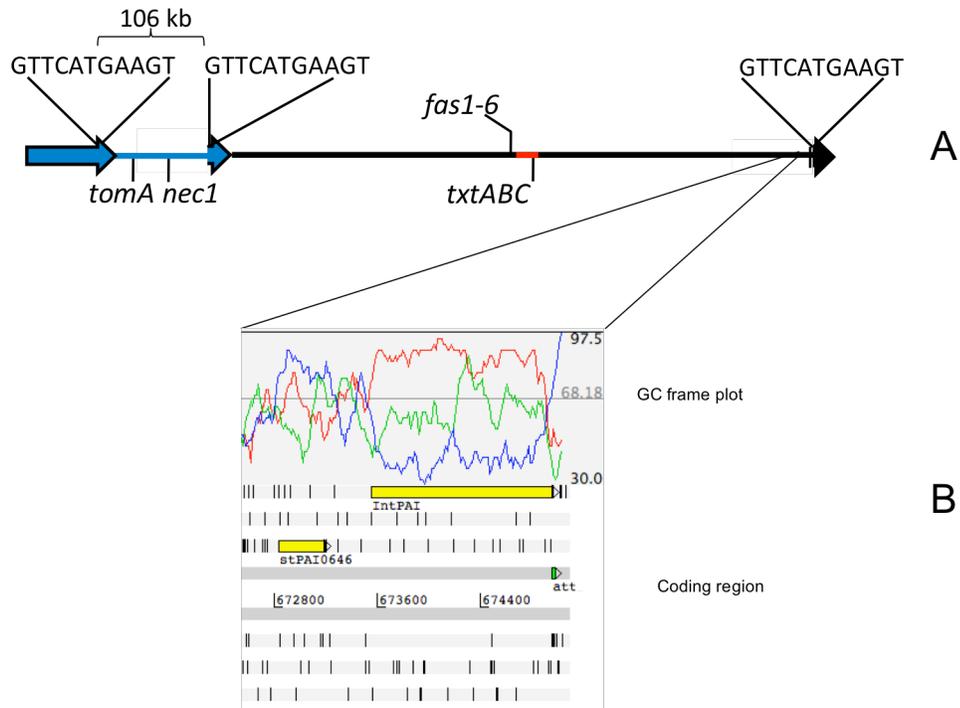


Figure 4-2. The *S. turgidiscabies* pathogenicity island, PAISt. Organization of the PAISt (674,225 bp) A). Regions within PAISt containing the virulence factors, *tomA*, *nec1*, *fas1-6*, *txtABC* and the *att* recombination site (TTCATGAA) (A) Localization of *IntPAI* at the 3' end of the PAISt (B). The GC frame plot shows the GC content of the first (green), second (blue) and third (red) position in all codons simultaneously. The GC plot indicates the presence of a putative coding region consistent with the location of the *IntPAI*. Each line represents the reading frames. Stop codons are depicted as black bars.

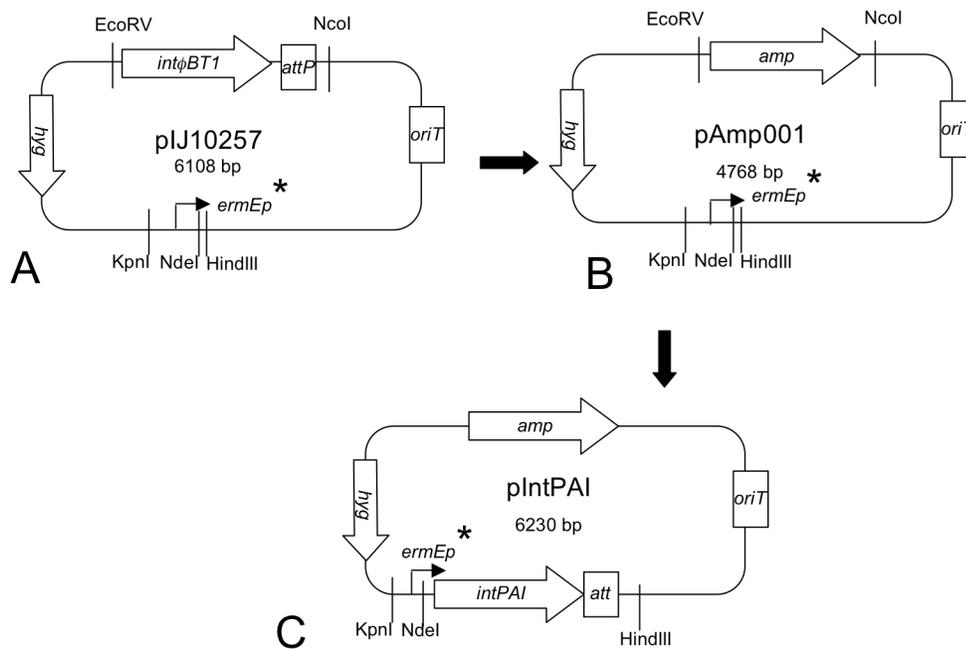


Figure 4-3. Maps of the plasmids used for functional analysis of IntPAI. Vector pIJ10257 (A). Vector pAmp001 (B). Vector pIntPAI (C) contains the tyrosine recombinase *intPAI* and the *att* site. The gene is under control of the promoter *ermEp** (arrow). *OriT* (origin of transfer). *Amp* (Ampicillin gene resistance). *IntφBT1* (Recombinase gene of phage φBT1). *hyg* (hygromycin resistance gene). Restriction sites are indicated in each plasmid.

Our results suggest as well that an integrated pIntPAI can provide new *att* sites for recombination of additional plasmids pIntPAI (Figure 4-7B). The sequential integration events observed in *S. coelicolor* could result in a large and modular genomic island in the genome of the host, not unlike the genomic structure of PAIS_t in *S. turgidiscabies* described in chapter III.

IntPAI is a novel tyrosine-type recombinase

We searched for conserved motifs in the IntPAI using the Pfam (Bateman *et al.*, 2004) and COG (Tatusov *et al.*, 2000) databases. Our results indicate that IntPAI lacks an annotated protein motif. We conducted a PSI-BLAST search (Position Specific Iteration- BLAST), which is a more sensitive BLAST strategy. After three iterative searching in the Genbank non-redundant database searches we obtained a list of recombinases with reliable scores and e-values (best e-value= $1e^{-88}$, score =360). Lack of conserved motifs in IntPAI and the need to use a more sensitive BLAST search strategy suggest that IntPAI represents an undefined group of the tyrosine recombinases. In order to expand the definition of this protein family we propose a new profile protein model. The model describes the features of IntPAI and related recombinases retrieved in our BLAST search strategy.

The protein model contains 430 positions and displays the three conserved tyrosine recombinases boxes. Box A of the model is located between positions 224-248 (Figure 4-8). The region contains two highly conserved amino acid residues. The first residue is the arginine (position 234) that forms the predicted catalytic core of the recombinase. The second conserved residue in Box A is a glutamic acid residue (position 237). The remainder of the amino acid residues that comprise Box A are aliphatic and may form a structural alpha helix.

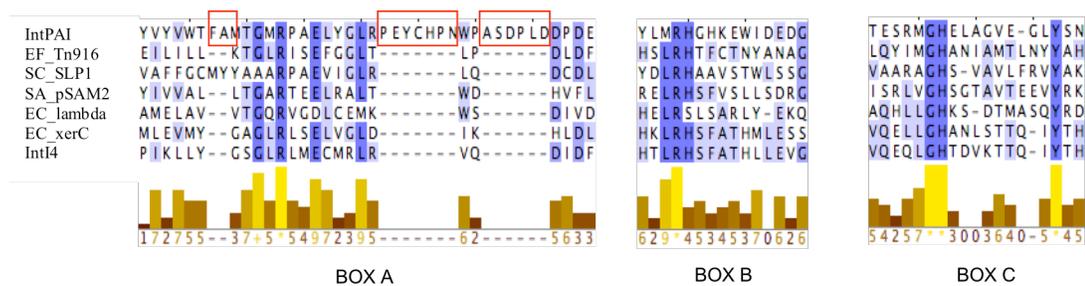


Figure 4-4. Alignment of IntPAI with characterized tyrosine recombinases. The three conserved boxes are indicated. The histogram indicates the level of conservation at each amino acid position. Red rectangles within IntPAI sequence indicate the insertion that make it unique in comparison with others in the alignment. intPAI: recombinase of the PAISt, EF_Tn916, recombinase of the integrative and conjugative element Tn916 found in *Enterococcus faecalis* (YP_133692); SC_SLP1, recombinase from *Streptomyces coelicolor* integrative plasmid (NP_628777). SA_pSAM2, recombinase from Integrative plasmid found in *S. ambofaciens* (CAA33029); EC_lambda: recombinase from phage lambda (NP_040609); EC_xerc, recombinase XerC from *E. coli* involved in chromosome dimer resolution (NP_418256); and IntI4, *Vibrio cholerae* recombinase (ZP_01682564).

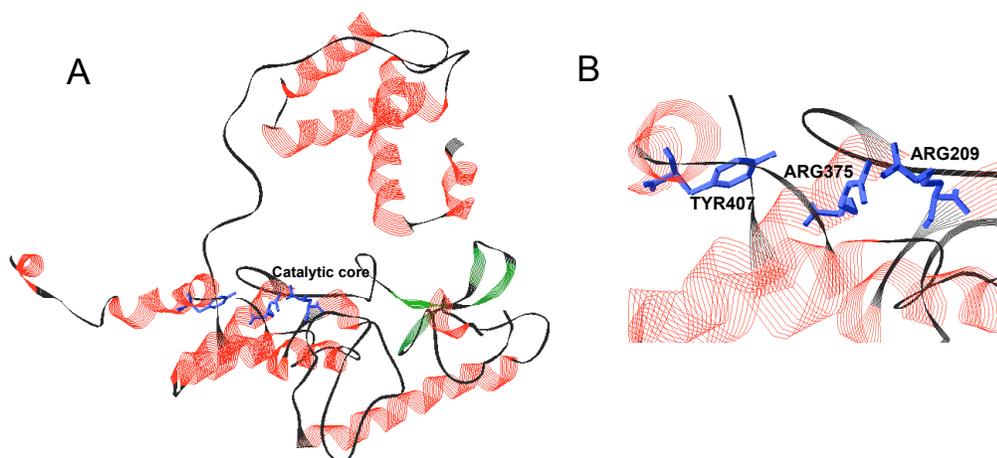


Figure 4-5. Structure analysis of intPAI Predicted 3D structure of IntPAI. The structure was predicted using *Vibrio cholerae* integrase structure (PDB code: 3a2V) (A). In red predicted alpha helices. In green predicted beta sheets. Predicted position of the catalytic amino acid triad in the IntPAI recombinases (B).

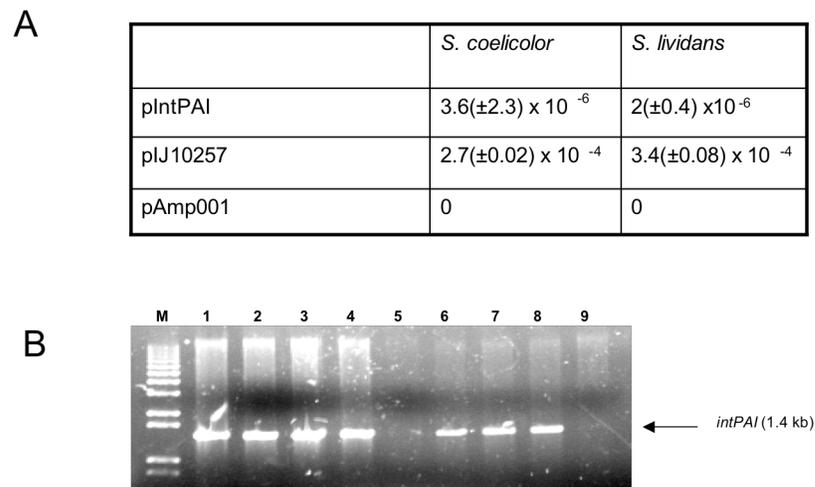
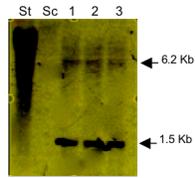


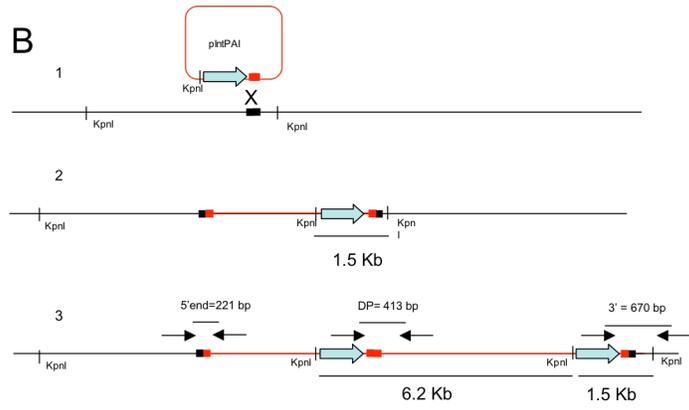
Figure 4-6. Functional analysis of IntPAI. Frequencies of transconjugants obtained in *S. coelicolor* and *S. lividans* (A). PIJ10257 was used as control of successful conjugation while pAmp001 was used as control of illegitimate recombination. Results are the average of three independent conjugation assays. Detection of integration of pIntPAI in *S. coelicolor* and *S. lividans* transconjugants (B). Primers JH69 and JH125 amplify the *IntPAI* (1.4 Kb). (Lane 1), Wild type *S. turgidiscabies*, (Lane 2-4) *S. coelicolor* transconjugants. (Lane 5) *S. coelicolor* wild type. (Lane 6-8) *S. lividans* transconjugants. (Lane 9) *S. lividans* wild type.

Figure 4-7. Detection of mutimeric integration of pIntPAI in *S. coelicolor*. Southern blot showing three *S. coelicolor* transconjugants with the integration structures (A). St indicates genomic DNA of *S. turgidiscabies*, which contains a copy of *intPAI*, and SC indicates genomic DNA of *S. coelicolor* wild type. The blot was probed with a PCR product of *intPAI*. Two hybridization signals are observed (1.5 kb and 6.2 kb) that suggest a multiple recombination event. Suggested recombination events that resulted in mutimeric structures in *S. coelicolor* (B). Integration of plasmid pIntPAI in *S. coelicolor* chromosome at the *att* site (1). Formation of single integration event and duplication of the *att* sites at the end of the integrated pIntPAI (2). Second integration event of an additional pIntPAI and formation of mutimeric structures (3). PCR amplification of the three structures formed during the multiple integration of pIntPAI (C). 5' end is the junction formed at the 5' end of the integration, 3' end is the junction formed at the 3' end of the integration. The 5' end structure was identified as a PCR product of 221 bp using primers JH139 and JH140. The 3' end structure was identified as a PCR product of 670 bp using primers JH134 and JH106. DB is the double integration event. This structure was identified as a PCR product of 413 bp, using primers JH107 and JH126. C- is DNA of a wild type *S. coelicolor* strain (negative control of the PCR reaction). Sequences of the PCR products showing the presence of the 8 bp site of recombination (D).

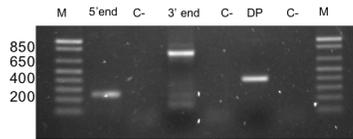
A



B



C



D

5' end
3' end
Double

```

10 20 30 40 50
CACCGGTACGTGGTCATCGCCTGCATGAAATTTCATCTCGACCAAGTCGTTTCATGCC
CCAAGGCTAAGTGGTCATAGGTTTCATGAAATTTCATCTCGACCAAGTCGTTTCATGCC
CCAAGGCTAAGTGGTCATAGGTTTCATGAAATTTCATCTCGACCAAGTCGTTTCATGCC

```

8 bp

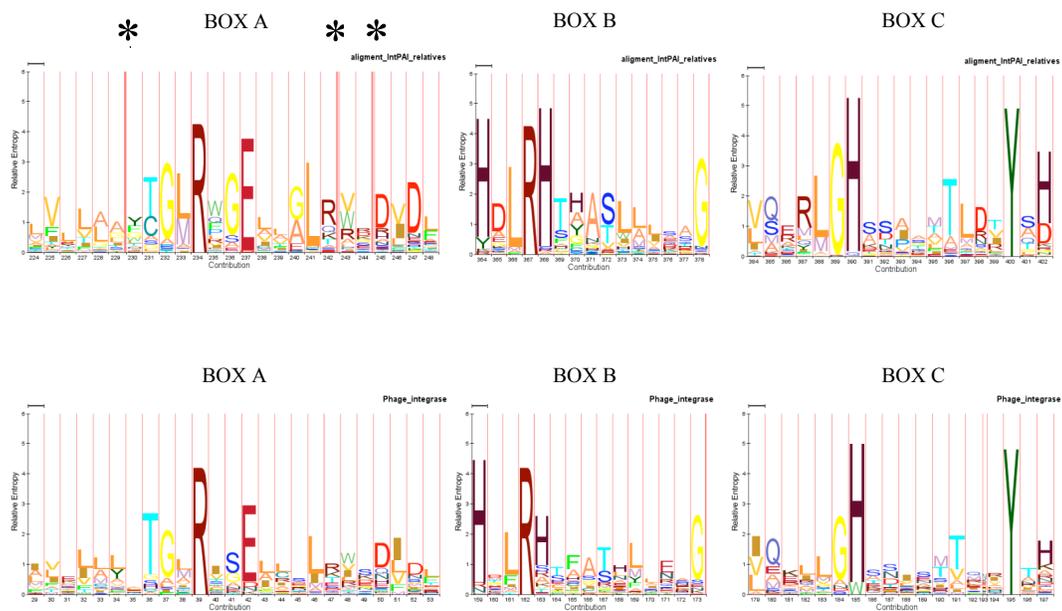


Figure 4-8. Conserved boxes of the proposed model for IntPAI and recombinases obtained with *intPAI* as a query (top) and general model of phage integrases reported in the Pfam databases (bottom). The relative size of a letter expresses its emission probability from a state's distribution. Insert states (position in which inserted amino acid are frequent) are shown in pink bars and labeled with (*).

Figure 4-9. Alignment of recombinases in public databases obtained with intPAI as a query. Five experimentally characterized tyrosine recombinases were incorporated into the alignment. Conserved boxes are shown. Accession numbers are indicated at the left side of the alignment. The level of conservation for each position is indicated at the bottom of each alignment.

The Box B of the model is located at positions 364 – 378 (Figure 4-8). Within Box B there are three highly conserved, positively charged amino acid residues: two histidine residues (positions 364 and 368) and one arginine residue (position 367). This arginine residue is predicted to be part of the catalytic core. The third region, Box C, contains a highly conserved tyrosine residue at position 400 (Figure 4-8), which is the third component of the catalytic core. Two histidine residues (positions 390-402) also are conserved in this region and might play a complementary function in the catalytic core.

Comparison of the conserved boxes in our model with an existing model for curated and functionally characterized integrases in the Pfam database (Bateman *et al.*, 2004) confirms slight variations in these regions (Figure 4-8). The most important differences are located in Box A. Three insertions states are observed in Box A of our model in comparison to the Pfam model (Figure 4-8). The insertion states represent additional amino acid residues that are located in some of the recombinases that were used to create the model. IntPAI is one of the recombinases with insertions states within Box A. The Insertions are FA, PEYCHPN and PASDPLD. Small insertions in the Box A are observed in the characterized recombinase found in plasmid SLP1 of *S. coelicolor* (Figure 4-4) and in other recombinase sequences retrieved from the Genbank to created the model Figure (4-9).

In Box B both models show that a histidine residue is the most representative amino acid at the position 364. However, the site contains some variations. IntPAI contains a tyrosine residue in this position. Additional to IntPAI, two experimentally characterized recombinases found in plasmids pSAM2 and SLP1 display variations at this site. The recombinase of pSAM2 has a tyrosine residue and the recombinase of SLP1 has an arginine residue. The position where these changes occur is proximal to

the catalytic site, and it is tempting to suggest that it could slightly alter the mechanism of recombination.

CONCLUSIONS

DNA recombination is an important process in genome maintenance and evolution. In bacteria, DNA recombination allows the interchange of genetic information across species. Tyrosine recombinases play an important role in the process of recombination and have been shown to be the key factors responsible for integration of mobile elements such as phage and ICEs. Here we report the sequence of a novel tyrosine recombinase that is involved in the recombination and consequently the maintenance of the PAISt in the *Streptomyces* genomes, in some case transforming the host into a plant pathogen (Kers *et al.*, 2005).

We demonstrate that the IntPAI is functional in the non-pathogens *S. coelicolor* and *S. lividans*. Both *Streptomyces* spp. have been described as recipients of the PAISt transferred from *S. turgidiscabies* (Kers *et al.*, 2005). The IntPAI requires only the presence of the *att* sites to recombine the mobile element into the genome of the host. Finally, intPAI represents a new tyrosine recombinase and reveals additional diversity in this family of recombinases. Slight differences in the regions that contain the residues for the catalytic core might reflect changes in the chemistry of the recombination reaction. Additional experiments are necessary to elucidate the biological meaning of the sequence variability in intPAI.

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.
- Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., and Eddy, S.R. (2004) The Pfam protein families database. *Nucleic Acids Res* **32**: D138-141.
- Bentley, S.D., Chater, K.F., Cerdeno-Tarraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C.W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C.H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabinowitsch, E., Rajandream, M.A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B.G., Parkhill, J., and Hopwood, D.A. (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**: 141-147.
- Blakely, G., Colloms, S., May, G., Burke, M., and Sherratt, D. (1991) Escherichia coli XerC recombinase is required for chromosomal segregation at cell division. *New Biol* **3**: 789-798.
- Bukhalid, R.A., Chung, S.Y., and Loria, R. (1998) nec1, a gene conferring a necrogenic phenotype, is conserved in plant-pathogenic *Streptomyces* spp. and linked to a transposase pseudogene. *Mol Plant Microbe Interact* **11**: 960-967.

- Burrus, V., Marrero, J., and Waldor, M.K. (2006) The current ICE age: biology and evolution of SXT-related integrating conjugative elements. *Plasmid* **55**: 173-183.
- Cornet, F., Mortier, I., Patte, J., and Louarn, J.M. (1994) Plasmid pSC101 harbors a recombination site, psi, which is able to resolve plasmid multimers and to substitute for the analogous chromosomal *Escherichia coli* site dif. *J Bacteriol* **176**: 3188-3195.
- Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.
- Esposito, D., and Scocca, J.J. (1997) The integrase family of tyrosine recombinases: evolution of a conserved active site domain. *Nucleic Acids Res* **25**: 3605-3614.
- Gopaul, D.N., and Duyne, G.D. (1999) Structure and mechanism in site-specific recombination. *Curr Opin Struct Biol* **9**: 14-20.
- Grainge, I., and Jayaram, M. (1999) The integrase family of recombinase: organization and function of the active site. *Mol Microbiol* **33**: 449-456.
- Gregory, M.A., Till, R., and Smith, M.C. (2003) Integration site for *Streptomyces* phage phiBT1 and development of site-specific integrating vectors. *J Bacteriol* **185**: 5320-5323.
- Groth, A.C., and Calos, M.P. (2004) Phage integrases: biology and applications. *J Mol Biol* **335**: 667-678.
- Guex, N., and Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**: 2714-2723.
- Guex, N., Peitsch, M.C., and Schwede, T. (2009) Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. *Electrophoresis* **30 Suppl 1**: S162-173.

- Joshi, M.V., and Loria, R. (2007) *Streptomyces turgidiscabies* possesses a functional cytokinin biosynthetic pathway and produces leafy galls. *Mol Plant Microbe Interact* **20**: 751-758.
- Kers, J.A., Cameron, K.D., Joshi, M.V., Bukhalid, R.A., Morello, J.E., Wach, M.J., Gibson, D.M., and Loria, R. (2005) A large, mobile pathogenicity island confers plant pathogenicity on *Streptomyces* species. *Mol Microbiol* **55**: 1025-1033.
- Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* **235**: 1501-1531.
- Lambert, C., Leonard, N., De Bolle, X., and Depiereux, E. (2002) ESyPred3D: Prediction of proteins 3D structures. *Bioinformatics* **18**: 1250-1256.
- Loria, R., Bignell, D.R., Moll, S., Huguet-Tapia, J.C., Joshi, M.V., Johnson, E.G., Seipke, R.F., and Gibson, D.M. (2008) Thaxtomin biosynthesis: the path to plant pathogenicity in the genus *Streptomyces*. *Antonie Van Leeuwenhoek* **94**: 3-10.
- Miyajima, K., Tanaka, F., Takeuchi, T., and Kuninaga, S. (1998) *Streptomyces turgidiscabies* sp. nov. *Int J Syst Bacteriol* **48 Pt 2**: 495-502.
- Sauer, B. (1994) Site-specific recombination: developments and applications. *Curr Opin Biotechnol* **5**: 521-527.
- Schuster-Bockler, B., and Bateman, A. (2005) Visualizing profile-profile alignment: pairwise HMM logos. *Bioinformatics* **21**: 2912-2913.
- Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**: 33-36.