Chapter 5

Experimental validation of the predictions

5.1 Introduction

So far I have discussed three different methodologies for the prediction of Genomic Islands (GIs), i.e. a compositional-based (chapter 2), a comparative-based (chapter 3) and a structural-based (chapter 4) approach. For each method I have used an *in silico* derived, manually curated test-dataset in order to validate the results and benchmark the prediction accuracy. However, what I have not yet discussed is a "real-life", combined application of these methods on un-annotated datasets, derived from very early stages in the annotation pipelines; this reveals the true strengths/weaknesses of this multifactorial, integrative approach in aiding and/or guiding (rather than extending pre-existing) annotation methodologies, especially when the genome sequences of closely related strains are not available to identify horizontally acquired regions.

This challenge forms the focus of this chapter; using a newly sequenced, un-annotated bacterial genome, the aim is to make *in silico* predictions of horizontally acquired regions, exploiting an integrative compositional and structural-based approach, and use experimental, rather than *in silico*, protocols to confirm the putative origin (vertical or horizontal) of the predicted genomic regions. Applying a Polymerase Chain Reaction (PCR) protocol, the presence and absence of the predicted islands will be probed in 17 un-sequenced closely and distantly related strains and the true borders of these islands will be confirmed by sequencing across the boundary site in strains lacking the island.

At the time that this project was conceived, the genome sequence of *Stenotrophomonas maltophilia*, strain K279a became available. *S. maltophilia*, previously taxonomically classified as *Xanthomonas maltophilia* or *Pseudomonas maltophilia*, is a gram-negative, aerobic, nonfermentative bacillus (Denton and Kerr, 1998). *S. maltophilia*, is an

important nosocomial pathogen, especially in immunocompromised patients, it has an unclear route of acquisition, little is known about its virulence properties (Denton and Kerr, 1998) and it shows resistance to broad-range antimicrobial agents, including ß-lactam (Saino et al., 1982) and aminoglycoside antibiotics (Muder *et al.*, 1996). Clinical manifestations related to *S. maltophilia*, include, but are not limited to, endocarditis (Mehta *et al.*, 2000), bacteremia (Muder *et al.*, 1996), meningitis (Libanore *et al.*, 2004) and pneumonia (Fujita *et al.*, 1996).

Therefore the genome sequence of *S. maltophilia* K279a forms an excellent test-dataset for the purposes of this analysis; *S. maltophilia* is an important life-threatening pathogen, with unknown virulence properties, and there is only one complete genome sequence of this species available, rendering benchmarking based on *in silico* comparative genomics inapplicable.

5.2 Methods

Given my very limited previous experience in lab-based techniques and protocols, the experimental methodology followed in this analysis was designed to be effective and at the same time simple and easy to implement, without requiring special training and extensive supervision. The aim was to validate the *in silico* predictions by exploiting the PCR protocol, using primers designed to flank the borders of the candidate islands predicted in the sequenced genome; this methodology made it feasible to sample the presence/absence of those GIs in closely and distantly related un-sequenced *S. maltophilia* clinical isolates, draw conclusions about their phylogenetic distribution and estimate the accuracy of the predicted boundaries.

The *in silico* and experimental methods pursued in this analysis are described in the following sections. It should be noted that the conclusions drawn will be purely based on the results confirming both the presence of the candidate islands in some strains and their absence from at least one of the remaining strains; in case the data cannot confirm these two

requirements, I will not make any inferences about the possible phylogenetic distribution and the origin of those predicted regions (see discussion section).

5.2.1 *In silico* prediction of GIs

The genome sequence (size: 4.85Mb, G+C%: 66.32) of *S. maltophilia* strain K279a (http://www.sanger.ac.uk/Projects/S_maltophilia/) was used as input to the Alien_Hunter (Vernikos and Parkhill, 2006) software (see chapter 2) and candidate GIs were predicted exploiting only compositional-based information. In a second step, the predicted candidate GIs were structurally annotated as discussed in chapter 4 and their structural annotation was used as input to the relevance vector machine (RVM) classifier (Tipping, 2001); RVM assigned a score to each prediction, quantifying our posterior belief that those structures are likely to be true GIs.

For the classification purposes, the three genus-specific structural GI models of *Salmonella*, *Staphylococcus* and *Streptococcus* described in chapter 4, as well as a model trained on all three datasets (Table 5.1, Table 5.2) were exploited. A sample of eight predictions with both highly and less probable GI structures with a score range of 0.2371–0.9997 formed the test-dataset of this analysis.

5.2.2 Comparative analysis

For the *in silico* sequence comparisons between the predicted boundaries of the putative GIs in the reference strain and the sequenced DNA fragments across the predicted insertion point in the un-sequenced *S. maltophilia* strains, a BLASTN (Altschul *et al.*, 1997) comparison was implemented and the results were visualized using ACT (Carver *et al.*, 2005).

Table 5.1: Structural annotation of eight genomic regions predicted as candidate GIs in the genome of *S. maltophilia*, strain K279a. Eight structural features were evaluated: The Interpolated Variable Order Motif (IVOM) score that measures both low and high order compositional deviation from the backbone composition and is expressed as the relative entropy between the query and the genome-backbone (variable order) compositional distribution, the insertion point (INSP) of each genomic region; two states were (binary) evaluated: insertion point within a CDS locus (disrupting the corresponding CDS) or insertion within an intergenic part of the chromosome, the size (SIZE) of each genomic region (bp), the gene density (DENS = number of genes per kb) of each region, presence or absence (binary) of direct/inverted repeats (REPEATS) flanking the boundaries of each genomic region, presence or absence (binary) of integrase and/or integrase-like (INT) protein domains, presence or absence (binary) of phage-related protein domains (PHAGE) and presence or absence (binary) of non-coding RNA (RNA) genes in the proximity of each region.

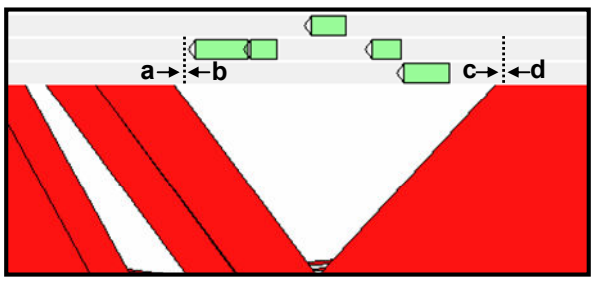
Location	Region	IVOM	INSP	SIZE	DENS	REPEATS	INT	PHAGE	RNA
60416..70829	R1	0.38128	1	10,413	1.3444	1	1	1	0
3089398..3127169	R16	0.74458	0	37,771	1.0060	1	1	1	1
299814..335480	R4	0.32642	0	35,666	1.2897	1	1	1	1
1323939..1367750	R12	0.55018	0	43,811	1.2325	1	1	1	0
1720046..1724493	R14	0.72176	0	4,447	1.7986	1	0	0	1
1945379..2002745	R15	0.28154	0	57,366	1.1854	1	1	1	1
3913072..3931089	R20	0.16626	0	18,017	0.6666	1	0	0	0
631285..661659	R7	0.27377	0	30,375	0.8559	0	0	0	0

Table 5.2: Posterior probability of being a true GI, for eight predicted genomic regions, exploiting four GI models, i.e. *Salmonella*-specific (Salm), *Staphylococcus*-specific (Staph), *Streptococcus*-specific (Strep) and the all-three (all3) genera model.

Region	Salm model	Staph model	Strep model	all3 model
R1	0.9918	0.9991	0.9994	0.9997
R16	0.9995	1.0000	0.9965	0.9992
R4	0.9944	0.9959	0.9804	0.9948
R12	0.9851	0.9997	0.9922	0.9903
R14	0.9978	0.9999	0.9005	0.9890
R15	0.9786	0.9826	0.9765	0.9835
R20	0.5023	0.2109	0.4742	0.4983
R7	0.3070	0.5223	0.1368	0.2371

5.2.3 Principle of the experimental approach

The principle of the experimental approach followed throughout this study is based on the analysis of the presence or absence of the amplified products for each set of primers, designed to flank the two boundaries of the predicted GIs (primers “a” and “b” for the left boundary; primers “c” and “d” for the right boundary), as well as for the “a” and “d” primers (Figure 5.1).



strain	PCR product			Inferred GI distribution
	ab	cd	ad	
1	+	+	-	Present
2	-	-	+	Absent
3	+	-	-	Ambiguous
4	-	+	-	Ambiguous

Figure 5.1: Screenshot summarizing the principle of the experimental approach followed in this study.

This experimental approach exploits the following three assumptions: A. If both the "a+b" and "c+d" products for a given GI-strain set are successfully amplified, then the predicted GI is inferred to be present in the corresponding strain; B. If only the "a+d" product is amplified, then the predicted GI is inferred to be absent from the corresponding strain; in this case the true boundaries can be determined by generating sequence from this product across the boundary site in strains lacking the island; C. Finally, amplified products for any other different combination of primers (e.g. only "a+b" or only "c+d" products) are inferred to be ambiguous results.

5.2.4 DNA purification

17 un-sequenced *S. maltophilia* clinical strains (Figure 5.2) were kindly provided by Dr Matthew Avison at the Department of Cellular and Molecular Medicine, University of Bristol. The 17 strains were grown overnight on Luria-Bertani broth (LB) media, at 37°C. Purification of genomic DNA was carried out using the *Wizard Genomic DNA*

Purification Kit of *Promega* according to the protocol for isolating genomic DNA from Gram negative bacteria (pages 16-17, *Promega* manual).

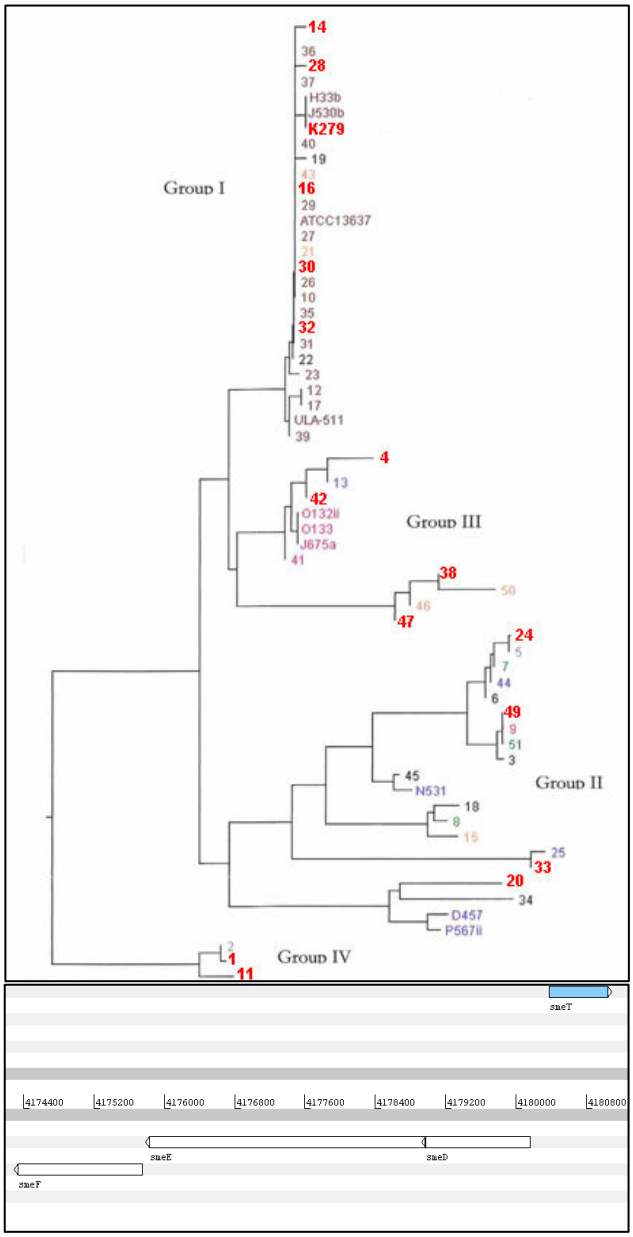


Figure 5.2: Phylogenetic tree of *S. maltophilia* isolates based on the *smeT–smeD* intergenic sequence (top); figure modified from (Gould *et al.*, 2006). The name of the strains used in this analysis, is highlighted in bold, red-coloured font. Three CDSs (smeD, smeE and smeF – accession number AJ252200), encode components of a multidrug efflux pump (Alonso and Martinez, 2000) present in *S. maltophilia*. The expression of the *smeDEF* locus (bottom), is regulated by a putative transcriptional repressor (smeT, belonging to the TetR and AcrR transcriptional regulator family), located upstream of the *smeDEF* locus (Sanchez *et al.*, 2002). The *smeT–smeD* intergenic region consists of a highly conserved and a hypervariable untranslated region (Gould *et al.*, 2004) and contains the putative promoters of *smeT* and *smeDEF*. The grouping of the *S. maltophilia* strains in the four (I, II, III and IV) phylogenetic groups has been based on the analysis of the 16s rRNA locus.

The concentration of the genomic DNA extracted from the 17 strains and the genomic DNA of the reference K279a strain was measured using the NanoDrop ND-1000 spectrophotometer; the results are shown in Table 5.3.

Table 5.3: Genomic DNA concentration of the 18 *S. maltophilia* strains used in this study.

Strain	Concentration (ng/µl)
K279a	200 (diluted to a final concentration of 20 ng/µl)
K279(1)	9.1
K279(2)	13.6
30	6.1
1	12.6
4	8.4
47	4.6
28	7.9
20	11.1
11	6.8
33	4.8
16	6.1
14	4.4
32	8.1
42	24.2
38	16.5
24	7.5
49	9.4

5.2.5 Primer design

For each of the eight candidate GIs, two sets of primers, one flanking the upstream and one flanking the downstream boundary were designed implementing the Primer3 software (Rozen and Skaletsky, 2000), available at http://frodo.wi.mit.edu/, using the default parameters. The 16 designed primers (Table 5.4) were ordered from SIGMA GENOSYS (http://www.sigmaaldrich.com/Brands/Sigma_Genosys.html).

Table 5.4: Primer sequences used in this analysis.

Genomic Region	Left boundary primer set	Right boundary primer set
R1	5′-gcagtgactcctgcagatcc-3′	5′-tcccccattacagcaggtag-3′
	3′-aggcttggtcttgcgaatag-5′	3′-ggagatccgaacatgcaatc-5′
R4	5′-ggcctgagcgactactacatc-3′	5′-gcaactccagctcatgctc-3′
	3′-ctgaaacatcggggaatcac-5′	3′-gcaagggctttcaagagttg-5′
R7	5′-agaagaccgagctgttcacc-3′	5′-cggtttcgaatatccagtgc-3′
	3′-gtttgacgtagctggcattg-5′	3′-ggatctgtttgcgatcctg-5′
R12	5′-cttcaagagctcgaccaacc-3′	5′-gactccatctcctggactgc-3′
	3′-tcgttcttgggctattatgg-5′	3′-accgtggccaatatcaagtc-5′
R14	5′-aatggtcgcgataccagttc-3′	5′-tacttgcttccctgccagac-3′
	3′-ctcgttcctcggcttcatag-5′	3′-atgacttcgggaatgcagac-5′
R15	5′-gagcgtagttgtcgtcgttg-3′	5′-acaggccttcgcagacatag-3′
	3′-gtttagccagagccgcatag-5′	3′-gcacgccaatactgagactg-5′
R16	5′-tgatccatccattctgcaag-3′	5′-atgcttgacgaaaggtttgc-3′
	3′-cctcccagattcgtgaaacc-5′	3′-tgtgcacgatgatctcaacc-5′
R20	5′-ggtggatgagaagccgatg-3′	5′-atctggccggagaagtacac-3′
	3′-cgtgtgctcaacgagaagg-5′	3′-acgagatcatgggctaccac-5′

5.2.6 Polymerase Chain Reaction – PCR

The purpose of PCR is the amplification of specific DNA fragments to a very large number of copies. The PCR protocol consists of three major steps (i.e. denaturation, annealing and extension), each of which is repeated 30-40 times.

The DNA fragments of interest were PCR amplified using the following reaction mixture (total volume 10μl): 0.2μl of genomic DNA, 0.1μl (100μM initial concentration) forward and reverse primers, 1μl (2 μM) dNTPs (dATP, dCTP, dGTP, and dTTP), 1μl PCR buffer (10x), containing 15mM of $MgCl_2$, 7.4μl of double-distilled water and 0.2μl (5units/μl) *Taq* polymerase (Amplitaq). PCR amplification was carried out using a PTC-225 peltier thermal cycler (MJ Research), implementing the program detailed in Table 5.5.

Table 5.5: The PCR protocol used in this analysis. At step 3 the optimal annealing temperature for each primer set was initially determined (and subsequently applied) using a gradient PCR protocol with a range of annealing temperature of 53-68 °C.

Step	Temperature (°C)	Time
1	95	10min
2	95	30sec
3	53-68	30sec
4	72	3min
5	goto step 2 (x39)	
6	72	10min
7	10	0min (for ever)

5.2.7 Gel electrophoresis

DNA fragments were separated on an agarose gel exploiting the electrophoresis protocol. The principle of this protocol is the separation of nucleic acids or proteins based on their charge and mass. Using an electric field, the macromolecules can be separated on a gel, with a rate of migration that depends on many factors, including the applied voltage, the hydrophobicity, size and shape of the molecules, the agorose gel concentration and the ionic strength of the buffer solution.

The agarose gel (1% w/v) was prepared by dissolving 0.5g of agarose (Sigma) in 50ml of Tris-acetate-EDTA (TAE) buffer (1x). The samples were loaded using 2μl of ficoll loading dye (0.25% bromophenol blue, 0.25% xylene cyanol FF, 15% Ficoll 400 in water) and run on the gel for 45 minutes applying a voltage of 60V. The samples were stained for 10-20 minutes by adding 5μl of ethidium bromide (10mg/ml) to the running buffer; the DNA bands on the gel were viewed under ultraviolet light. The size of the DNA fragments was determined by comparison with 1kb (Invitrogen) DNA ladder (1μg/μl).

5.2.8 Sequencing

In order to confirm the true borders of the predicted islands the boundary site in strains lacking the islands was sequenced (sequences are listed in Appendix J). All sequencing was performed by the core sequencing teams at the Sanger Institute, according to the protocols of the sequencing facility; briefly the templates were sequenced using AB BigDye terminator chemistry, and run on AB3730 machines. The resulting traces were base-called with in-house software (ASP), which also recognised and trimmed cloning vector and poor quality sequences.

5.3 Results

5.3.1 Genomic Island candidates

5.3.1.1 Genomic Island 1

The first candidate GI (R1), is a 10.5kb genomic region of low G+C content (63.82% – genome average 66.32%) and high gene density (1.34 – genome average 0.904) inserted within a coding sequence (CDS) (Smlt0055) encoding a putative alcohol dehydrogenase; this CDS is now disrupted by the integrated GI with the two CDS fragments flanking the 18bp direct repeats (DRs) of the island (Figure 5.3). R1 consists of 14 CDSs, the majority of which encode products with unknown function (Appendix K) and has the highest RVM score (0.9997, under the all3 model – Table 5.2), representing a highly probable GI structure.

R1 seems to represent a very recent acquisition in the *S. maltophilia* K279 lineage (Figure 5.4) since it is present only in the three *S. maltophilia* K279 strains (namely K279a, K279(1) and K279(2)). The absence of this GI was confirmed for strains 30, 28, 20, 14, 32, 24 and interestingly for strains K279(1) and K279(2) too (Figure 5.4); these results seem to contradict the presence of this GI in the latter two strains. However, sequencing across the boundaries of R1 in strain 28 (which lacks the island) and K279(1) shows that the two sequences are 97.5% identical

(711/729 identical residues – Appendix L) suggesting that the insertion point of R1 is present in all eight strains.

Figure 5.3: ACT screenshot: Predicted genomic island R1. Top: BLASTN comparison between *S. maltophilia* strain K279a and the sequenced fragment across the boundary site of R1 in strain 28. Regions within the two sequences with similarity are joined by red coloured bands that represent the matching regions. The G+C% content with a window size of 1kb is shown at the top of this screenshot. R1 is shown as green-coloured feature flanked by a set of 18bp DRs (grey coloured joined features). The DRs of R1 are flanked by the two fragments of Smlt0055 (brown-coloured joined features). Bottom: Higher resolution ACT screenshot showing the sequence similarity of the left and the right boundaries of R1 and the sequenced fragment of strain 28. The two sets of primers used to amplify the left and the right boundaries of R1 are shown as red-coloured features flanking the left and right attachment sites of this island.

Given that both the left and the right boundaries of R1 are present in all three K279 strains and, at the same time, the PCR results across the predicted insertion point of R1 suggest that R1 is also absent from K279(1) and K279(2) (Figure 5.4), it is likely that the insertion point (i.e. Smlt0055) of R1 has been duplicated in the latter two strains; the first copy has been disrupted by R1 while the second is intact.

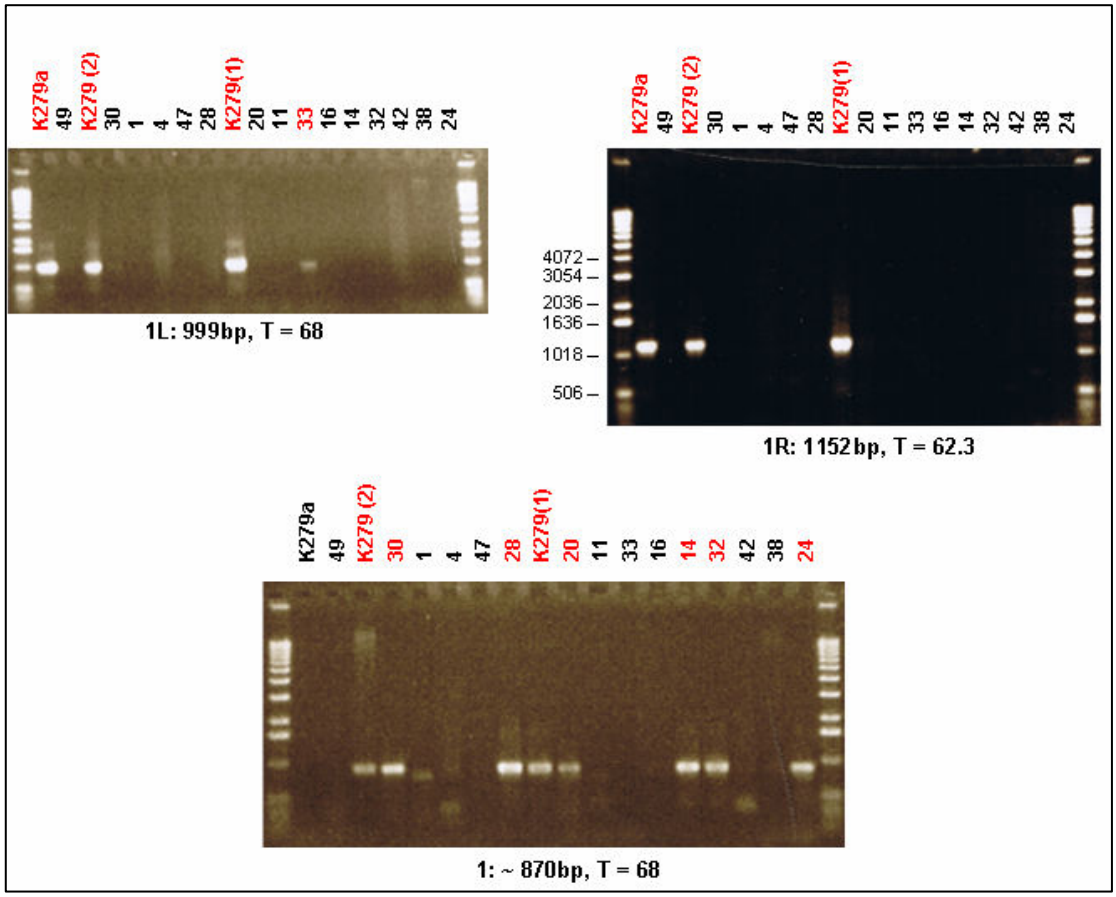


Figure 5.4: PCR amplification of the left (1L) and the right (1R) boundaries (top) of genomic island R1 and of the region across the boundary site of R1 in strains lacking the island (bottom). The name of each strain is provided at the top of each lane and strains with amplified product, of the expected size, are highlighted in red. For each amplified product the expected sequence size (bp) and the optimal annealing temperature (T) is provided below each gel screenshot.

The global alignment between the insertion point of R1 in the reference strain K279a and the corresponding sequenced fragments in K279(1) and strain 28 (S28) shows that the three sequences are highly similar (K279a-K279(1): 99% identical – 723/730 identical residues; K279a-S28: 97.1% identical – 709/730 identical residues). An alternative

hypothesis that might well explain the above ambiguity, is that a fraction of the K279(1) and the K279(2) populations, used to extract the genomic DNA for those two strain types, might have R1 inserted within Smlt0055 while the remainder of the population has the corresponding CDS intact (e.g. via a putative deletion event of R1).

Frequent deletion of GIs during population growth has been seen in other organisms (Buchrieser *et al.*, 1998; Bueno *et al.*, 2004; Nair *et al.*, 2004) and appears to occur via homologous recombination between the flanking DRs.

5.3.1.2 Genomic Island 16

R16, is a 37.7kb island of low G+C content (62.21% – genome average 66.32%) and similar gene density (1.006) to the genome average (0.904), inserted at the 3' end of a $tRNA^{Ser}$ locus. R16 is flanked by a set of 21bp DRs with the terminal 13bp corresponding to the disrupted 3' end of the tRNA gene (Figure 5.5). R16 consists of 41 CDSs, the majority of which encode products of unknown function while three CDSs (Smlt3051, Smlt3053 and Smlt3069) encode two putative conjugal transfer proteins (traA and traD) and a putative plasmid partitioning protein, respectively (Appendix K). Based on the RVM score (0.9992, Table 5.2) R16 also represents a highly probable GI structure.

Similar to R1, R16 probably represents a recent acquisition in *S. maltophilia* K279 strains (Figure 5.6). The PCR results confirm that the same form of R16 is present in all three K279 strains, leaving open however the possibility that a variation of R16, with a different left boundary, might also be present in at least seven other strains (that gave amplified product, of the expected size, for the right boundary of R16) (Figure 5.6). Sequencing across the insertion site of R16, confirmed the complete absence of this island in at least one strain (strain 24), (Figure 5.6, Appendix J).

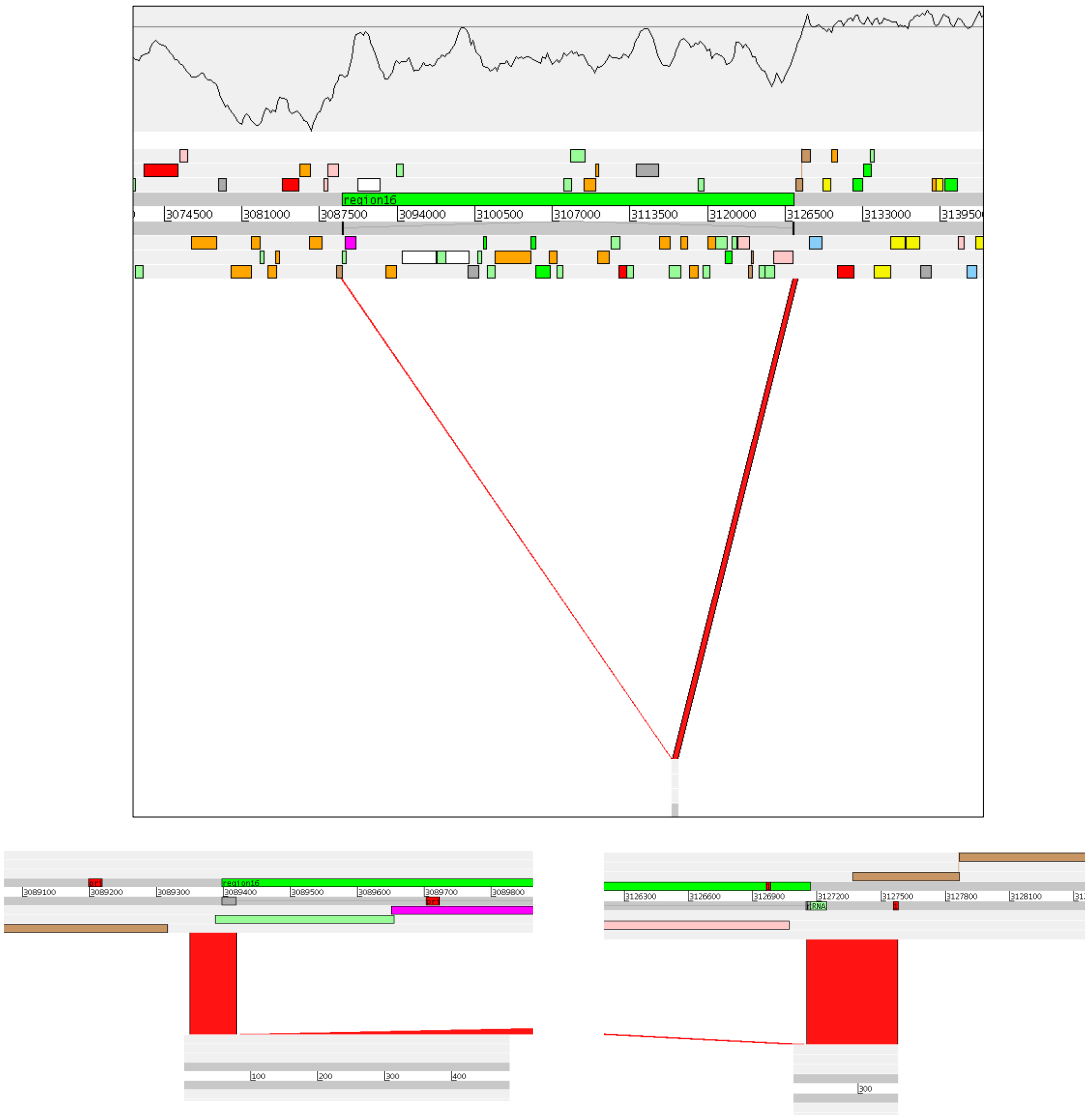


Figure 5.5: ACT screenshot: Predicted genomic island R16. Top: BLASTN comparison between *S. maltophilia* strain K279a and the sequenced fragment across the boundary site of R16 in strain 24. Bottom: Higher resolution ACT screenshot; R16 is shown as green-coloured feature flanked by a set of 21bp DRs (grey coloured joined features). The disrupted $tRNA^{Ser}$ gene is shown as a light-green coloured feature overlapping with the DR at the right boundary of R16 (bottom-right screenshot). The two sets of primers used to amplify the left and the right boundaries of R16 are shown as red-coloured features flanking the left and right attachment sites of this island.

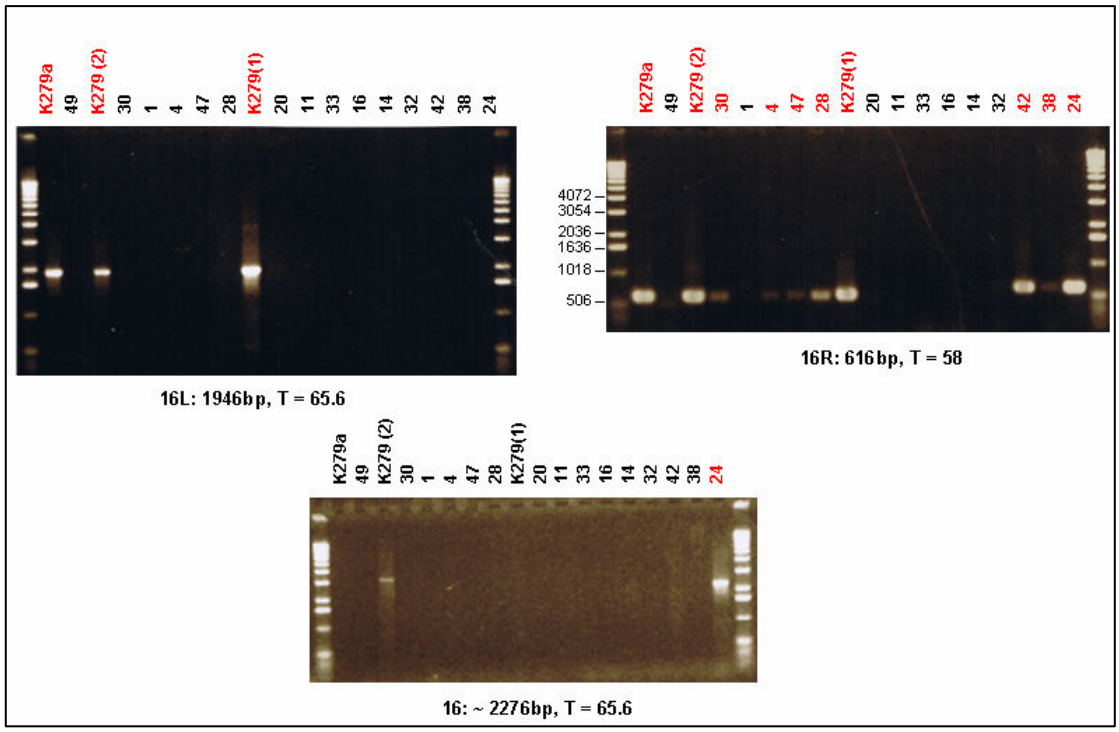


Figure 5.6: PCR amplification of the left (16L) and the right (16R) boundaries (top) of genomic island R16 and of the region across the boundary site of R16 in strains lacking the island (bottom).

5.3.1.3 Genomic Island 4

R4 is a 35.7kb island of low G+C content (63.42%) and high gene density (1.29) inserted at the 3' end of a $tRNA^{Thr}$ locus (Figure 5.7). R4 is a putative prophage flanked by a set of 31bp DRs that correspond to the 3' end of the tRNA gene. R4 has a very high RVM score (0.9948) and consists of 45 CDSs, over half of which have sequence similarity to annotated phage-related CDSs (Appendix K).

R4 is present in the three *S. maltophilia* K279 strains (Figure 5.8) and a variation of this island with a different right boundary cannot be excluded from being present in strains 28, 32 and 24. PCR across the insertion point of R4 did not confirm the absence of this island in any of the 17 strains (see benchmarking section below); however the fact that for nine *S. maltophilia* strains the left boundary of R4 gave an amplified product of the expected size, and that the left primer set corresponds to the 3' end of Smlt0285 encoding a phage-related integrase (which is often

conserved amongst related phages), leaves open the possibility that different or similar prophages might also be present in these strains.

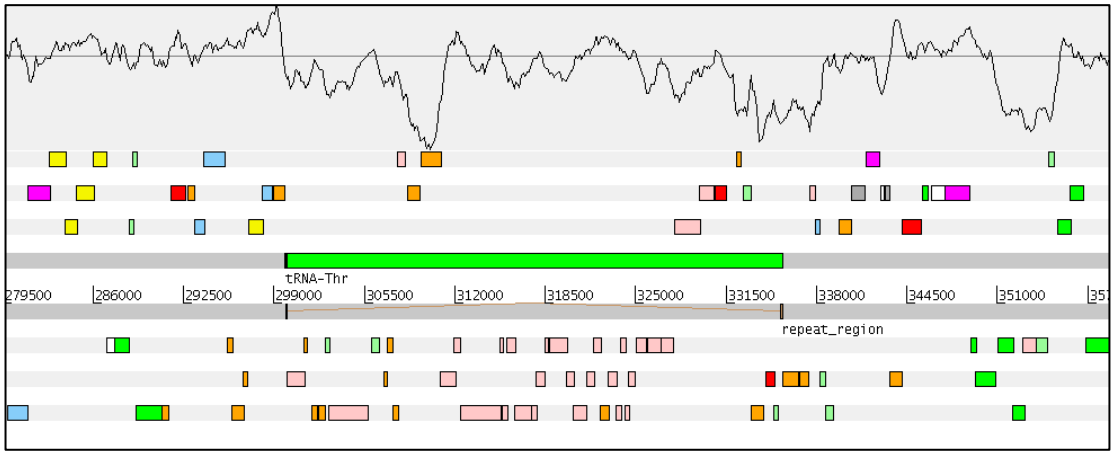


Figure 5.7: Artemis (Rutherford *et al.*, 2000) screenshot: Predicted genomic island R4, present in *S. maltophilia* strain K279a. The disrupted $tRNA^{Thr}$ gene is shown immediately upstream of R4. The 31bp DRs are shown as brown-coloured joined features flanking the island. The G+C% content with a window size of 1kb is shown at the top of this screenshot.

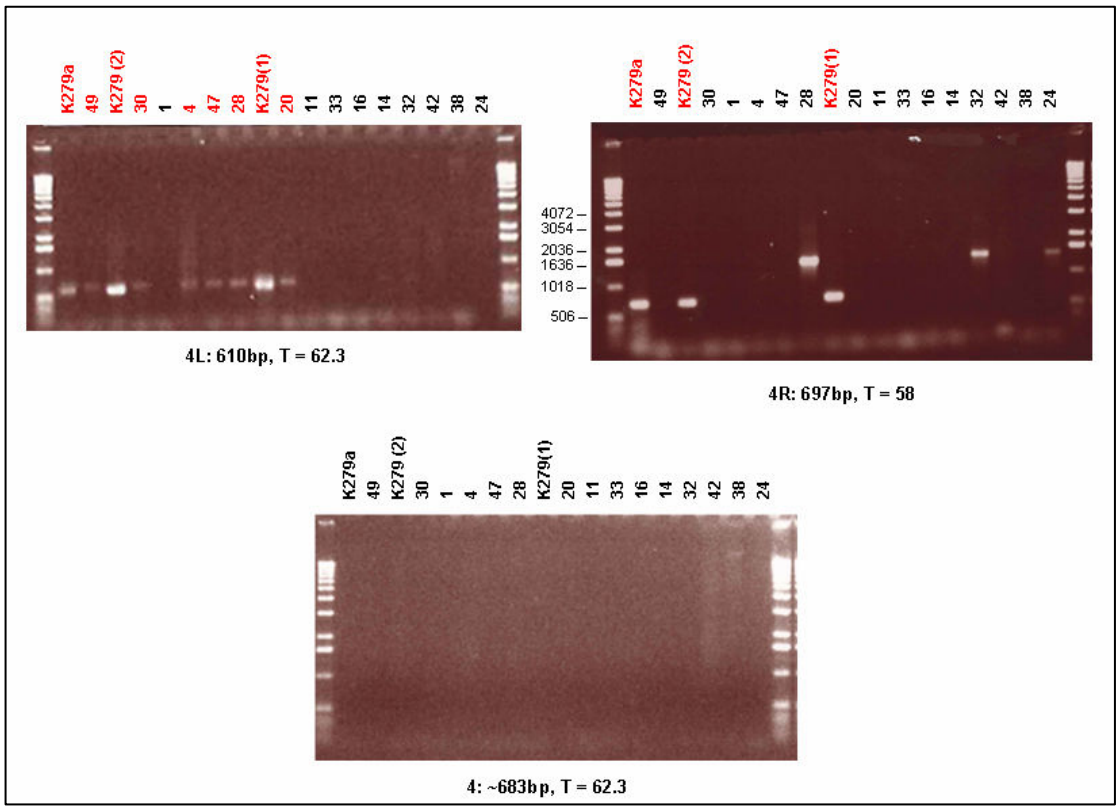


Figure 5.8: PCR amplification of the left (4L) and the right (4R) boundaries (top) of genomic island R4 and of the region across the boundary site of R4 in strains lacking the island (bottom).

5.3.1.4 Genomic Island 12

Similar to R16, R12 carries at least 10 CDSs encoding putative conjugal transfer proteins (Appendix K).

Figure 5.9: ACT screenshot: Predicted genomic island R12. Top: BLASTN comparison between *S. maltophilia* strain K279a and the sequenced fragment across the boundary site of R12 in strain 49. Bottom: Higher resolution ACT screenshot; R12 is shown as green-coloured feature flanked by a set of 22bp DRs (brown coloured joined features). The two sets of primers used to amplify the left and the right boundaries of R12 are shown as red-coloured features flanking the left and right attachment sites of this island.

R12 (Figure 5.9) is a 43.8kb island of low G+C content (62.69%) and high gene density (1.23) flanked by a set of 22bp DRs. R12 consists of 53 CDSs and it is inserted within a locus of three ribosomal protein coding genes (*smlt*1278 and *smlt*1279 encoding two putative 50S ribosomal proteins L21 and L27, located upstream of the left R12 boundary; *smlt*1337, encoding a putative 30S ribosomal protein S20, located downstream of the right R12 boundary). The posterior probability of this genomic region of being a true GI, under the all3 model, is 0.9903.

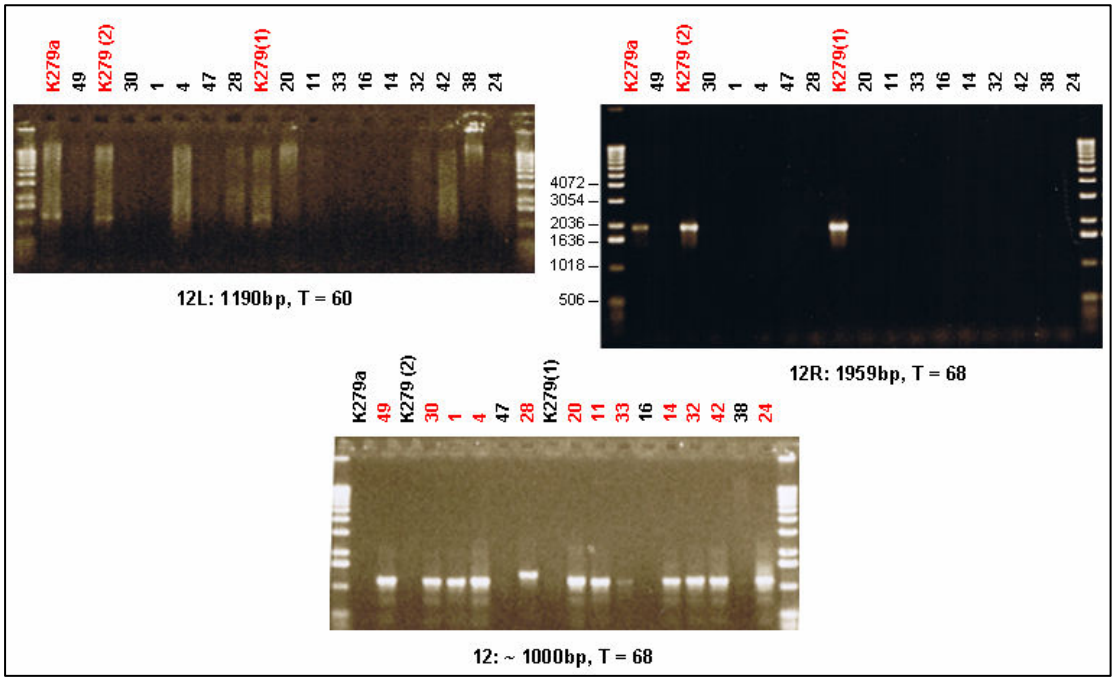


Figure 5.10: PCR amplification of the left (12L) and the right (12R) boundaries (top) of genomic island R12 and of the region across the boundary site of R12 in strains lacking the island (bottom).

R12 represents a very recent insertion, present in all three K279 strains (Figure 5.10), while its absence is confirmed, by PCR and sequencing across its insertion point, in 12 of the 17 *S. maltophilia* strains; these data suggest that most likely R12 is a K279-specific island and its insertion point is unoccupied in the majority of the un-sequenced isolates.

5.3.1.5 Genomic Island 14

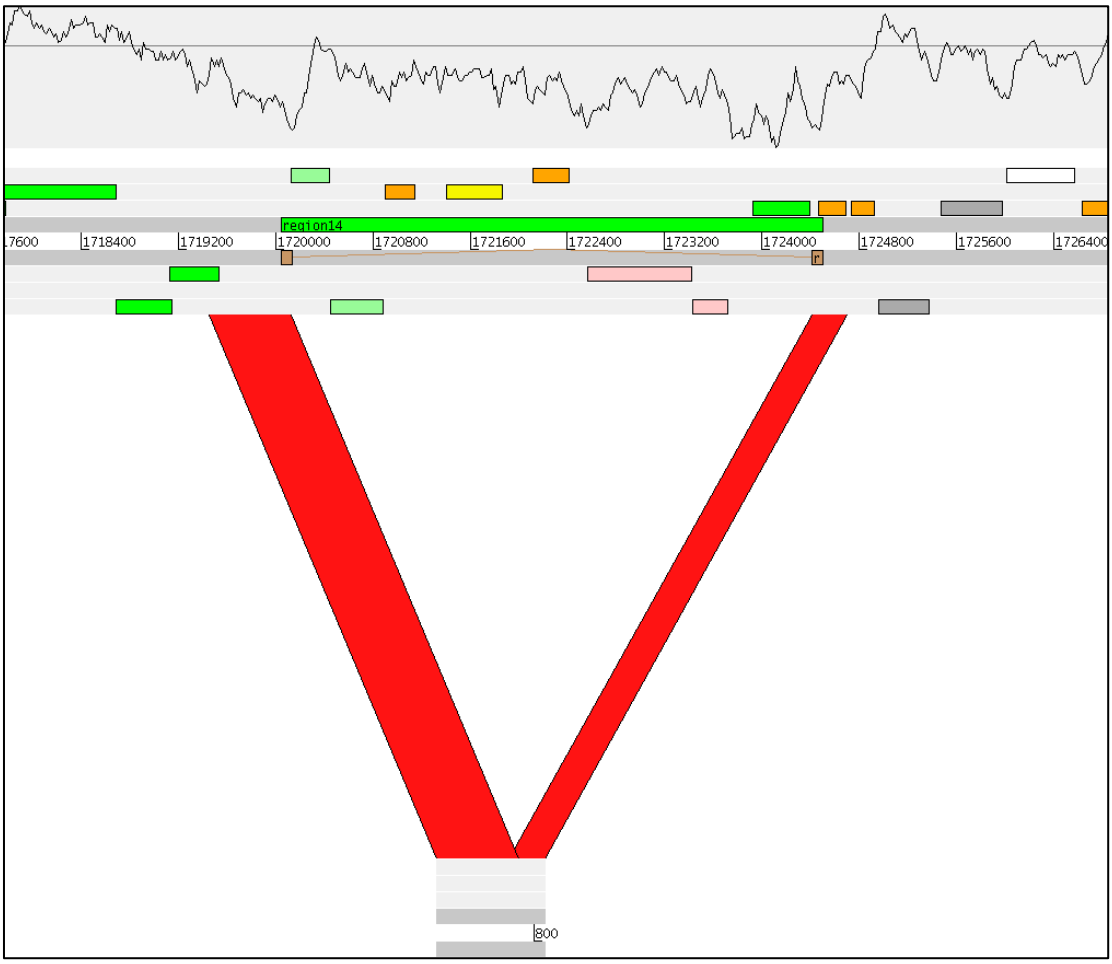
R14 is a small island of 4.4kb and very high gene density (1.8), a value that is double the average gene density (0.904) characterising the genome of *S. maltophilia*, strain K279a. R14 has a very low G+C content (58.7%) and consists of eight CDSs, three of which (Smlt1662, Smlt1663 and Smlt1660) encode two insertion sequence (IS) Xac3-like transposases and a putative modification methylase, respectively (Appendix K). R14 is flanked by a set of very large (81bp) DRs, that overlap with 65% of the entire $tRNA^{Cys}$ gene, located upstream of the left boundary of R14 (Figure 5.11); the insertion point of this island corresponds to the 3' end of this tRNA locus. The right boundary of R14 overlaps for 28bp with the 5' end of Smlt1665 (conserved hypothetical protein). The RVM score of R14 is 0.989.

R14 is present in the three K279 strains, while the PCR results confirmed its absence in at least eight *S. maltophilia* strains (Figure 5.12). However only two of those strains (30 and 14) gave the expected product size (~900bp) corresponding to the sequence across the insertion point of R14, while the remaining six strains (4, 47, 28, 20, 42 and 24) gave a product of slightly larger size (~1,200-1,300bp); these data leave open the possibility of a putative internal sequence variation of the corresponding R14 insertion site in the latter six strains, given that the sequencing of the two different products confirmed the same left and right boundaries of this island (Figure 5.11).

It is worth mentioning, that the sequencing of the corresponding region in strains 14 and 28 would, in theory, reveal (see R15 in the following section) the gene content of the ~400bp size difference between the amplicons; however the entire sequence of the amplified region was successfully determined only for strain 14, whereas in the case of strain 28, the sequence is missing a fragment from the left end of the corresponding amplicon. Based on the gene content information, showing three tRNA genes located immediately upstream of R14 (Figure 5.11b), it

is likely that the ~400bp size difference might be due to the presence of extra copies of tRNA genes in strain 28.

Interestingly this size variation is consistent with the phylogenetic tree of the *S. maltophilia* lineage (Figure 5.2); indeed strains 14 and 30 (product size ~900bp) are members of the same taxonomic group (i.e. group I) with the three K279 strains. On the other hand, the remaining six strains (with the exception of strain 28) are more distantly related isolates and belong to the taxonomic groups II and III (Figure 5.2).



A.

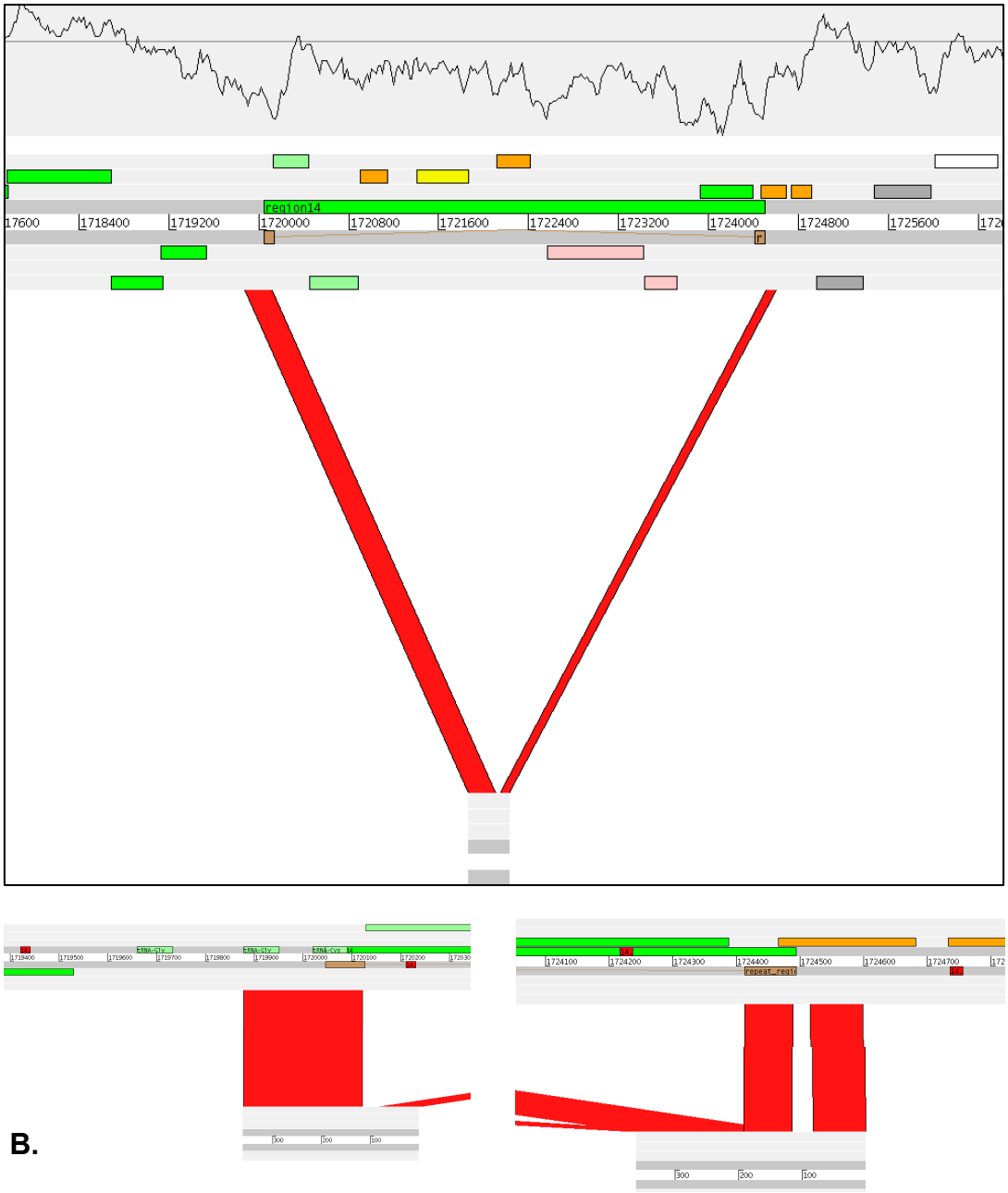


Figure 5.11: ACT screenshot: Predicted genomic island R14. Top: BLASTN comparison between *S. maltophilia* strain K279a and the sequenced fragment across the boundary site of R14 in strain 14 **(A)** and strain 28 **(B)**. Bottom: Higher resolution ACT screenshot; R14 is shown as green-coloured feature flanked by a set of 81bp DRs (brown coloured joined features). The two sets of primers used to amplify the left and the right boundaries of R14 are shown as red-coloured features flanking the left and right attachment sites of this island; the $tRNA^{Cys}$ gene, upstream of R14 is shown as a light-green coloured feature overlapping with the left boundary of this island (bottom-left).

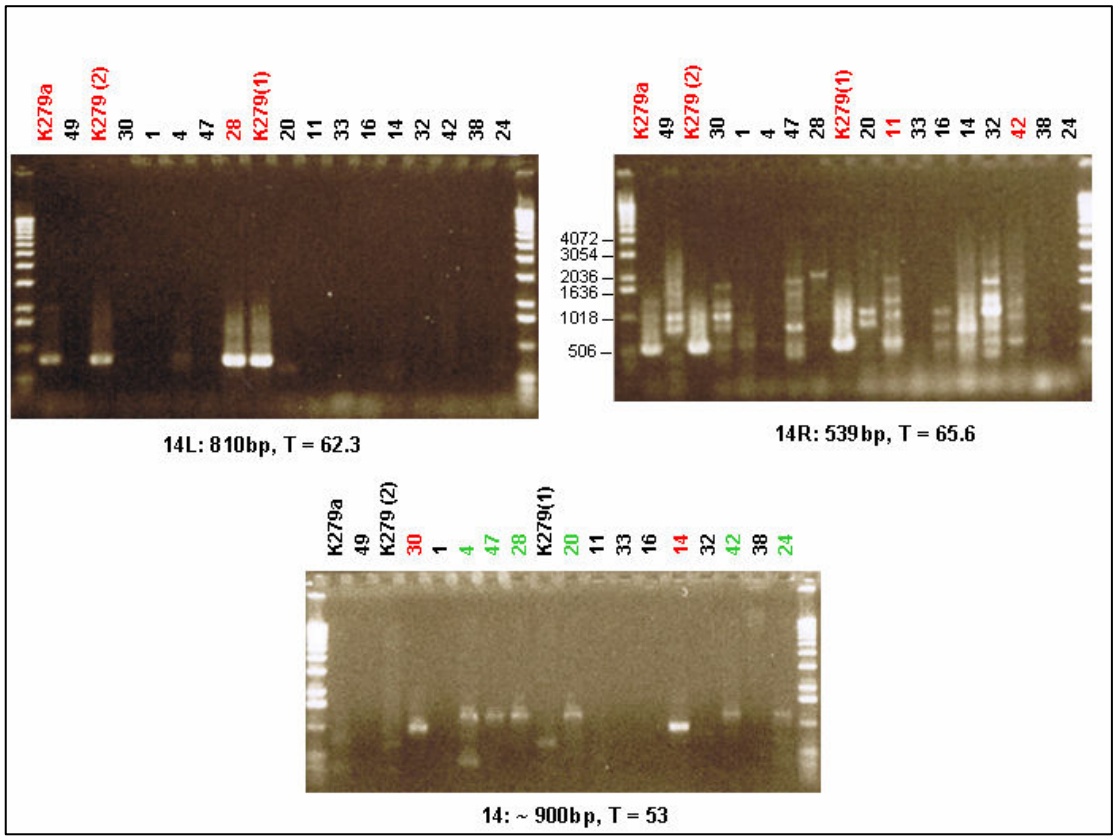


Figure 5.12: PCR amplification of the left (14L) and the right (14R) boundaries (top) of genomic island R14 and of the region across the boundary site of R14 in strains lacking the island (bottom). Colour scheme (bottom): Strains whose product size is ~900bp (expected size) are red coloured while strains with a larger product size ~ 1,200-1,300bp are green coloured. Note: for a higher annealing temperature (T = 65.6 °C), only strains 30 and 14 gave an amplified product for the sequence fragment across the boundary site of R14.

5.3.1.6 Genomic Island 15

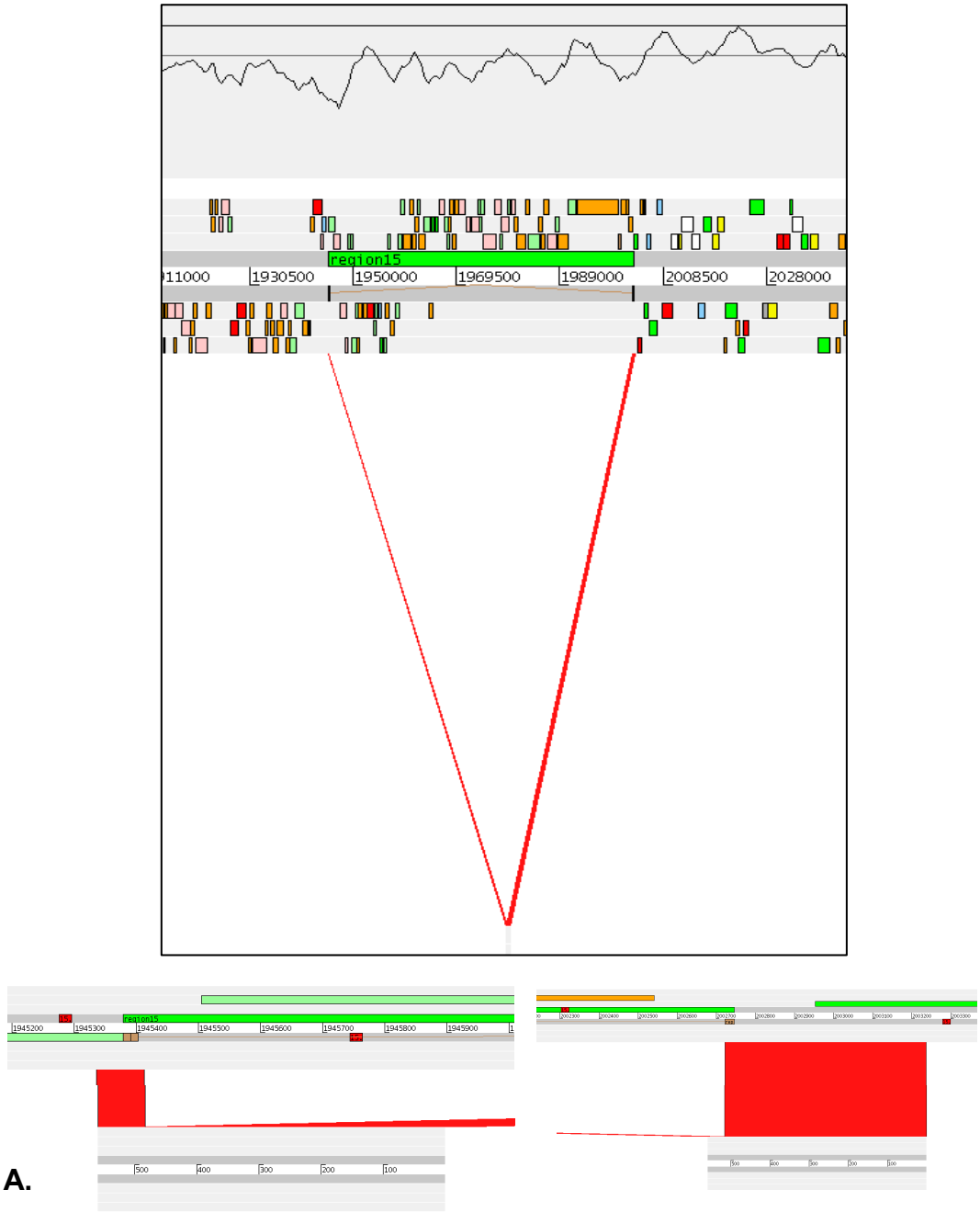
R15 is a 57.4kb island of low G+C content (64.8%) and high gene density (1.18) inserted at the 5' end of a tmRNA (also known as 10Sa RNA) gene (Figure 5.13). R15 carries 68 CDSs, 11 of which have sequence similarity to annotated phage-related CDSs while the majority of the remaining CDSs encode for proteins of unknown function (Appendix K). R15 has a high RVM score (0.984) and is flanked by a set of 24bp DRs that overlap with the first 12 bases of the tmRNA locus; this tmRNA gene seems to represent an insertion site hot-spot, since its 3' end forms the insertion

site of a second (52.9kb) genomic element of putative phage origin (data not shown), flanked by a set of (8bp) DRs, that is located immediately upstream of R15 in a head-to-head orientation; for these reasons, the left primer set for R15 was designed within the tmRNA locus to avoid possible problems with the differential presence of the other island. Interestingly R15 carries a $tRNA^{Met}$ gene located (internally) 20.4kb downstream of the left boundary of this island; it is worth noting that overall there are five copies (including the R15 copy) of $tRNA^{Met}$ genes present in the genome of *S. maltophilia*, strain K279a.

Unlike the previously discussed GIs, the presence of R15 was confirmed in four *S. maltophilia* strains, namely K279a, K279(1), K279(2) and strain 32 (Figure 5.14); the latter belonging to the same taxonomic group (group I) as the three K279 strains (Figure 5.2). The absence of R15 was confirmed in at least seven strains and, similarly to R14, there are two different product sizes that are phylogenetically consistent with the *S. maltophilia* phylogenetic tree; for strains 30, 28, 16, 14 and 24 the PCR amplified products had the expected size (~656bp) while strains 4 and 42 gave a product size of ~1,500bp (Figure 5.14). With the exception of strain 24 (taxonomic group II) all four strains that gave the expected product size belong to the taxonomic group I, while strains 4 and 42 are members of the taxonomic group III (Figure 5.2).

It is worth mentioning that the ~800bp size difference between the PCR products is almost exclusively attributed to the presence of two predicted CDS fragments present in strain 42 (and presumably in strain 4); those two CDS fragments, named herein CDS1 and CDS2 are very similar (Figure 5.13c) to SmalDRAFT_1529 (encoding a putative uncharacterized protein) and SmalDRAFT_1530 (encoding a GCN5-related N-acetyltransferase) present in *S. maltophilia* R551-3 ctg153 (Accession Number: AAVZ01000019). CDS1 and CDS2 along with CDS3 (encoding a putative transmembrane protein, similar to the 5' end of SmalDRAFT_1530 and Smlt1982 – present in K279a) are also sequentially located in the same orientation in *S. maltophilia* R551-3;

however based on the BLAST comparison, CDS1 is probably a remnant of SmalDRAFT_1529 (Figure 5.13c). These data confirm the absence of R15 from the corresponding predicted insertion site present in the available sequence of *S. maltophilia* R551-3 and further suggest that the gene content of this locus is conserved and unoccupied (by a GI) in at least three *S. maltophilia* strains.



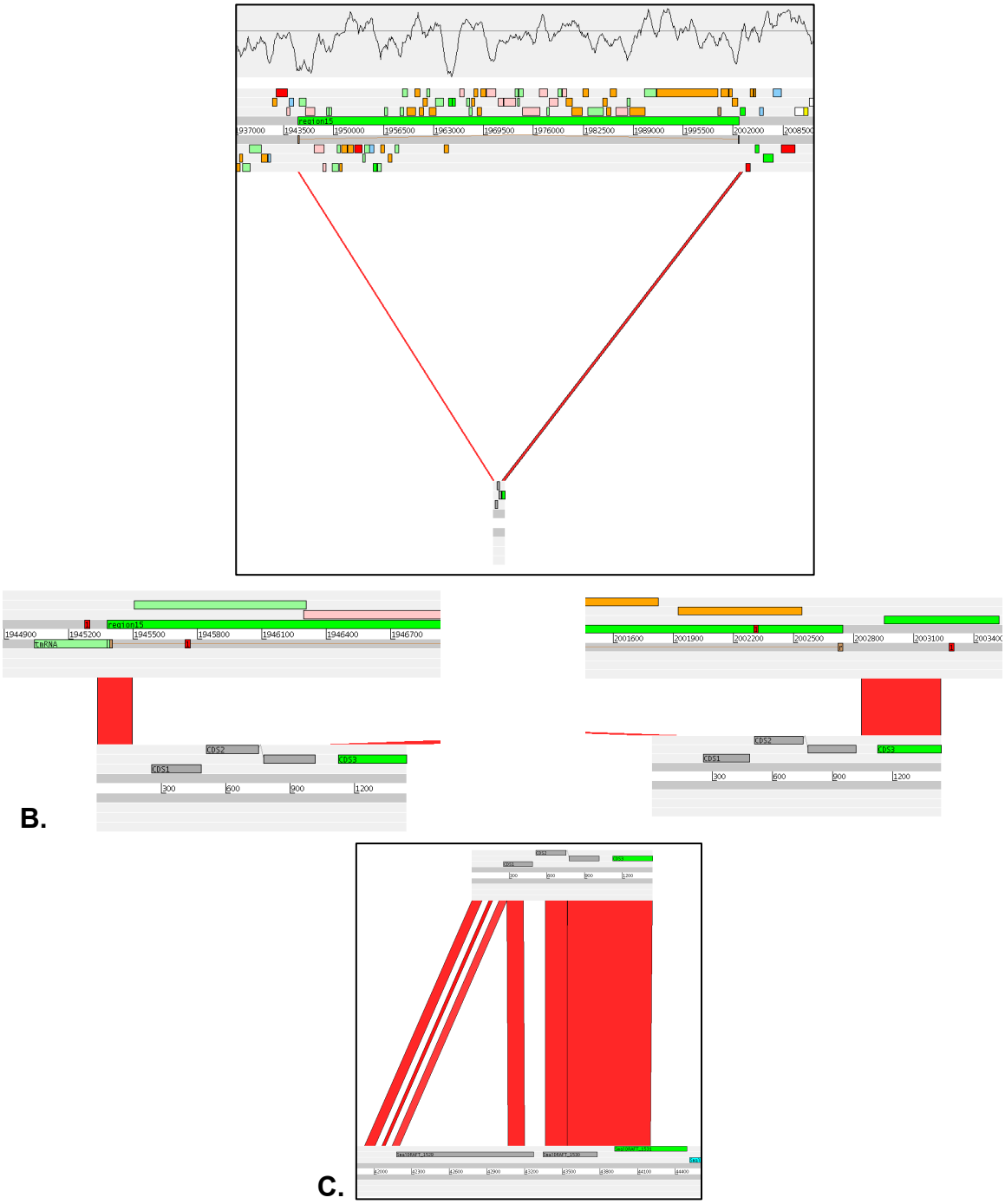


Figure 5.13: ACT screenshot: Predicted genomic island R15. Top: BLASTN comparison between *S. maltophilia* strain K279a and the sequenced fragment across the boundary site of R15 in strain 24 **(A)** and strain 42 **(B)**. Bottom: Higher resolution ACT screenshot; R15 is shown as green-coloured feature flanked by a set of 24bp DRs (brown coloured joined features). The two sets of primers used to amplify the left and the right boundaries of R15 are shown as red-coloured features flanking the left and right attachment sites of this island; the tmRNA gene, upstream of R15 is shown as a light-green coloured feature overlapping with the left boundary of this island. **C.** ACT comparison between the sequence across the boundary site of R15 in stain 42 and the corresponding sequence in *S. maltophilia* R551-3 ctg153 (Accession Number: AAVZ01000019).

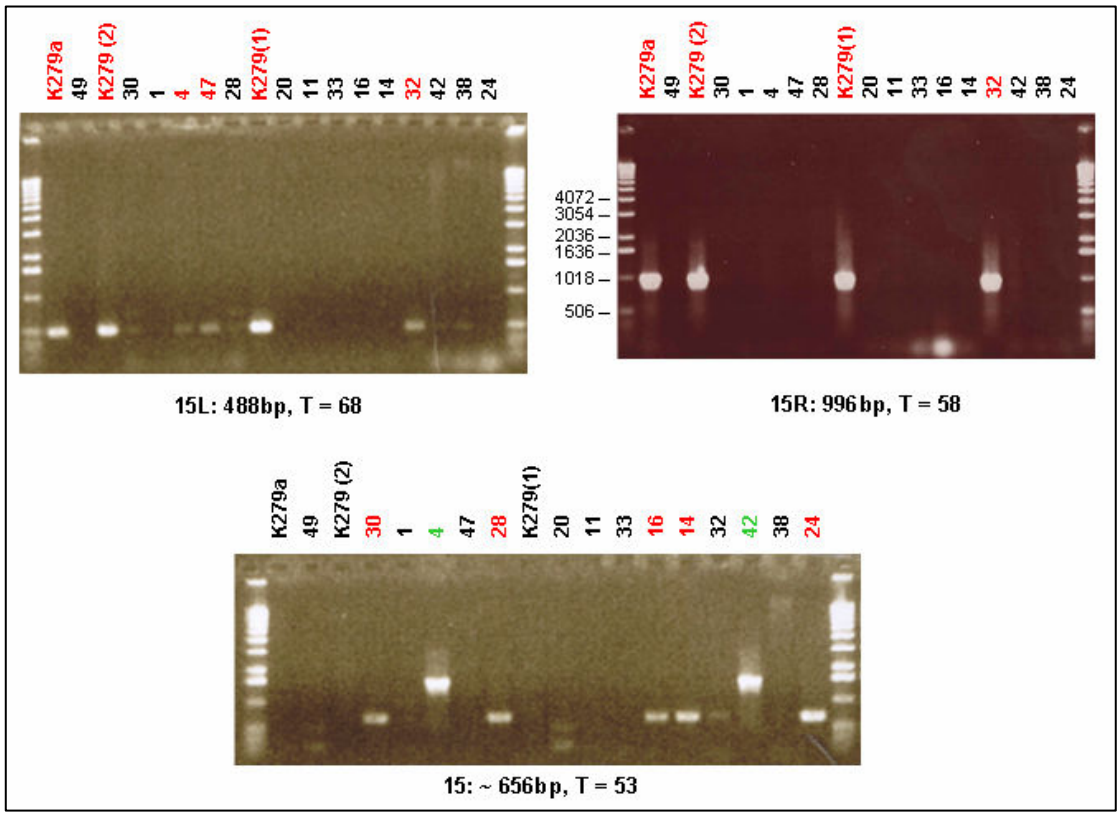


Figure 5.14: PCR amplification of the left (15L) and the right (15R) boundaries (top) of genomic island R15 and of the region across the boundary site of R15 in strains lacking the island (bottom). Colour scheme (bottom): Strains whose product size is ~656bp (expected size) are red coloured while strains with a larger product size ~ 1,500bp are green coloured.

5.3.1.7 Genomic Island 20

R20 is a medium size putative island of 18kb, low gene density (0.67) and very similar G+C content (65.8%) to the genome average (66.32%). R20 consists of 12 CDSs (Appendix K), encoding, among others, an autotransporter haemagglutinin-related protein (Smlt3829), two putative giant cable pilus-related proteins (Smlt3830 and Smlt3833), a putative outer membrane usher protein (Smlt3832) and a putative 50S ribosomal protein L31 (Smlt3836). R20 is flanked by a set of 24bp DRs (Figure 5.15) with the left DR being located immediately downstream of the termination codon of Smlt3827 (conserved hypothetical protein). The posterior probability of R20 of being a true GI is quite low (0.498).

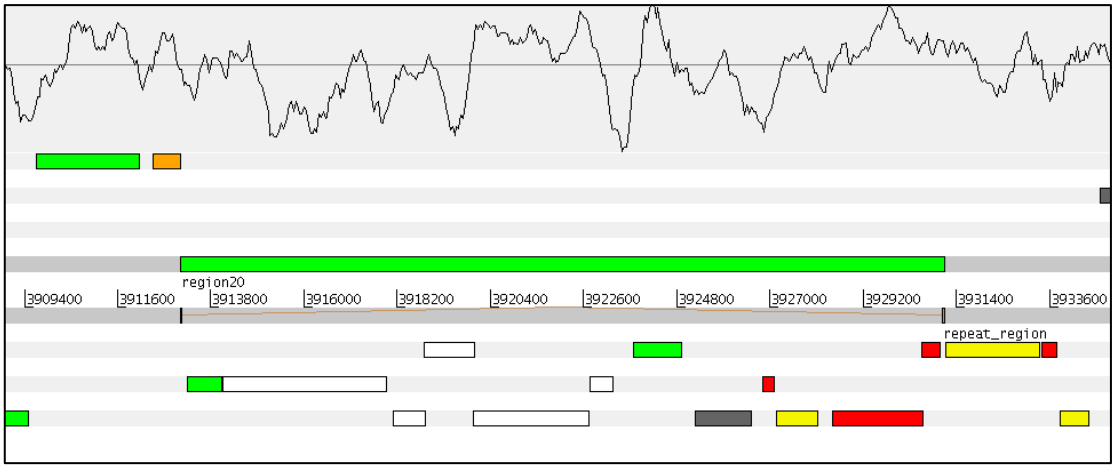


Figure 5.15: Artemis screenshot: Predicted genomic island R20, present in *S. maltophilia* strain K279a. The 24bp DRs are shown as brown-coloured joined features flanking the island. The G+C% content with a window size of 1kb is shown at the top of this screenshot.

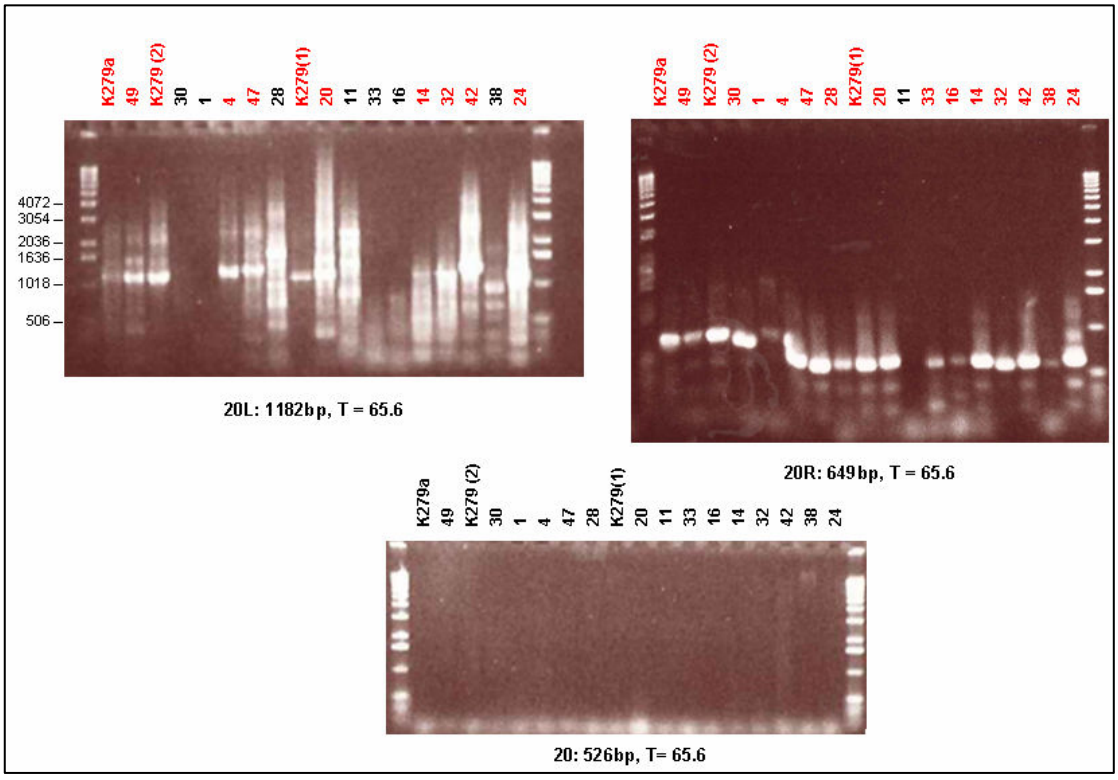


Figure 5.16 : PCR amplification of the left (20L) and the right (20R) boundaries (top) of genomic island R20 and of the region across the boundary site of R20 in strains lacking the island (bottom).

The fact that R20 region encompasses a major component (ribosomal protein L31) of the translation machinery (ribosome), in combination with its very low RVM score, makes it unlikely that this predicted region represents a true GI that has been horizontally acquired. Indeed, the PCR results suggest that this genomic region is present in the majority of the *S. maltophilia* strains used in this study (Figure 5.16), while the PCR across the insertion point of R20 did not indicate its absence in any of the 17 strains.

5.3.1.8 Genomic Island 7

R7 is a 30.4kb predicted island (Figure 5.17) of low G+C content (62.4%) and low gene density (0.86). R7 consists of 26 CDSs that encode proteins mainly involved in the lipopolysaccharide (LPS) biosynthesis (Appendix K) and represents a very unlikely GI structure with a very low posterior probability of 0.237.

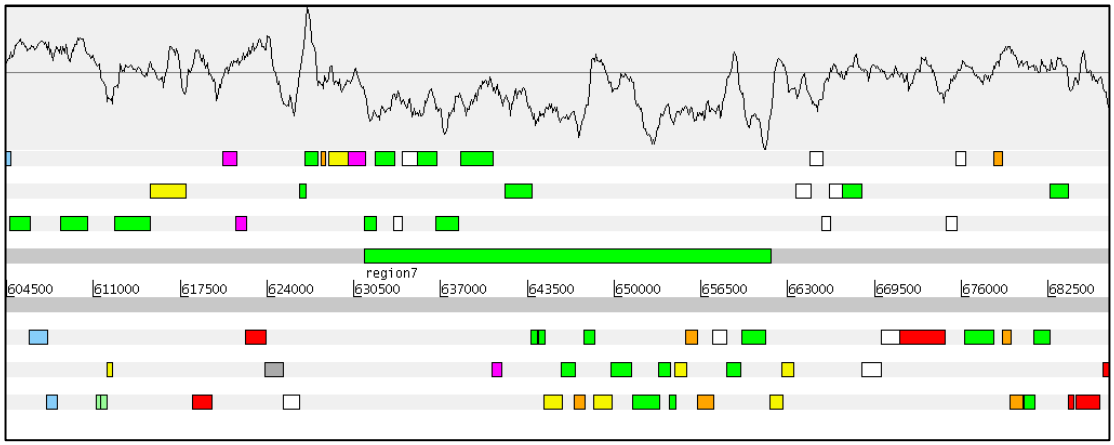


Figure 5.17: Artemis screenshot: Predicted genomic island R7, present in *S. maltophilia* strain K279a. The G+C% content with a window size of 1kb is shown at the top of this screenshot.

The presence of R7 was confirmed for the three K279 strains (Figure 5.18), although a small size variation of the left boundary of this island cannot be excluded in the case of strains 30 and 32; the absence of R7 was not identified in any of the 17 strains.

Failure to identify the left and right boundaries of this island in other strains, suggests that there is variation in the genes present in this locus; extensive variation in these types of loci is well known in other organisms (Bentley *et al.*, 2006).

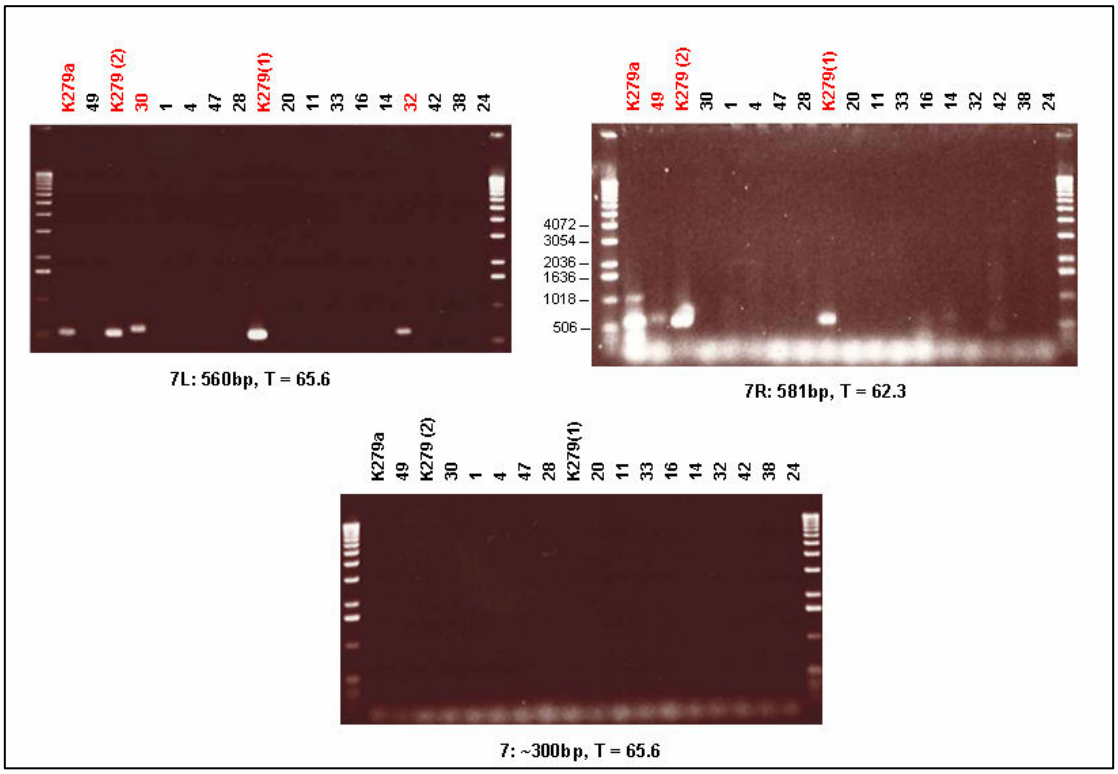


Figure 5.18: PCR amplification of the left (7L) and the right (7R) boundaries (top) of genomic island R7 and of the region across the boundary site of R7 in strains lacking the island (bottom).

5.3.2 Performance benchmarking

5.3.2.1 Prediction accuracy

In order to estimate the accuracy of this GI prediction pipeline I will make the following, three-fold assumption; a predicted region will be considered a true positive (TP) prediction if the following three conditions are met: A. For a predicted candidate GI a PCR product of the expected size, for the left and the right predicted boundary, is observed in at least one of the 17 un-sequenced strains; B. For the same candidate GI a PCR product, of the expected size, for the sequence across the predicted insertion point of this

GI, is observed in at least one of the 17 un-sequenced strains; C. The same predicted GI structure has a posterior probability (of being a true GI) higher than an arbitrarily determined threshold of 0.5.

If only conditions A and C are met, the predicted regions will be considered false positives (FP). If only condition A is met the predicted regions will be considered true negatives (TN). Finally if conditions A and B are met but condition C is not, the predicted regions will be considered false negatives (FN).

Exploiting the above rationale, we can get a naïve estimation of the predictive accuracy of the current pipeline, relying purely on an experimentally validated dataset of eight candidate GIs; five (R1, R12, R14, R15 and R16) of the eight candidate GIs represent TP, one region (R4) is a FP prediction and two (R7 and R20) are TN predictions, yielding a specificity of 0.83 (= TP/(TP+FP) = 5/6), a sensitivity of 1.0 (= TP/(TP+FN) = 5/5) and an overall accuracy of 0.875 (= (TP+TN)/(TP+TN+FP+FN) = 7/8).

5.3.2.2 Boundary accuracy

Assuming that the correct (observed) boundaries of a predicted candidate GI are the ones determined by sequencing across its insertion point in strains lacking the island, we can estimate the prediction accuracy of this methodology in terms of boundary optimization. In the current evaluation, I have used the absolute error defined as $\delta x = | x - x_0 |$, where x is the observed boundary determined by sequencing across the predicted insertion point in strains lacking the island (if applicable) and x_0 is the predicted one; the results (Table 5.6) show that the current methodology that integrates compositional-based (Alien_Hunter) and structural-based (RVM) prediction approaches gives a very small average, absolute error of 21bp; this number is significantly lower than the absolute error (3830bp) of Alien_Hunter (see section 2.3.3 of chapter 2) that relies purely on compositional information.

Table 5.6: Absolute error of the GI prediction pipeline (Alien_Hunter + RVM) for the predicted boundaries of the eight candidate islands.

Region	Boundaries				Absolute error (bp)	
	Left		Right		Left	Right
	Predicted	Observed	Predicted	Observed		
R1	60416	60293	70829	70894	123	65
R16	3089418	3089419	3127149	3127153	1	4
R4	299814	–	335480	–	–	–
R12	1323960	1323958	1367729	1367727	2	2
R14	1720126	1720130	1724413	1724413	4	0
R15	1945402	1945412	2002722	2002722	10	0
R20	3913072	–	3931089	–	–	–
R7	631285	–	661659	–	–	–
ALL (left/right)					28	14.2
ALL (left+right)					21.1	

5.4 Discussion

The aim of this analysis was three-fold. First, a blind-test exploiting an experimentally derived test-dataset of a single sequenced and 17 un-sequenced reference strains was carried out in order to sample the presence or absence of the predicted candidate islands in closely and distantly related *S. maltophilia* isolates; this approach would make it feasible to draw conclusions about the phylogenetic distribution of those putative GIs that in return would confirm or reject their horizontal origin.

Second, the integrative GI prediction pipeline described in this chapter was applied on the newly sequenced, un-annotated genome of *S. maltophilia*, strain K279a and used as a complementary methodology to the annotation pipelines developed in the pathogen sequencing unit (PSU) at the Sanger Institute. Predictions of putative GI structures were used to infer the likely origin of the initially un-annotated CDSs, overlapping with these predictions, as well as to more accurately determine the true boundaries of partially annotated putative horizontally acquired regions. Furthermore, while this anylisis was still in progress, the gene-content information derived from the ongoing annotation of K279a genome put the

predicted insertion point of GIs into context; for example, in the case of R1 a set of 18bp DRs were predicted to flank the boundaries of this GI, and based on the gene prediction and subsequent manual curation, it was inferred that the insertion point of R1 was within the coding sequence of Smlt0055; this further suggests that *in silico* predictions and experimental protocols can mutually benefit from each other.

Third, the generalization properties of this prediction pipeline, which integrates compositional-based and structural-based techniques, in making accurate predictions for previously unseen examples of a newly sampled genomic dataset were evaluated, relying purely on an experimental rather than an *in silico* based benchmarking approach. This analysis evaluated two specific properties of the current GI prediction approach; how reliably this methodology predicts GIs in newly sequenced genomes and, for the predictions that are true positives, how accurately their boundaries can be determined.

For a sample of eight candidate GIs with a posterior probability range of 0.2371–0.9997, the data confirm that over half (5/8) of the predictions are likely to be true GI structures that have been probably acquired very recently in the lineage of the three *S. maltophilia* K279 strains. Moreover, the experimental validation of two, very low scoring (0.4983 and 0.2371) predicted GI structures (R20 and R7) suggests that those regions are probably not real GIs, in line with their very low posterior probability; these data confirm the increased specificity of the proposed method in reliably predicting true GIs.

Although the experimental methodology described in this chapter, along with the performance benchmarking, gives results showing a very good overall prediction accuracy for the described approach, even in the case of a previously unseen genomic dataset, there are several obvious limitations affecting the conclusions drawn from this analysis, that have to be taken into account.

Overall the experimental PCR protocol as implemented in the current analysis suffers from low resolution. Firstly, probing the presence

or absence of the putative GIs, under the given methodology, is feasible only if the sequence of their predicted boundaries is highly conserved among the reference K279a strain and the 17 un-sequenced *S. maltophilia* strains.

Theoretically speaking, in the case of more distantly related strains this methodology would not necessarily give amplified products for the sequence that corresponds to the predicted GI boundaries since the low level of sequence similarity would prohibit the binding of the corresponding primer set to its genomic DNA template; however, because of the second assumption (section 5.2.3) of the experimental methodology exploited in this analysis, the requirement for an "a+d" amplicon acts as a control, since in the case of distantly related strains, the "a" and "d" primers will also fail to bind to the DNA template and give an amplified product.

An alternative PCR approach that could overcome this limitation, would involve the design of degenerate primers that would allow sequence ambiguity between the primers and the template. However, the results of this analysis suggest that this is probably not the case for the given genomic dataset of the 18 *S. maltophilia* strains since PCR amplified products, of the expected size, are successfully produced even in the case of distantly related isolates. For example the results in Figure 5.10 show that for phylogenetically distantly related strains (Figure 5.2), e.g. strain 11 (group IV) and 20 (group II) a PCR product of the expected size for the sequence across the insertion point of R12 was successfully obtained.

Secondly, this methodology does not provide any information about the actual gene content, size and internal structural variation of GIs inferred to be present in any of the 17 un-sequenced strains. For example a predicted GI of putative phage origin might have similar bacteriophage integrase and tail protein coding CDSs at the two boundaries with an inferred "identical" GI structure present in some of the un-sequenced genomes; clearly prophages of different type or family can have high sequence similarity at those flanking CDSs but do not necessarily

represent the same prophage. In other words sequence similarity at the predicted boundaries between genomic regions present in different strains neither guarantees that those regions are of the same origin, or gene content, nor does it exclude internal size variation, e.g. in the case of GI remnants, or deletions.

An alternative, more sophisticated approach that would overcome those limitations is the Southern blotting protocol (Southern, 1975) that exploits a probe hybridization principle; however such a methodology is out of the scope of this analysis, for reasons discussed at the beginning of this chapter; it is worth mentioning that the protocol used in this analysis was only devised to check the predicted boundaries of GIs and not to completely explore the content of the GIs.

Thirdly, in the case of probing the absence of a given candidate GI in some of the un-sequenced strains by seeking to amplify the sequence across the predicted insertion point of this GI, again this methodology will fail to give an amplified product if a different GI has been inserted at the corresponding insertion point in the target strains. In that case, we will not be able to infer that the reference GI is absent from the target strains, although this is clearly the case. An alternative methodology would involve a long-range PCR protocol that could amplify longer genomic regions; however the results of this approach would still be conditional on the size of the intervening sequence between the left and the right ends of the corresponding insertion point.

Clearly the current experimental methodology exploits very simple concepts and principles and as such it provides a very rough evaluation of the true strengths and weaknesses of the discussed *in silico* pipeline. Nonetheless, this analysis forms a proof of concept that the *in silico* prediction of GIs can be integrated successfully in experimental methodologies and gives data suggesting that some of the *in silico* predictions have probably limited phylogenetic distribution and represent putative recent horizontal gene transfer events in the *S. maltophilia* lineage.

Moreover, the data presented in the current analysis show that prediction pipelines that merge compositional-based (low-level) with structural-based (high-level) approaches can yield more reliable predictions of putative GIs compared to methodologies exploiting either of those approaches. Overall it can be concluded that *in silico* prediction methods, relying on and exploiting a minimum level of pre-existing annotation, can be very powerful tools in aiding or guiding, in a high-throughput fashion, the annotation pipelines of microbial genomes (see next chapter).