



Strathclyde Institute for Pharmaceutical and Biomedical Sciences
Faculty of Science

Investigating the Genome Complexity of
Streptomyces clavuligerus

Lis Algora Gallardo

Thesis presented in fulfilment of the requirement for the degree of
Doctor of Philosophy

2020

Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50.

Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Signed:

A handwritten signature in blue ink that reads "Lis Algora". The signature is written in a cursive style with a large, sweeping flourish at the end.

Date:

24th August 2020

Acknowledgments

The work presented on this thesis was achieved thanks to the support of numerous people that I would like to acknowledge here. Thanks everyone for being part of this adventure.

First, I would like to express my sincere gratitude to my supervisor Dr Paul Herron, who has been a constant source of support and inspiration, giving me plenty of opportunities to grow both professionally and personally. Thank you for trusting and believing in me during these four years and developing my passion for microbiology.

I thank my funders IBioIC and GSK for making this project possible. I would like to thank Ben, Steve, Nic and Andrew (GSK) particularly for their constant interest and valuable insight towards my work.

I would like to acknowledge our collaborators at the John Innes Centre: Dr Barrie Wilkinson and Dr Juan Pablo Gómez Escribano for all the help and material that heavily contributed to some of my experiments.

I thank the current and former members of the Herron group: Jana, Tiago, Gillian, David, Gaetan and Elmira for all the help and fun times in the lab bay. I would like to mention the rest of the *S. clavuligerus* group Anna, John and Kirsty, it has been a pleasure working with you, thanks for all the good times, I have learnt so much from you. I would also like to acknowledge the rest of the people that have been involved in the micro-lab during these years including PIs, postdocs and students. Special thanks to Nadine, Jordan T., Adam, Walid, Emily, Derek, Gordon, Ally, Charlie, Josie, Parra, Adriana, Eilidh and Molly for all the amazing moments, you have made these the best years of my life. I would like to give special mention to my year colleagues and lovely friends Ainsley, Becca and Liam. Thanks for all the support, laughs, help and great memories, I will always cherish our friendship.

Me gustaría agradecer todo el apoyo recibido por parte de mi familia y amigos en España. Gracias a mis padres por todo el amor incondicional, libertad y valores inculcados que se han mantenido en la distancia. A mi hermana Camelia y mi hermano David por inspirarme de maneras tan diversas durante toda mi vida. Y a mis tres madrileñas favoritas Gis, Lidia y Lorena, gracias por vuestro amor y comprensión durante estos años, me quedo siempre con vosotras.

Lastly, my most sincere thanks to Jonny. Or maybe I should thank the PhD for bringing you to my life. You have been constant source of support and positivity that has made this thesis possible. Thanks for your unconditional love and believing so much in me. I would only hope to do the same in your journey to become a doctor. Thanks for being my chosen family.

Contents

Abstract	1
1. Introduction	2
1.1 Clavulanic acid: an agent against antimicrobial resistance	3
1.2 <i>Streptomyces clavuligerus</i> : the clavulanic acid producer	4
1.2.1 The genus <i>Streptomyces</i>	7
1.2.2 Biosynthesis of clavulanic acid	7
1.2.3 The genome of <i>S. clavuligerus</i>	8
1.3 The role of terminal proteins in replication of <i>Streptomyces</i> telomeres.	13
1.4 Scope of the project.....	20
1.5 Specific aims.....	21
2. Materials and Methods	22
2.1 Media and antibiotics	22
2.2 Microbial strains	24
2.3 Plasmids	25
2.4 Primers	26
2.5 Bacterial cultivation	27
2.6 Pulsed-field Gel Electrophoresis (PFGE).....	28
2.6.1 Plug preparation	28
2.6.2 Plug digestion with restriction enzymes	28
2.6.3 Loading the gel.....	29
2.6.4 Electrophoresis and gel Imaging	29
2.6.5 Alternative PFGE protocol using a phenol-chloroform pre-treatment.....	29
2.6.6 Alternative PFGE protocol using HEPES buffers.....	29
2.6.7 PFGE gel images process and fragment size estimation.....	30
2.7 Genomic DNA isolation	30
2.8 Whole genome sequencing of <i>S. clavuligerus</i> strains and bioinformatics	31
2.9 Purification and sequencing of telomeric sequences	34
2.10 Bioinformatic analysis of terminal protein sequences	34
2.11 Southern blotting	35
2.11.1 Probe labelling	35

2.11.2 DNA digestion and electrophoresis	35
2.11.3 DNA transfer to nylon membrane	36
2.11.4 Probe hybridisation.....	36
2.11.5 Stringency washes.....	36
2.11.6 Detection.....	36
2.12 Molecular biology and cloning.....	37
2.12.1 DNA amplification	37
2.12.2 Purification of DNA amplicons and plasmids	37
2.12.3 Agarose gel electrophoresis.....	38
2.12.4 Ligation.....	38
2.12.5 Transformation of <i>E. coli</i> competent cells	38
2.12.6 Restriction digestion	38
2.12.7 Preparation and transformation of chemical competent cells.....	39
2.12.8 Intergenic conjugation from <i>E. coli</i> to <i>S. clavuligerus</i>	39
2.12.9 <i>In silico</i> cloning.....	39
2.13 Bacterial Two-Hybrids experiments.....	39
2.14 Microscopy.....	40
2.15 Growth curves.....	41
2.16 Clavulanic acid production assay	42
2.17 Bioactivity plug assay	42
3. Optimisation of Pulsed-field Gel Electrophoresis Method for the Study of the Genomic Architecture of <i>S. clavuligerus</i>	44
3.1 Establishment of a PFGE protocol for <i>S. clavuligerus</i>	46
3.2 Lysis Time is not a significant factor for PFGE of <i>S. clavuligerus</i>	48
3.3 Phenol-chloroform pre-treatment of cells does not reduce DNA degradation on <i>S. clavuligerus</i> samples.	50
3.4 Substitution of Tris by HEPES in buffers composition significantly increases the quality of PFGE analyses on <i>S. clavuligerus</i>	50

3.5. Higher amounts of mycelium give stronger signal bands on PFGE analyses but less specificity for Southern blotting.	53
3.6 Mycelium harvested within 24 hours of inoculation gives the best results on PFGE analyses of <i>S. clavuligerus</i> samples.....	56
3.7 Southern Blot analyses of PFGE gels confirm the band of 1.8 Mb as pSCL4.	58
3.8 The final physical map of <i>S. clavuligerus</i> DSM 738 genome.....	62
3.9 Summary.....	64
4. Closure of the <i>S. clavuligerus</i> DSM 738 Genome Sequence	67
4.1 Combining Next Generation technologies Illumina and PacBio obtains closed sequences of <i>S. clavuligerus</i> DSM 738 chromosome and plasmids.	70
4.2 <i>S. clavuligerus</i> DSM 738 carries several copies of the archetypal telomere found in F613-1 strain sequence.....	76
4.3 The telomeres of <i>S. clavuligerus</i> DSM 738 chromosome and plasmids are identified using the self-ligation method.....	80
4.4 The genome of <i>S. clavuligerus</i> DSM 738 carries four different sets of archetypal telomeres.....	86
4.5 The read coverage depth indicates a higher copy number of pSCL2 and lower copy number of pSCL4 and pSCL3 in reference to the chromosome.....	92
4.6 General characteristics and annotation of the genome of <i>S. clavuligerus</i> DSM 738 ..	95
4.7 The genome of <i>S. clavuligerus</i> carries 43 putative biosynthetic gene clusters, 17 of them are located on the megaplasmid.....	100
4.8 Whole genome sequence comparison suggest <i>S. clavuligerus</i> DSM 738 carries the same chromosome and megaplasmid as the ATCC 27064 Strain.....	107
4.9 The genetic architecture of <i>S. clavuligerus</i> suggests chromosomal gene complementation by plasmids.....	109
4.10 Summary	113
5. The Terminal Proteins of <i>S. clavuligerus</i>	117
5.1 Characterisation of three sets of terminal proteins in the genome of <i>S. clavuligerus</i>	118
5.2 A copy of pSCL4 <i>tap-tpg</i> operon was incorporated on <i>S. clavuligerus</i> chromosome.....	124
5.3 Homologous recombination did not allow elimination of <i>tap</i> and <i>tpg</i> from pSCL4..	128
5.4 BACTH experiments confirm interaction of the Tap protein encoded on pSCL4 with itself.....	136
5.5 Interaction of the Tpg protein with itself was confirmed by BACTH experiments using the pSCL2 and pSCL3 proteins.	142

5.6 The terminal proteins from the three plasmids pSCL4, pSCL3 and pSCL2 do not interact with each other.	147
5.7 Summary.	151
6. Essentiality of the Megaplasmid pSCL4 in <i>S. clavuligerus</i>	155
6.1 Elimination of the megaplasmid pSCL4 from <i>S. clavuligerus</i>	156
6.2 Whole genome sequencing of pSCL4-free strains confirms loss of plasmid and chromosome ends.	164
6.3 Translocation of 50 kb region from pSCL4 to pSCL3 on strain M2.	171
6.4 Identification of <i>attB</i> site on the <i>S. clavuligerus</i> chromosome for ϕ BT1 integration.	173
6.5 Strains M1 and M2 carry circular chromosomes.	176
6.6 Identification of the missing annotations on the pSCL4-free strains.	182
6.7 Characterisation of the growth of plasmid-free strain in liquid cultures.	185
6.8 Microscopic characterisation of pSCL4-free strain confirms they present shorter, thinner, and more frequent branches.	187
6.9 The pSCL4-free strains M1 and Mc3 exhibit decreased clavulanic acid production.	198
6.10 Is the megaplasmid conjugative?.....	203
6.11 Summary.	206
7. Discussion	211
7.1 Combined physical and genetic map of the <i>S. clavuligerus</i> genome.....	211
7.2 Three sets of terminal proteins for five sets of telomeres.	213
7.3 pSCL4, a dispensable, yet essential plasmid of <i>S. clavuligerus</i>	217
8. Conclusions & Future Work.....	219
8.1 Conclusions.	219
8.2 Future work.....	221
9. References	222
10. Appendix.....	a
10.1 Custom scripts.....	a
10.2 Supplementary figures.....	b
10.3 Supplementary tables	h

ABSTRACT

Streptomyces clavuligerus is the producer of clavulanic acid; a β -lactamase inhibitor that is used in combination with the antibiotic amoxicillin to treat many bacterial infections. The unique genome of *S. clavuligerus* includes a chromosome, and four linear plasmids pSCL1, pSCL2, pSCL3 and the megaplasmid pSCL4, which is the largest linear plasmid ever identified at 1.8 Mb in size. In order to study the genomic architecture of *S. clavuligerus* strains, we established an optimised method of pulsed-field gel electrophoresis, which allowed visualisation of the three giant linear plasmids pSCL2, pSCL3 and pSCL4, as well as chromosomal fragments in order to establish the physical map of the *S. clavuligerus* genome. Furthermore, to establish a foundation on which the genomic analyses of industrial strains of this species could be carried out, we closed the genome sequence of the *S. clavuligerus* type strain DSM 738, using a combination of short-read Illumina sequencing and long-read PacBio technology. Additionally, the telomeres of the five replicons were purified by a self-ligation method, and identification of the sequences confirmed that the five replicons carry archetypal telomeres, with the chromosome and megaplasmid sharing identical telomeric sequences. Streptomyces telomeres are maintained by the terminal proteins Tap and Tpg, however, while copies of the genes encoding these proteins were identified in pSCL4, pSCL3 and pSCL2, the chromosome carries none, making these the only known essential elements missing from the chromosome. Moreover, in order to investigate the essentiality of the megaplasmid in *S. clavuligerus*, we cured pSCL4 by targeting *parB* with CRISPR/Cas9. This caused deletion of the chromosomal ends and circularisation, confirming that pSCL4 Tap-Tpg are required for maintaining the chromosome telomeres, and cannot be complemented by the proteins encoded on the other two plasmids. In addition, pSCL4-free strains were confirmed to be non-sporulating as well as presenting a phenotype of short and thin hyphae, in addition to higher branching frequency. Some mutant strains also exhibited reduced clavulanic acid production which confirmed that the loss of pSCL4 has important effects on morphological development and secondary metabolism of *S. clavuligerus*.

Overall, with this study we have closed the genome sequence of *S. clavuligerus* type strain and optimised a method for rapid screening of the *S. clavuligerus* mutant strains that will greatly benefit the industry for development of production strains. Additionally, we have confirmed that pSCL4 is a dispensable plasmid, but is yet necessary for maintaining optimum fitness of this bacterium.

1. INTRODUCTION

Bacterial infections have been one of the leading causes of morbidity and mortality throughout history. Despite the development of modern medicine, they still represent a great concern for public health worldwide, with infectious diseases such as tuberculosis, meningitis and pneumonia that impair or take countless lives every year. Antibiotics have been the principal form of therapy for bacterial infections since the 1940s, reducing significantly the severity and mortality of infectious diseases ever since. Antibiotics originate naturally from the secondary metabolism of other microorganisms as a mechanism of defence, and represent a key factor in the evolution of prokaryotes (Gupta, 2011). The discovery of the first antibiotic penicillin by Alexander Fleming in 1928, and its clinical success against Gram-positive bacteria, led to the study of other antibiotic-producing organisms that prompted the discovery of many clinically important compounds and their subsequent development to an industrial scale. Members belonging to the phylum of Actinobacteria are highly prolific in producing bioactive metabolites, to the point of producing two-thirds of the available antimicrobials (Watve *et al.*, 2001), establishing actinomycetes as “nature’s pharmacists” for their large contribution to control and prevent many infectious disease during the past few decades. Nonetheless, the high rate of antibiotic discovery added to the extended overuse of antimicrobial drugs during the last few decades have resulted in emergence of antibiotic resistance, in such species as *Staphylococcus aureus*, *Klebsiella pneumoniae* and *Pseudomonas aeruginosa*, which caused severe disease outbreaks in recent decades (Boucher *et al.*, 2009). Antibiotic resistance is a global problem that is responsible for 700,000 deaths per year, and it is expected to rise to over ten million deaths by 2050 (O’Neill, 2016), representing one of the greatest threats to human health at the moment. In order to face the problem of multi-drug resistant infections, there are several factors that can be taken into consideration: firstly, the demand of antibiotic drugs should be reduced by taking actions to promote rapid diagnostics and the use of vaccines, increasing public awareness, and reducing unnecessary use in agriculture; and secondly, by increasing the number of effective antibiotics by promoting the investment of discovery of new compounds and improving the effectivity of existing ones. Unfortunately, the discovery of naturally occurring antimicrobial compounds has diminished over the last few decades and researchers have to find ways to stimulate the production of new and better bioactive compounds by, for example, metabolic engineering of producing strains (Weber *et al.*, 2015). Furthermore, the study of antibiotic-producing microorganisms, such as streptomycetes has

led to the discovery of mechanisms that counteract the action of antibiotic resistance, such as the production of β -lactamase inhibitors. The exhaustive study of organisms able to inhibit antibiotic-resistance mechanisms is a promising approach to contribute to the fight against multi-drug resistant pathogens.

1.1 Clavulanic acid: an agent against antimicrobial resistance.

β -lactam antibiotics are considered one of the most important biotechnological products. They represent around 65% of the total number of clinically used antibiotics in the world and include penicillins, cephalosporins, carbapenems and monobactams (Elander, 2003). They all share a common structural feature: the β -lactam ring, which confers a three dimensional shape to the molecule, necessary for the antimicrobial activity. The mechanism of action is based on the inhibition of enzymes necessary for the synthesis of the bacterial peptidoglycan wall (Goffin and Ghuysen, 1998). This creates wall permeability and loss of cellular homeostasis, which leads to activation of autolysins and eventually, death of the cell (Heidrich *et al.*, 2001).

The structure of β -lactam antibiotics has undergone a continuous evolution since their first introduction, driven by the emergence of bacteria resistant to each generation of novel β -lactam antibiotics. The introduction of changes in the structure, enabled pharmaceutical companies to stay ahead in the race with resistant bacteria for a brief period of time (Kong *et al.*, 2010). However, the great selective pressure placed by antimicrobial agents led to appearance of highly resistant microorganisms, which endangers the continuation of this class of drug in clinical medicine, as well as posing a serious risk to public health worldwide.

Resistance to β -lactam antibiotics is mediated by the ability to produce **β -lactamases**, which are enzymes that inactivate these compounds irreversibly by hydrolysing the β -lactam ring (Abraham and Chain, 1940). More than 300 types of β -lactamases are produced by clinically important microorganisms such as *S. aureus*, *K. pneumoniae*, *P. aeruginosa* and *Acinetobacter baumannii* (Ambler, 1980; Majiduddin *et al.*, 2002). However, the action of β -lactamases can be counteracted by utilising β -lactamase inhibitors that bind irreversibly to the active site of the enzyme (Bush, 1988), such as the commercially available clavulanic acid, tazobactam or sulbactam.

Clavulanic acid was the first β -lactamase inhibitor to be discovered, and also the first to be introduced to the clinic (Reading and Cole, 1977). It was isolated from ***Streptomyces***

clavuligerus in the 1970s as it inhibited the growth of a β -lactamase-producing strain of *Klebsiella aerogenes* in the presence of Benzylpenicillin (Brown *et al.*, 1976). Clavulanic acid belongs to the family of clavams and differs from the penicillin by having an oxygen instead of a sulphur in the β -lactam ring (**Figure 1.1**). The mechanism of action is related to the unique 3R, 5R stereochemistry of the molecule that enables the irreversible binding of the β -lactam ring to the hydroxylic group of a serine residue in the active site of the enzyme, resulting in a stable acylated metabolite that is inactive against β -lactams (Liras and Rodríguez-García, 2000). It is particularly effective against penicillinases and cephalosporinases (Neu and Fu, 1978).

Although clavulanic acid alone exhibits low antimicrobial activity, when combined with β -lactam antibiotics it enables the restoration of the antibiotic action against β -lactamase-producing pathogens (**Figure 1.2**). Amoxicillin-clavulanic acid combinations have been evidenced to effectively inhibit the growth of *S. aureus*, *Escherichia coli*, *Klebsiella pneumoniae* and *Haemophilus influenza* among others (Bush and Johnson, 2000). Combination of amoxicillin-clavulanate has been commercially available since 1981, under the name of **Augmentin**[®] by GSK (GlaxoSmithKline), and it is commonly used to treat respiratory tract infections such as community-acquired pneumonia, otitis media and chronic bronchitis (White *et al.*, 2004).

1.2 *Streptomyces clavuligerus*: the clavulanic acid producer.

The extension of antimicrobial resistance has driven the investigation of strains able to produce inhibitors. To date, clavulanic acid is produced by three species of *Streptomyces*: *S. clavuligerus*, *S. jumonjinensis* and *S. katsurahamanus*, although *S. clavuligerus* is considered the model system as it is the most widely studied and utilised bacterium for industrial production of clavulanate (Jensen and Paradkar, 1999). *S. clavuligerus* was first described in 1971 by Higgins and Kastner, who were conducting a study of β -lactam-producing strains, and revealed its ability to produce penicillin N and cephamycin C (Higgins and Kastner, 1971). Since then, it has been the subject of many studies for its antibiotic, antifungal activities, as well as β -lactamase-inhibition, which was first observed in 1976 as previously mentioned (Brown *et al.*, 1976). Images of *S. clavuligerus* mycelium are presented on **Figure 1.3**.

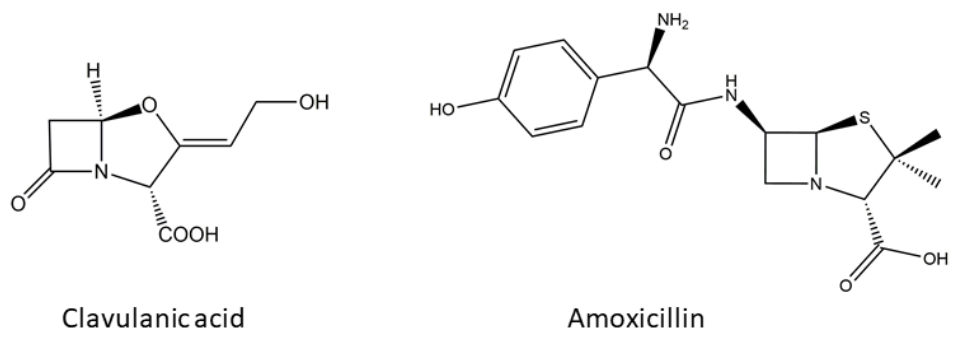


Figure 1.1. Chemical structures of the β -lactam antibiotics clavulanic acid and amoxicillin

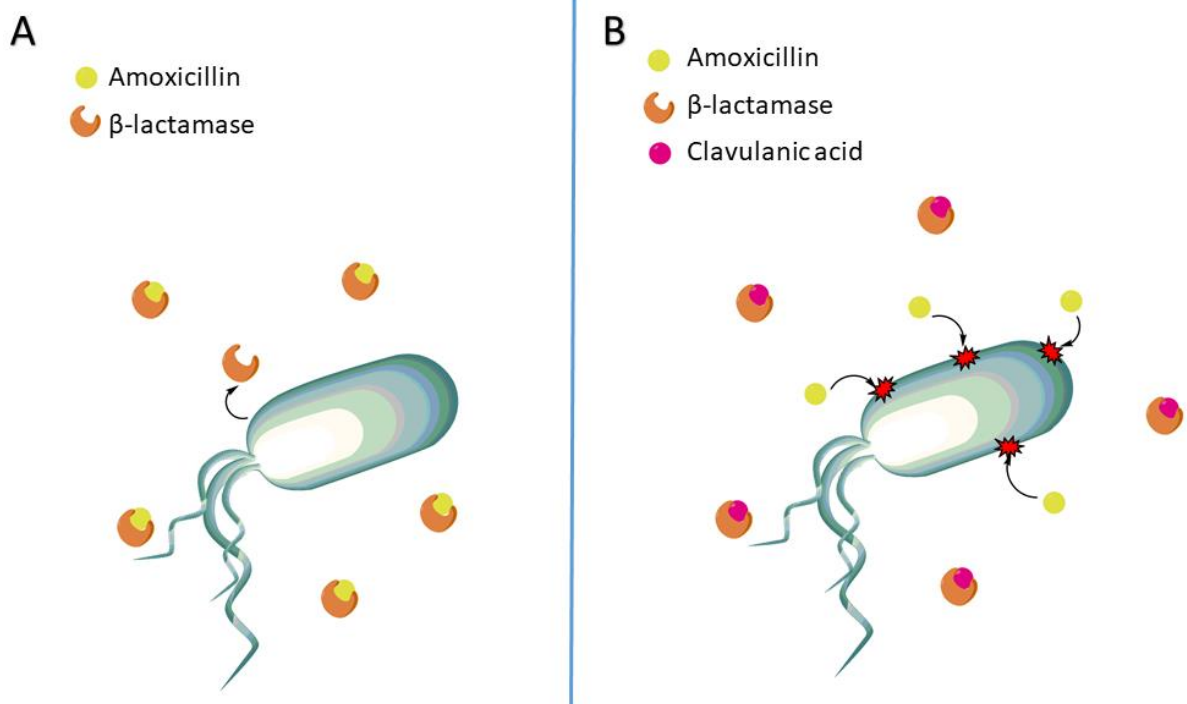


Figure 1.2. Schematic overview of the treatment with amoxicillin alone (A) and amoxicillin + clavulanic acid (B) on β -lactam resistant bacteria. **A:** The bacteria are able to inactivate amoxicillin molecules by the production of β -lactamases that bind to the antibiotic. **B:** Clavulanic acid restores amoxicillin activity by inhibiting the β -lactamases.

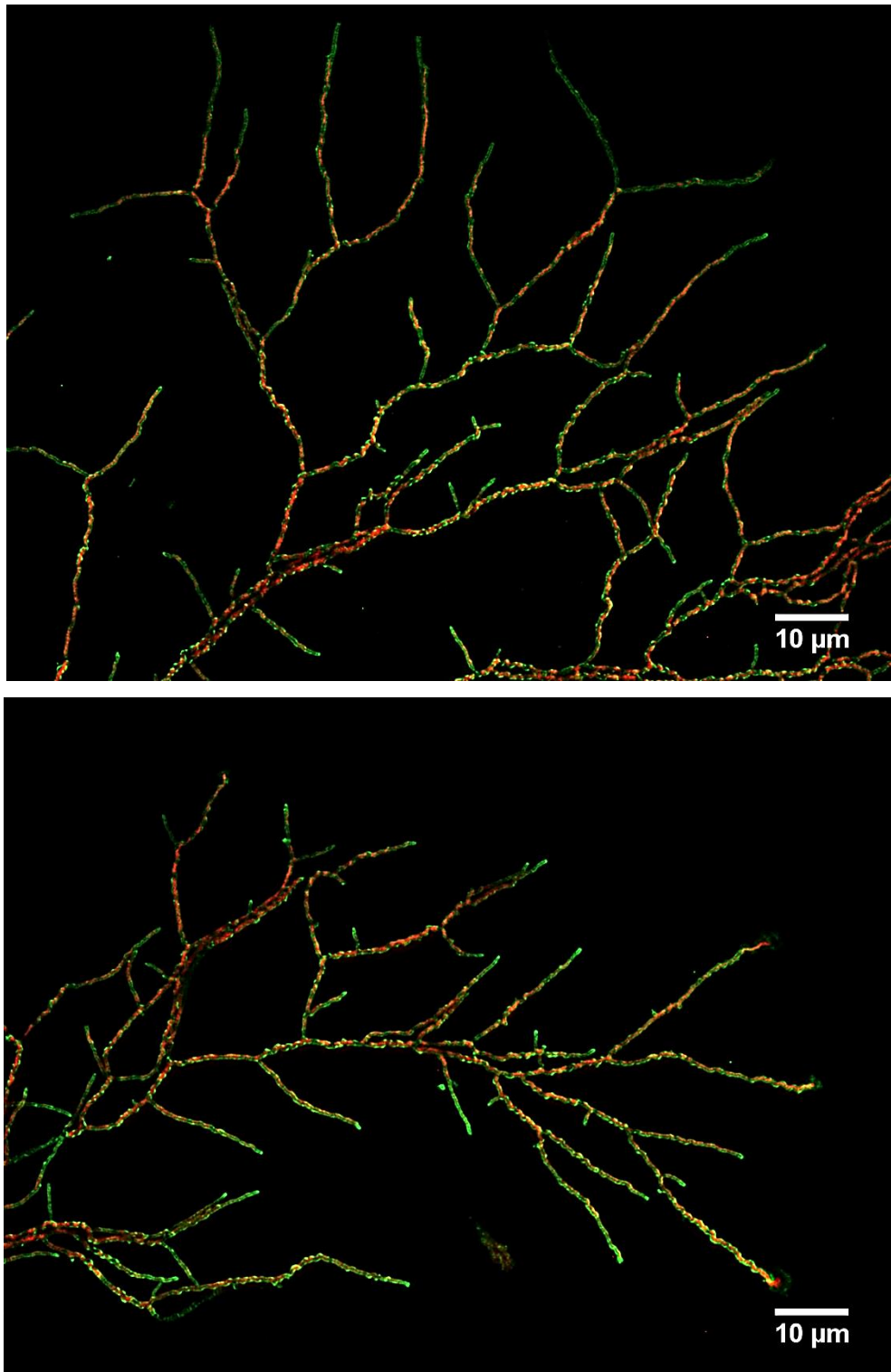


Figure 1.3. Fluorescence microscopy images of *S. clavuligerus* LEON strain mycelium. Peptidoglycan was stained with fluorescein isothiocyanate-wheat germ agglutinin (FITC-WGA) and is presented in green under a FITC emission filter. Nucleic acids were stained with propidium iodide and are presented in red under a 600 nm emission filter.

1.2.1 The genus *Streptomyces*.

S. clavuligerus belongs to the phylum Actinobacteria, family *Streptomycetaceae*. Being the largest genus of the actinobacterial phylum, *Streptomyces* include heterogeneous Gram-positive bacteria with linear chromosomes rich in guanine and cytosine (G+C) (Labeda *et al.*, 2012). They are multicellular, filamentous, sporulating organisms commonly found in soil where they exhibit a complex life cycle (Kämpfer, 2006). The life cycle starts by germination of dormant spores when nutritional and environmental conditions are favourable, developing the vegetative or substrate mycelium by apical extension of the hyphae and branching. During cell division, hyphal cross-walls are formed that separate compartments with multiple copies of the chromosome. These compartments are elongated by synthesis of peptidoglycan cell wall material at the lateral walls. Aerial mycelia develop when nutrients start to deplete, a process controlled by the *bld* genes, that give colonies a white appearance. Sporulation eventually takes place by the expression of *whi* genes that induce segmentation and coiling of the hyphae into chains of spores (Piette *et al.*, 2005; Chater, 1993).

bld and *whi* are among the genes responsible for the transition from primary to secondary metabolism. The primary metabolism involves the essential processes for the growth of the bacteria that are common in streptomycetes, including the uptake, transport and conversion of diverse metabolites and nutrients such as carbon and nitrogen (Barka *et al.*, 2016). However, the importance for the study of *Streptomyces* lies on their extensive secondary metabolism, which, unlike primary metabolism, is specialised and species-specific, and responsible for the production of clinically-relevant bioactive metabolites such as antibiotics whose expression is regulated by global primary regulators and coordinated with the morphological differentiation. These secondary metabolites are encoded in gene clusters within the genome (Liu *et al.*, 2013). *S. clavuligerus* in particular is estimated to harbour 48 of these putative secondary gene clusters (Medema *et al.*, 2010), which include those coding for bioactive compounds such as cephamycin C, a β -lactam antibiotic, the antitumor agent holomycin, as well as, clavulanic acid and other 5 types of clavams (Baggaley *et al.*, 1997; Li and Walsh, 2010).

1.2.2 Biosynthesis of clavulanic acid.

The clavulanic acid biosynthetic gene cluster is located on the chromosome, adjacent and downstream to that which codes for cephamycin C. Expression of both metabolites is controlled by the same type of regulators, SARP (*Streptomyces* antibiotic regulatory proteins)

(Paradkar, 2013). The biosynthesis of clavulanic acid, in particular, is controlled by a series of genes that are organised hierarchically (Liras, *et al.*, 2008). The first level genes, *ccaR* and *claR*, are regulated by global cellular mechanisms such as those involved in morphological differentiation including production of aerial mycelium and sporulation in response to nutritional conditions and environmental imbalances. *bld* genes, and production of autoregulatory γ -butyrolactones are involved in the regulation of these genes (Paradkar, 2013). *ccaR* located in the cephamycin C cluster encodes the SARP regulator and regulates the expression of “early” genes and, at the same time, the transcription of *claR* into a LysR-type regulator that activates the “late” genes in the clavulanic acid biosynthesis pathway (**Figure 1.4B**). The early genes are duplicated in the genome of *S. clavuligerus* as they are also involved in the biosynthesis of a 3S, 5S clavam metabolite.

The precursors of clavulanic acid synthesis are arginine and D-glyceraldehyde-3-phosphate, their concentration, as well as the concentration of the arginine intermediate ornithine, stimulate the biosynthesis of clavulanate (Romero *et al.*, 1986). The early steps start with the condensation reaction of the C5 of arginine and C3 of D-glycerldehyde-3-P to form CEA (N2-(2-carboxyethyl)-arginine), catalysed by CEA synthase encoded in *ceaS* gene (Pérez-Redondo *et al.*, 1999) (**Figure 1.4A**). The synthesis of the β -lactam ring is catalysed by β -lactam synthase β -LS (*bls* gene) (Bachmann *et al.*, 1998). The following reactions involve hydroxylation, hydrolysis and cyclization into clavaminic acid, which are catalysed by CAS (clavaminic synthase) and PAH (proclavaminic amidinohydrolase) enzymes (Baldwin *et al.*, 1993; Wu *et al.*, 1995). Clavaminic acid is known to be also an intermediate of the 5S clavams (Mosher *et al.*, 1999). The late steps involve the conversion into clavulanic acid, which is mediated by GCAS (glycylclavaminic synthase encoded in *orf-17*) and CAD (clavuldehyde dehydrogenase) (Jensen *et al.*, 2004; Pérez-Redondo *et al.*, 1998).

1.2.3 The genome of *S. clavuligerus*.

Streptomycete chromosomes are highly unstable and so the evolution of their genomes is very dynamic and has been the subject of many studies (Chen *et al.*, 2002). Their chromosomes are linear and range in length from 8 to 10 Mb, being some of the largest genomes within bacteria (Ventura *et al.*, 2007). The analysis of the genome sequence of *Streptomyces coelicolor* and *Streptomyces avermitilis* established a broader knowledge about the chromosome organisation, where most of the essential genes for primary metabolism, cell division, DNA replication, transcription and protein biosynthesis, lie in the conserved core

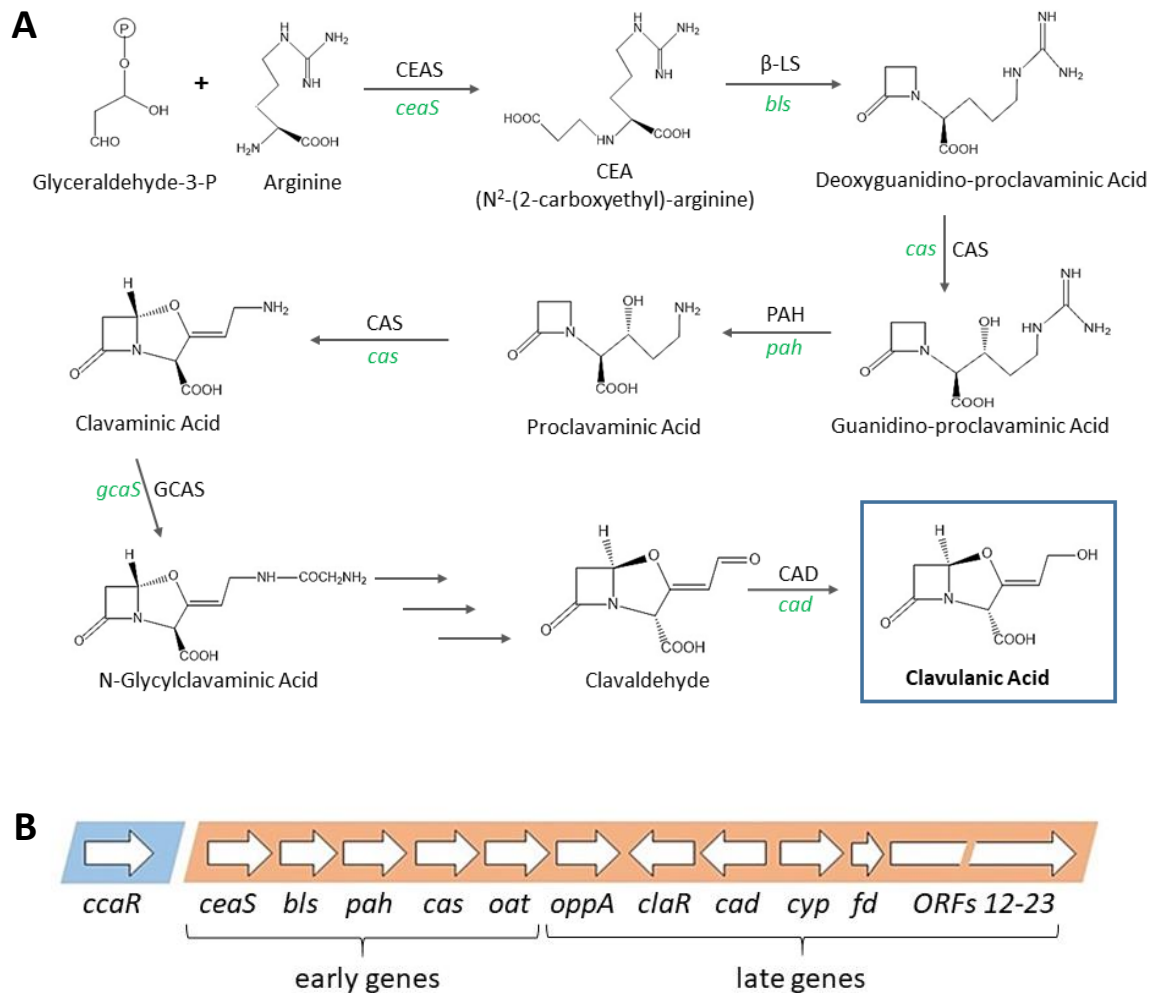


Figure 1.4. Biosynthesis of clavulanic acid. **A:** Biosynthetic pathway from precursors glyceraldehyde-3-phosphate and arginine, indicating the enzyme that catalyses the reaction on each step and the gene that codes for the enzyme (in green). **B:** Genes involved in the regulation and biosynthesis of clavulanic acid located in the cephamycin C cluster (in blue), and the clavulanic acid cluster (in orange).

region of the chromosome (Bentley *et al.*, 2002; Ikeda *et al.*, 2003). The terminal arms in the other hand are species-specific and more variable and susceptible to rearrangements, harbouring genes responsible for secondary metabolite production and other adaptive responses, they are usually exchangeable, with the exception of the extreme ends where the machinery for terminal replication is located (Bentley *et al.*, 2002; Hopwood, 2006).

Extrachromosomal elements, such as circular or linear plasmids, are also common in *Streptomyces* strains, and are often responsible for the transfer of exchangeable genes from variable regions of the chromosome. They present a similar G+C content to the chromosome and sizes that range between 10 and 1,800 kb, carrying their own origin of replication (*oriC*) and ParA and ParB partitioning proteins, they display replicative autonomy (Kieser *et al.*, 2000; Medema *et al.*, 2010). In addition, they often harbour secondary metabolite clusters, which is often the case for giant linear plasmids such as SCP1 of *S. coelicolor* and pSLA2-L of *S. rochei* that encode methylenomycin A and lankamycin production respectively (Kirby and Hopwood, 1977; Mochizuki *et al.*, 2003).

The genome of *S. clavuligerus* type strain ATCC 27064 has been the subject of many studies. Physical and next-generation sequencing analyses established a linear chromosome of around 6.8 Mb which was predicted to carry 23 putative secondary metabolite clusters, including the clavulanic acid cluster (Chen *et al.*, 1994; Song *et al.*, 2010; Medema *et al.*, 2010; Hwang *et al.*, 2019). Additionally, four linear plasmids have been confirmed to belong to the genome of *S. clavuligerus*, which are known as pSCL1, pSCL2, pSCL3 and pSCL4. The smallest one pSCL1 of 11.7 kb was first purified in the 80's by sucrose gradients and its complete sequence was determined by Sanger sequencing (Keen *et al.*, 1988; Wu and Roy, 1993). Evidence of the presence of pSCL2 and pSCL3 was first obtained with pulsed-field gel electrophoresis, a technique commonly used to detect large DNA molecules that cannot be detected using normal gel electrophoresis; but it was not until 2010 that sequencing data was obtained for both plasmids by Sanger and 454 sequencing technologies estimating the sizes of 150 and 444 kb respectively (Netolitzky *et al.*, 1995; Song *et al.*, 2010). Also in 2010, it was first discovered the presence of pSCL4 by SOLiD sequencing, and it was estimated to have a size of 1.8 Mb (Medema *et al.*, 2010). A table compiling general characteristics of the five replicons of *S. clavuligerus* ATCC 27064 is presented on **Table 1.1**. Additionally, two other *S. clavuligerus* strains with industrial origin have been sequenced: strains F613-1 and F1D-5, which present a larger chromosome and smaller pSCL4 (Cao *et al.*, 2016). These differences

from the type strain are probably due to genomic rearrangements during the strain development.

pSCL4, commonly referred as the megaplasmid, is the largest linear plasmid so far sequenced and is predicted to harbour 25 putative secondary metabolite gene clusters that include the antibiotics staurosporine, monoenomycin and other β -lactams (Medema *et al.*, 2010). Additionally, sequencing results revealed that pSCL4 presents genetic material that is characteristic from *Streptomyces* chromosomes and was likely acquired by horizontal gene transfer (Medema *et al.* 2010).

Comparison of the sequence of *S. clavuligerus* chromosome and megaplasmid with other *Streptomyces* genomes showed that the sum of genomic features of the chromosome and plasmid matches the values found in other sequenced *Streptomyces* species, such as the size (6.760 Mb + 1.796 Mb= 8.556 Mb) and coding sequences (5700 + 1581= 7281) as indicated on **Table 1.2** (Medema *et al.*, 2010). This, in conjunction with the small size of the *S. clavuligerus* chromosome, suggests that the megaplasmid is derived from the chromosome, endowing some of the chromosomal regions. Given the central location of the origin of replication of pSCL4, a single recombination event could not explain the evolution of the megaplasmid, for this reason, Medema *et al.* proposed three hypotheses that involved a double crossover of the core regions of the chromosome and a small plasmid, two recombination events that took place consecutively and resulted in a symmetrical replicon, or integration and followed by breaking off of a plasmid.

The studies of Medema *et al.*, 2010 also confirmed that the chromosome harbours all the housekeeping genes responsible for primary metabolism and therefore the megaplasmid is not considered essential for the growth of the organism, although this theory is yet to be confirmed. Regarding secondary metabolism, apart from the numerous putative metabolites encoded on the megaplasmid, there is some evidence of potential cross-linked regulation between the plasmid and the chromosome, as pSCL4 encodes genes such a γ -butyrolactone receptor (ScaR/Brp) that is involved in repressing the synthesis of clavulanic acid and cephamycin C (Kim *et al.*, 2004; Santamarta *et al.*, 2005). In addition, the studies of Medema *et al.*, 2010 identified the terminal proteins-coding genes ***tap*** and ***tpg*** in pSCL4, however, they were not identified in the sequence of the chromosome, predicting that the latter might rely on the megaplasmid to provide these proteins.

Table 1.1. General characteristics of the 5 replicons found in *Streptomyces clavuligerus* ATCC 27064, including the main chromosome, and the four plasmids pSCL1, pSCL2, pSCL3 and pSCL4 (Wu and Roy, 1993; Song *et al.*, 2010; Medema *et al.*, 2010).

Replicon	Size (kb)	GC (%)	Coding sequences	2 ^{ary} metabolites clusters	Examples of 2 ^{ary} metabolites
Chromosome	6,760	72.69	5,700	23	clavulanic acid, cephamycin C, holomycin
pSCL1	10.5	71.96	10	-	-
pSCL2	149.4	70.07	159	-	-
pSCL3	444.2	70.77	423	-	-
pSCL4	1,796	71.85	1,581	25	staurosporine, moenomycin, β -lactams

Table 1.2. Comparison of the genome architecture of *S. clavuligerus* with *Streptomyces coelicolor*, *Streptomyces avermitilis* and *Streptomyces griseus*. Adapted from Medema *et al.* 2010. Mb: megabasepairs. Chrom: chromosome.

Replicon	<i>S. clavuligerus</i>			<i>S. coelicolor</i>	<i>S. avermitilis</i>	<i>S. griseus</i>
	Chrom	pSCL4	Chrom + pSCL4	Chrom	Chrom	Chrom
Sequence length (Mb)	6.760	1.796	8.556	8.668	8.546	9.026
Coding sequences	5,700	1,581	7,281	7,825	7,574	7,138
2 ^{ary} metabolite clusters	23	25	48	20	30	34

1.3 The role of terminal proteins in replication of *Streptomyces* telomeres.

Streptomyces, along with other actinomycetes such as *Kitasatopora* and *Rhodococcus*, are among the few types of bacteria that carry linear chromosomes and plasmids (Lin et al., 1993; Chen et al., 1994). Although the benefit of having linear chromosomes is unclear, it has been hypothesised that linearity might increase adaptability in their vegetative lifestyle by facilitating genome rearrangements that provides an accelerated evolution (Chater and Kinashi, 2007). Linearity implies the presence of chromosomal or plasmid ends, known as telomeres, which is a very rare aspect within prokaryotes.

Streptomyces chromosomes and linear plasmids carry telomeric sequences that are densely packed with palindromic sequences and are almost identical at both ends. These are known as terminal inverted repeats, and range in size from 14 bp (*Streptomyces hygroscopicus*) to over 1 Mb (*S. coelicolor*) (Lin et al., 1993; Weaver et al., 2004). Most *Streptomyces* species carry conserved telomeres known as **archetypal telomeres** that carry up to seven palindromes in the first 180 bp. Examples of these are the telomeres of *S. coelicolor* chromosome, and the *Streptomyces lividans* chromosome and plasmid SLP2 (Huang et al., 1998). These and other examples of archetypal telomeres are illustrated in the alignment on **Figure 1.5**. These telomeres were extracted from streptomyces genomic DNA using glass beads-binding, and cloning into vectors to obtain the telomeric sequence (Huang et al., 1998). The characterisation of the telomere sequences allowed identification of the seven conserved palindromes, as shown in the alignment, with particular relevance of Palindrome I, containing the first and last 13 nucleotides of the molecule, which is identical in all archetypal telomeres (Huang et al., 1998). Interestingly, this work included the telomeres of *S. clavuligerus* plasmids pSCL1 and pSCL2, confirming these replicons carry archetypal telomeres, although the telomeres of *S. clavuligerus* chromosome and other two plasmids have not been characterised yet.

The ability of telomeres to bind to glass beads lies on the presence of **terminal proteins** (TP) that cap the chromosome ends (Lin et al., 1993). These proteins are necessary for the conservation of telomeres in streptomyces acting like the telomerase system from eukaryotes. Two terminal proteins have been confirmed to be essential for the maintenance of telomeres: **Tap** and **Tpg**. Archetypal TPs are highly conserved in sequence and size among different *Streptomyces* species, and are encoded by the *tap* and *tpg* genes located on the same operon towards the chromosome terminus (Yang et al., 2002). These proteins carry

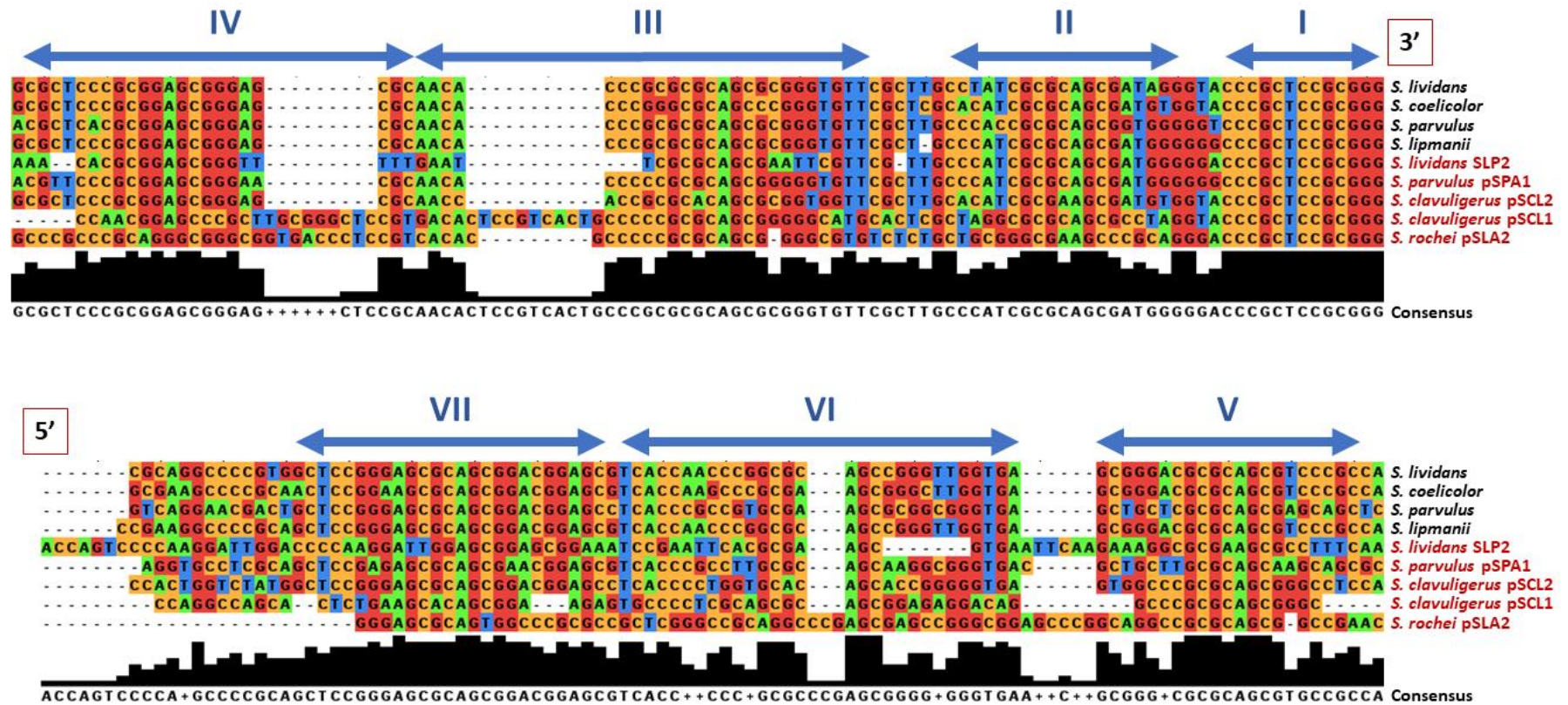


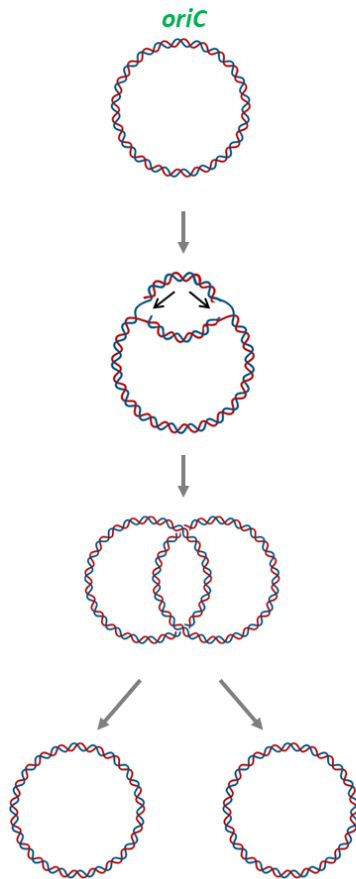
Figure 1.5. Alignment of the archetypical telomeres (180 nucleotides) of *Streptomyces* chromosomes (in black) and plasmids (in red). The palindromic sequences are indicated in roman numerals and blue arrows. Adapted from Huang *et al.*, 1998.

helix-turn-helix domains that allow interaction with specific telomeric sequences during replication.

Upon DNA replication, whereas most bacteria yield two circular replicons, streptomycetes produce two linear molecules proceeding bidirectionally from *oriC*, located internally on the central region, towards the two ends of the chromosome (Musialowski *et al.*, 1994). Given the bidirectional nature of the replication, the maturation of the Okazaki fragments of the lagging strand leaves single-stranded overhangs of around 300 nucleotides at the 3' end that cannot be replicated by the replisome (**Figure 1.6**) (Chen *et al.*, 1994). Due to the presence of multiple palindromic sequences at the telomeres, the single-stranded overhang folds into compact secondary structures of hairpin loops that are thought to be resistant to the attack of nucleases (Huang *et al.*, 1998). The structures formed by the first four palindromes are thought to be essential for telomere replication and have been studied in detail by the group of Dr Carton W. Chen (National Yang-Ming University, Taipei), who established two models that represent the most thermodynamically stable forms of single-stranded archetypal telomeres. The Clover Leaves model, in which the first four palindromes form independent hairpins or loops; and the Rabbit Ears model, in which palindromes I and IV share one loop forming a duplex of 13 base-pairs (**Figure 1.7**) (Yang, *et al.*, 2015; Yang *et al.*, 2017).

After replication, if the lagging strand is not completed, it would risk the loss of genetic material with each replication round. In order to avoid so, *Streptomyces* replicons undergo a mechanism called **end patching**, in which terminal proteins recognise the overhangs and allow filling the gap. (Chang and Cohen, 1994; Qin and Cohen, 1998). The Tap protein recognises and binds to the second and third palindromes on the telomere, and then recruits the Tpg protein forming the Tap-Tpg heterodimer (Bao and Cohen, 2003; Yang *et al.*, 2015). The Thr¹¹⁴ residue of Tpg provides the hydroxyl group that binds covalently to a first dCTP and then Tap adds the following nucleotides up to the 13th (Palindrome I) using the 3' overhang as template (Yang *et al.*, 2006; Yang *et al.*, 2015). This 13-nt oligomer acts as a primer for the DNA polymerase I that extends the rest of the gap yielding a full-length double-stranded molecule and the proteins remain attached covalently to the 5' end of the replicon (**Figure 1.8**) (Yang *et al.*, 2015). In addition, it has been evidenced how Tap can repair truncated telomeric sequences, restoring and protecting the telomeres during replication. However, this repair was established to only be possible following the Clover Leaves model,

Circular chromosome



Linear chromosome

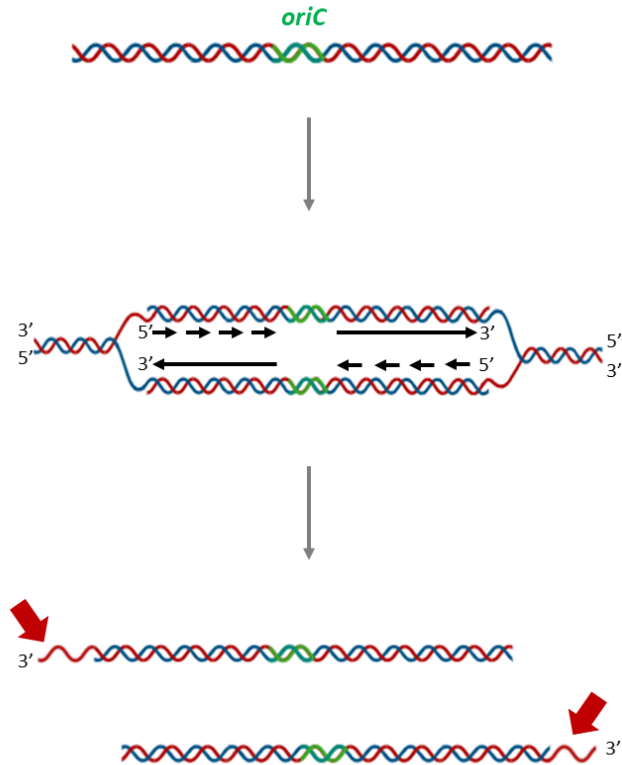


Figure 1.6. Schematic representation of the replication of circular and linear chromosomes. Replication of circular replicons proceeds from the origin of replication (*oriC*) towards the opposite point, yielding two complete circular molecules. The replication of linear replicons has a central origin and continues bidirectionally towards the ends. The maturation of the Okazaki fragments of the lagging strands leaves single-stranded overhangs at the 3' end of each new DNA molecule (indicated by red arrows).

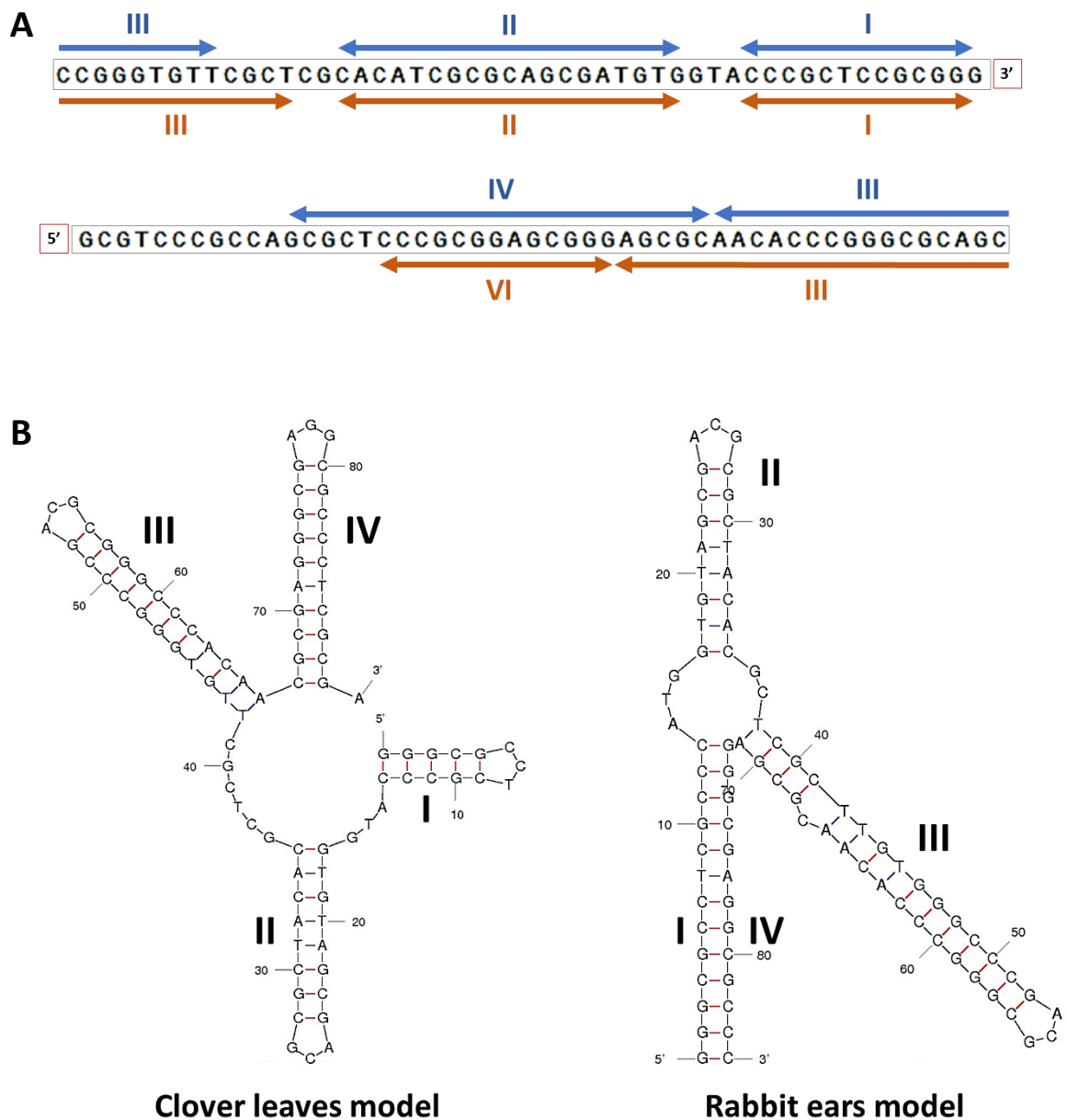


Figure 1.7. Two models were established by Yang *et al.*, 2017 to explain the secondary structures of archetypal telomeres. **A:** Telomere of *S. coelicolor* chromosome (100 nucleotides) indicating the palindromes in roman numerals and arrows: in blue the Clover leaves model palindromes and in orange the Rabbit ears model palindromes. **B:** Folded structures of the 3' single-stranded overhangs of *S. coelicolor* following the Clover leaves model and the Rabbit ears model. The palindromes are indicated in roman numerals. Structures constructed using the mfold web server (University of Albany)

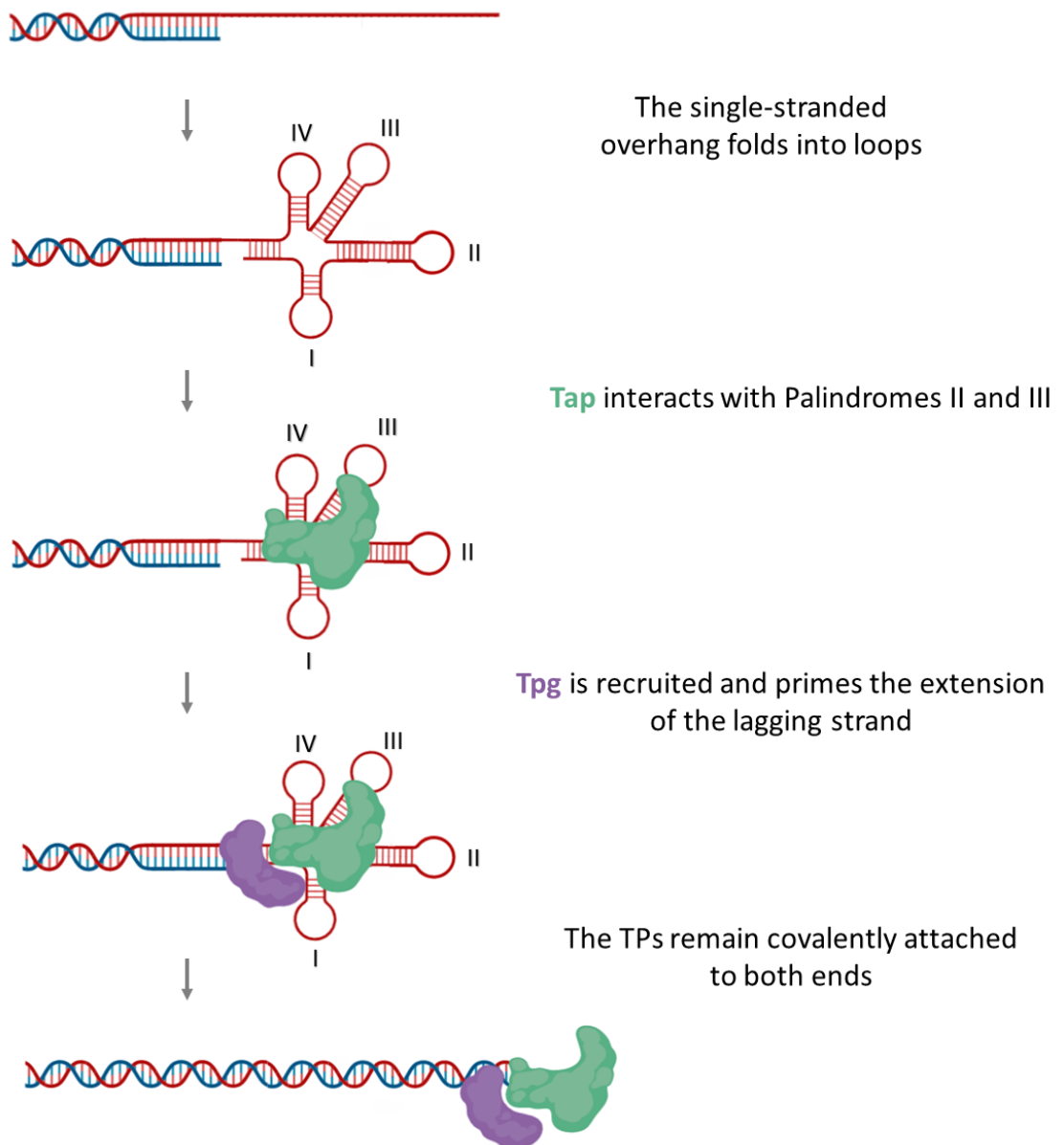


Figure 1.8. End patching by terminal proteins (TPs) achieves maintenance of complete telomeric sequences with each replication round. The palindromes are indicated in roman numerals.

which indicates that this is the most probable conformation of single-stranded archetypal telomere (Yang *et al.*, 2017).

Although cases of almost identical telomeres between streptomycetes chromosome and plasmids in the same organism have been evidenced, which is the case of *S. lividans* chromosome and plasmid SLP2 (Huang *et al.*, 1998), it is also common to find different telomeres in replicons of the same organism, and so each replicon usually encodes its own copy of *tap* and *tpg*. Different telomeres in the same strain were first observed in the *S. coelicolor* chromosome and plasmid SCP1. Unlike the chromosome, SCP1 carry non-archetypal telomeres that start with six G nucleotides instead of three Cs, and are maintained by the non-archetypal Tap and Tpg homologous: Tac and Tpc (Bentley *et al.*, 2004; Huang *et al.*, 2007). Another example of non-archetypal telomeres that are also different from SCP1 are the ones on the *Streptomyces griseus* chromosome, although they are also packed with multiple palindromic sequences (Goshi *et al.*, 2002).

In the case of *S. clavuligerus*, there is evidence of how the telomeres of pSCL1 and pSCL2 that were characterised by Huang *et al.*, 1998 carry archetypal telomeres, nevertheless, their sequences differ greatly after the second palindrome. In contrast, the telomeres of *S. clavuligerus* chromosome, pSCL4 and pSCL3 have not been purified and characterised, however, the whole genome sequencing data from the industrial strain F613-1 chromosome, as well as ATCC 27064 pSCL4 and pSCL3 show sequences highly similar to the first palindromes of archetypal telomeres towards the ends implying that all *S. clavuligerus* replicons might carry conserved archetypal telomeres (Cao *et al.*, 2016; Song *et al.*, 2010). These putative telomeres showed low similarity to each other and to the known ones from pSCL1 and pSCL2 beyond the first palindrome and were only found at one end and not the other of the molecules, which evidences again the lack of completeness of the published genome sequences. In addition, only one copy of the *tap-tpg* operon has been identified so far in *S. clavuligerus* genome, which is the one located on pSCL4 (Medema *et al.*, 2010). This suggests that either the same terminal proteins are promiscuous enough to interact with different telomeric sequences, or other unknown Tap and Tpg genes are encoded in the genome of *S. clavuligerus*.

1.4 Scope of the project

The action of clavulanic acid combined with amoxicillin (Augmentin®-GSK) against β -lactamase-producing bacteria is currently an essential measure to treat severe infectious diseases worldwide. With an estimated sales value of up to \$1.3 billion per annum during the past few decades, Augmentin® is still considered one of the largest selling antibacterial formulations (Elander, 2003). *S. clavuligerus* is the strain used to manufacture clavulanic acid at scale in a fermentation plant in Irvine (Scotland). Unfortunately, the clavulanic acid yield by *S. clavuligerus* is low and overseas generic competitors for Augmentin® are exerting intense commercial pressure that endangers its position in the market. Consequently, finding strategies to improve the performance of *S. clavuligerus* is vital for the continuation of the drug as a choice for antimicrobial therapy by GSK.

The genome sequence of *S. clavuligerus*, including four extrachromosomal elements, is yet to be completed, however, the published data predicted the presence of more than 7000 coding sequences, 20% of which encoded in the extrachromosomal megaplasmid pSCL4. It is thought that this plasmid does not carry genes essential for primary metabolism and therefore it is, in principle, dispensable for the growth of the organism. Elimination of the plasmid would represent a significant reduction of metabolic burden by saving energy and metabolic precursors, improving the stability of the strain and potentially increasing the production of clavulanic acid. In addition, curing pSCL4 would help in elucidating the function of plasmid genes and the energy expenditure difference between essential and secondary metabolites.

Nevertheless, following the theory of Medema *et al.*, 2010 regarding the presence of the only copy of *tap-tpg* being in pSCL4, removal of the plasmid would deprive the chromosome from the action of terminal proteins on telomere replication risking chromosomal shortening with each replication round. For this reason, characterisation of the role of Tap and Tpg in the replication of chromosomal telomeres is a key step prior to the elimination of the megaplasmid from the total genome of *S. clavuligerus*.

1.5 Specific aims

- Study of the genome architecture of *S. clavuligerus*, including obtaining physical evidence of the three giant linear plasmids and chromosomal restriction fragments by Pulsed-field gel electrophoresis.
- Obtaining a high quality complete genome sequence of *S. clavuligerus* DSM 738 using the latest whole genome sequencing technology.
- Extraction and characterisation of the *S. clavuligerus* chromosomal and plasmids telomeres.
- Investigation of the terminal proteins to elucidate their role in telomere replication and protein-protein interactions.
- Construction of *tap-tpg*-less mutants and study the repercussion of these deletions on chromosomal linearity.
- Elimination of the pSCL4 plasmid to study its implications on the physiology and metabolism of *S. clavuligerus* and assess the clavulanic acid production in the absence of the megaplasmid.

2. MATERIALS AND METHODS

All the reagents and chemicals utilised in this project were provided by the following suppliers: Promega, New England Biolabs (NEB), Fisher, Roche, Sigma Aldrich, Qiagen, BioLine, Life Technologies, GE Healthcare, Oxoid, Alfa Aesar, Biotium and Bio-Rad.

2.1 Media and antibiotics

The culture media and the respective recipes are listed on **Table 2.1**. Each medium was sterilised in an autoclave at 121°C for 15 minutes. The media used for the culture of *S. clavuligerus* were adjusted to pH6.8. The antibiotics used for strain selection are indicated in **Table 2.2** and they were sterilised using 0.22 µm filters.

Table 2.1. List of the media and their respective recipes used to culture microbial strains in this study.

Medium (Reference)	Recipe
GYM	4 g Glucose 4 g Yeast extract 10 g Malt extract 2 g CaCO ₃ 12 g Agar Add 1000 ml distilled water
L3M9 (GSK)	3 g Dextrin 10 g Trehalose dihydrate 0.5 g K ₂ HPO ₄ 1 g NaCl 1 g MgSO ₄ 0.5 g CaCl ₂ 2 g Casamino acids 10.5 g MOPS buffer 1 ml Agar trace salts* 30 g Agar Add 1000 ml distilled water
LB broth (Bertani, 1951)	10 g Tryptone 5 g Yeast extract 10 g NaCl Add 1000 ml distilled water
LB agar (Bertani, 1951)	10 g Tryptone 5 g Yeast extract 20 g NaCl 20 g Agar Add 1000 ml distilled water
TSB (Kieser <i>et al.</i> , 2000)	17 g Casein peptone 3 g Soya peptone 5 g NaCl

	2.5 g K ₂ HPO ₄ Add 1000 ml distilled water Adjust to pH6.8
2x YT (Kieser <i>et al.</i> , 2000)	16 g Tryptone 10 g Yeast extract 5 g NaCl Add 1000 ml distilled water
CM5 (GSK)	40 g Rape seed oil 11.5 g Maltodextrin 18 g Soya protein concentrate 3 g KH ₂ PO ₄ 19.8 g MES 1.6 g MgSO ₄ .7H ₂ O 0.3 g (NH ₄) ₂ SO ₄ 10 ml trace element solution Add 1000 ml distilled water Adjust to pH6.8
Nutrient agar	4g Difco nutrient broth powder 10g Agar Add 1000 ml distilled water
Soft nutrient agar	4g Difco nutrient broth powder 5g Agar Add 1000 ml distilled water

*trace element solution: 8.78 g/l FeCl₃, 2.04 g/l ZnSO₄, 1.02 g/l MnCl₂*4 H₂O, 0.43 g/l CuSO₄*2 H₂O, 0.42 g/l NaI, 0.31 g/l H₃BO₃, 0.24 g/l CaCl₂*6 H₂O and 0.24 g/l Na₂MoO₄*2 H₂O

Table 2.2. List of the antibiotic compounds used for strain selection, indicating the concentration of the stocks as well as the final concentration in the media.

Antibiotic	Stock [mg/ml]	Final concentration [µg/ml]
Apramycin	50 in dH ₂ O	50
Kanamycin	25 in dH ₂ O	25
Chloramphenicol	25 in 100% Ethanol	25
Nalidixic acid	25 in 0.15M NaOH	25
Carbenicillin	100 in dH ₂ O	100
Hygromycin	50 in dH ₂ O	50
Thiostrepton	50 in 100% DMSO	50
Streptomycin	100 in dH ₂ O	100

2.2 Microbial strains

All the microbial strains utilised, and their respective characteristics are shown on **Table 2.3**. The *S. clavuligerus* strains DSM 738 and DSM 41826 were obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ).

Table 2.3. List of microbial strains used and constructed in this study, their genotype and reference.

Strain	Genotype	Reference
<i>S. coelicolor</i> M145	SCP1-, SCP2-	Kieser <i>et al.</i> , 2000
<i>S. clavuligerus</i> DSM 738	Type strain. Wild type	Higgins & Kastner 1971
<i>S. clavuligerus</i> DSM 41826	Wild type	Higgins & Kastner 1971
<i>S. clavuligerus</i> Sc2	Industrial production strain	GSK
<i>S. clavuligerus</i> Sc6	Industrial production strain	GSK
<i>S. clavuligerus</i> LEON	pSCL3-	Gomez-Escribano
<i>E. coli</i> ET12567/pUZ8002	<i>dam-13::Tn9, dcm-6, hsdM, hsdR, recF143, zij201::Tn10, galk2, galT22, ara14, lacYI, xylS, leuB6, thi-1, tonA31, rpsL136, hisG4, tsx78, mtli, glnV44, F-</i>	MacNeil <i>et al.</i> 1992
<i>E. coli</i> competent cells JM109	<i>recA1, endA1, gyrA96, thi, hsdR17 (rK-,mK+), relA1, supE44, Δ(lac-proAB), [F', traD36,proAB, lacI^qZΔM15]</i>	Promega
SCLA22	738, pLIS22 integrated in chromosome	This study
SCLA23	41826 pLIS22 integrated in chromosome	This study
SCLA24	Sc2, pMS82 integrated in chromosome	This study
SCLA82	738, pMS82 integrated in chromosome	This study
SCLA405	738, pLIS405 integrated in pSCL4	This study
SCLA22-405	SCLA22, pLIS405 integrated in pSCL4	This study
SCLA6902	738, pIJ6902 integrated in chromosome	This study
<i>E. coli</i> BTH101	<i>F-, cya-99, araD139, galE15, galK16, rpsL1 (Str'), hsdR2, mcrA1, mcrB1.</i>	Karimova <i>et al.</i> 1998
SCLA-M1	pSCL4-	This study
SCLA-M2	pSCL4-, larger pSCL3	This study
SCLA-M5	pSCL4-, pSCL3-	This study
SCLA-Mc3	pSCL4-, pSCL3-, pLIS22 integrated in chromosome	This study
SCLA405-Mc3	pLIS22 integrated in chromosome, pLIS405 integrated in pSCL4	This study

2.3 Plasmids

All the plasmid utilised or constructed in this project are indicated in **Table 2.4**.

Table 2.4. List of plasmids and constructs used in this project. Amp: ampicillin, Hyg: hygromycin, Apr: apramycin, Thio: thiostrepton, Kan: kanamycin, BACTH: bacterial two-hybrid, MCS: multiclonal sequence site.

Plasmid	Description	Antibiotic resistance	Reference
pGEM-T easy	Cloning vector	Amp	Promega
pMS82	Integrative vector in <i>Streptomyces</i>	Hyg	Gregory & Smith 2003
pIJ6902	Integrative vector in <i>Streptomyces</i>	Apr, Thio	Huang <i>et al.</i> , 2005
pLIS22	pMS82 plus <i>tap-tpg</i> operon	Hyg	This study
pBGS19	Cloning vector	Kan	Spratt <i>et al.</i> 1986
pIJ773	Cloning vector	Apr	Gust <i>et al.</i> 2002
pGEM-SP-Right	pGEM-T easy plus SP-Right	Amp	This study
pGEM-SP-Left	pGEM-T easy plus SP-Left	Amp	This study
pLIS401	pBGS19 plus SP-Right	Kan	This study
pLIS402	pBGS19 plus SP-Right and SP-Left	Kan	This study
pLIS404	<i>tap-tpg</i> knock-out vector	Kan, Apr	This study
pWHM3	Shuttle vector <i>E. coli</i> / <i>Streptomyces</i>	Thio, Amp	Vara <i>et al.</i> 1989
pLIS405	<i>tap-tpg</i> knock-out vector	Thio, Apr, Amp.	This study
pKT25	BACTH vector with T25 domain of CyaA, MCS at the 3' end of T25	Kan	Karimova <i>et al.</i> 1998
pKNT25	BACTH vector with T25 domain of CyaA, MCS at the 3' start of T25	Kan	Karimova <i>et al.</i> 1998
pUT18	BACTH vector with T18 domain of CyaA, MCS at the 3' start of T18	Amp	Karimova <i>et al.</i> 1998
pUTC18	BACTH vector with T18 domain of CyaA, MCS at the 3' end of T18	Amp	Karimova <i>et al.</i> 1998
pKT25-zip	pKT25 plus leucine zipper of GCN4	Kan	Karimova <i>et al.</i> 1998
pUT18C-zip	pUT18 plus leucine zipper of GCN4	Amp	Karimova <i>et al.</i> 1998
pLIS611	pKT25 plus pSCL4 <i>tap</i>	Kan	This study
pLIS631	pUT18 plus pSCL4 <i>tap</i>	Amp	This study
pLIS731	pKT25 plus pSCL4 <i>tpg</i>	Kan	This study
pLIS711	pUT18 plus pSCL4 <i>tpg</i>	Amp	This study This study
pLIS612	pKT25 plus pSCL2 <i>tap</i>	Kan	This study
pLIS632	pUT18 plus pSCL2 <i>tap</i>	Amp	This study
pLIS732	pKT25 plus pSCL2 <i>tpg</i>	Kan	This study
pLIS712	pUT18 plus pSCL2 <i>tpg</i>	Amp	This study
pLIS613	pKT25 plus pSCL2 <i>tap</i>	Kan	This study
pLIS633	pUT18 plus pSCL2 <i>tap</i>	Amp	This study

pLIS733	pKT25 plus pSCL2 <i>tpg</i>	Kan	This study
pLIS713	pUT18 plus pSCL2 <i>tpg</i>	Amp	This study
pGE240	CRISPR/Cas9 plasmid, targets pSCL4_ <i>parB</i>	Apr	Gomez-Escribano
pUC19	Cloning vector	Amp	Yanisch-Perron, Vieira and Messing, 1985

2.4 Primers

The primers used for PCR amplification can be found in **Table 2.5**. Primers were designed using SnapGene Software version 4.2.11, and they were synthesised using the services of Integrated DNA Technologies (IDT).

Table 2.5. List of primers designed and used for this study. The restriction sites used for cloning are highlighted in bold. GGATCC, *Bam*HI; TCTAGA, *Xba*I; AAGCTT, *Hind*III; GAATTC, *Eco*RI; ATGCAT, *Nsi*I. Fw: forward; Rv: reverse. Tm: melting temperature.

Name	Gene(s)/Region	Direction	5' - 3' Sequence	Tm (°C)
PTP-F-HindIII	pSCL4 <i>tap-tpg</i>	Fw	GTAA GGCTTGACCGTGGCGGGATGT	63.4
PTP-R-NsiI		Rv	TAAT GCAT GCAGGGCAGGACCGTT	63.3
Apr_BamHI_Fw	Apramycin resistance gene	Fw	ATGGATCCT GTAGGCTGGAGCTGCTTC	63.6
Apr_XbaI_Rv		Rv	GTTCTAGA ATTCCGGGGATCCGTCGACC	63.3
SP-R-EcoRIXbaI	Right flanking region	Fw	GAGAATTCTCTAGAT GTAGGTATTCGCCCTTGCCGTT	64.0
SP-R-BamHI		Rv	TAGGATCC ATACTCCGCGCAGAACACTA	64.1
SP-L-BamHI	Left flanking region	Fw	GCGGATCCC CAGAAATATTCCTTGACGCGAA	62.3
SP-L-SphIXbaI		Rv	TAGCATGCTCTAGAT TGTACCGTCCATCGATTCA	62.6
CaoChr-L-end-F	Chromosome left end	Fw	CCCTGGAAGTGACTCCGACTCCGCTAACG	66.2
CaoChr-L-end-R		Rv	GCCTTCGGATCAATAGAGCATAAGCGGCAATCCG	65.7
CaoChr-R-end-F	Chromosome right end	Fw	ACGGACCGCATCTGCTGGACATGCTC	66.4
CaoChr-R-end-R		Rv	GTTGTCTGTCGACCGCAGCCTGAAGCC	66.6
Tpg-Fw	pSCL4 <i>tpg</i>	Fw	TCAGTAGGACAGGTGCGAGGTAGTTCGATGTCC	66.3
Tpg-Rv		Rv	AGTGAGTAGCCGATGGGCGACATCGA	66.5
SCL4-L-end-F	pSCL4 left end	Fw	TGCGGCACCTTGCCAGTACAAGACCC	66.7
SCL4-L-end-R		Rv	GGTGGTCTGTCCCACGCGCACTTCACAC	66.4
SCL4-L-end-R2		Rv	ATTTTATGGCGGAAGGTGCGGGTTCATGGGG	66.6
SCL4-R-end-F	pSCL4 right end	Fw	GGGGTCTGGAGGAAGGCGTAGTTCTTGT	66.1
SCL4-R-end-R		Rv	TACACGAAGAAGCCCTTCGACGGGGGC	67.3
Chrom-iPCR-L1	Chrom left telomere	Fw	TTTATCCGATACGGCCTCGCTGAC	60.6
Chrom-iPCR-L2		Rv	GCATCACTCTGGAGCGACTTCAAT	59.2
Chrom-iPCR-R1	Chrom right telomere	Fw	TCATGACACGGCTGGAGAAGCTCGGT	65.3
Chrom-iPCR-R2		Rv	GCCGTGGATCAAGACATACGTTGCCATGTT	63.9
pSCL1-iPCR-L1	pSCL1 left telomere	Fw	TATTCACACTCGGGCATTCTCTGAAAT	58.1
pSCL1-iPCR-L2		Rv	AAAATCCGTTCCGGGTCGCTGTATTT	60.8
pSCL1-iPCR-R1	pSCL1 right telomere	Fw	CCTGTCCAGCCGTTTCAAGGAAGAAA	61.1
pSCL1-iPCR-R2		Rv	TGTCCACGGTGAAGTCGTTCTTCTT	60.1
pSCL4-iPCR-L1	pSCL4 left telomere	Fw	AAGCCTTCGAGGGGGTATACGGACAG	63.4
pSCL4-iPCR-L2		Rv	AAAAAGGCTCCGACAGGATTTACCCGGAAA	63.2
pSCL4-iPCR-R1	pSCL4 right telomere	Fw	TGTGACGGGGTCTGACCGGGTAAA	67
pSCL4-iPCR-R2		Rv	CCCGACCTTCCGCCATAAAAATCGGTTAGGTTA	67.4
pSCL2-iPCR-L1	pSCL2 left telomere	Fw	TTCGACCTCGCGATCGTGGATGAGG	64.1
pSCL2-iPCR-L2		Rv	TGGTACTGAGCCACCTTCGGATGGAAT	64.5

pSCL2-iPCR-R1	pSCL2 right	Fw	TTCCCACTTCATAATTCTCGCGTCAAGTGGG	64.3
pSCL2-iPCR-R2	telomere	Rv	GGCATGGTCGTGTCAGCGTCGAAAAGATCCA	65.3
pSCL3-iPCR-L1	pSCL3 left	Fw	GCGGAATTGACGGTCATTCCGTCCTCT	63.8
pSCL3-iPCR-L2	telomere	Rv	TTCCACCTAGCCAGCCATTTCAGTGACCACAA	66.9
pSCL3-iPCR-R1	pSCL3 right	Fw	GTCGTGCTTCCACTCGCGCTTCT	63.8
pSCL3-iPCR-R2	telomere	Rv	TCTGTAGACGCTCTGCGAGACAGCGGTAGA	66
FtsZ-Fw	<i>ftsZ</i>	Fw	GTGGCAGCACCGCAGAACTACCTCG	67
FtsZ-Rv		Rv	TCACTTCAGGAAGTCCGGGACATCCAGTTCC	67
BACTH-p4-Tap-Fw	pSCL4 <i>tap</i> for	Fw	ACTCTAGATATGTCCACGGCCAGGATGCCCTTTT	65.7
BACTH-p4-Tap-Rv	BACTH	Rv	ATGGATCCTCCTCCCCCTCGTCGAGGACG	64.9
BACTH-p4-Tpg-Fw	pSCL4 <i>tpg</i> for	Fw	ACTCTAGATATGGGCGACATCGACGACGCG	63.7
BACTH-p4-Tpg-Rv	BACTH	Rv	ATGGATCCCAGTAGGACAGGTCGAGGTAGTCGATG TCCGT	65.5
BACTH-p2-Tap-Fw	pSCL2 <i>tap</i> for	Fw	ACTCTAGATGTGTCCACGGAGAACGAGCTGTT	61.3
BACTH-p2-Tap-Rv	BACTH	Rv	TTGAATTCCTTCGCCGTTGTCGGTGCC	61.2
BACTH-p2- TapEND-Rv		Rv	TTGAATTCCTTCGCCGTTGTCGGTGCCGG	64.3
BACTH-p2-Tpg-Fw	pSCL2 <i>tpg</i> for	Fw	ACTCTAGATATGAGCAGCGAGTTCGGCGG	62.8
BACTH-p2-Tpg-Rv	BACTH	Rv	ATGGATCCAACAGCTCGAAGTCGAGGTGGACC	62.7
BACTH-p3-Tap-Fw	pSCL3 <i>tap</i> for	Fw	ACTCTAGATATGGGGAGCAGCGACCCGG	65.1
BACTH-p3-Tap-Rv	BACTH	Rv	TTGAATTCGGCCACGCGGACGAG	68.9
BACTH-p3- TapEND-Rv		Rv	TTGAATTCAGGCCACGCGGACGAGC	65.6
BACTH-p3-Tpg-Fw	pSCL3 <i>tpg</i> for	Fw	ACTCTAGATGTGGGACTGATCGAAGACCGCT	62.9
BACTH-p3-Tpg-Rv	BACTH	Rv	TTGAATTCGGGTGATGCTGAAGCGGATGTG	62.6
BACTH-p3- TpgEND-Rv		Rv	TTGAATTCAGGTGATGCTGAAGCGGATGTGGT	63.9
M1-junction-F	Chromosomal	Fw	AACGTCGAGGACTGGAAGCCCAAGAA	62.6
M1-junction-R	junction M1	Rv	CTGTTGCCGTGCGTACGGCTGC	65.7
M2-junction-F	Chromosomal	Fw	ATGAGGAGTTGTTGCCGATCTCGACGATC	63
M2-junction-R	junction M2	Rv	AAGATCCTGGCGGACCCTCGTTTCA	63.9
Mc3-junction-F	Chromosomal	Fw	ATTCTTACCAGGACCGACTACACCA	63.6
Mc3-junction-R	junction Mc3	Rv	GGGCACGACCTGTTTCGTCGAGATCA	63.2

2.5 Bacterial cultivation

S. clavuligerus strains were grown on GYM agar and L3M9 agar for 5-10 days at 26°C. Spores were harvested from the surface using sterile cotton buds after adding 2 ml of 20% (v/v) glycerol. Spore suspensions were stored in cryovials at -20°C. Spore stocks were used to inoculate plates containing SFM, GYM, L3M9 as well as 25 ml of GYM broth and TSB media. For liquid cultures, spores were pre-germinated at 50°C for 10 minutes.

E. coli strains were grown overnight at 37°C at 200 rpm (revolutions per minute) in LB broth or in LB agar for solid cultures and supplemented with the respective antibiotic (concentration indicated in **Table 2.2.**). 500 µl of overnight culture were used to prepare glycerol stocks, adding an equal volume of 50% glycerol solution and stored at -80°C.

2.6 Pulsed-field Gel Electrophoresis (PFGE)

PFGE analyses were performed using the Bio-Rad CHEF-DR® II Pulsed Field Electrophoresis System. The programme is based of electrophoresis of DNA embedded in small blocks or plugs of agarose and using alternating electrodes of different orientation to obtain separation of DNA molecules of sizes up to 2 Mb.

Table 1.7. List of solutions and buffers used for PFGE and their components.

Solution	Components
TE25-Sucrose	25 mM Tris-HCl pH8, 25 mM EDTA pH8, 0.3 M sucrose
NDS	0.5 M EDTA pH8, 10 mM Tris-HCl pH8, 1% SDS, or 1% lauryl sarcosine
Lysozyme solution	2 mg/ml of TE25-Sucrose
Protein K solution	1 mg/ml of NDS
TE Buffer	10 mM Tris-HCl pH8, 1 mM EDTA pH8
10X TBE Buffer	1 M Tris Base 1 M Boric Acid 0.02 M EDTA pH8
0.5X TBE Buffer	Diluted from 10X TBE stock

2.6.1 Plug preparation

Agarose plugs containing DNA were prepared following the protocol described by Kieser *et al.* (2000). *Streptomyces* mycelium grown in liquid media was harvested by centrifugation and washed in 25 ml TE25-Sucrose buffer. After centrifugation, the mycelium was re-suspended in TE25-Sucrose to give an OD₆₀₀ of 1.9-2.0. 500 µl of mycelium suspension was added to 500 µl of 1.5% agarose. The sample was mixed and poured using a pipette into plug moulds. Once settled, the plugs were removed from the mould and incubated with lysozyme solution at 37°C for 30 minutes in case of *S. clavuligerus* and up to 2 hours for other *Streptomyces* species. Then the plugs were treated with proteinase K solution at 50°C overnight. The digestion was stopped by washing the plugs in 10 ml of TE buffer containing 0.1 mM PMSF (Phenylmethylsulfonyl fluoride) for one hour. Then the plugs were washed three times in TE buffer for 1 hour each.

2.6.2 Plug digestion with restriction enzymes

The plugs were placed in 1.5 ml tubes and covered in 1x restriction enzyme buffer. After 15 minutes at room temperature, the liquid was replaced with fresh buffer and 1 µl of bovine

serum albumin and 50 U of restriction endonuclease *Asel*, *DraI* or *AflII* was added. The mix was incubated overnight at 37°C. In case of *Asel* digestion, a second sample of 10 U of enzyme was added after 2 hours from the start. The next day, the liquid was replaced with TE buffer.

2.6.3 Loading the gel

100 ml of 1% agarose prepared in 0.5x TBE buffer was melted and cooled down. The plugs were carefully placed on the wells of the comb and most of the agarose was poured on the gel casting unit with the comb inserted. Once the gel is settled the comb was removed and wells were filled with the rest of the 1% agarose.

2.6.4 Electrophoresis and gel imaging

The gels were placed in the electrophoresis chamber with 2.2 l of 0.5x TBE buffer circulating at 14°C with the pump set at 70%. The electrophoresis conditions were set at 6 V/cm, with an initial switch time of 70 seconds and a final switch time of 130 seconds. The electrophoresis was performed for 24 hours. The gel was then stained for 30 minutes in 1 µg/ml ethidium bromide solution and de-stained with dH₂O for an hour. The gel was finally imaged with a UV trans-illuminator.

2.6.5 Alternative PFGE protocol using a phenol-chloroform pre-treatment

An alternative protocol published by Beyazova and Lechevalier, 1993 and described by Kieser (Kieser *et al.*, 2000) includes a pre-treatment of the mycelium with phenol and chloroform to reduce DNase activity. The mycelium is washed twice with 0.5 M EDTA pH8 and the pellets are dispersed using an autoclaved tissue grinder. The mycelium is then incubated with equal volume of water-saturated phenol at 65°C for 15 minutes. The cells are then washed three times with chloroform-isoamyl alcohol at room temperature. The cells in the aqueous phase are collected and incubated in 2% SDS-5mM EDTA at 65°C for 30 minutes. The cells are then washed three times in 5 mM EDTA pH8 + 1 mg/ml BSA and three more times with 10 mM EDTA pH8. The mycelium is then resuspended in TE buffer and agarose plugs are prepared and treated following the protocol described previously.

2.6.6 Alternative PFGE protocol using HEPES buffers

An alternative protocol described by Evans & Dyson 1993 was used to obtain gels with less DNA degradation, that substitutes TRIS for HEPES in buffers. For this the mycelium was washed in HES buffer (25 mM HEPES-NaOH, 25 mM EDTA, 0.3 M Sucrose pH:8) and digested in 1 mg/ml lysozyme in HES buffer. After lysis, plugs were washed three times with NDS (1%

N-laurylsarcosine, 0.5M EDTA, 10mM Glycine, pH:9.5). The plugs were incubated overnight with NDS containing 1mg/ml Proteinase K and 1mM CaCl₂. The next day, the plugs were washed following the same steps as described in 2.9.1 substituting TE buffer for HE buffer (10mM HEPES-NaOH, 1mM EDTA, pH:8). For the electrophoresis TBE buffer was substituted by HEPES buffer (16mM HEPES-NaOH, 16mM Sodium acetate, 0.8mM EDTA, pH:7.5) and the voltage was reduced to 4 V/cm.

2.6.7 PFGE gel images process and fragment size estimation

PFGE Images were processed and edited using ImageJ software.

In silico digest of genome sequences of *S. coelicolor* and *S. clavuligerus* was performed using SnapGene.

2.7 Genomic DNA isolation

25 ml of *S. clavuligerus* cultures grown in GYM broth or TSB at 26°C for around 40 hours at 200 rpm were centrifuged and the pelleted mycelium was re-suspended in 500 µl of lysozyme buffer (25 mM Tris, 25 mM EDTA, 0.3M sucrose, pH=8) containing lysozyme (2.5 mg per 50 mg mycelium) and RNase A (0.06 mg per 50 mg mycelium). The mixture was then incubated at 37°C for 60-90 minutes under constant shaking until the sample is clear. 10 µl of 10% (w/v) SDS was added and samples were incubated at 55°C for 15 minutes and then transferred to 37°C for 30 more minutes. 0.25 mg/50 mycelium of Proteinase K was then used to digest the samples for 2 – 2.5 hours at 55°C until clear of clumps. 200 µl of KCl (250 mM) was added to precipitate SDS, followed by 500 µl of phenol:chloroform:isoamyl alcohol (25:24:1). Samples were vortexed and centrifuged for 5 minutes at 14,000 rpm and the upper phase was transfer to fresh tubes. The phenol-chloroform extraction was repeated twice, as well as two more times with 500 µl of pure chloroform. 60 µl of 3M NaAc was added to 500 µl of upper phase, as well as 300 µl of isopropanol and mixed by inversion. The samples were then centrifuged for 10 minutes at maximum speed and the pellet was washed with 70% (v/v) ethanol and centrifuged. The ethanol was discarded, and samples were left to dry before being re-suspended in distilled H₂O. The DNA concentration was measured with a NanoDrop 2000c Spectrophotometer at 260 nm (Thermo Fischer Scientific). The absorbance ratio 260/280 was used to quantify the DNA purity, considering a ratio >1.8 as pure DNA.

2.8 Whole genome sequencing of *S. clavuligerus* strains and bioinformatics

2.8.1 MiSeq Illumina sequencing of *S. clavuligerus* DSM 738

The genome of *S. clavuligerus* DSM 738 was sequenced using the Illumina MiniSeq System.

S. clavuligerus DSM 738 genomic DNA was quantified using the Qubit and 500 ng in 30 µl of ddH₂O was used to prepare the indexed paired-end library using the Illumina Nextera™ DNA Flex Library Prep Kit which is based on an on-bead tagmentation process. We followed the Nextera™ DNA Flex Library Prep manual instructions (October 2018 version) and used 5 cycles of DNA amplification.

The library was then quantified using Qubit and the quality was checked using the Agilent 2100 Bioanalyzer System.

The library was then diluted to 1 nM in 100 µl and denatured following the manual's specifications. The library was finally loaded into the Illumina MiniSeq sequencer.

2.8.2 PacBio sequencing and contig assembly of *S. clavuligerus* DSM 738

The genome of *S. clavuligerus* DSM 738 was also sequenced using Pacific Biosciences long-read sequencing (PacBio) in Northumbria University. 9 µg of genomic DNA in 100 µl of ddH₂O was sent to be sequenced, with prior quality check by performing electrophoresis on genomic DNA on a 1% agarose gel.

The genomic DNA was further size sheared to ~9 Kbp using Covaris g-Tube and visualised using Agilent Bioanalyzer. Single-stranded DNA was removed by Exonuclease VII treatment and the double-stranded DNA was blunt ended. The DNA fragments were then ligated to blunt hairpin SMRTbell adapters and sequencing primers v3 were annealed to the SMRTbell template. PacBio Polymerase was bound to the SMRT template using the Sequel Binding kit 2.1 and DNA was then sequenced on the PacBio Sequel instrument using 10-hour movie capture.

High accurate consensus reads were generated using Circular Consensus Sequencing (CCS) tools and assembled into contigs using HGAP4.

2.8.3 HiSeq Illumina sequencing of *S. clavuligerus* strains

The genomes of *S. clavuligerus* DSM 738 and mutants M1 M2 and Mc3 were sequenced using the services of MicrobesNG (Birmingham), which uses the Illumina HiSeq sequencing platform.

14 to 18 ng of genomic DNA in 90 µl dH₂O was sent to MicrobesNG. Briefly, the DNA was quantified by a Quantit dsDNA HS assay in an Ependorff AF2200 plate reader and DNA libraries were constructed using the Nextera XT Library Prep Kit (Illumina, San Diego, USA) using 2 ng of genomic DNA and a PCR elongation time of 1 minute. Both the DNA quantification and library preparation were carried out with a Hamilton Microlab STAR automated liquid handling system. Furthermore, the DNA libraries were quantified with the Illumina Kapa Biosystems Library Quantification Kit on a Roche light cycler 96 qPCR machine. Sequencing of the libraries was then performed with the Illumina HiSeq using a 250bp paired end protocol. Following sequencing, reads were trimmed with Trimmomatic 0.30 with a sliding window quality cut-off of Q15 (Bolger, Lohse and Usadel, 2014).

2.8.4 Bioinformatic analyses of the genome of *S. clavuligerus* DSM 738

De novo assembly of *S. clavuligerus* DSM 738 was achieved by combining Illumina and PacBio data. The Illumina reads and PacBio contigs were assembled using the SPAdes assembler (Galaxy version 3.12.0+galaxy1) (Bankevich *et al.*, 2012). The PacBio contigs were manually mapped to the SPAdes contigs and assembled using SnapGene (**Figure 2.1**). The obtained closed sequence was polished using Pilon (Galaxy version 1.20.1) (Walker *et al.*, 2014) with the original Illumina reads. The assembly quality was assessed using Quast (Galaxy version 5.0.2+galaxy0) (Gurevich *et al.*, 2013).

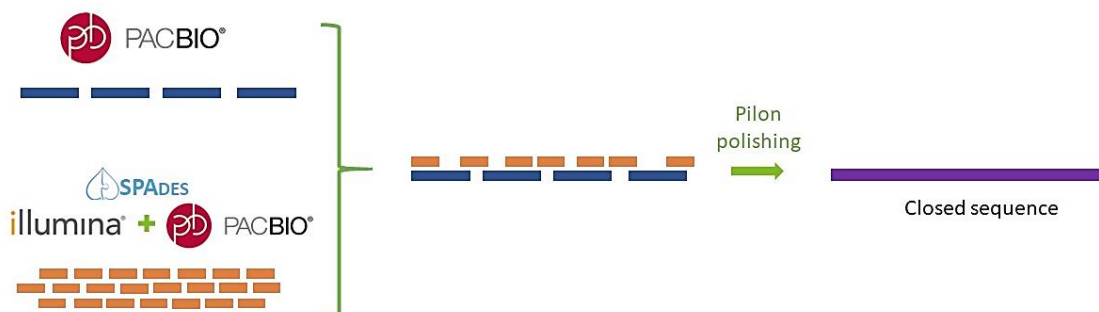


Figure 2.3. Graphical representation of the assembly pipeline followed to obtain a closed sequence from Illumina contigs (orange blocks) and PacBio contigs (blue blocks).

The genome was annotated using the RAST (Rapid annotation using subsystem technology) server version 2.0 and annotations subsystems were identified using The SEED Viewer version 2.0. Furthermore, secondary metabolite clusters were predicted by antiSMASH version 5.1.0 (Blin *et al.*, 2019).

Mapping of Illumina reads against assemblies was performed with BWA (Burrows-Wheeler Aligner) version 0.6.1-r104 using the bwasw algorithm with default options. The alignments were then converted to BAM format, sorted and indexed using SAMtools version 0.1.18. This was achieved using the following commands described by Gomez-Escribano *et al.* 2015:

```
bwa index assembly.fasta  
  
bwa bwasw assembly.fasta illumina_reads_pair1.fastq illumina_reads_pair2.fastq >  
  alignment.sam  
  
samtools view -b -S -o alignment.bam alignment.sam  
  
samtools sort alignment.bam alignment-sort  
  
samtools index alignment-sort.bam
```

For further coverage assessment, BAM files were converted into BIGWIG using bamCoverage (Galaxy version 3.3.0.0.0) (Ramírez *et al.*, 2016), and visualised using Integrated Genome Browser (IGB) version 9.1.2 (Freese *et al.*, 2016). Mean coverage data was obtained using Qualimap version 2.2.1 (Okonechnikov *et al.*, 2016).

Genome alignments were performed using Mauve version 20150226 (Darling *et al.*, 2004) and Easyfig version 2.2.2 (Sullivan *et al.*, 2011). Ortholog gene clusters were identified using OrthoVenn2 web server (Xu *et al.*, 2019). Genomic maps were obtained with CGView Server (Grant and Stothard, 2008), including Prokka annotations and GC skew.

A core genome BLAST database was created with amino acid sequences, and protein BLAST searches were performed using the following command lines on the SIPBS microbiology Linux server:

```
makeblastdb -in Core_genome.fsa -parse_seqids -blastdb_version 4 -title  
  "Core_Genome" -dbtype prot  
  
blastp -db Core_genome.fsa -query Query.fsa -evalue 1e-200 -out Blast_results.txt
```

2.9 Purification and sequencing of telomeric sequences

The purification of the telomeric sequences was performed following an adaptation of the protocol described by Fan *et al.*, 2012.

1 to 5 µg of genomic DNA of *S. clavuligerus* DSM 738 was digested overnight with the blunt-end restriction enzymes: *EcoRV*, *SmaI* or *ZraI*. The products were then purified using the Promega PCR-Clean Up kit and eluted in a final volume of 90 µl with dH₂O. 10 µl of 1 M NaOH was added and the mixture was incubated for 1 hour at 37°C. The samples were neutralised using 2 M HCl, and 1 M Tris pH8 was added to a final concentration of 0.1 M. 20X SSC solution was then added to final concentration of 2X and the samples were incubated at 68°C for 1 hour. The samples were purified once again with the Promega PCR-Clean Up kit and ligated overnight using Promega T4 DNA ligase. Inverted PCR analyses were performed on the ligation products and the amplified products were purified and sequenced by Eurofins Genomics.

The obtained telomeric sequences were added to the assembled genome sequences using SnapGene. Furthermore, alignments (MUSCLE algorithm) and maximum likelihood phylogenetic trees of telomeric sequences were performed using MEGA version 7.0.18 (Kumar, Stecher and Tamura, 2016) and Jalview version 2.10.5 (Waterhouse *et al.*, 2009). Phylogenetic trees were constructed using 1000 bootstraps and further edited using iTOL web server 5.5 (Letunic and Bork, 2019).

The folded forms of the telomeres were obtained using mfold web server (Zuker, 2003) with a folding temperature of 30°C.

2.10 Bioinformatic analysis of terminal protein sequences

The amino acid sequences of Tap and Tpg from different *Streptomyces* strains were pulled from genomes obtained from NCBI (<https://www.ncbi.nlm.nih.gov/genbank/wgs/>). Amino acid sequence alignments (ClustalO algorithm) and maximum likelihood phylogenetic trees were performed with MEGA version 10.1.7 (Kumar *et al.*, 2018) for Tap and Tpg proteins using 1000 bootstraps.

2.11 Southern blotting

The protocol followed is based on vacuum-based transference of digested DNA to a nylon membrane using the buffers described by Sambrook *et al.*, 1989 (**Table 2.6**). The probe was labelled with digoxigenin (DIG) using the DIG DNA Labelling and Detection Kit (Roche Life Science) which is based on the random priming of hexanucleotides dATP, dCTP, dGTP and DIG-labelled dUTP by Klenow Enzyme. Detection was achieved with a chromogenic substrate as described by Roche Life Science.

2.11.1 Probe labelling

The probe was prepared using 15 µl containing 2 µg of amplified gene/region of interest. The amplicons were boiled at 95°C for 10 minutes and then cooled down in ice for 10 more minutes. 2 µl of DIG DNA labelling mix, 2 µl of Hexanucleotide mix 10x and 1 µl of Klenow enzyme were added and the mix was incubated at 37°C overnight. The reaction was stopped by adding 2 µl of 0.2 M EDTA (pH 8.0) and stored at -20°C.

Table 2.6. List of solutions and buffers used for Southern Blotting. SSC: Saline-sodium citrate; SDS: Sodium dodecyl sulphate; NBT: Nitro blue tetrazolium chloride; BCIP: 5-Bromo-4-chloro-3-indolyl phosphate.

Application	Solution	Constituent
For DNA transfer	Depurination solution	19 ml 13 M HCl in 1 L dH ₂ O
	Denaturation solution	87 g NaCl, 20 g NaOH in 1 L dH ₂ O
	Neutralisation solution	77.8 ammonium acetate in 1 L dH ₂ O
	20x SSC	175.3 g NaCl, 88.2 g Sodium citrate in 1 L dH ₂ O
	2x SSC	20x SSC diluted 10 x in dH ₂ O
For Pre-hybridisation and hybridisation	Maleic acid buffer	22.2 g Maleic acid, 17.6 g NaCl in 1 L dH ₂ O
	10% Blocking solution	5 g blocking reagent in 50 ml maleic acid buffer
	Standard Hybridisation Buffer (SHB)	5x SSC, 0.1% N-laurylsarcosine (w/v), 0.02% SDS (w/v), 1% Blocking solution (v/v)
Stringency washes	Stringency Buffer A	2 x SSC, 0.1% SDS
	Stringency Buffer B	0.2 x SSC, 0.1% SDS
For detection	Detection buffer	12.1 g Tris-HCl, 5.85 g NaCl in 1 L dH ₂ O
	Washing buffer	0.3% (w/v) Tween-20 in Maleic acid buffer
	Antibody solution	Anti-DIG-AP diluted 1:5000 in 1% Blocking solution
	Colour-substrate solution	200 µl of NBT/BCIP, 10 ml Detection buffer

2.11.2 DNA digestion and electrophoresis

2.5 µg of genomic DNA of *S. clavuligerus* was digested with 50 units of restriction enzyme overnight at 37°C. The products were loaded on a 0.7% agarose gel and electrophoresis was

performed for 2 to 3 hours at 70 volts. Lambda DNA digested with *Hind*III was used as size marker.

2.11.3 DNA transfer to nylon membrane

3 sheets of Whatman paper soaked in 2x SSC buffer were placed on the base of the vacuum apparatus. The Hybond N+ nylon membrane (GE Healthcare) was then placed on top, followed by the plastic sheet which was prepared by cutting a transfer window smaller than the size of the membrane. The transfer tray was then clipped into the base of the apparatus and the gel was placed on top of the plastic sheet, centred with the membrane. 50 ml of Depurination Buffer was added onto the gel, and vacuum pressure of 50 mbar was applied for 20 minutes. This process was then repeated with 50 ml of Denaturation Buffer, and 50 ml of Neutralisation Buffer. The transfer was finalised by adding 50 ml of 20x SSC Buffer and applying vacuum for 1 hour or 4 hours in case of PFGE gels. The apparatus was then disassembled and the membrane was exposed to UV for 2 minutes in a UV Crosslinker (UVP). The membrane was washed in 2x SSC for 2 minutes.

2.11.4 Probe hybridisation

The membrane was wrapped in mesh and introduced in a Hybridization Tube. 20 ml preheated SHB was added and the tube was incubated at 65°C for 1 hour in a rotating Hybridization Oven (UVP). The probe was then denatured at 100°C for 15 minutes, and then chilled in ice for 10 more minutes. The SHB was decanted and substituted by 20 ml of fresh SHB containing the denatured probed. The membrane was then left overnight rotating at 65°C to hybridise.

2.11.5 Stringency washes

The membrane was incubated twice with 150 ml of Stringency Buffer A at 65°C for 15 minutes. The same procedure was carried out with 150 ml of Stringency Buffer B.

2.11.6 Detection

The membrane was rinsed with Washing Buffer for 30 seconds and washed for 30 minutes in 100 ml Blocking Solution with gentle shaking. The Blocking Solution was poured off and replaced with 30 ml of Antibody Solution and the membrane was incubated for 30 minutes with gentle shaking. The membrane was then washed twice with 100 ml of Washing Buffer for 15 minutes each and then equilibrated for 2 minutes in 20 ml Detection buffer. 10 ml of freshly prepared Colour Substrate Solution was added to the membrane and sealed in a

plastic bag. The membrane was left in the dark until colour developed and the reaction was stopped by rinsing thoroughly with water.

2.12 Molecular biology and cloning

Plasmids were constructed following the method of restriction cloning including PCR amplification of target genes, ligation to cloning and shuttle vectors, transformation of competent cells. *S. clavuligerus* mutants were constructed using intergenic conjugation of plasmids from *E. coli* to *S. clavuligerus*.

2.12.1 DNA amplification

Polymerase Chain Reaction (PCR) was used to amplify genetic material from genomic DNA sample of *S. clavuligerus* DSM 738. Either the Bioline Standard My Taq, NEB GoTaq or the Promega Q5 High-Fidelity Polymerase protocol was followed using the primers listed on **Table 2.5** with the addition of 2% of Dimethyl sulfoxide solution (DMSO). The Veriti™ Thermal Cycler of Thermo Fisher Scientific was used following 30 to 35 cycles of denaturation at 95°C for 45 seconds, primer annealing at the respective T_m (**Table 2.5**) for 30 seconds, and extension at 72°C for 1-2 minutes (30 seconds per 1 kb of DNA to amplify).

E. coli colony PCR analyses were performed by adding a bacterial colony instead of DNA template into the PCR mix and incubating the samples at 95°C for 10 minutes before the PCR cycles.

Streptomyces colony PCR were performed by suspending mycelium from an overnight culture in 50% DMSO solution and incubating at 50 °C for 30 min. 5 µl of the mycelium was added to the PCR mix.

2.12.2 Purification of DNA amplicons and plasmids

PCR amplicons were purified using QIAGEN QIAquick®, Bioline PCR Purification and Promega Wizard® SV Gel and PCR Clean-Up System kits which follow the spin column-based DNA isolation method. Plasmids were purified using QIAGEN QIAprep Spin Miniprep Kit and Wizard® Plus SV Minipreps DNA Purification System from a 5 ml overnight culture containing LB and the selective antibiotic. Purified DNA was eluted in 50 µl dH₂O and the concentration was measured with a NanoDrop 2000c Spectrophotometer at 260 nm (Thermo Fischer

Scientific). The absorbance ratio 260/280 was used to quantify the DNA purity, considering a ratio >1.8 as pure DNA.

2.12.3 Agarose gel electrophoresis

Bioline 5x DNA Loading Buffer Blue was added to the DNA samples and loaded on a 1% agarose gel in TAE buffer (40 mM Tris Base, 20 mM glacial acetic acid, 1 mM EDTA). Electrophoresis was performed at 90-110 V for 45-60 minutes in TAE buffer. The gel was stained in 30% (v/v) GelRed™ Nucleic Acid Gel Stain 0.1 M NaCl solution for 30 minutes and bands were imaged using Azure c200 Gel Imaging Workstation (Azure Biosystems). Otherwise, 1-5 μ l of 10 mg/ml Ethidium Bromide solution was added to the agarose and the gel was imaged using the VWR GenoSmart UV Transilluminator. Gel bands were usually excised with a scalpel and DNA was purified using the Promega Wizard® SV Gel and PCR Clean-Up System

2.12.4 Ligation

Purified DNA genes were ligated to cloning or shuttle vectors overnight at 4°C with 3 units of T4 DNA Ligase (Promega) and T4 DNA Ligase Rapid Ligation Buffer. The amount of insert and vector DNA followed a ratio of 1:5 (vector:insert).

2.12.5 Transformation of *E. coli* competent cells

E. coli cells were transformed by adding 3-10 μ l of ligation products to 50 μ l of JM109 Ready-to-use Competent Cells (Promega). The mix was incubated in ice for 20 minutes and then heat-shocked 45 seconds at 42°C, followed by a 1-hour incubation (3 hours in the case of *E. coli* ET12567/pUZ8002) at 37°C with 750 μ l of LB broth. The cells were then plated on LB agar containing the selective antibiotic and incubated at 37°C overnight. In the case of vectors containing *lacZ* gene at cloning site, such as pGEM-T Easy and pBGS19, transformation plates also contained 100 μ M Isopropyl β -D-1-thiogalactopyranoside (IPTG) and 40 μ g/ml X-Gal for subsequent blue-white screening of colonies.

2.12.6 Restriction digestion

Digestion with restriction enzymes was carried out following the instructions described by Promega or NEB. In general, 10 units of restriction enzymes were used to digest 500 ng of DNA, with the addition of 10% Restriction Enzyme 10X Buffer and 1% of Acetylated Bovine serum albumin (BSA) 10 μ g/ μ l. The digestion was performed at 37°C for 4 hours or overnight. Negative controls were prepared with the same components, except for the restriction

enzyme. Following digestion, vectors were often dephosphorylated at the 5' by adding 1 unit of Shrimp Alkaline Phosphatase and incubating for further 30 minutes at 37°C.

2.12.7 Preparation and transformation of chemical competent cells

In the case of *E. coli* ET12567/pUZ8002 and *E. coli* BTH101, chemical competent cells were prepared manually. A single colony of *E. coli* was inoculated in 5 ml of LB broth containing the selection antibiotic and incubated overnight at 37°C at 200 rpm. 1 ml of culture was then inoculated in 100 ml of LB Broth and grown until reaching an Optical Density at a wavelength of 600 nm (OD₆₀₀) value of 0.5. The cells were collected by centrifugation for 10 minutes at 4000 rpm and gently re-suspended in 1 ml of cold 0.1 M CaCl₂. The suspension was kept in ice for 20 minutes and centrifuged for 10 minutes at 4000 rpm. The supernatant was discarded, and the cells were re-suspended in 1 ml of 0.1 M CaCl₂. Samples were frozen using liquid nitrogen and kept at -80°C.

2.12.8 Intergenic conjugation from *E. coli* to *S. clavuligerus*

E. coli ET12567/pUZ8002 containing the complementation vectors were conjugated with *S. clavuligerus* DSM 738 following the protocol described by Kieser *et al.* 2000. The ET strains were grown in LB broth in the presence of kanamycin, chloramphenicol and the respective antibiotic for plasmid selection overnight at 37°C. The culture was diluted 1:50 in fresh LB plus antibiotics to an OD₆₀₀ of around 0.5 and washed twice with an equal volume of fresh LB for finally being suspended in 0.1 volume of LB broth. 500 µl of 2x YT broth were added to 200 µl of *S. clavuligerus* spores and heat-shocked at 50°C for 10 minutes. 500 µl of ET cells were added to the heat-shocked spores and the mixture was plated on SFM, GYM or L3M9 agar containing 10 mM MgCl₂ and incubated at 26°C for 16 hours. The plates were then overlaid with 1 ml of dH₂O containing 0.5 mg nalidixic acid and 1 mg of the selective antibiotic. Incubation was continued at 26°C for 4 more days. Transconjugants were streaked on GYM or L3M9 agar and spore stocks were prepared.

2.12.9 *In silico* cloning

Plasmids were constructed *in silico* with SnapGene Software using restriction cloning.

2.13 Bacterial Two-Hybrids experiments

The BACTH system is based on reconstitution of the adenylate cyclase activity mediated by protein-protein interactions in *E. coli* (Karimova *et al.*, 1998).

The *tap* and *tpg* genes from pSCL4, pSCL3 and pSCL2 were cloned into the pKT25, pKNT25, pUT18 and pUT18C vectors following the methodology described on Section 2.12. Forward primers were designed in order to maintain the genes in frame with the AC subunits (in pKT25 and pUT18C vectors) or the AC promoter (pKNT25 and pUT18). Reverse primers were designed disrupting the *tap/tpg* stop codon, so only the AC subunit stop codon would be expressed, unless this latter one was eliminated by the cloning. Sets of T25 and T18 plasmids were co-transformed into chemically competent *E. coli* BTH101 cells, and grown at 30°C for 1 to 2 days in LB containing Kan, Carb and X-Gal (40 µg/ml).

2.14 Microscopy

2.14.1 Schwedock Staining of samples

The Schwedock Staining method is based on staining streptomycetes mycelium grown on a coverslip with fluorescein isothiocyanate-wheat germ agglutinin (FITC-WGA), which stains peptidoglycan, and propidium iodide (PI), which stains nucleic acids (Schwedock *et al.*, 1997).

Square glass coverslips (Type 1.5, 22 mm x 22 mm) were sterilised by being soaked in ethanol and flamed, and were then inserted into L3M9 agar, leaving an acute angle between the coverslip and the agar. Streptomycete spores or liquid cultures were inoculated at low density at the interface between the agar and the coverslip and grown in the incubator at 26°C.

The coverslips containing the mycelium were pulled from the agar and treated with 500 µl fixative solution (2.8% paraformaldehyde, 0.0045% glutaraldehyde in PBS, phosphate-buffered saline) at room temperature for 15 minutes. They were then washed twice with PBS and left to air dry thoroughly. The coverslips were then rehydrated with PBS for 5 minutes and treated with GTE (50 mM glucose, 20 mM Tris-HCl, pH 8, 10 mM EDTA) for 1 minute at room temperature. They were then washed with PBS and incubated with PBS containing 2% of BSA (bovine serum albumin) for 5 minutes. The coverslips were then incubated with 500 µl of staining solution (2 µg/ml FITC-WGA, 10 µg/ml PI, 2% BSA in PBS) for 3 hours at room temperature in the dark. They were then washed eight times with a second staining solution containing 10 µg/ml PI in PBS. The coverslips were left to air dry and mounted onto the slides using 10 µl of 40% glycerol solution. The coverslips were sealed using nail varnish and stored at -20°C before being thawed prior to imaging.

Slides of liquid *S. clavuligerus* cultures were prepared by adding 50 μ l of culture onto coverslips and they were then left at room temperature until fully dry. Fixation and staining were then performed following the same procedure as used for solid cultures.

2.14.2 Imaging stained samples

The slides were visualised using a TE-2000 inverted widefield epi-fluorescence microscope with a x100/1.30 numerical aperture oil immersion objective lens (Nikon, Japan). Phase contrast and fluorescence images were acquired using a Hamamatsu ORCA-100 CCD camera. Illumination was sourced from a mercury arc lamp which was coupled into the read epi-port of the microscope. A source blocking filter was used for fluorescence imaging. FITC signal ($\lambda_{\text{excitation}}=490$ nm, $\lambda_{\text{emission}}=512$ nm) was detected using a FITC emission filter (Nikon, Japan) and PI signal ($\lambda_{\text{excitation}}=560$ nm, $\lambda_{\text{emission}}=617$ nm) was detected using a 600 nm long-pass emission filter (Nikon, Japan).

2.14.3 Processing and analysis of microscopy images

The images were processed using FIJI (Schindelin *et al.*, 2012), using a custom macro script (in Appendix) to separate composite images, scale them, automated scale contrast, and merge fluorescence and phase contrast channels into one image for further downstream analyses.

Branch length and inter-branch distance measurements were taken using the Simple Neurite Tracer plugin for FIJI and the hyphae width was also measured using FIJI straight line selection. Violin plots were created using PlotsOfData web server (Postma and Goedhart, 2019). One-way ANOVA followed by Tukey's multiple comparison test were performed using GraphPad Prism (version 8.3.1).

2.15 Growth curves

S. clavuligerus strains were grown in TSB media shaking for 30 hours at 26°C. OD₆₀₀ measurements were taken and the strains were standardised to an OD₆₀₀ of 0.4. 150 μ l of culture was added to 1.5 ml of TSB in a 24-well plate in triplicates. The plate was incubated in the Spectramax 190 plate reader at 26°C with fast continuous shaking and OD₆₀₀ was measured every 30 minutes for 70 hours.

The Specific Growth Rate (μ) was calculated by plotting the $\ln(\text{OD}_{600})$ over time exclusively at log phase (when cells double in a certain period of time), which yields a linear correlation where the slopes equals μ .

Data was plotted using GraphPad Prism (version 6.00 for Mac, GraphPad Software, La Jolla California USA), and statistics were analysed by performing One-way ANOVA followed by Tukey's multiple comparison test.

2.16 Clavulanic acid production assay

For liquid culture fermentations, *S. clavuligerus* strains were grown in TSB media shaking for 30 hours at 26°C. OD_{600} measurements were taken and the strains were subcultured into 25 ml of CM5 media for a final OD_{600} of 0.1. The flasks were incubated in a 26°C shaking incubator for 72 hours. The cultures were then centrifuged, and the aqueous supernatant was used for the clavulanic acid assay.

For solid media fermentations, *S. clavuligerus* strains were grown on CM5 agar for 21 days at 26°C. Plugs of cultured agar were carved using a cork borer and placed into 7 ml Bijou tubes. 2 ml of dH_2O per g of agar was added and the tubes were incubated for 24 hours at 4°C. The aqueous extract was then used for the clavulanic acid assay.

8 μl of sample or standard was added to 200 μl of imidazole reagent (10% Imidazole pH7) in a 96-well plate in triplicates. The plate was then incubated at room temperature for 30 minutes and the absorbance was measured at 324 nm using a Spectramax 190 plate reader. The standards were prepared by dissolving lithium clavulanate in MilliQ water and diluting to the following concentrations: 400, 350, 300, 250, 200 and 100 $\mu\text{g}/\text{ml}$.

Data was plotted using GraphPad Prism and One-way ANOVA followed by Tukey's multiple comparison test was performed for statistical analysis.

2.17 Bioactivity plug assay

Square plates were prepared by adding 25 ml nutrient agar. 10 ml of soft nutrient agar containing OD_{600} 0.1 *E. coli* JM109 or *E. coli* JM109::pUC19 was added on top of the nutrient agar. In the case of pUC19 plates, the agar was complemented with carbenicillin. *S. clavuligerus* strains were grown on CM5 agar for 21 days at 26°C. Plugs of cultured agar were

carved using a cork borer and placed over the *E. coli* lawn. Uncultured CM5 agar and clavulanic acid standards (as described as in clavulanic acid assay section) were used as controls. Plates were incubated at 37°C overnight and photos were taken.

3. OPTIMISATION OF PULSED-FIELD GEL ELECTROPHORESIS METHOD FOR THE STUDY OF THE GENOMIC ARCHITECTURE OF *S. CLAVULIGERUS*

Gel electrophoresis is a method used to separate charged macromolecules such as DNA, RNA and proteins based on their size. For this, an electric field is applied that forces the molecules to move through a matrix or gel of specific porosity allowing smaller molecules to migrate further while larger molecules stay behind. Although nowadays this method is mostly utilised for analytical purposes, such as detection of PCR amplification products, gel electrophoresis has played an important role in characterisation of bacterial genomes, allowing visualisation of chromosomal and extrachromosomal elements, as well as being a key step in chain-termination sequencing (Sanger and Coulson, 1975). With this work we proposed to use a type of gel electrophoresis called pulsed-field gel electrophoresis (PFGE) for the study of the *S. clavuligerus* genome.

While conventional electrophoresis techniques fail to separate DNA fragments larger than 50 kb, PFGE allows separation of DNA molecules between a few kb to over 10 Mb. This technique is based on the concept that large molecules can be separated using alternating electric fields, which was first described by Schwartz and Cantor, 1984. More specifically, during PFGE, DNA is subjected to alternating electrical fields at different angles for a given period of time or pulse. This being so, with each change in the voltage angle, DNA molecules rearrange their orientation before migration, with the reorientation time being proportionate to the molecule size. In this way, large molecules take longer to rearrange and have less time to move during each pulse than smaller molecules (Cantor *et al.*, 1988).

For this project, PFGE was performed using a Contour-Clamped Homogeneous Electric Fields (CHEF) instrument, which has 24 electrodes located on an hexagonal contour that generate uniform electric fields in alternating positions with a reorientation angle of 120° as indicated on **Figure 3.1** (Levene, 1992).

The ability to separate large DNA molecules is of great importance as it makes possible the acquisition of physical evidence of genomic elements with the visualisation of entire or fractionated chromosomes, as well as plasmids. This is particularly relevant in *Streptomyces* as many strains carry giant linear plasmids (GLP) that carry important functions necessary for expression and regulation of specialised metabolites.

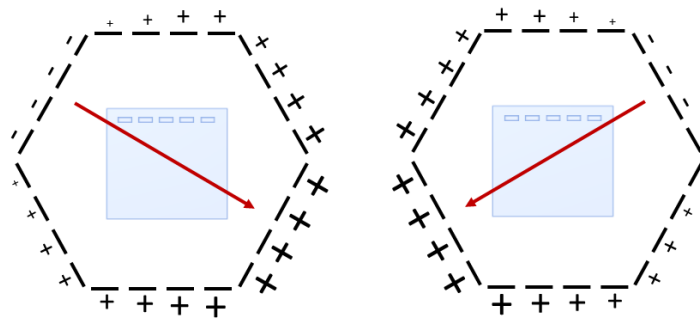


Figure 3.1. Schematic representation of the alternating orientations of the electric field during PFGE using a Contour-Clamped Homogeneous Electric Fields (CHEF) instrument. The red arrows indicate the orientation of the voltage, and the – and + indicate the position and charge of the electrodes.

Since its introduction, PFGE has been used for the characterisation of the genomes of *Streptomyces* strains. For example, PFGE allowed the construction of physical maps for the genome of *Streptomyces coelicolor*, which complemented the genetic maps, and allowed comparison to similar species (Kieser *et al.*, 1992; Leblond *et al.*, 1993). In addition, PFGE analyses made possible identification of giant linear plasmids in many species, such as *Streptomyces venezuelae*, *Streptomyces rochei*, *Streptomyces violaceoruber*, *Streptomyces fradiae*, *Streptomyces parvulus*, *Streptomyces lasaliensis* and *Streptomyces ambofaciens* (Kinashi and Shimaji, 1987; Leblond *et al.*, 1990). Furthermore, the presence of terminal proteins attached to the chromosome's ends of *Streptomyces lividans* and *Streptomyces griseus* was confirmed by PFGE analyses when comparing samples treated with and without proteinases (Lin *et al.*, 1993; Lezhava *et al.*, 1995).

In the case of *S. clavuligerus*, the first evidence of PFGE analyses published was an article focused on identification of giant linear plasmids in β -lactam antibiotic producing streptomycetes. In this article they identified two linear plasmids with sizes 120 and 430 kb in *S. clavuligerus* NRRL 3585, corresponding to pSCL2 and pSCL3 respectively (Netolitzky *et al.*, 1995). These two plasmids were again visualised using PFGE, as well as genomic DNA fragmented with *AseI* that showed five fragments of sizes 400 to 100 kb in an article studying the telomeres of *Streptomyces* chromosomes (Huang *et al.*, 1998). Another study used this technique to separate chromosomal DNA from plasmids DNA in *S. clavuligerus* ATCC 27064 in order to identify genes responsible for the regulation of clavulanic acid production (Song *et al.*, 2009). Despite these efforts, DNA molecules larger than 400 kb, such as the megaplasmid pSCL4, have never been visualised in *S. clavuligerus* strains. This has delayed the construction of a physical map of the whole genome of this bacterium and, as a result, the characterisation of its genomic architecture in relation to its genome sequence.

The aim of this chapter is to obtain a PFGE method optimised for *S. clavuligerus*, in order to obtain physical evidence for establishing the genome architecture of this organism.

3.1 Establishment of a PFGE protocol for *S. clavuligerus*.

A PFGE analysis was first performed on the widely studied model organism *S. coelicolor* M145 (Table 2.3 in Chapter 2) following the protocol described by Kieser *et al.*, 2000 in order to become familiarised with the procedure. As the genome of this strain is solely composed of a chromosome of 8 Mb, which is too large to be seen by PFGE, the samples were digested with the restriction enzyme *AseI* (AT'TAAT) that cuts infrequently in the GC rich DNA of streptomycetes. PFGE of *AseI* digested samples showed 10 bands of around 1500, 1300, 960, 900, 780, 600, 450, 360, 280, 200 kb (**Figure 3.2.A**). In order to compare the PFGE bands, a table of the expected fragment sizes was constructed using the sequences of *S. coelicolor* (published by Hsiao and Kirby, 2008) digested *in silico* with *AseI* (**Table 3.1**). The PFGE band sizes correlate with the expected fragment sizes and the undigested sample did not present any band inferring the chromosome did not migrate, as was expected. These results demonstrate that the PFGE protocol was operating successfully for *S. coelicolor*.

The same protocol was performed on the *S. clavuligerus* type strain DSM 738 and the industrial strain Sc6 (Table 2.3 in Chapter 2), except for the lysozyme treatment which was reduced from 2 hours to 5 minutes as suggested for *S. clavuligerus* by Kieser *et al.*, 2000. Undigested samples were loaded on the gel, as well as samples digested with *AseI*, *DraI* (TTT'AAA) and *AfIII* (C'TTAAG) enzymes. The PFGE performed on undigested *S. clavuligerus* 738 showed two bands at around 365 and 150 kb which could represent the plasmids pSCL3 and pSCL2 respectively. The digested samples of 738 with *AseI* and *DraI* showed the same bands as the uncut samples which suggest a failed restriction digestion, while the sample treated with *AfIII* shows additional bands at around 600 kb and 200-150 kb, however the expected larger bands are missing (**Figure 3.2.B**). In the case of Sc6, the uncut sample showed the same bands as the type strain and the samples treated with *AseI* and *DraI* showed a different band pattern while *AfIII* digest showed no bands at all (**Figure 3.2.C**). Estimated fragment sizes were obtained using sequences of *S. clavuligerus* published previously as reference (**Table 3.1**) (Cao *et al.*, 2016; Medema *et al.*, 2010; Song *et al.*, 2010). In the case of the *AseI* digest, the gel showed 3 bands, the two bigger bands could represent the 235 and 227 kb fragments of digested pSCL3. The *DraI* digest showed the same bands as the uncut sample which contributes to the fact that those bands represent pSCL3 and pSCL2 as neither

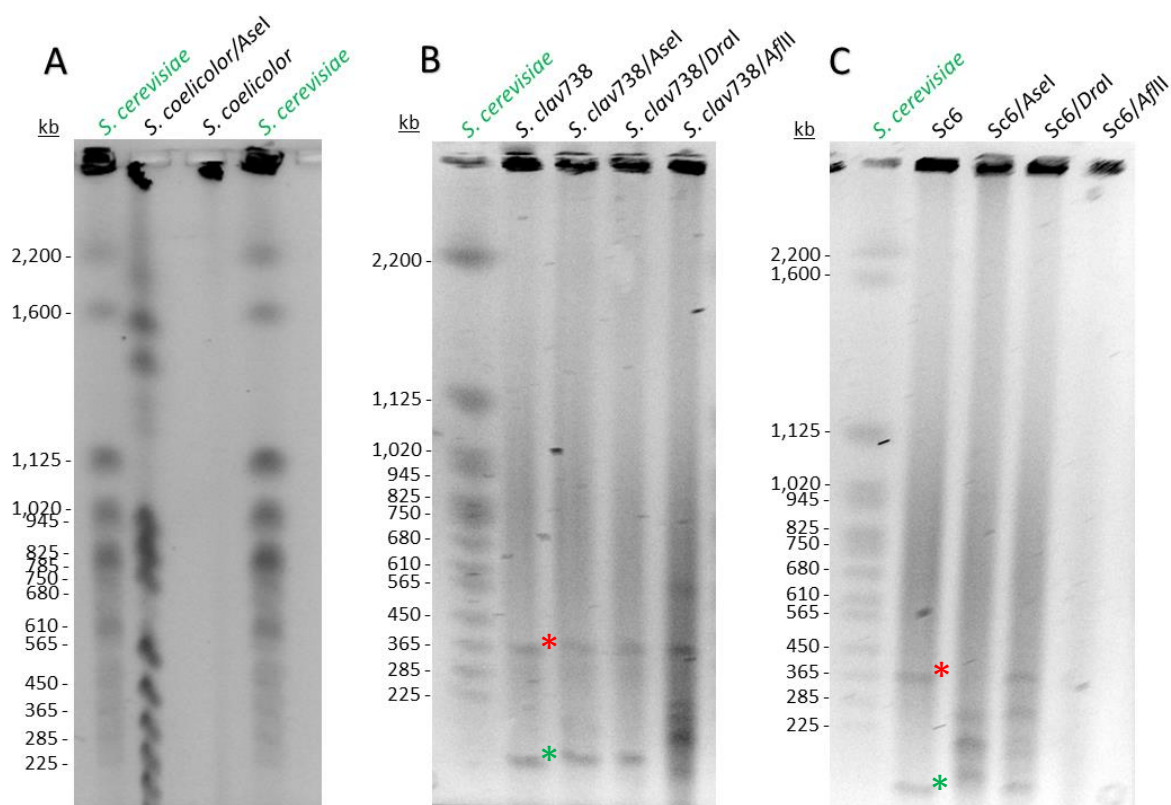


Figure 3.2. PFGE of the strains *S. coelicolor*, *S. clavuligerus* DSM 738 and Sc6, performed at 6 V/cm for 24 hours. *Saccharomyces cerevisiae* YNN295 was used as DNA size standard (Bio-Rad). **A:** PFGE of *S. coelicolor* digested with *AseI* and undigested. **B:** PFGE of *S. clavuligerus* DSM 738 undigested, and digested with *AseI*, *DraI* and *AflIII*. **C:** PFGE of Sc6 strain undigested, and digested with *AseI*, *DraI* and *AflIII*. Asterisks indicate pSCL3 band (red) and pSCL2 band (green).

Table 3.1. List of expected band sizes (kb) of the *S. coelicolor* M145 and *S. clavuligerus* genomes when digested with *AseI* (AT'TAAT), *DraI* (TTT'AAA) and *AflIII* (C'TTAAG) nucleases. Chromosomal fragments are indicated in black, while pSCL4 fragments are indicated in blue and pSCL3 in red.

<i>S. coelicolor</i>	<i>S. clavuligerus</i>		
	<i>AseI</i> digests (kb)	<i>AseI</i> digests (kb)	<i>DraI</i> digests (kb) <i>AflIII</i> digests (kb)
1,500	1,898	2,707	1,337
1,300	1,612	1,067	833
960	929	1,033	831
900	684	805	637
820	442	598	441
610	406	594	334
480	355	526	300
380	339	506	293
290	328	444	284
200	275	279	255
190	250	238	242
140	234	200	241
110	233	115	228
45	227		225
16	211		200
11	187		194

of these are cut by this enzyme. The two other bands of around 250 and 200 kb could belong to the chromosome or pSCL4, but the expected larger fragments are still missing.

Overall, these results suggest that while this particular protocol worked well for *S. coelicolor*, it does not allow migration of bands larger than 500 kb in *S. clavuligerus*; a fact that impedes the visualisation of large chromosomal fragments and pSCL4. In addition, *S. clavuligerus* samples on PFGE show a large amount of smearing which could be due to the action of nucleases in the samples and makes evident the need for optimisation of the sample preparation protocol for *S. clavuligerus*.

3.2 Lysis Time is not a significant factor for PFGE of *S. clavuligerus*.

One of the explanations considered to be responsible for the lack of visualisation of large bands by PFGE was that DNA release was hindered due to a poor cell lysis. For this reason, in order to study the optimum lysozyme treatment time, samples from *S. clavuligerus* DSM 738 were treated with lysozyme over 13 different times from 5 minutes, the time recommended for *S. clavuligerus* (Kieser *et al.*, 2000), until 2 hours, which is the time recommended for most *Streptomyces* species, and with a 10 minute difference in between each time point. Moreover, all samples were digested with *AseI*.

PFGE results indicate that all the treatment time point samples showed the same band pattern: 5 clear bands in between 400 and 150 kb (**Figure 3.3**). No significant difference was seen in between the samples, with the exception of samples treated for 5 and 10 minutes that showed slightly weaker bands, for this reason it was preferred to use a 30-minute lysozyme treatment for the following PFGE analyses.

In addition, all the samples still showed a great deal of smearing and no large bands were found; bands that were expected after *AseI* digestion. All these results suggest that the lysis time is not a significant factor for the performance of PFGE on *S. clavuligerus* and further conditions should be studied.

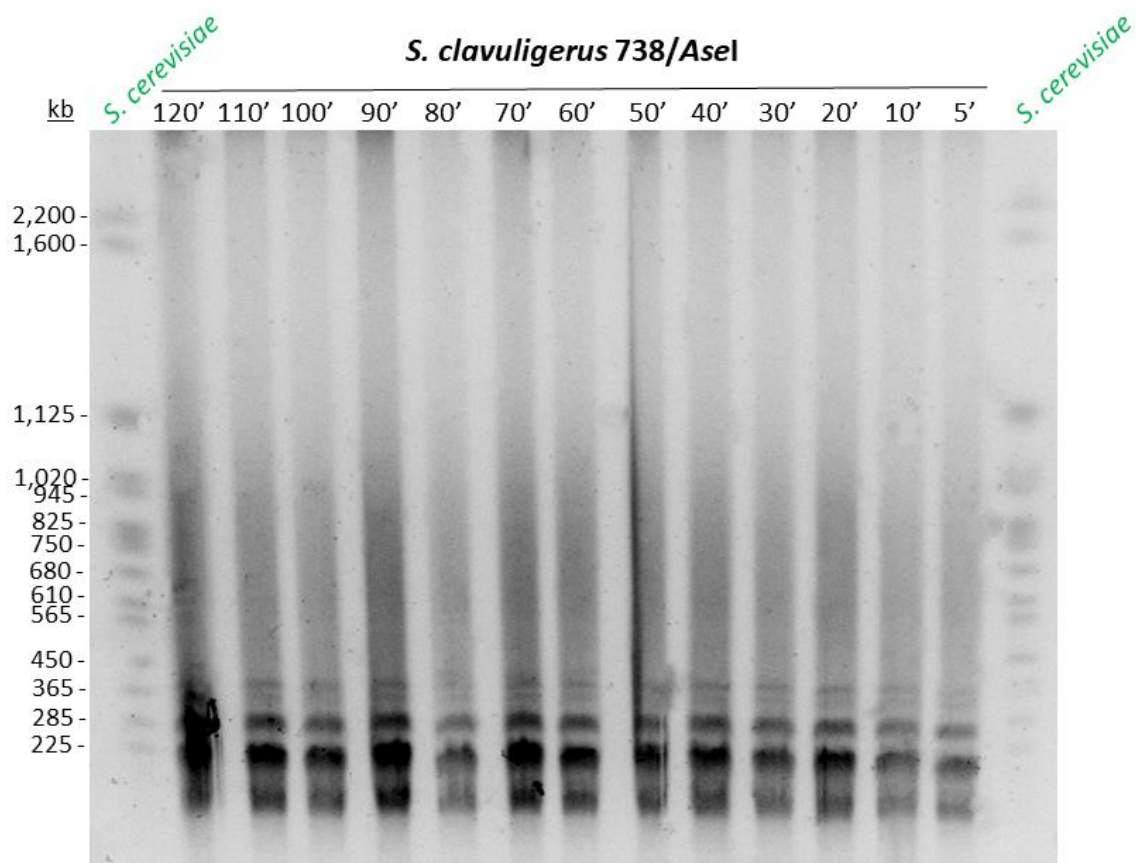


Figure 3.3. PFGE of *S. clavuligerus* DSM 738 treated with lysozyme in between 120 and 5 minutes and digested with AseI restriction enzyme. Electrophoresis was performed at 6 V/cm for 24 hours. *S. cerevisiae* YNN295 was used as DNA size standard (Bio-Rad).

3.3 Phenol-chloroform pre-treatment of cells does not reduce DNA degradation on *S. clavuligerus* samples.

In order to decrease the amount of smearing on the PFGE samples, a protocol published by Beyazova and Lechevalier in 1993 and described by Kieser (Kieser *et al.*, 2000) was performed on *S. clavuligerus* strains DSM 738, DSM 41826, Sc2 and Sc6. This protocol includes a cell pre-treatment using phenol and chloroform prior to the agarose embedment to help inhibit nuclease activity. In addition, the samples were digested with *AseI* or *DraI* restriction enzymes.

The PFGE results exhibited the same two bands at around 400 and 150 kb viewed in **Figure 3.2** on the undigested 41826, Sc2 and Sc6 samples, that correspond to pSCL3 and pSCL2, although no bands corresponding to pSCL4 (around 1.8 Mb) were seen (**Figure 3.4**). The samples digested with *AseI* presented a few bands in between 500 and 200 kb but no clear bands were identified on the *DraI* samples. These results confirm that either the phenol-chloroform treatment is not effective at inhibiting nucleases in *S. clavuligerus* or the smearing is due to other reasons.

3.4 Substitution of Tris by HEPES in buffers composition significantly increases the quality of PFGE analyses on *S. clavuligerus*.

An article published in 1993 showed how Tris-containing buffers induce DNA cleavage and therefore cause smearing of PFGE analyses of *S. lividans*. The authors also showed that this DNA degradation was significantly decreased when using HEPES instead of Tris in the electrophoresis running buffer (Evans and Dyson, 1993). In order to confirm if the smearing on our analyses was due to Tris-dependent DNA degradation, PFGE was performed using this alternative protocol on undigested samples of *S. clavuligerus* strains DSM 738, DSM 41826, Sc2 and Sc6. The voltage was reduced to 3 V/cm to keep the current below 200 mA as HEPES buffer has an increased ionic strength, and the running time increased to 48 hours.

The PFGE showed bands for 41826 Sc2 and Sc6 at around 1,800, 450 and 150 kb, sizes that corresponded to the predicted size of the replicons pSCL4, pSCL3 and pSCL2 respectively. This was the first time that we could visualise large band sizes (1,800 kb) (**Figure 3.5**). In addition, DNA degradation was significantly decreased, which suggests that indeed *S. clavuligerus* is susceptible to Tris buffers during PFGE. For this reason, the subsequent gels were performed using exclusively HEPES-based buffers.

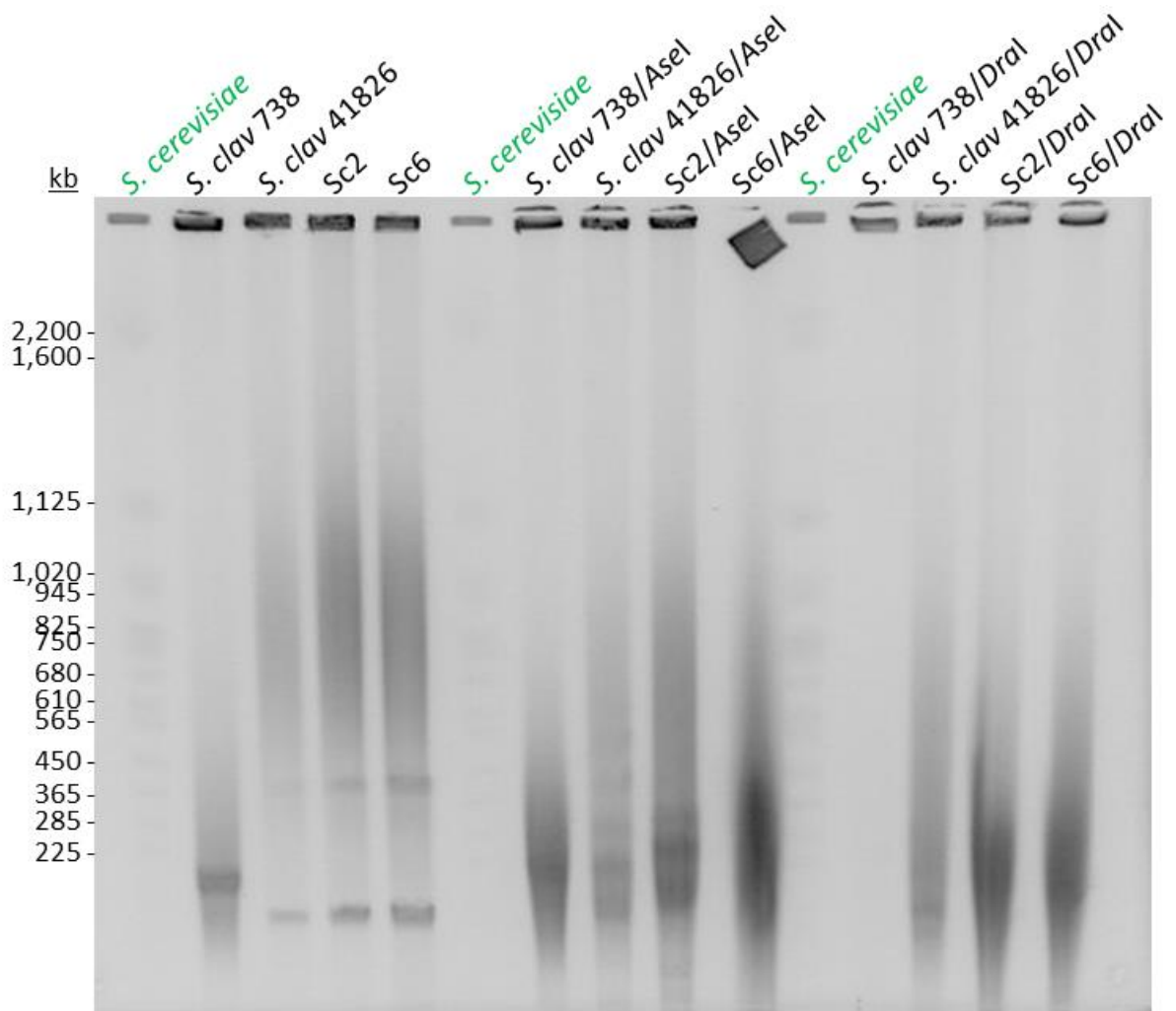


Figure 3.4. PFGE of *S. clavuligerus* samples DSM 738, DSM 41826, Sc2 and Sc6 pre-treated with phenol and chloroform. Samples were either loaded undigested (left part), digested with *AseI* (middle part) or digested with *DraI* (right part). Electrophoresis was performed at 6V/cm for 24 hours. *Saccharomyces cerevisiae* YNN295 was used as DNA size standard (Bio-Rad).

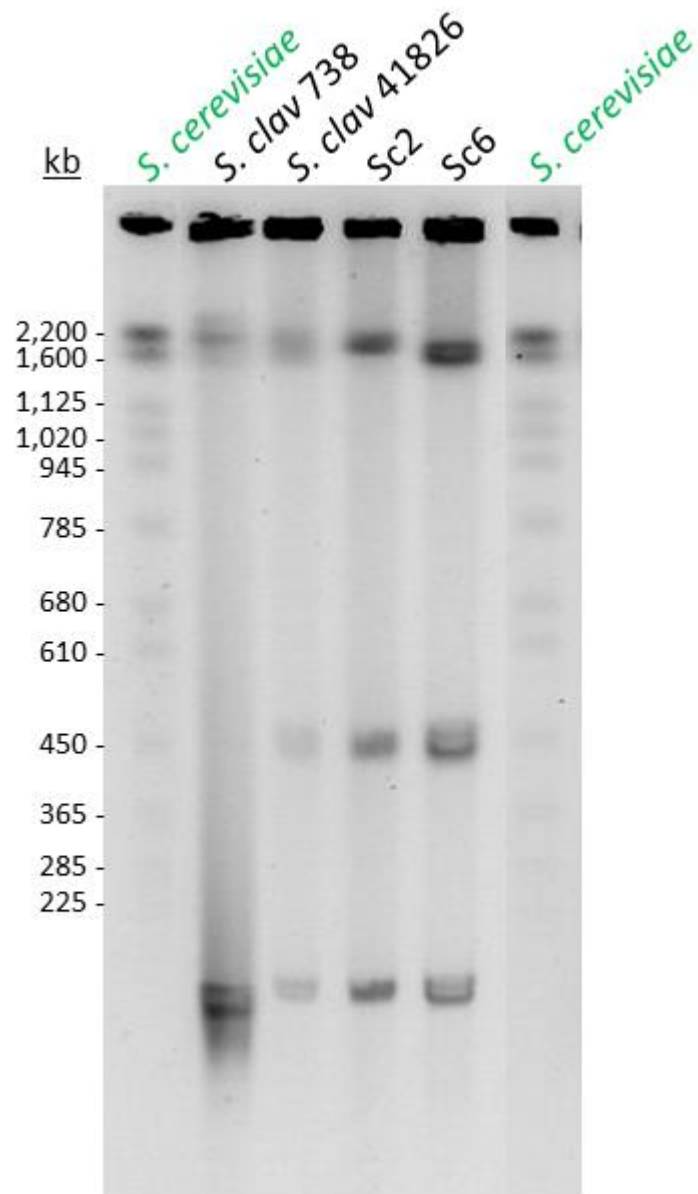


Figure 3.5. PFGE of undigested samples of *S. clavuligerus* strains DSM 738, DSM 41826, Sc2 and Sc6. Electrophoresis performed using HEPES buffer at 3 V/cm for 48 hours. *Saccharomyces cerevisiae* YNN295 was used as DNA size standard (Bio-Rad).

3.5. Higher amounts of mycelium give stronger signal bands on PFGE analyses but less specificity for Southern blotting.

In order to establish the optimum concentration of mycelium to be loaded into the agarose plugs, a PFGE analysis was performed on 6 samples of *S. clavuligerus* DSM 738 with different concentrations of mycelium in HE buffer. The samples were designated C1 to C6, each one being the 1:2 dilution of the previous one when C1 is the most concentrated and C6 is the least concentrated. In addition, the optical density at 600 nm (OD₆₀₀) was read for each of the samples measuring 2.2 for C3, 1.8 for C4, 1.2 for C5 and 0.6 for C6. Moreover, the samples were loaded either undigested or digested with *AseI*.

The PFGE results of the undigested samples showed the three bands of 1,800, 450 and 150 kb that correspond to pSCL4, pSCL3 and pSCL2, as well as an additional large band lower than 100 kb that could represent fragmented DNA. Bands were visible at all concentration except from C6, being C1 the one with the strongest signal (**Figure 3.6**). The samples digested with *AseI*, showed a similar signal strength pattern in between the 6 different concentrations. With bands at around 1,800, 1,600 (only C1), 930, 450, 365, 300, 200 and <100 kb, sizes that correspond to some of the expected fragment sizes previously indicated in **Table 3.1**. Moreover, this analysis was performed using HEPES instead of Tris on the buffers composition which correlates with the previous PFGE results for the lack of degradation on the samples.

Furthermore, in order to confirm which bands represent megaplasmid DNA, this PFGE gel was blotted to obtain a Southern blot using the genes *tap* and *tpg* as probe. These genes are exclusive of the megaplasmid pSCL4 as described by Medema *et al.*, 2010. The probe, as expected, hybridised to the 1,800 kb band on the undigested samples and the 930 kb band on the *AseI* digested ones, however, nearly all bands were highlighted on the blot, especially on the higher concentration samples (**Figure 3.7**). This could be due to unspecific hybridisation of the probe to the bands, which is particularly observable at higher DNA concentration.

These results suggest that although samples with higher mycelium concentration are better visualised on PFGE analyses, they increase the risk for unspecific probe hybridisation on Southern blots. For this reason, concentrations of mycelium giving OD₆₀₀ of 2 to 1.8 are preferred for the following analyses.

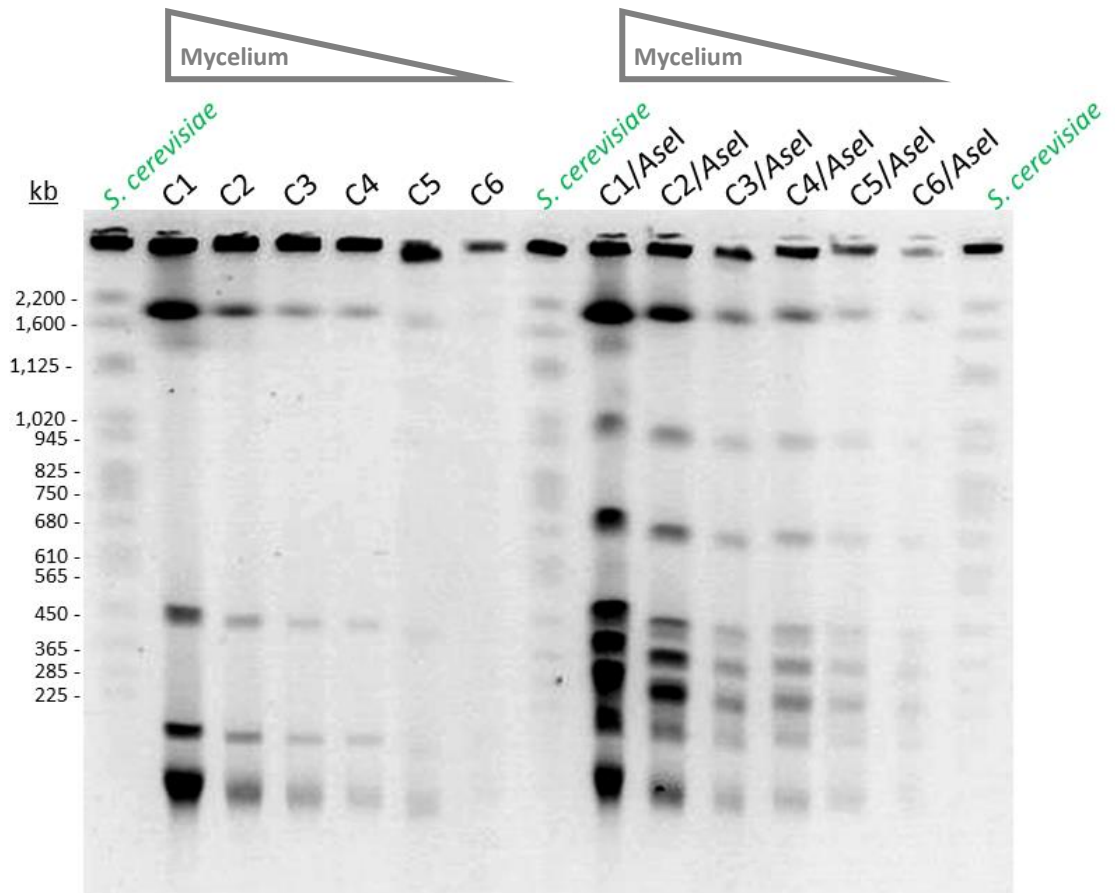


Figure 3.6. PFGE of *S. clavuligerus* DSM 738 samples at 6 different concentrations (C1 to C6) either uncut (left part) or digested with *AseI* (right part). Electrophoresis was performed using HEPES buffer at 4 V/cm for 24 hours. *Saccharomyces cerevisiae* YNN295 was used as DNA size standard (Bio-Rad).

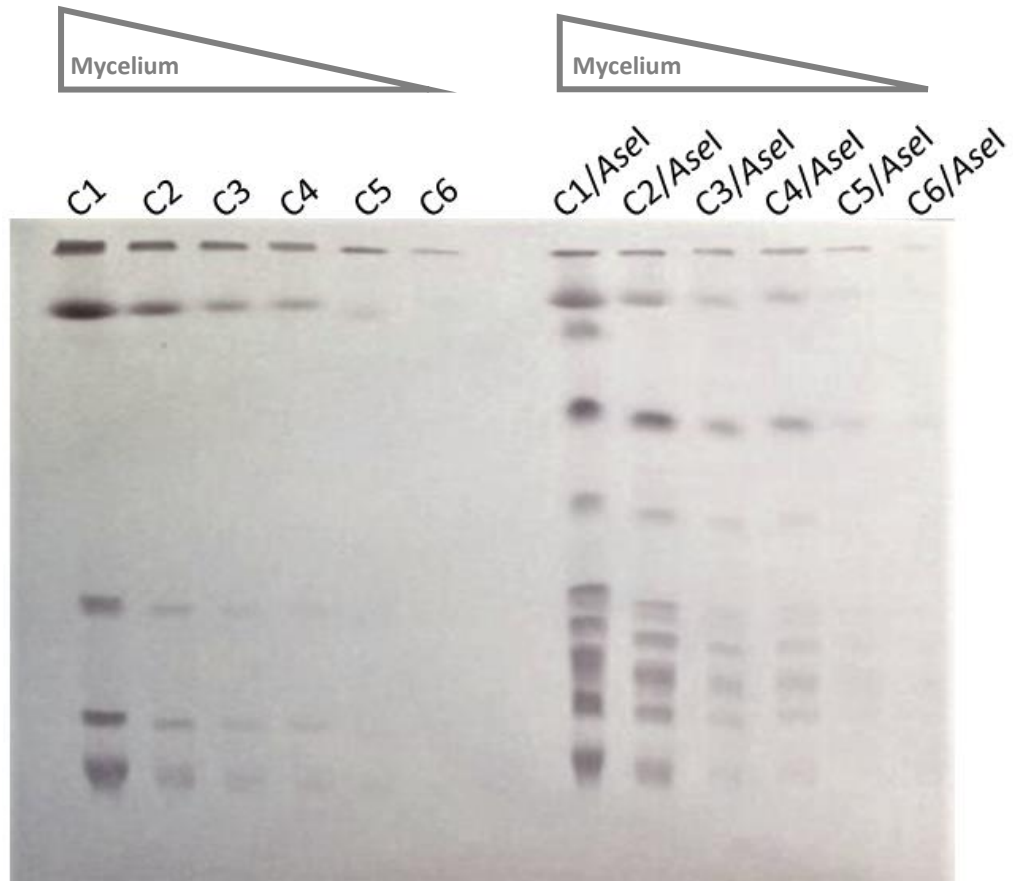


Figure 3.7. Southern blot of PFGE gel shown in **Figure 3.6**, using pSCL4 genes *tap-tpg* as probe. *S. clavuligerus* DSM 738 samples at 6 different concentrations uncut (left part) or digested with *AseI* (right part).

3.6 Mycelium harvested within 24 hours of inoculation gives the best results on PFGE analyses of *S. clavuligerus* samples.

In order to characterise the optimum time point to harvest mycelium for PFGE analyses in *S. clavuligerus*, electrophoresis was performed at 4 V/cm for 34 hours using mycelium of *S. clavuligerus* DSM 738 grown in TSB medium and collected at 6 different time points: 24, 48, 72, 96, 120 and 168 hours from inoculation. The samples were loaded either undigested or treated with *AseI* restriction enzyme.

The obtained gel exhibited bands exclusively at time points 24, 48 and 72 hours in both undigested and digested samples (**Figure 3.8A**). The undigested samples showed bands of sizes that correspond to plasmids pSCL4, pSCL3 and pSCL2, while the *AseI* digest included higher number of bands that might represent fractionated chromosomal DNA. These results suggest that intact genomic DNA can only be extracted from mycelium grown up to 3 days in TSB media. In addition, the band pattern did not show significant change between the first three time points which confirms that the genome architecture remains stable during vegetative growth. Moreover, the sample harvested at 24 hours exhibited the least DNA degradation which suggests a lower exposure to endonucleases during this time.

Moreover, in order to further characterise the fitness and integrity of *S. clavuligerus* mycelium during those time points to correlate to the PFGE data, microscopy images were taken. The mycelium samples were stained following Schwedock staining (Schwedock *et al.*, 1997), and imaged with an epi-fluorescence microscope. This method allowed the peptidoglycan stained with fluorescein isothiocyanate-wheat germ agglutinin (FITC-WGA) to be presented in green under a FITC emission filter, and nucleic acids stained with propidium iodide to be presented in red under a 600 nm emission filter. These images showed *S. clavuligerus* hyphae progressively lysing and generally losing integrity with each day (**Figure 3.8B**). The 24 hours sample exhibited the least damaged mycelium which could imply a higher availability of intact genomic DNA.

These results suggest that the best time point to collect mycelium for PFGE analyses is around 24 hours from inoculation, a fact that was followed for the subsequent gels performed.

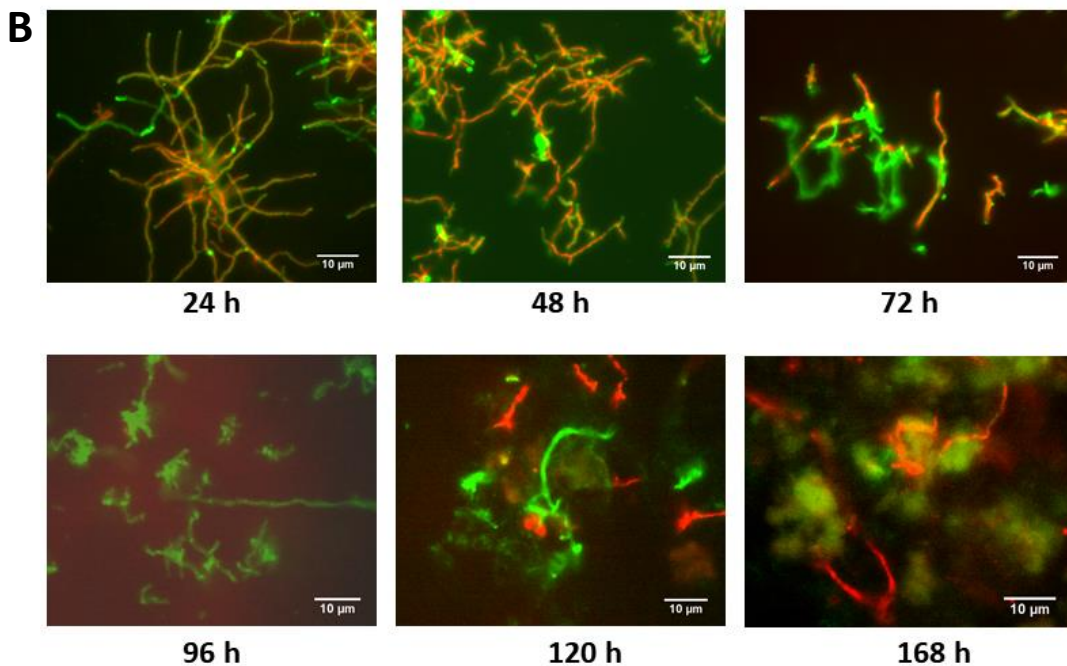
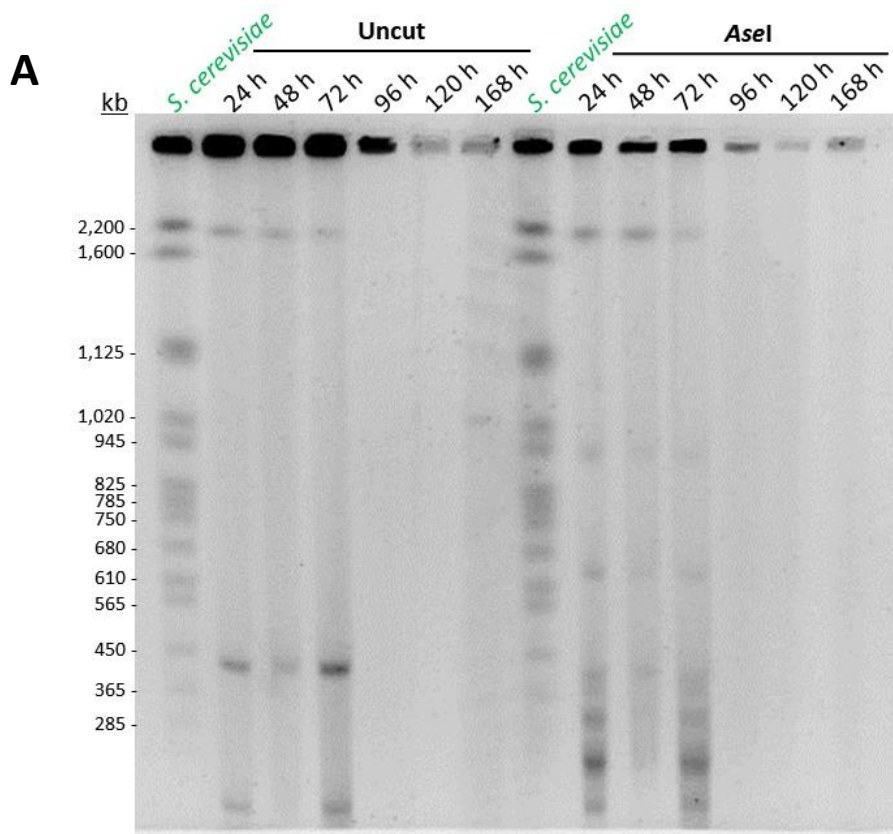


Figure 3.8. A: PFGE of samples of *S. clavuligerus* DSM 738 uncut or digested with *AseI* collected at 6 time points 24 to 168 hours. Electrophoresis was performed at 4 V/cm for 34 hours *Saccharomyces cerevisiae* YNN295 was used as DNA size standard (Bio-Rad). **B:** Fluorescence microscopy images of mycelium at each time point. Peptidoglycan was stained with fluorescein isothiocyanate-wheat germ agglutinin (FITC-WGA) and is presented in green under a FITC emission filter. Nucleic acids were stained with propidium iodide and are presented in red under a 600 nm emission filter.

3.7 Southern Blot analyses of PFGE gels confirm the band of 1.8 Mb as pSCL4.

In order to confirm the genome architecture of *S. clavuligerus*, PFGE was performed using the optimised conditions on the two DSM strains DSM 738 (type strain) and DSM 41826, and the two industrial strains Sc2 and Sc6; as well as two different *Streptomyces* species *S. coelicolor* M145 and *S. rimosus* G7. The samples were loaded either undigested or treated with *Asel* nuclease and the electrophoresis was performed at 4 V/cm for 36 hours using HEPES-based buffers.

The PFGE gel showed bands at around 450 and 150 kb for all undigested *S. clavuligerus* strains and 738, Sc2 and Sc6 also presented a band at 1,800 kb that could represent pSCL4 (**Figure 3.9**). The digested samples of *S. clavuligerus* showed a similar band pattern, with the addition of a band at 1,600 kb in 41826. Sc2 and Sc2 lacked the band at around 930 kb present on the other *S. clavuligerus* strains. This band, according to the predicted fragment sizes (**Table 3.2**), was part of pSCL4. In addition, Sc2 also lacked a band at around 380 kb. On the other hand, Sc6 had additional bands at around 1,100, 1,050 and 900 kb. Moreover, the bands from the digested *S. clavuligerus* DSM 41826 sample show the best correlation to the expected fragment sizes. The samples belonging to *S. coelicolor* and *S. rimosus* showed a similar band pattern to the expected fragments when digested with *Asel*, however the undigested samples presented unexpected bands and smearing between 2,000 and 1,125 kb, which could be due to cellular debris from the plugs.

Furthermore, in order to confirm the band at 1,800 kb on the undigested *S. clavuligerus* samples represents the megaplasmid pSCL4 and not chromosomal DNA, a Southern blot was performed on the PFGE gel using the pSCL4 *tap-tpg* as probe. The blot showed how the probe hybridised to the 1,800 kb band on the undigested samples and to the 930 kb band on the *Asel* digested sample, fragment that contains the *tap-tpg* operon on pSCL4 (**Figure 3.10**). In addition, no hybridisation was shown on the *S. coelicolor* and *S. rimosus* samples, corroborating the specificity of this probe to *S. clavuligerus* pSCL4.

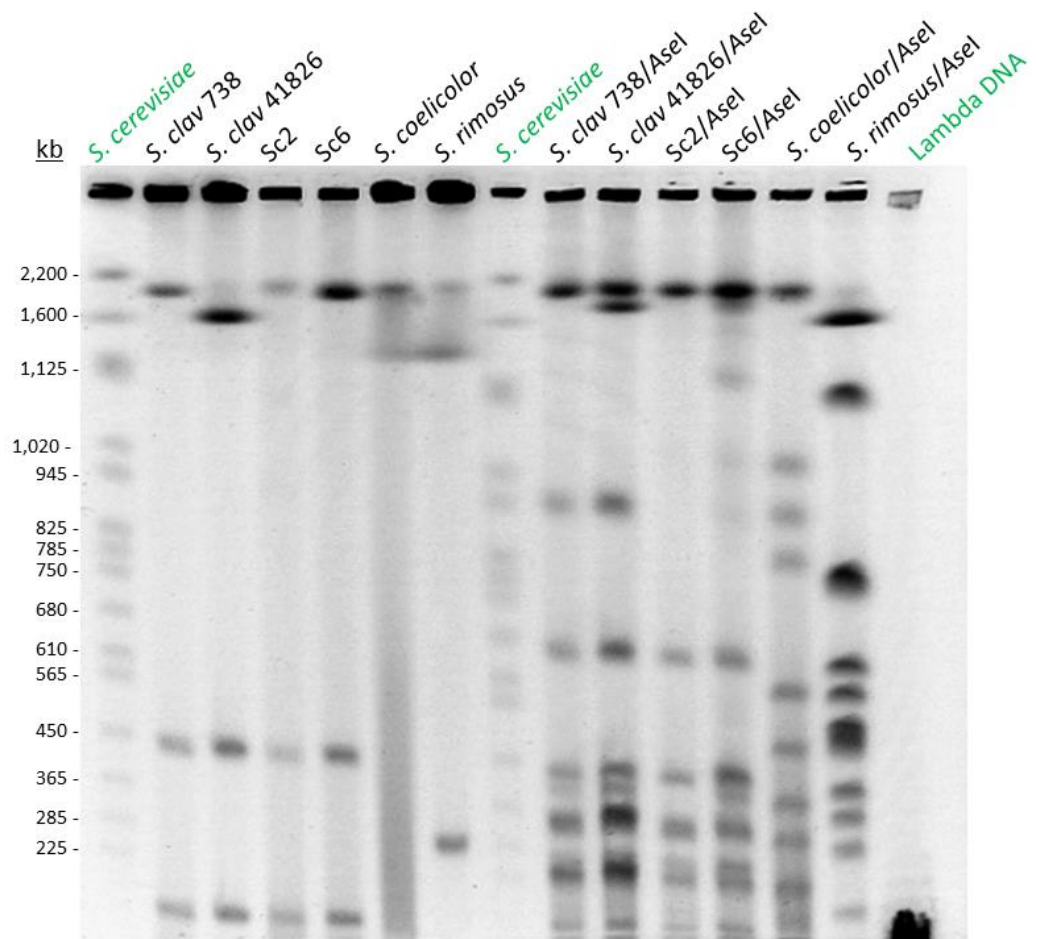


Figure 3.9. PFGE of samples of *S. clavuligerus* DSM 738, DSM 41826, Sc2, Sc6, *S. coelicolor* and *S. rimosus* undigested or digested with *AseI*. Electrophoresis was performed at 4 V/cm for 36 hours *Saccharomyces cerevisiae* YNN295 and Lambda PFG Ladder were used as DNA size standard (Bio-Rad).

Table 3.2. List of expected band sizes (in kilo-bases) in *S. clavuligerus* *S. coelicolor* M145 and *S. rimosus* G7 genomes when digested with *AseI*. In the case of *S. clavuligerus* chromosomal fragments are indicated in black, while pSCL4 fragments are indicated in blue and pSCL3 in red.

<i>S. clavuligerus</i>	<i>S. coelicolor</i>	<i>S. rimosus</i>
1,898	1,500	1800
1,612	1,300	1200
929	960	850
684	900	680
442	820	630
406	610	580
355	480	500
339	380	430
328	290	400
275	200	300
250	190	180
234	140	80
233	110	40
227	45	
211	16	
187	11	
159	10	

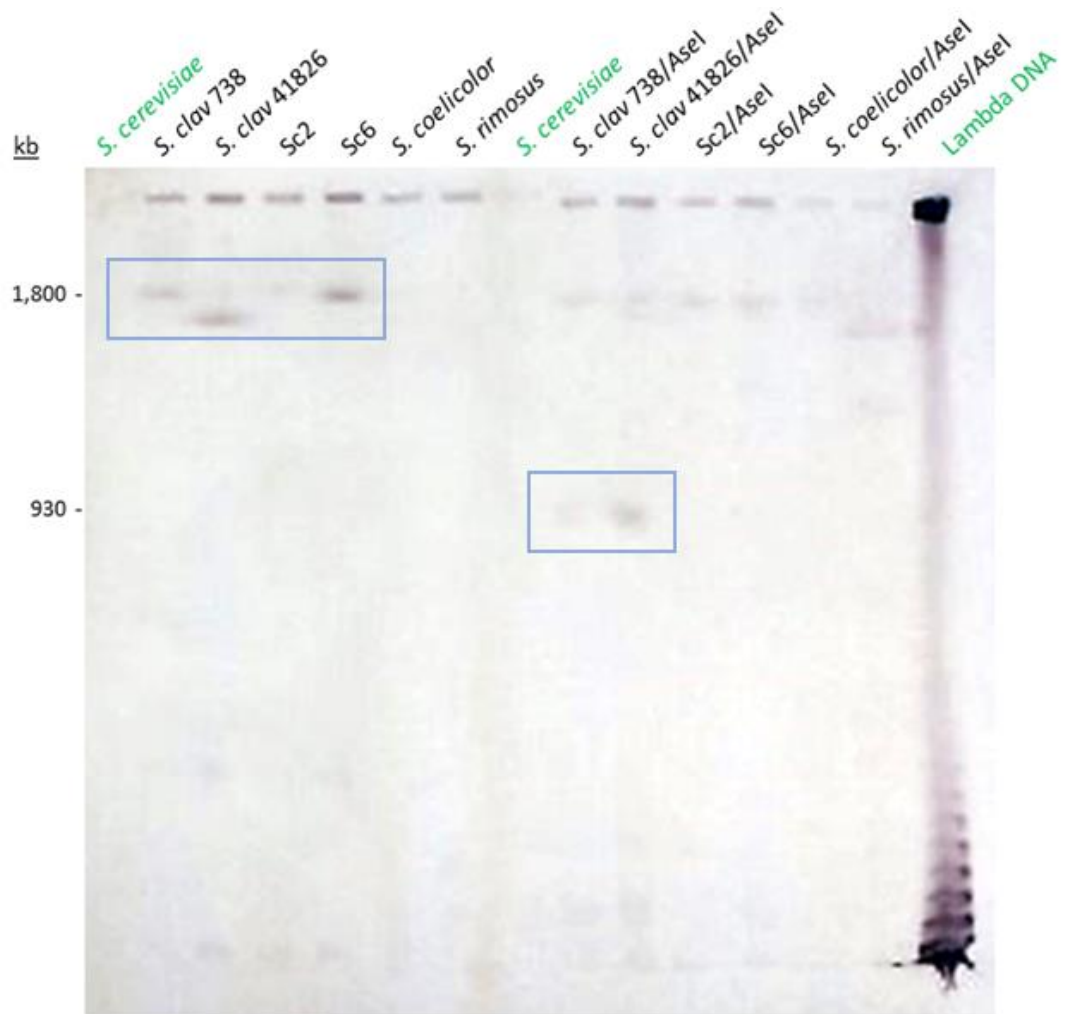


Figure 3.10. Southern blot of PFGE gel shown in **Figure 3.9** using *tap-tpg* as probe. The probe hybridises to the band of 1.8 Mb in uncut samples, and 930 kb in *AseI* digests in *S. clavuligerus* samples (indicated in the blue boxes). Both bands correspond to pSCL4 DNA.

Moreover, in order to detect chromosome-specific bands in PFGE gels we performed a Southern blot using *S. clavuligerus* DSM 738 *ftsZ* as probe. PFGE of undigested samples of *S. clavuligerus* DSM 738, Sc2 and Sc6 was performed at 4 V/cm for 34 hours and the gel was blotted and hybridised. As it was expected, the probe did not hybridise to the bands corresponding to the plasmids, as this gene is only present in the chromosome (**Figure 3.11**).

This leads us to the conclusion that the chromosome is unable to migrate in a PFGE unless fractionated with restriction nucleases, thus the visualised bands on undigested samples correspond to extrachromosomal DNA.

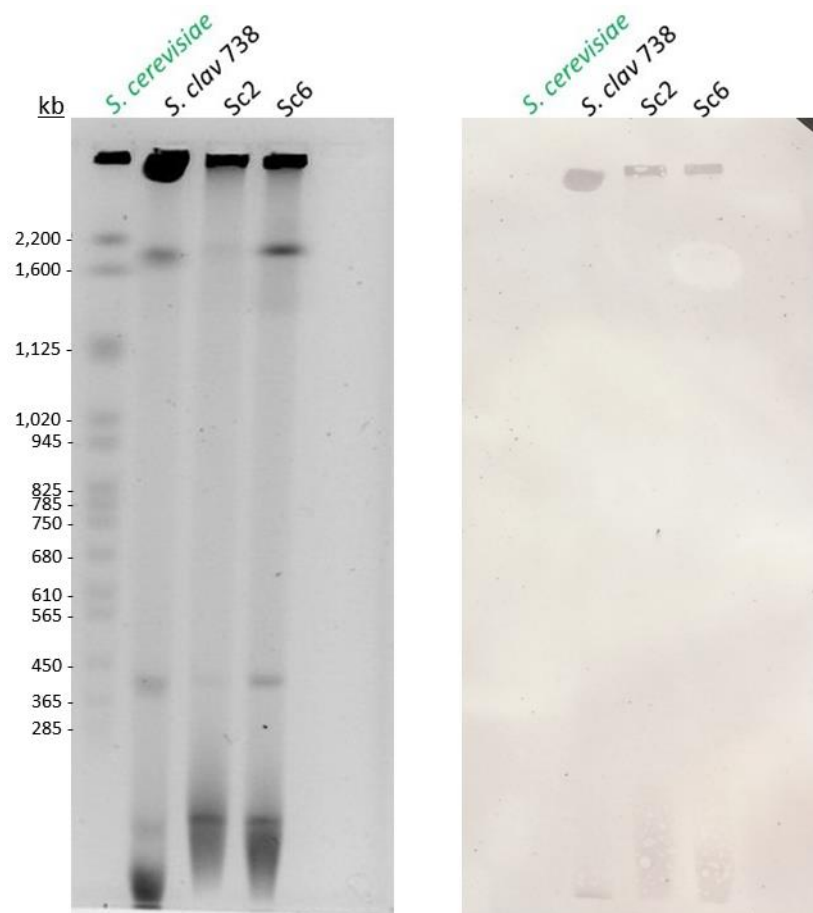


Figure 3.11. PFGE of undigested *S. clavuligerus* DSM 738, Sc2, Sc6 (left). Electrophoresis was performed at 4 V/cm for 34 hours. *Saccharomyces cerevisiae* YNN295 was used as DNA size standard (Bio-Rad). Southern blot of the PFGE using *S. clavuligerus* DSM 738 *ftsZ* as probe (right).

3.8 The final physical map of *S. clavuligerus* DSM 738 genome.

Once the optimised method for PFGE analyses of *S. clavuligerus* samples was obtained, and the presence of pSCL4 confirmed, a PFGE was performed using *S. clavuligerus* DSM 738 samples digested with four restriction enzymes: *AseI*, *DraI*, *AflIII* and *SspI* (AAT'ATT), with the intention of constructing a final physical map of the genome of said strain. The electrophoresis was performed at 4 V/cm for 33 hours and an uncut sample of *S. clavuligerus* was loaded along with the digested ones.

The gel showed bands patterns similar to the estimated fragments from *in silico* digest of previously published sequences (**Figure 3.12; Table 3.3**) (Cao *et al.*, 2016; Medema *et al.*, 2010; Song *et al.*, 2010). This suggests that the genome of *S. clavuligerus* DSM 738 presents similar genetic architecture to the genome of *S. clavuligerus* strains previously sequenced. In addition, the presence of 3 extra-chromosomal bands at sizes 1,800, 440 and 150 kb confirms that *S. clavuligerus* DSM 738 carries the plasmids pSCL4, pSCL3 and pSCL2.

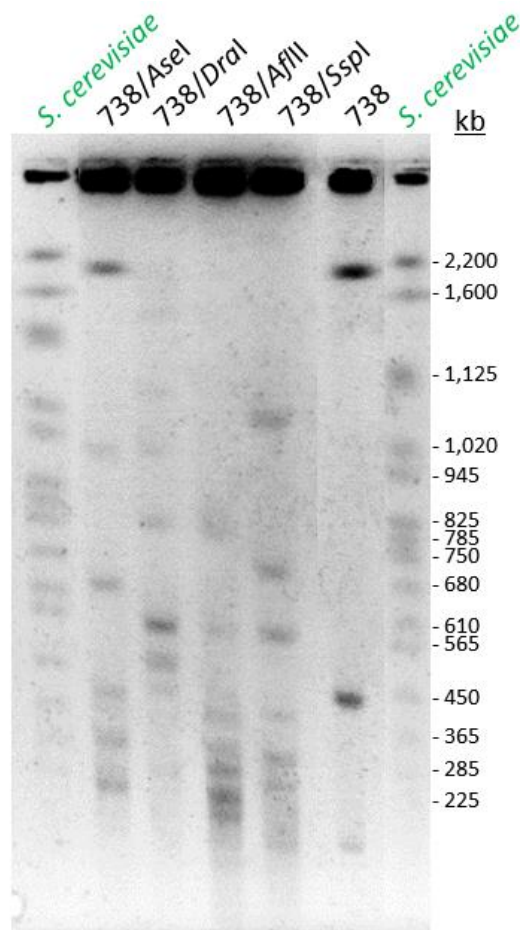


Figure 3.12. PFGE of *S. clavuligerus* DSM 738 digested with *AseI*, *DraI*, *AflII* and *SspI*, as well as undigested. Electrophoresis was performed at 4V/cm for 33 hours. *Saccharomyces cerevisiae* YNN295 was used as DNA size standard (Bio-Rad).

Table 3.3. List of expected band sizes in *S. clavuligerus* genome when digested with *AseI*, *DraI*, *AflII* and *SspI*. Chromosomal fragments are indicated in black, while pSCL4 fragments are indicated in blue and pSCL3 in red.

<i>AseI</i> digests (kb)	<i>DraI</i> digests (kb)	<i>AflII</i> digests (kb)	<i>SspI</i> digests (kb)
1,898	2,707	1,337	993
1,612	1,067	833	733
929	1,033	831	602
684	805	637	590
442	598	441	415
406	594	334	328
355	526	300	324
339	506	293	307
328	444	284	306
275	279	255	269
250	238	242	261
234	200	241	233
233	115	228	214
227		225	203
211		200	183
187		194	180

3.9 Summary

The optimisation experiments described in this chapter have allowed us to develop an effective method for pulse-field gel electrophoresis analyses on *S. clavuligerus* samples. This method is not only valid for publicly available strains such as the ones provided by DSMZ, but also for industrial strains, as seen for Sc2 and Sc6 strains. This fact is particularly beneficial for GSK as it would facilitate rapid screening of interesting mutants, as well as characterisation of their already established strain lineage. In addition, this method has also been proven effective for other *Streptomyces* species, such as *S. coelicolor*, *S. rimosus*, as well as unknown streptomycetes isolates (data not shown), which proves this method to be considered as a good alternative to the established protocol described by Kieser *et al.*, 2000.

These results show how *S. clavuligerus* is susceptible to Tris-dependent DNA degradation such as seen in PFGE of other species of the genus like *S. lividans* (Evans and Dyson, 1993). This would also explain the missing bands from the gels published by Netolitzky *et al.*, 1995 and Huang *et al.*, 1998, who used Tris-based buffers for their analyses. The increased quality of PFGE gels performed using HEPES-based buffers was also confirmed for microorganisms of different phyla, like *Salmonella enterica* (Koort *et al.*, 2002). The reason behind Tris-dependent DNA cleavage was first determined using conventional electrophoresis of *S. lividans* DNA. It was suspected that this was due to formation of nucleolytic free radical species from electric field-activated Tris, as this DNA cleavage was significantly decreased in the presence of the radical scavenger agent thiourea (Zhou *et al.*, 1988; Ray, Weaden and Dyson, 1992). This DNA degradation (Dnd) phenotype of *S. lividans* is now known to be a consequence of incorporation of sulphur between guanine nucleotides (phosphorothioation) by the action of Dnd proteins (Wang *et al.*, 2007). This sulphur-modified sites suffer double-stranded cleavage during electrophoresis in the presence of Tris, as an oxidative peracid derivative of this molecule is formed at the anode which induces amine-catalysed DNA scission (Ray, Mills and Dyson, 1995).

Three genes were identified in *S. clavuligerus* chromosome sequences as putative *dndA*, *dndB* and *dndC*, annotated on the *S. clavuligerus* ATCC 27064 sequence (Medema *et al.*, 2010) as SCLAV_2520, SCLAV_2515 and SCLAV_2517 respectively (**Figure 3.13**). The presence of these genes in *S. clavuligerus* implies the possibility of a Dnd phenotype in this organism which could be susceptible to activated Tris derivatives during electrophoresis explaining the exhibition of DNA degradation on the gels performed with Tris-based buffers.

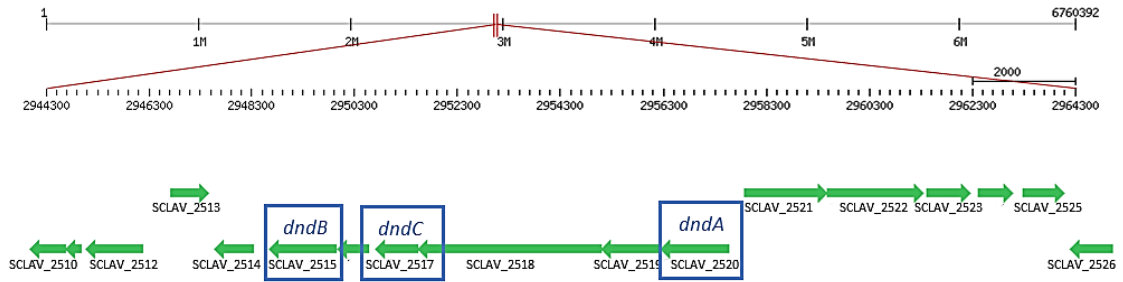


Figure 3.13. *S. clavuligerus* ATCC 27064 chromosome sequence (view from StrepDB) annotated genes SCLAV_2515, SCLAV_2527 and SCLAV_2520 identified as putative *dndB*, *dndC* and *dndA* respectively.

Our PFGE results also evidence the sensitivity of *S. clavuligerus* to lysozyme treatment, as we could observe DNA bands after only 5 minutes of lysis. This sensitivity was also exhibited during Schwedock staining of *S. clavuligerus* mycelium for microscopy, as the 1-minute lysozyme treatment proved to yield greatly damaged mycelium as compared to *S. coelicolor* mycelium that remained intact (data not shown). In addition, rapid lysis of *S. clavuligerus* mycelium has also been evidenced in our work during protoplast formation, when large number of protoplasts were visualised only 5 minutes after the addition of lysozyme. Although no specific studies have been published on lysozyme effect on *S. clavuligerus*, it is an important observation and fact to consider for further experiments involving protoplast formation, such as DNA extraction and protoplast transformation.

These experiments have also proved how DNA availability significantly decreases after three days in liquid culture which could be due mainly to generalised mycelium lysis and nuclease activity. Previous articles showed how *S. clavuligerus* enters stationary phase after 24 hours of culture (Hung *et al.*, 2007), a fact that was also demonstrated in our growth curves (Chapter 6), and how biomass significantly decreases after three days of culture (Fu *et al.*, 2019), which correlates to our microscopy images. This suggest that once the stationary phase is reached, the microorganism starts to starve and begins autolysis. The increased presence of cytosolic enzymes and nucleases as a consequence of cell lysis after the third day of submerged culture has been confirmed for other species, like *S. coelicolor* and *Streptomyces antibioticus* (Manteca *et al.*, 2008; Manteca *et al.*, 2006). This fact could be extrapolated to *S. clavuligerus* and emphasises the importance of using young mycelium for experiments involving DNA extraction.

The PFGE analyses performed with the optimised protocol have allowed us to obtain physical evidence of the genome of *S. clavuligerus* DSM 738 containing DNA fragments between 2,200 and 100 kb of either intact or digested genomic DNA. The PFGE of intact DNA has confirmed the presence of the plasmids pSCL2 and pSCL3 with sizes around 150 and 450 kb respectively. This data correlates to the PFGE results published by Netolitzky *et al.*, 1995 and Huang *et al.*, 1998. Furthermore, our analyses have made possible the first visualisation of the entire megaplasmid pSCL4 on a gel, which was confirmed by Southern blotting and, with a size of around 1.8 Mb, is the largest linear plasmid ever identified in PFGE. The PFGE of digested samples has also allowed identification of chromosomal fragments, that together build up the genomic map. In general, these bands correlate well with the *in silico* digests of the available sequences of *S. clavuligerus* strains ATCC 27064 and F613-1 (Cao *et al.*, 2016; Medema *et al.*, 2010; Song *et al.*, 2010), which suggests that the genome sequence of DSM 738 is similar to these. Moreover, we have constructed a table with the fragment sizes estimated from these PFGE analyses (**Table 3.4**). This will be compared to the whole genome sequencing results described on Chapter 4, in order to establish a final combined genome map of *S. clavuligerus* DSM 738.

Table 3.4. List of estimated fragment sizes (kb) of the genome of *S. clavuligerus* DSM 738, unfragmented or digested with the enzymes: *AseI*, *DraI*, *AflIII* and *SspI* as predicted from PFGE analyses.

Uncut	<i>AseI</i>	<i>DraI</i>	<i>AflIII</i>	<i>SspI</i>
1,800	1,800	1,800	1,300	1,000
450	950	1,300	800	700
150	680	1,050	600	550
	430	940	450	400
	400	780	300	300
	355	550	280	250
	260	470	230	150
	230	450	200	
	150		180	

4. CLOSURE OF THE *S. CLAVULIGERUS* DSM 738 GENOME SEQUENCE

Since the discovery of the ability of *Streptomyces* strains to produce interesting secondary metabolites, these organisms have been intensely exploited for the mass production of many molecules used currently in the clinic. The advance in genomics has allowed identification of the biosynthetic gene clusters responsible for these metabolites, as well as other silenced clusters that demonstrate the great potential of streptomycetes. The importance of obtaining a good quality genome sequence is evident, not only for genome mining studies, but also in order to acquire better understanding of the physiology of these complex microorganisms.

The advances in whole genome sequencing during the last few decades have allowed researchers to obtain genome sequences of *Streptomyces* strains in a much faster and affordable way. The most common next generation sequencing technology used is short read Illumina sequencing, which has high nucleotide accuracy. However, *Streptomyces* genomes carry repetitive sequences and regions of high homology within the same molecule which hinders the assembly of short reads into complete sequences yielding genomes split into hundreds of contigs. The use of long read sequencing technologies such as Single molecule real time (SMRT) sequencing by PacBio (Pacific Biosciences) facilitates assembly yielding fewer contigs. SMRT technology also provides more even coverage for organisms with high GC content, such as *Streptomyces*, than Illumina sequencing (Shin *et al.*, 2013). Using PacBio technology for sequencing genomes and correcting the assembly with high-quality short reads by Illumina has resulted in the optimum method to obtain a closed and accurate genome sequence of *Streptomyces* strains. An example of this is the whole genome sequencing project of *S. leeuwenhoekii* by Gomez-Escribano *et al.*, 2015, where they successfully obtained a closed sequence of the chromosome and two plasmids.

The genome of *S. clavuligerus* type strain ATCC 27064 has been sequenced using next generation sequencing technologies on three occasions (**Table 4.1**). The first one, published by Medema *et al.*, 2010, was constructed into two scaffolds representing the chromosome and pSCL4, which confirmed the presence of this giant linear plasmid for the first time in *S. clavuligerus*. The second publication presented 5 scaffolds: chromosome, pSCL4, pSCL3, pSCL2 and pSCL1; and this is the only one presenting all replicons (Song *et al.*, 2010). The third article reports the only published sequence of an *S. clavuligerus* type strain genome that is closed, and by closed sequence we mean a single contig for each replicon, however,

this publication only presents sequences for the chromosome and pSCL4 (Hwang *et al.*, 2019). Furthermore, two more closed sequences for *S. clavuligerus* are publicly available, which are the ones belonging to the industrial strains F613-1 and F1D-5 (**Table 4.1**). These strains only present sequences for the chromosome and pSCL4; and in the case of F1D-5 also pSCL2, which means that either these strains do not carry the rest of the replicons, or they were ignored during sequencing (Cao *et al.*, 2016). The discrepancy between these sequences makes it impossible to correlate them to our type strain, which makes evident the necessity of obtaining a high-quality complete sequence of *S. clavuligerus* genome that we can work with.

Table 4.1. List of publicly available genome sequences of *S. clavuligerus* strains obtained from NCBI. *No published article available for F1D-5 strain. **Closed instead of complete: they present one contig per replicon but they lack telomeric sequences.

Reference	Strain	Technology	Replicons	Assembly level
Medema <i>et al.</i> , 2010	ATCC 27064	SOLiD	Chromosome, pSCL4	Scaffolds
Song <i>et al.</i> , 2010	ATCC 27064	Sanger & 454	Chromosome, pSCL4, pSCL3, pSCL2, pSCL1	Scaffolds
Cao <i>et al.</i> , 2016	F613-1	Illumina & PacBio	Chromosome, pSCL4	Closed**
Cao <i>et al.</i> , 2018*	F1D-5	Illumina & PacBio	Chromosome, pSCL4, pSCL2	Closed**
Hwang <i>et al.</i> , 2019	ATCC 27064	Illumina & PacBio	Chromosome, pSCL4	Closed**

Moreover, combining Next Generation Sequencing technologies can allow construction of closed sequences but considering the linear topology of *Streptomyces* genomes, closed assemblies do not necessarily involve sequence completeness. This is because these sequencing technologies often fail to obtain the complete telomeres of linear chromosomes or plasmids. In addition, the presence of long terminal inverted repeats on these organisms hinders assembly at the sequence ends. This is the case for the published *S. clavuligerus* sequences; for example, an archetypal telomere sequence is found at one of the ends of the chromosome of the F613-1 strain, while the other end is missing. The same situation is observed for the sequences of pSCL4 and pSCL3 published by Song *et al.*, 2010. For this reason, we consider that in order to refer a sequence as complete, it should contain both telomeres, and so none of the current *S. clavuligerus* sequences are fully complete.

Telomere identification has traditionally been done by cloning. This involved removal of covalently-linked terminal protein by alkali treatment, restriction enzyme digestion and ligation to a vector prior to cloning in *E. coli* (Goshi *et al.*, 2002). Two adaptations of this protocol were also successfully used, one involved utilisation of glass beads to bind the protein-DNA complex and removing proteins with a piperidine solution (Huang *et al.*, 1998). Interestingly, this publication presented the sequence of the telomere of *S. clavuligerus* ATCC 27064 pSCL2 plasmid amongst other species. The other adaptation of this method used integration of an *E. coli* vector by homologous recombination into the telomere sequence prior to enzyme digestion and self-ligation before cloning (McLeod *et al.*, 2006). A non-cloning method for telomere identification was described by Ohnishi *et al.*, 2008, where they incorporated polyA tails to the 3' ends and then amplified the telomere twice (nested PCR) using a primer containing a polyT region. A final method for telomere purification was published by Fan *et al.*, 2012, that did not involve cloning and reduced the amplification steps. This is the self-ligation method, in which genomic DNA is digested with a restriction enzyme that leaves a blunt end, the terminal proteins are removed by alkali treatment and the DNA fragments are then self-ligated at the blunt ends (**Figure 4.1**). The telomeres are then amplified by inverted PCR using a primer from the sequence close to the restriction site and a primer close to the sequence terminus. The self-ligation method appears to be the best option for rapid purification of the telomeres of *Streptomyces* strains, whose genome sequence is at least partially known.

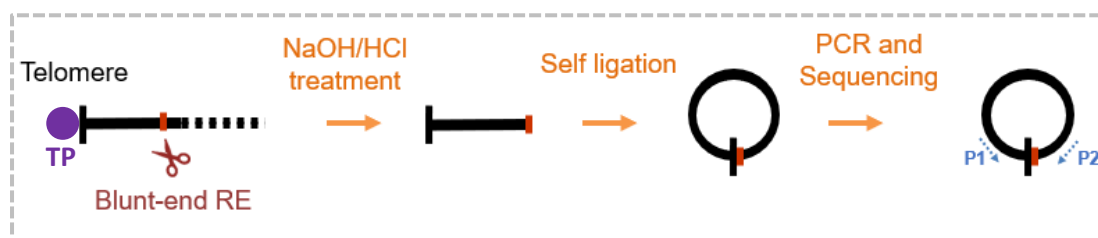


Figure 4.1. Diagram of the self-ligation method for telomere purification, adapted from Fan *et al.*, 2012. TP: terminal protein, RE: restriction enzyme. P1 and P2: primers for inverted PCR.

This chapter focuses on obtaining the whole genome sequence of *S. clavuligerus* DSM 738 using a combination of Illumina and PacBio data. This assembly is complemented with the identification of the telomeres of chromosome and plasmids. In addition, the final complete sequence is characterised and compared to relevant strains of *Streptomyces*.

4.1 Combining Next Generation technologies Illumina and PacBio obtains closed sequences of *S. clavuligerus* DSM 738 chromosome and plasmids.

In order to obtain accurate short-read sequencing data of the genome of *S. clavuligerus* DSM 738, genomic DNA was sequenced using the Illumina MiniSeq apparatus (University of Strathclyde). For this, genomic DNA of *S. clavuligerus* DSM 738 was extracted, quantified by Qubit and qualified with gel electrophoresis (**Figure A1** in Appendix). This DNA was used to prepare the indexed pair-end library using the Illumina Nextera™ DNA Flex Library Prep Kit. In order to check the quality of the library, we used the Agilent 2100 Bioanalyzer System. The Bioanalyzer results showed bands at migration times between 120 to 100 seconds with a mean migration time of 109.17 seconds (**Figure 4.2**). Comparison to the DNA ladder confirmed a mean library size of 600 bp. The genomic library was loaded into the MiniSeq and forward and reverse reads were obtained. Assembly of the reads yielded 1,135 contigs, with a total length of 9,139,898 bp and an N50 value of 48,890. Additionally, for comparison purposes, the genome of *S. clavuligerus* DSM 738 was also sequenced using the services of Microbes NG (Birmingham) that utilise Illumina HiSeq technology. The assembly of HiSeq reads yielded a total number of 8,999,860 bp split in 1,660 contigs, and a N50 value of 14,149.

In order to obtain long-read sequencing data, the genome of *S. clavuligerus* DSM 738 was sequenced in the Northumbria University using PacBio technologies. The assembly obtained a genome of 8,983,123 bp in 15 contigs with a N50 value of 6,729,334. Furthermore, a hybrid assembly was constructed using the MiniSeq and PacBio reads, which reduced the number of Illumina contigs to 504 and increased the N50 value to 194,924.

For the purpose of comparing the assembly results from all these sequencing technologies, a table was constructed with the Quast results of each assembly (**Table 4.2**). This table evidences the much higher assembly coverage of PacBio sequencing data with a significantly lower number of contigs and a large contig of 6,729,334 which resembles the size of the *S. clavuligerus* chromosome described in previous publications. Moreover, the assembly obtained from MiniSeq data exhibits a substantially higher quality than the assembled HiSeq data. This could be due to the Illumina platform itself or to the library preparation methodology used. Given these observations, we decided to use the PacBio contigs, along with the MiniSeq + PacBio contigs for further analysis and construction of the final closed genome assembly.

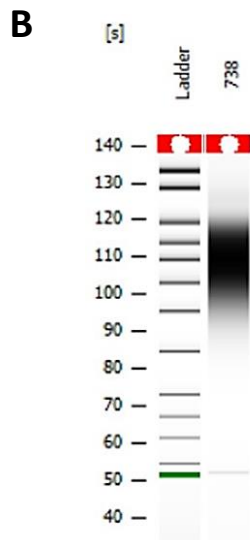
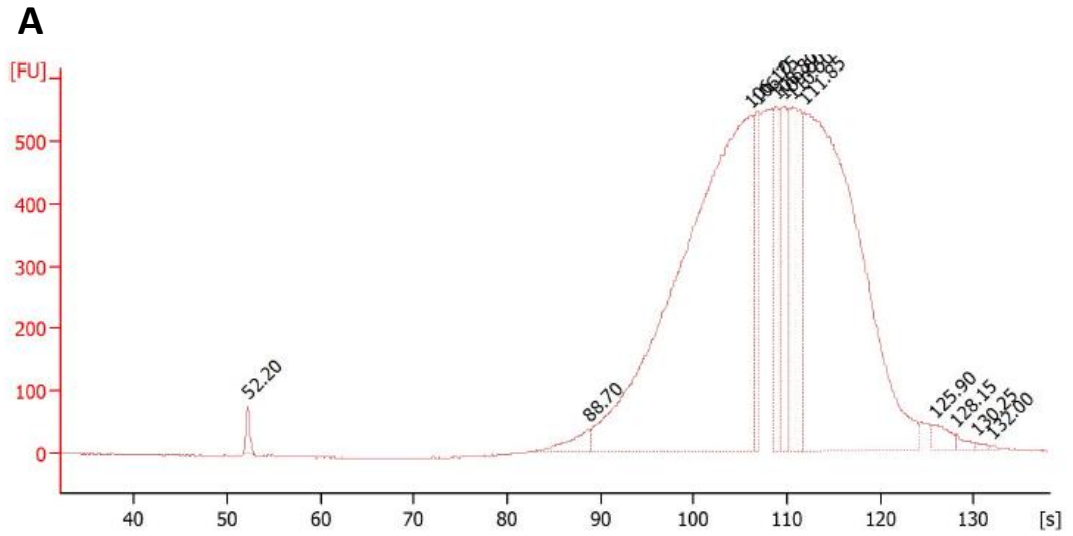


Figure 4.2. Bioanalyzer results for the genomic library of *S. clavuligerus* DSM 738. **A:** Electropherogram of Bioanalyzer electrophoresis shows a mean library migration time of 109.17 seconds. **B:** Representation of Bioanalyzer electrophoresis of library varying in migration times between 120 to 100 seconds.

Table 4.2. Quast results comparison of the four obtained assemblies of *S. clavuligerus* DSM 738 genome. The statistics are based on contigs bigger than 500bp, unless otherwise noted.

	PacBio	MiniSeq	HiSeq	MiniSeq + PacBio
# contigs (> = 0 bp)	15	1134	1660	504
# contigs (> = 1000 bp)	15	346	1052	86
Total length (> = 0 bp)	8983123	9139898	8999860	9186333
Total length (> = 1000 bp)	8983123	9027817	8787469	9142761
# contigs	15	380	1211	91
Largest contig	6729334	206614	65961	387419
Total length	8983123	9051120	8899421	9146339
GC (%)	72.4	72.4	72.32	72.36
N50	6729334	48890	14149	194924
N75	515261	25640	6905	114533
L50	1	59	197	18
L75	2	124	423	32
# N's per 100 kbp	0	0	0.01	0

In order to obtain a closed assembly for *S. clavuligerus* DSM 738 genome, the PacBio contigs were manually assembled using the MiniSeq + PacBio contigs (hereafter referred to as Illumina contigs to avoid confusion). For this, each end of every PacBio contig was identified using SnapGene and Artemis in the Illumina contigs. The PacBio contigs were then manually mapped to the Illumina contigs, extended and, in some cases, joined (**Figure 4.3**).

In the case of the chromosome, the Contig_0 of 6,729,334 bp of size was identified as chromosomal DNA and the ends were extended 12 kb upstream and 7 kb downstream using Illumina contigs, yielding a molecule of 6,748,317 bp of size, which is only 274 bp fewer than the chromosome sequence published by Hwang *et al.*, 2019.

The megaplasmid pSCL4 assembly was obtained joining 11 PacBio contigs: Contig_5, Contig_14, Contig_4, Contig_6, Contig_1, Contig_9, Contig_8, Contig_11, Contig_10, Contig_12 and Contig_3. This resulted in a sequence of 1,795,361 bp, 134 bp fewer than the Hwang plasmid sequence.

pSCL3 was almost entirely composed of the PacBio Contig_2, which was extended 6 kb upstream and 8 kb downstream using Illumina contigs to yield a sequence of 453,816 bp.

The pSCL2 plasmid was assembled by joining the PacBio contigs Contig_13 and Contig_7 using Illumina contigs into a sequence of 149,435 bp.

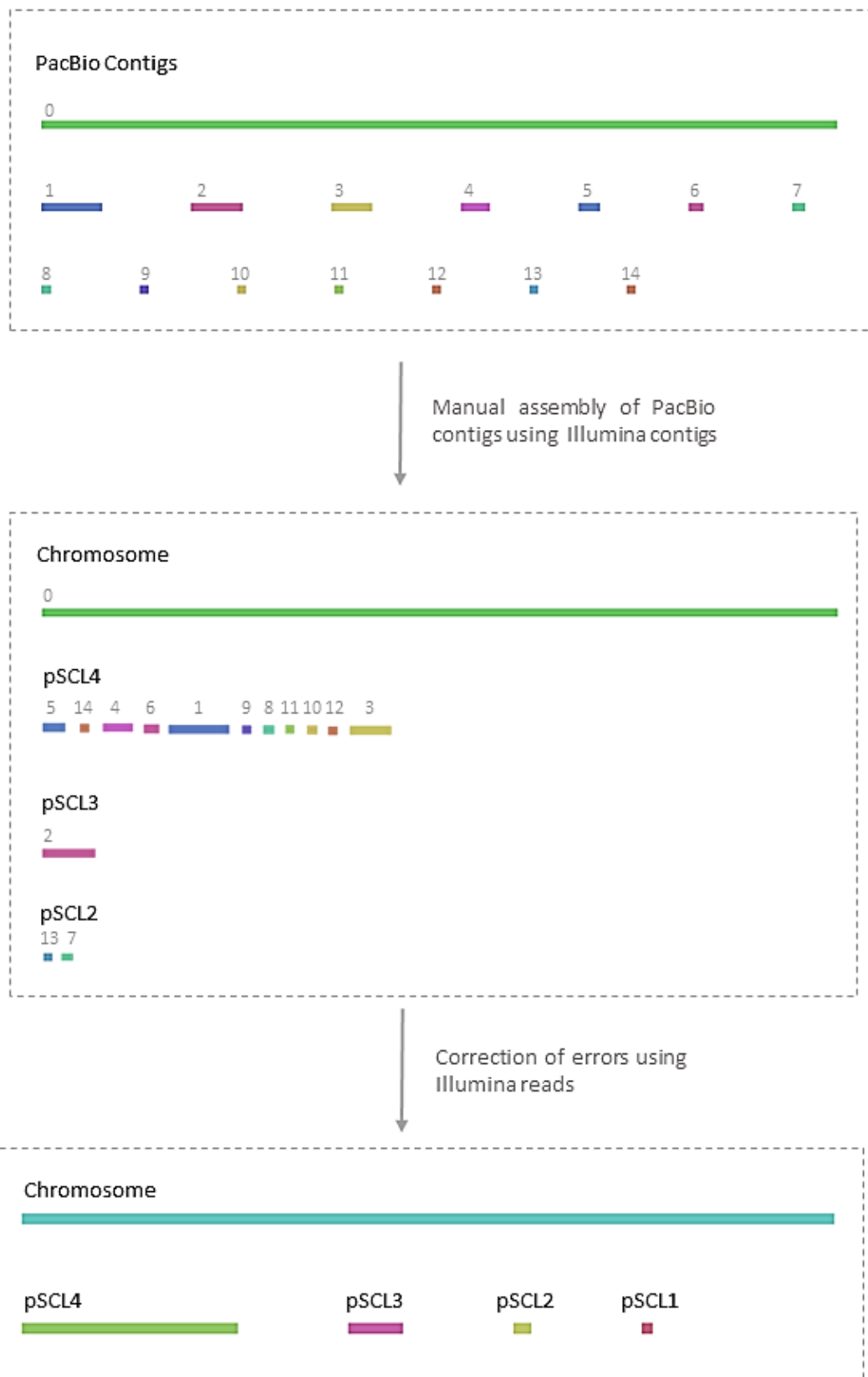


Figure 4.3. Diagram of the workflow followed to obtain a closed assembly for the genome of *S. clavuligerus* DSM 738, using PacBio and Illumina MiniSeq data.

In the case of pSCL1, no PacBio contigs were identified to correspond to this plasmid. However, the complete sequence of this plasmid was identified by Wu and Roy, 1993 from *S. clavuligerus* NRRL 3585, by cloning and Sanger chain-termination sequencing, and so we consider this as a trustworthy reference sequence. Illumina contigs were mapped to the Wu & Roy sequence and the contigs belonging to pSCL1 were identified and manually assembled. In addition, the MiniSeq reads were mapped against the Wu & Roy sequence showing an entire coverage of this sequence by our reads, which suggest this is an almost identical sequence to *S. clavuligerus* DSM 738 (**Figure 4.4A**). However, two SNPs (single nucleotide polymorphism) were identified in all the reads mapping those specific bp, which could represent mutations on our strain (**Figure 4.4B**). These two bp were corrected on our pSCL1 sequence.

Once obtained the closed sequences for all the replicons, we performed a Pilon analysis using the MiniSeq reads in order to polish the sequences. This analysis identified variants and corrected the most common inconsistencies on the assemblies.

After the polishing of the five different replicon sequences, we could confirm the closed, high quality sequences of *S. clavuligerus* DSM 738 chromosome, pSCL4, pSCL3, pSCL2 and pSCL1. However, other than pSCL1, the telomeres of these sequences remained missing and so, further physical analyses were necessary for their identification and to obtain the complete genome sequence.

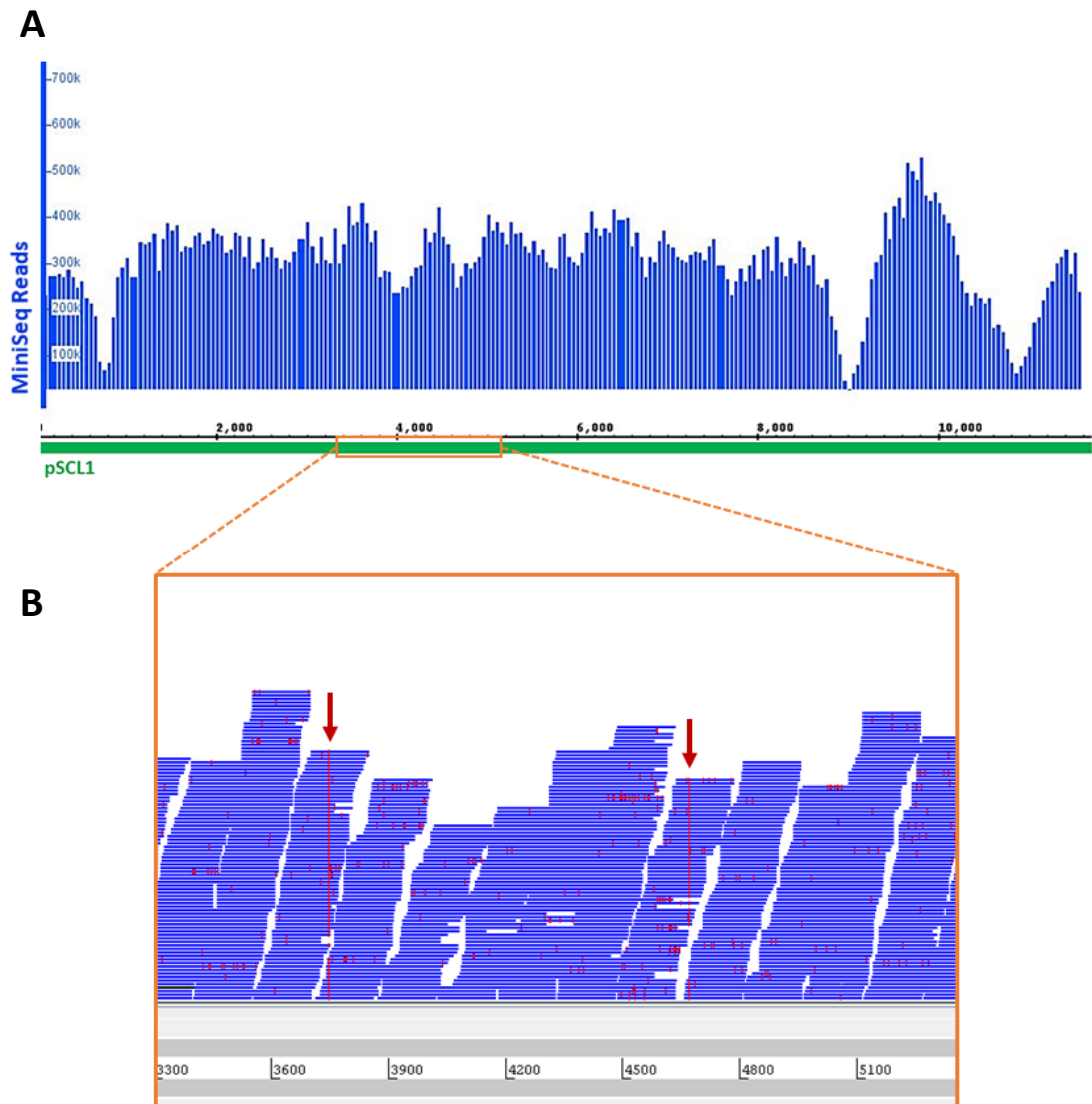


Figure 4.4. MiniSeq reads mapped against pSCL1 Wu & Roy sequence. **A:** IGB graph of entire pSCL1 sequence. **B:** Section of 3,300 to 5,200 bp of pSCL1 sequence against MiniSeq reads, showing SNPs in red (Artemis). Red arrows show the two SNPs found in all the reads.

4.2 *S. clavuligerus* DSM 738 carries several copies of the archetypal telomere found in F613-1 strain sequence.

In order to study the sequence ends and the presence of archetypal telomeres in *S. clavuligerus* DSM 738, genomic DNA of this strain was compared to sequence ends of the F613-1 strain. For this, a series of analyses by Southern blotting were performed.

The F613-1 strain contains an archetypal telomeric sequence at one of the ends. In order to detect the presence of this telomere in our strain, a probe was constructed using the first 500 bp of the F613-1 sequence containing this telomere. The probe was used for Southern hybridisation to genomic DNA of *S. clavuligerus* DSM 738. For this, genomic DNA was digested with the restriction enzymes *Bam*HI, *Sac*I, *Zra*I, *Eco*RV and *Bgl*II independently, and the digests were separated by electrophoresis. The gel was then blotted and hybridised to the probe. The blot shows many bands of different intensities (**Figure 4.5**). Strong bands were identified at the expected sizes for the F613-1 chromosome sequence at 0.75 (*Bam*HI), 1.2 (*Sac*I), 3 (*Zra*I), 4.3 (*Eco*RV) and 6.4 kb (*Bgl*II), which suggests that DSM 738 presents this telomere and carries a similar sequence at one of the ends. The presence of other strong bands at different sizes suggest the presence of this telomere at the other end of the chromosome, as well as the potential presence of similar telomere sequences on the plasmids.

Furthermore, the other end of the F613-1 sequence does not present a telomere-like sequence, thus we hypothesise that the real end is missing from this sequence. In order to study if this sequence is present on DSM 738 and confirm the missing material, a Southern blot using the last 500 bp of the F613-1 sequence as probe was performed. Firstly, genomic DNA was digested with the restriction enzymes *Bam*HI, *Sac*I, *Hind*III, *Eco*RV and *Bgl*II. Following electrophoresis, the gel was blotted and hybridised to the probe. The blot showed bands at around 8 (*Bam*HI), >23 (*Sac*I), >23 (*Hind*III), 15 (*Eco*RV) and 11 kb (*Bgl*II), which confirms the presence of this sequence on DSM 738 strain (**Figure 4.6**). Nevertheless, these bands are larger than the fragments obtained from *in silico* digest of the F613-1 sequence of 6.8, 25.6, >23, 12.1 and 9.3 kb respectively, which confirms that *in silico* sequence information is indeed missing from this sequence end.

Furthermore, another Southern blot was performed on a PFGE gel containing DNA of *S. clavuligerus* DSM 738 digested with *Afl*II and *Dra*I, using the non-telomere end probe. The blot exhibited bands at below 100 kb for *Afl*II digest and around 945 kb for *Dra*I (**Figure 4.7**).

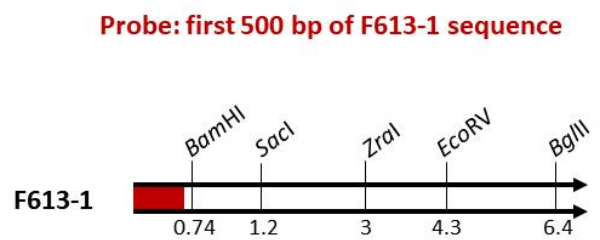
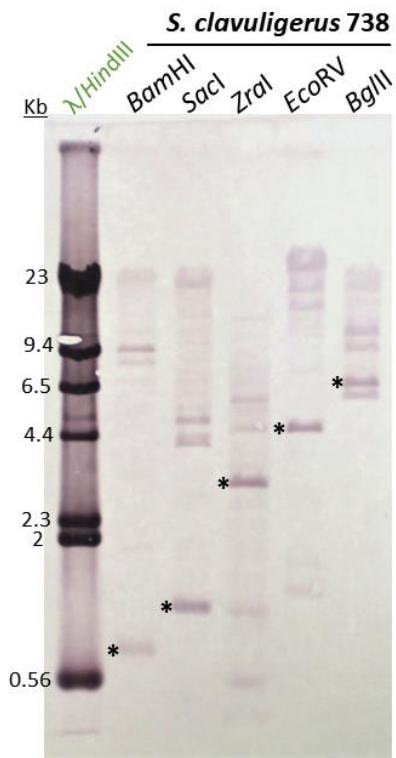


Figure 4.5. Southern blot of digested genomic DNA of *S. clavuligerus* DSM 738 hybridised to F613-1 telomere. The asterisks indicate the expected bands for the digested chromosome presented on the F613-1 end sequence diagram (probe represented as red box).

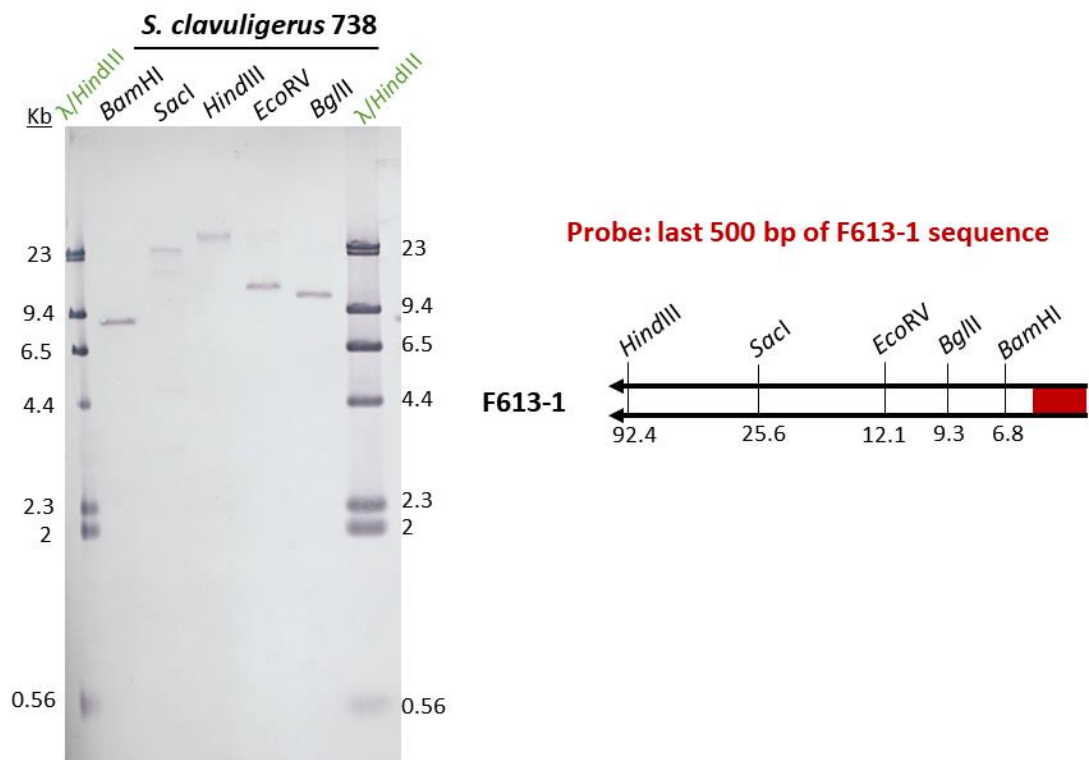


Figure 4.6. Southern blot of digested genomic DNA of *S. clavuligerus* DSM 738 hybridised to F613-1 non-telomeric end. The bands do not correlate to the expected sizes presented on the F613-1 end sequence diagram (probe represented as red box).

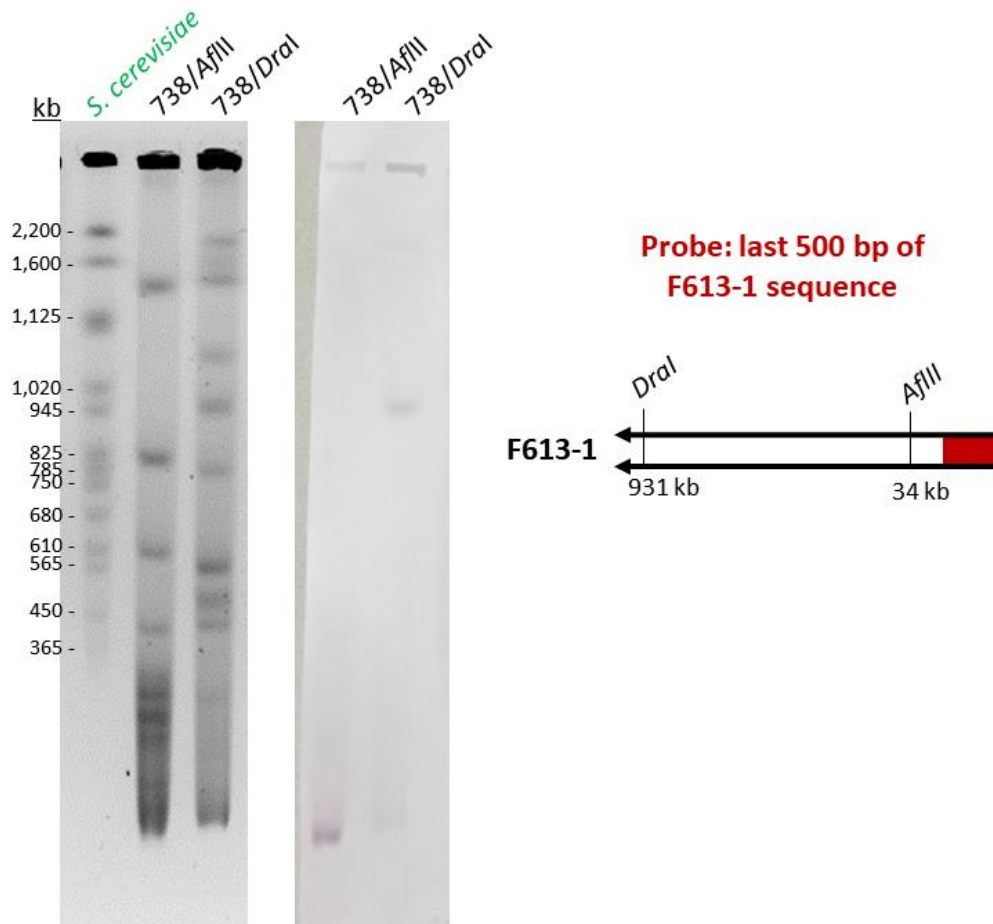


Figure 4.7. PFGE of *S. clavuligerus* DNA digested with *AflII* and *DraI* (left). Electrophoresis was performed at 4 V/cm for 30 hours. *S. cerevisiae* YNN295 was used as DNA size standard (Bio-Rad). Southern blot of the PFGE gel hybridised to non-telomeric end of F613-1 sequence (right). Bands correlate to the expected fragment sizes as indicated on the sequence diagram (probe represented as red box).

According to an *in silico* digestion of the F613-1 sequence, the fragments should show bands at 34 and 931 kb respectively. Our results suggest that, even though the right telomere is missing from the published sequence, this current sequence is not far from completion.

Overall, these results confirm the presence of more than one archetypal telomere in the genome of *S. clavuligerus* DSM 738. In addition, it is evidenced once again that even the most advanced sequencing technology available is not fully able to obtain complete linear sequences.

4.3 The telomeres of *S. clavuligerus* DSM 738 chromosome and plasmids are identified using the self-ligation method.

In order to identify the telomeres of the chromosome and plasmids of *S. clavuligerus* DSM 738, we performed the self-ligation method adapted from Fan *et al.*, 2012.

In the case of the chromosome telomeres, blunt-end restriction enzymes were chosen for each end depending on the restriction site found closest to the sequence end. *Sma*I (CCC'GGG) was chosen for both the left and right ends. Primers pairs were then designed so they would face outward on the sequence (divergent orientation); one of the primers anneals near the restriction site, while the other one anneals close to the end of the sequence. The self-ligation telomere purification method was then performed following the workflow presented on **Figure 4.8**. The primers Chrom-iPCR-L1 and Chrom-iPCR-L2 were used to amplify the left telomere, and the primers Chrom-iPCR-R1 and Chrom-iPCR-R2 for the right telomere. Gel electrophoresis of the amplified regions shows a band at 0.8 kb for the left telomere and two bands of sizes 2.7 and 1.1 kb for the right telomere (**Figure A2** in Appendix). The amplicons were finally sequenced using the previously mentioned primers. The obtained sequences exhibited the same telomere joint to half of the *Sma*I restriction site (CCC) (**Figure 4.9**). This confirms that the digested DNA fragment containing the end, self-ligated at the blunt ends (telomere and *Sma*I-cut) and so, the product could be amplified by divergent primers. This confirms that we have successfully purified and sequenced the telomeres of the *S. clavuligerus* DSM 738 chromosome using the self-ligation method.

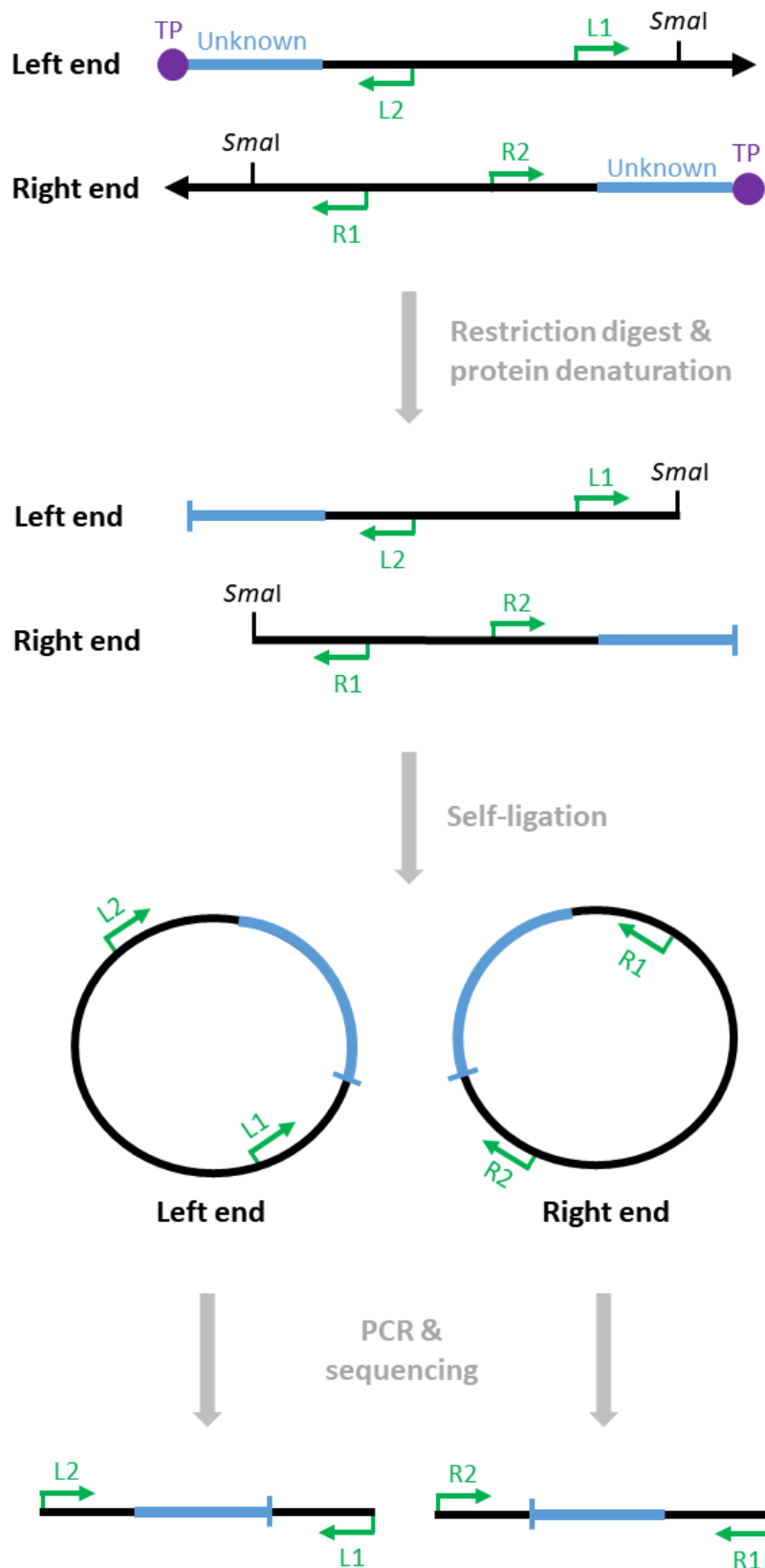


Figure 4.8. Diagram of the workflow followed to purify and identify the telomeric sequences of the chromosome of *S. clavuligerus* DSM 738. TP: terminal protein. Primers represented as L1/2 and R1/2.

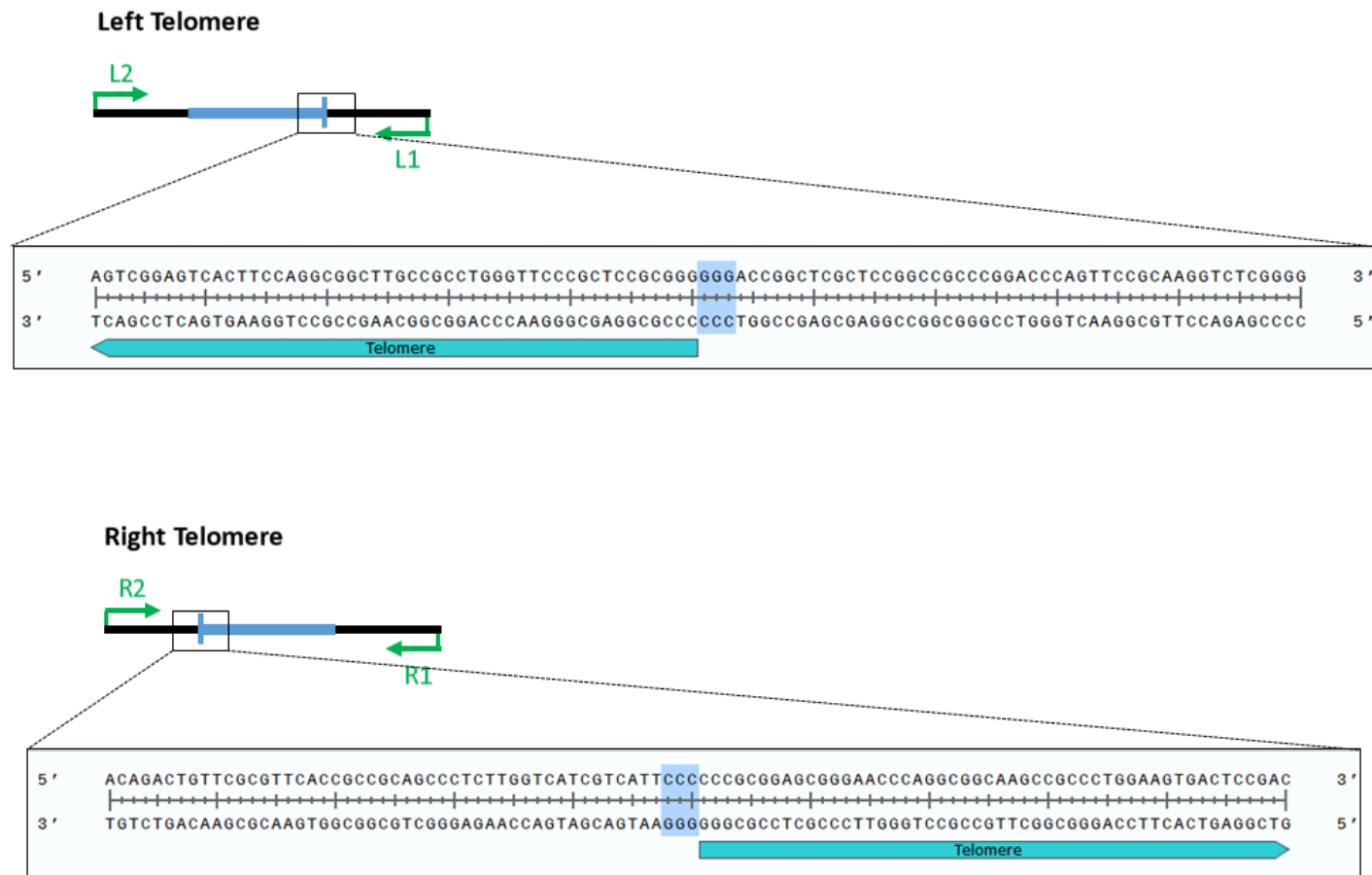


Figure 4.9. Sequences obtained for the chromosome left and right telomeres. The telomere sequence is joint to half of the *SmaI* restriction site (highlighted on the sequence by a blue box) due to self-ligation of digested fragments.

The same workflow was utilised for purification of the telomeres of the plasmids pSCL4, pSCL3 and pSCL2. For pSCL4, the enzymes *ZraI* (GAC'GTC) and *SmaI* were used for purifying the left and right telomeres respectively. Amplification of the left end was done using the primers pSCL4-iPCR-L1 and pSCL4-iPCR-L2, while the primers pSCL4-iPCR-R1 and pSCL4-iPCR-R2 were used for the right end, these exhibited bands at 0.6 and 0.7 kb respectively (**Figure A3** in Appendix). In the case of pSCL3, the enzymes *SmaI* and *EcoRV* (GAT'ATC) were used for the left and right telomeres, and the primers pSCL3-iPCR-L1- pSCL3-iPCR-L2, and pSCL3-iPCR-R1-pSCL3-iPCR-R2 that yielded a band of 0.3 kb for the left telomere and two bands of 0.7 and 1.3 kb for the right telomere. However, the sequencing results of the right end of pSCL3 did not extend to the sequence end, but since it was observed on the assembly that pSCL3 right end aligned well to the left telomere and was only missing the last 11 bp, the obtained left telomere was used also for completion of the right end. Finally, the enzyme *ZraI* was used for identification of both telomeres of pSCL2 and amplification was done using the primers pSCL2-iPCR-L1- pSCL2-iPCR-L2, and pSCL2-iPCR-R1-pSCL2-iPCR-R2 which exhibit amplicons of sizes 0.8 and 0.6 kb respectively.

Once obtained the telomeric sequences of all the replicons, these were aligned to the sequencing data, and the missing sequences were added to the assembly in order to construct the complete genome. 524 and 13 bp were added to the left and right ends of the chromosome assembly respectively. Furthermore, in order to observe the coverage of MiniSeq reads to the telomeres, the reads were mapped again to the final assembly. This shows how the Illumina reads cover up to the 9th bp from both ends of the chromosome sequence (**Figure 4.10**). Moreover, even though the first 524 bp were missing from the assembly, the sequencing data correlates to our purified telomere sequencing.

In the case of pSCL4, 61 bp were added to the left end, and 133 bp to the right end. When mapping the Illumina reads to the telomeres, it was observed that these covered up to the 9th bp from the sequence ends, just like the chromosome reads (**Figure 4.11**).

The sequence of pSCL3 was completed by adding 1,537 bp to the left end, and 11 bp to the right end. It was observed during assembly of pSCL3, the presence of larger terminal inverted repeats (TIR) sequences on this replicon, which explains the large missing region at the left end, as current automated assembly methods usually fail to properly obtain both TIR sequences. The final complete assembly of pSCL3 was mapped to the MiniSeq reads and showed how they mapped perfectly to the telomeres up to the last 12th bp (**Figure 4.11**).

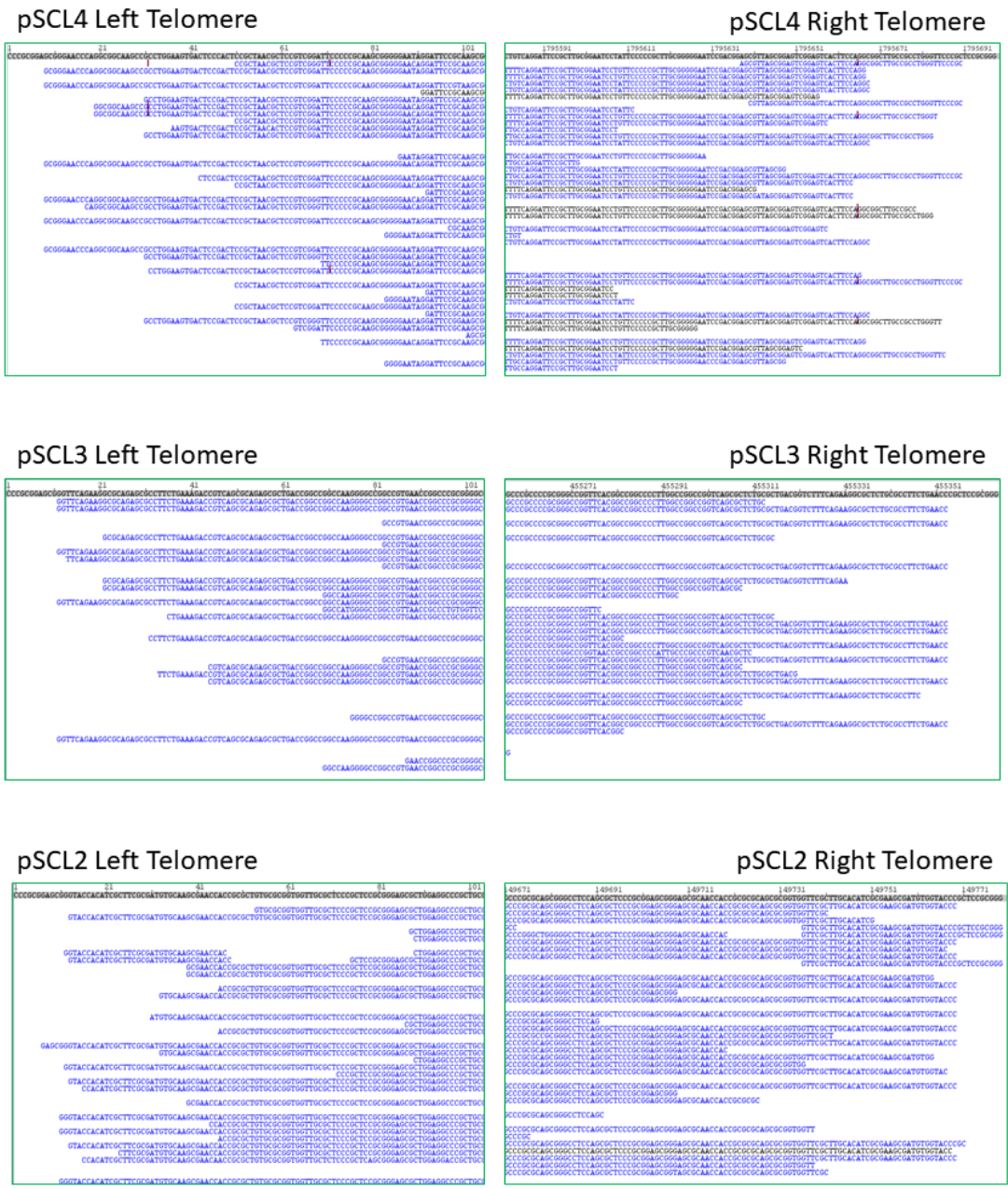


Figure 4.11. Illumina MiniSeq reads (in blue and black) mapped to the telomeric sequences (grey) of pSCL4, pSCL3 and pSCL2 as viewed on Artemis.

In the case of pSCL2, 334 bp and 10 bp were added to the left and right end of the assembly respectively. When mapping Illumina reads to the pSCL2 telomeric sequences, it is shown how they cover up to the 7th bp of the left telomere whilst for the right telomere, there are interestingly three reads that map to the very end of the sequence (**Figure 4.11**). Given that the aligned region of these reads is short, they were checked individually, and it was confirmed that the reads were incorrectly mapped to the end region given the presence of many repeated regions in the telomeres of *S. clavuligerus*. In consequence, correctly aligned reads are confirmed to map to up the 9th bp from the end.

In summary, we have obtained the complete sequences for the chromosome and plasmids of *S. clavuligerus* DSM 738 by adding the telomere sequences obtained using the self-ligation method to the closed assemblies. In addition, these results confirm good correlation between the telomeres sequences and the sequencing data and evidences the inability of Next-generation sequencing technologies to reach the end of linear sequences.

4.4 The genome of *S. clavuligerus* DSM 738 carries four different sets of archetypal telomeres.

In order to characterise the telomeres of the chromosome and plasmids in *S. clavuligerus* DSM 738, we compared the telomeric sequences between themselves and to the published telomeres of other Actinobacteria carrying linear chromosome and plasmids.

Firstly, the telomeres from the same replicons were compared to each other and it was confirmed that each molecule carries an almost identical telomere at both ends of the sequence. Furthermore, in order to compare these to other relevant species, we constructed an alignment using archetypal telomeres from 22 chromosomal and 20 plasmid sequences against the five telomeres of *S. clavuligerus*. This alignment shows the nucleotide identity of the 120 -130bp of the telomere at the 3' end (**Figure 4.12**). This comparison shows the presence of archetypal sequence in all the *S. clavuligerus* telomeres, this is evidenced by the observation of the 13 bp sequence CCCGCGAGCGGG that forms Palindrome I, which is characteristic of archetypal telomeres. Beyond these 13bp, the telomeric sequences start to differ but many conserved regions are still observable, as visualised on the consensus graph. These results suggest that all the *S. clavuligerus* DSM 738 telomeres are archetypal.

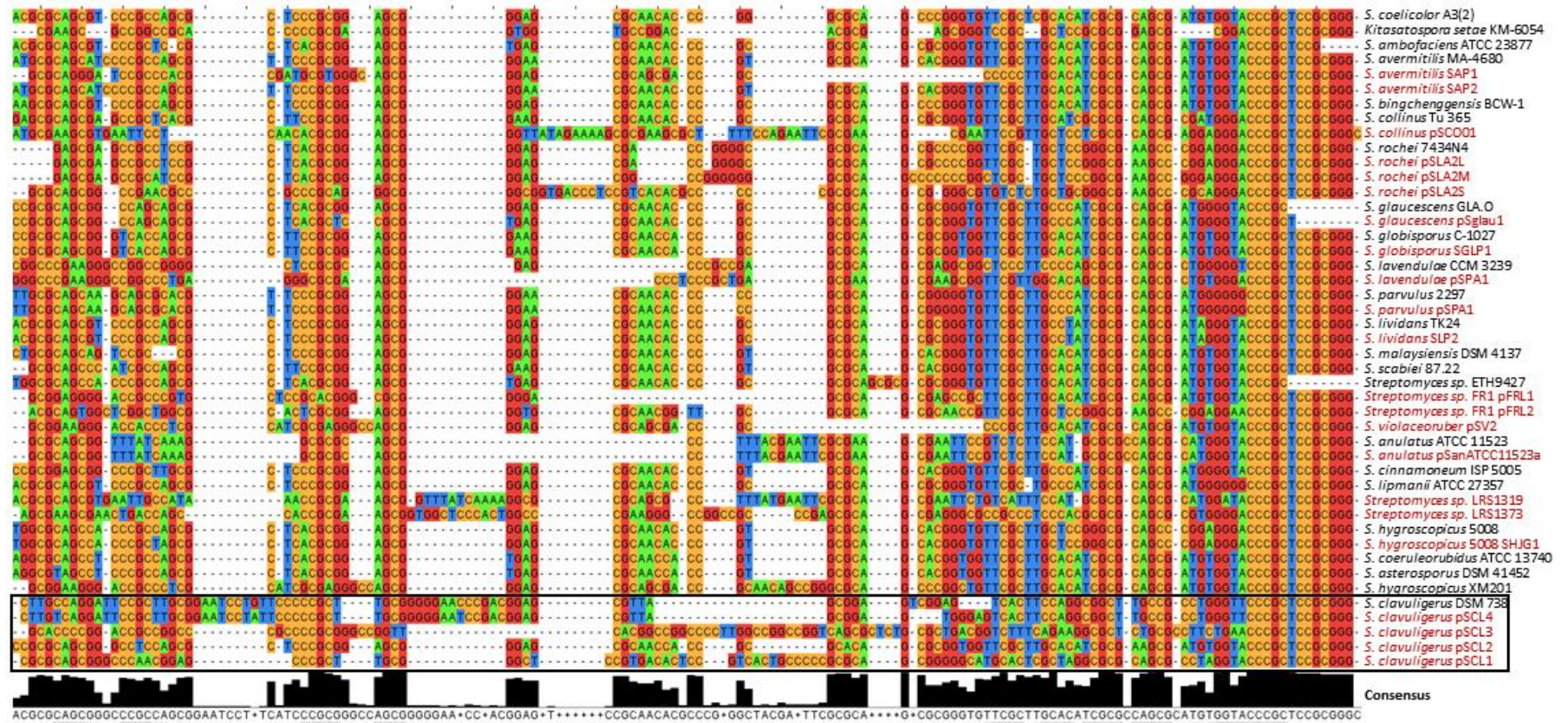


Figure 4.12. Alignment of telomeres of linear chromosomes and plasmids of Acinobacteria species, including the five telomeres of *S. clavuligerus* DSM 738 (highlighted in the back box). Telomeres from chromosome sequences are presented in black and plasmids in red. Alignment constructed in Jalview using the MUSCLE alignment. The consensus graph (in black) presents the % of the calculated consensus nucleotide for each position.

Moreover, a maximum likelihood phylogenetic tree was constructed in order to compare the last 150 bp of the *S. clavuligerus* telomeres to other Actinobacteria telomeres. The tree exposes how the telomeres are split in two main clusters (**Figure 4.13**). The smaller cluster contains the telomeres of *S. clavuligerus* chromosome, pSCL4, pSCL3 and pSCL1. The larger cluster harbours most of the telomeres analysed, including the telomere of *S. clavuligerus* pSCL2 and the chromosomal telomere of the model organism *S. coelicolor*. Furthermore, pairwise alignments of these telomeric sequences were calculated, which confirmed that the *S. clavuligerus* chromosome and pSCL4 telomeres are 96% identical (**Table 4.3**), which correlates to their close location on the tree. High similarity between chromosome and plasmid telomeres in the same strain is also visualised on the phylogenetic tree on many other species, such as, *S. anulatus*, *S. rochei*, *S. hygroscopicus*, *S. lavendulae*, *S. globisporus*, *S. avermitilis*, *S. parvulus*, *S. glaucescens* and *S. lividans*. However, this similarity substantially decreases when comparing the chromosome telomeres to the rest of the plasmids: 56.5% for pSCL3, 54.5% for pSCL2 and 61.36%. In the case of pSCL2, its telomeres are highly similar to other *Streptomyces* species such as *S. coelicolor* (80.5%), which correlates to its position on the phylogenetic tree. Moreover, low similarity is visualised between the telomeres of the *S. clavuligerus* plasmids.

Overall, these results suggest that the chromosome and the megaplasmid pSCL4 carry almost identical telomeres, while the rest of the telomeres in *S. clavuligerus* are substantially different. These results might help us gain better understanding on how these replicons evolved and how their telomeres are replicated and maintained.

Table 4.3. Identity percentage values calculated by pairwise alignment of *S. clavuligerus* 150 last bp of telomeres located on chromosome, pSCL4, pSCL3, pSCL2 pSCL1 and the *S. coelicolor* chromosome.

	Chrom	pSCL4	pSCL3	pSCL2	pSCL1	<i>S. coelicolor</i>
Chromosome	100%	96%	56.50%	54.50%	61.36%	57.39%
pSCL4	96%	100%	53.76%	55.68%	61.02%	58.86%
pSCL3	56.50%	53.76%	100%	55.23%	59.30%	59.28%
pSCL2	54.50%	55.68%	55.23%	100%	64.12%	80.50%
pSCL1	61.36%	61.02%	59.30%	64.12%	100%	63.35%
<i>S. coelicolor</i>	57.39%	58.86%	59.28%	80.50%	63.35%	100%

Tree scale: 0.1

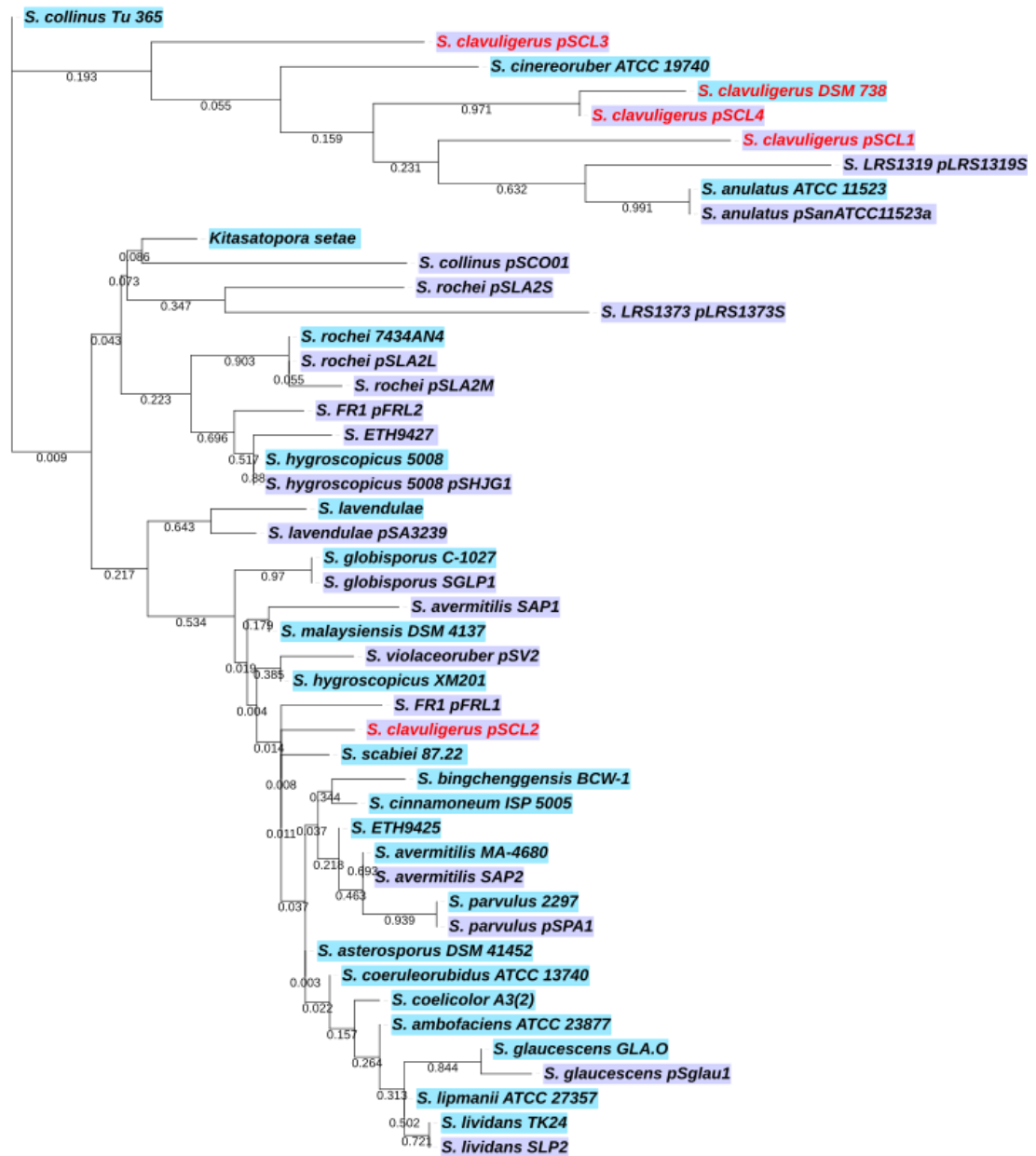


Figure 4.13. Maximum likelihood phylogenetic tree of the last 150 bp of archetypal telomeres of different Actinobacteria species. Chromosomal telomeres are highlighted in blue and plasmid telomeres in purple. *S. clavuligerus* telomeres are indicated by red font. Tree constructed on MEGA with 1000 bootstraps and edited on iTOL.

In addition, in order to characterise the secondary structures of *S. clavuligerus* telomeres, folded forms of the telomeres were obtained using the last 80 to 100 bp of the telomeres and compared to the secondary structure of the *S. coelicolor* chromosome telomere. The five folded telomeres exhibit four hairpin loops composed by the four Palindromes I to IV (**Figure 4.14**). These secondary structures follow the Clover Leaves model of four loops characteristic of archetypal telomeres. The five telomeres carry the same first loop (Palindrome I), which is an element common with the *S. coelicolor* chromosome telomere. Moreover, the telomeres on the chromosome, pSCL4, pSCL2 and pSCL1 carry a second loop of 19-20 nucleotides (Palindrome II) which is separated by 3 nucleotides from the first loop, just like in the *S. coelicolor* telomere. This is not the case in pSCL3's telomere, which exhibits a larger second loop that starts right after the first loop. A larger third loop of 19 to 24 nucleotides (Palindrome III) is present in all the structures, which are similar to that of *S. coelicolor*, apart from the *S. clavuligerus* chromosome and pSCL4, which show an aperture in the middle of the loop. Finally, they all present a long fourth loop of 22-23 nucleotides (Palindrome IV), except from pSCL1 whose fourth loop is shorter (15 nucleotides). The chromosome and the megaplasmid pSCL4 telomeres exhibit the same secondary structures, which is to be expected as the four loops carry exactly the same sequence, so we can consider that these two replicons carry the same archetypal telomere. It is also noticeable how the secondary structure of the pSCL2 telomere is highly similar to the *S. coelicolor* telomere, considering their sequences are different. However, this is consistent with their high pairwise alignment identity mentioned before. The telomeres of pSCL1 and pSCL3 are clearly distinct from the rest and *S. coelicolor* telomeres, while still presenting the archetypal conformation.

In summary, these results show how *S. clavuligerus* DSM 738 carries four types of archetypal telomeres, considering the chromosome and the megaplasmid carrying the same telomere.

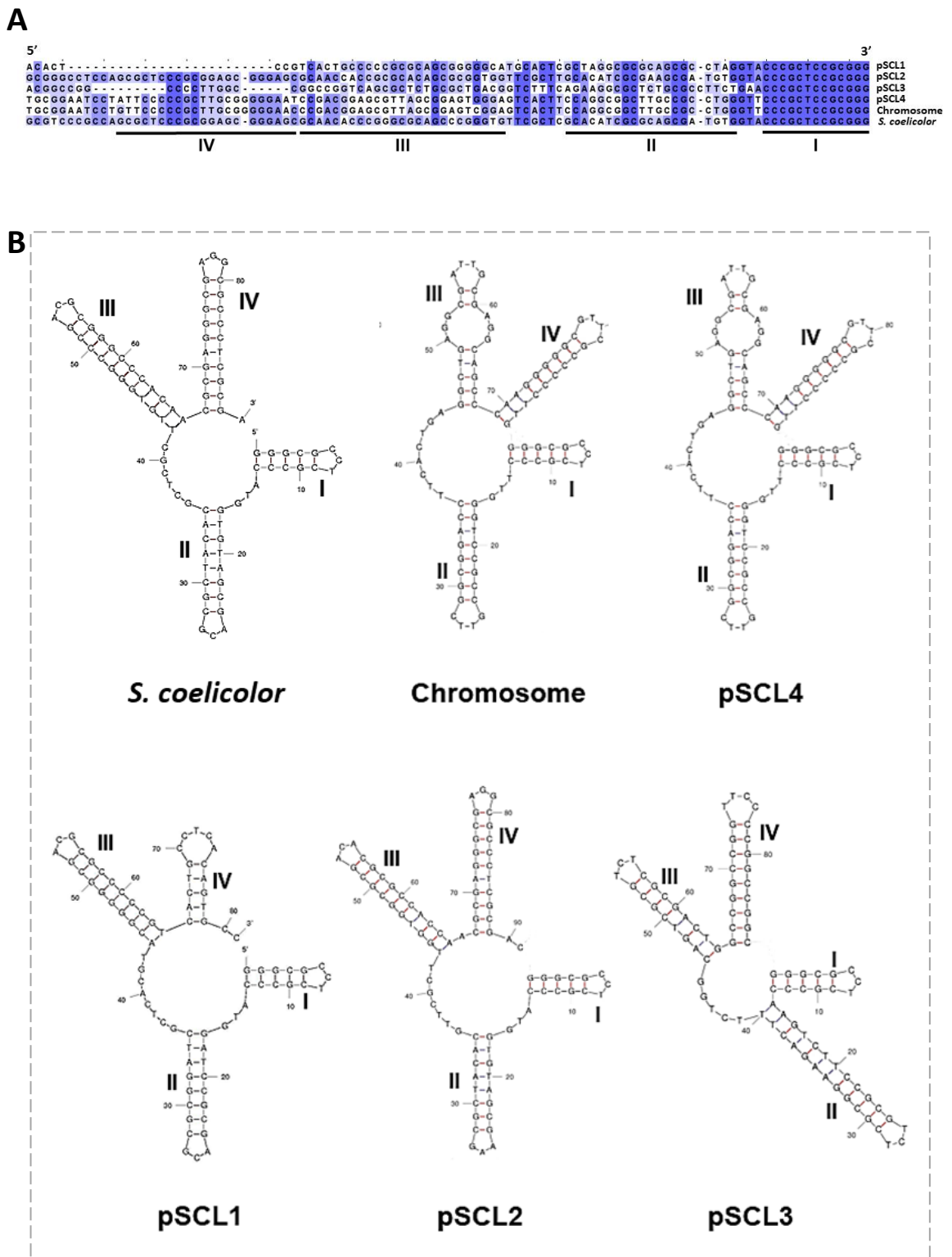


Figure 4.14. A: Alignment of the last 100 bp of the telomeres of *S. clavuligerus* chromosome and plasmids and the chromosome of *S. coelicolor* showing the similarity in blue. **B:** Folded forms of the telomeres showing a Clover Leaves structure. Palindromes are indicated in roman numerals.

4.5 The read coverage depth indicates a higher copy number of pSCL2 and lower copy number of pSCL4 and pSCL3 in reference to the chromosome.

Once the assemblies had been extended into the final complete sequences of the *S. clavuligerus* DSM 738 genome, we constructed maps of the sequenced reads aligned to the final assemblies, using the Illumina MiniSeq and PacBio raw data. The Illumina reads map presents full coverage of the genome by reads (**Figure 4.15A**) except for the last nucleotides of each sequence end as described in section 4.3. In the case of the PacBio reads map, reads align to the entire genome (**Figure 4.15B**), except for a few kb at the sequences ends, and the pSCL1 sequence that presents low coverage by these PacBio reads. This generalised high coverage of both Illumina and PacBio reads suggest a correct sequencing and assembly of our final genome sequences. Moreover, it is noticeable that the read number appears to be more consistent in the Illumina data than the PacBio across the genome, with only a few points of significantly higher number of reads that correspond to ribosomal RNA coding regions. Furthermore, the read number tends to increase at the middle of the sequence, which is observable especially on the chromosome and pSCL4, and it is probably due to DNA replication starting at the replicon centre, making these sequences more abundant than those at the telomeres. In addition, these maps allow visualisation of the difference in read coverage for each of the replicons. For example, the Illumina read coverage of pSCL4 and pSCL3 is slightly lower than the chromosome read coverage while the pSCL2 coverage is more than double of the chromosomes. A very similar pattern is observable on the PacBio reads map. This difference in coverage suggests a diversity in replicon copy number in *S. clavuligerus* mycelia.

In order to estimate the copy number of each replicon, the mean coverage depth was used as a proxy. Illumina reads coverage was used as it is more homogenous than the PacBio data, and the mean coverage was calculated individually for each molecule. The following mean coverage values were obtained for each replicon: chromosome 55.91, pSCL4 30.86, pSCL3 39.56, pSCL2 139.9 and pSCL1 59.73. Considering the chromosome copy number as 1, the replicon:chromosome ratio was calculated for each plasmid. pSCL4 shows an estimated copy number of 0.55 and pSCL3 0.71 (**Figure 4.16**). These numbers suggest the possibility of an inconsistent segregation of these two plasmids, and so, pSCL4 and pSCL3 might be less abundant than the chromosome in some mycelial compartments. In addition, the replicon:chromosome ratio of pSCL2 is 2.5, which implies that there are at least two copies

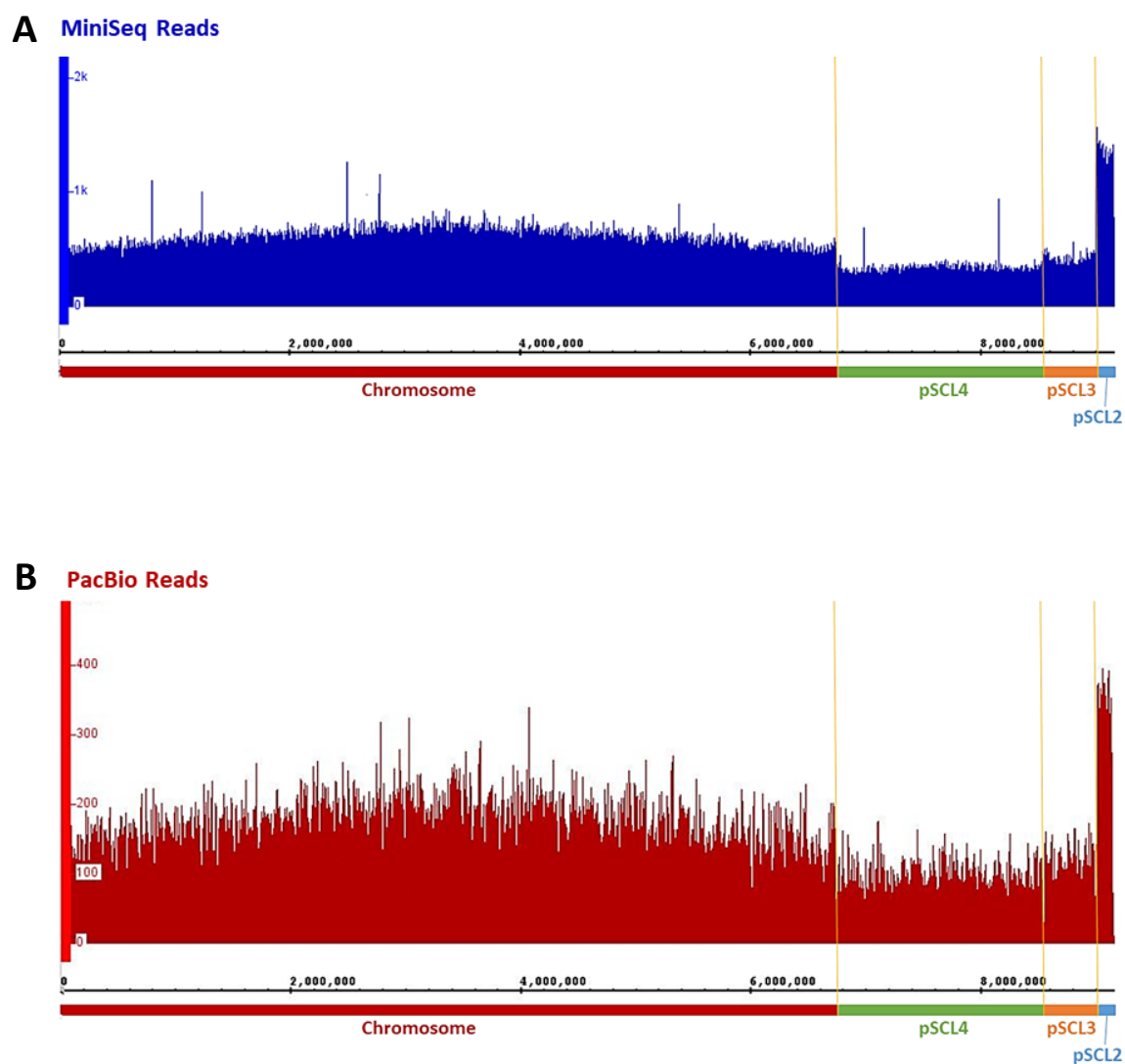


Figure 4.15. Whole genome sequencing reads mapped to the final assembly (view from IGB) **A:** Reads obtained from Illumina MiniSeq sequencing mapped to assembly **B:** Reads obtained from PacBio sequencing mapped to assembly.

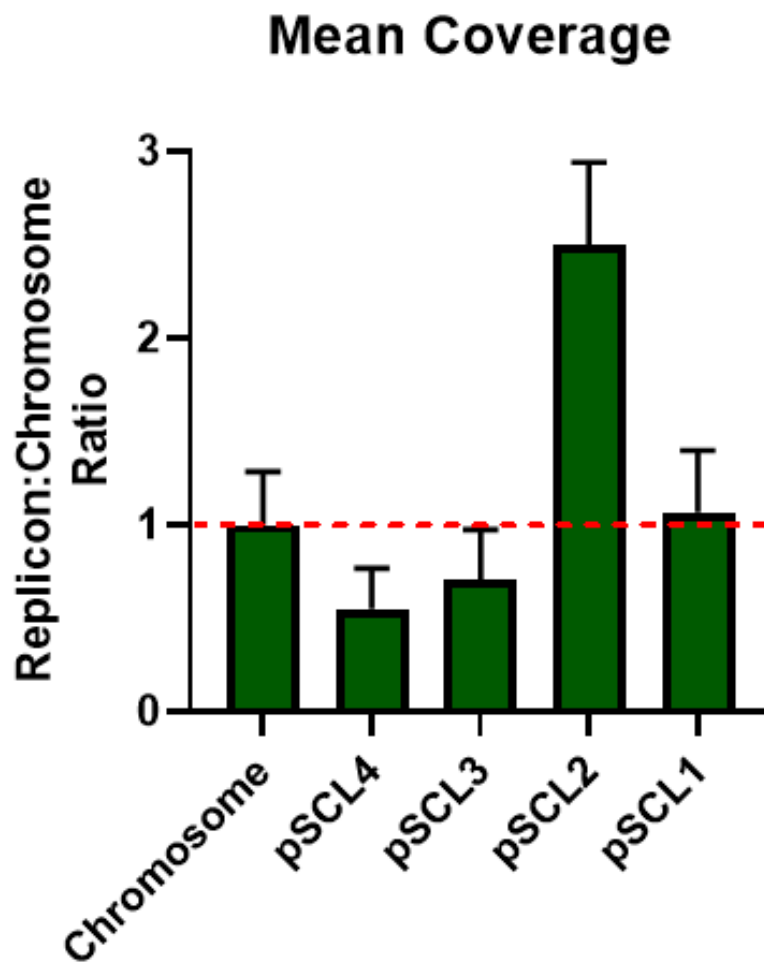


Figure 4.16. Illumina MiniSeq sequencing reads mean coverage for each replicon presented as replicon:chromosome ratio.

of this plasmid per copy of chromosome. The calculated ratio of pSCL1 is 1.07, which suggests a copy number of 1 in reference to the chromosome.

Overall, these results show how Illumina and PacBio read coverage maps confirm a correct assembly of the genome of *S. clavuligerus* DSM 738. In addition, mean coverage depth calculation allowed estimation of replicon copy number, suggesting a higher copy number of pSCL2 and lower copy number of pSCL4 and pSCL3 in reference to the chromosome.

4.6 General characteristics and annotation of the genome of *S. clavuligerus* DSM 738

In order to characterise the genome of *S. clavuligerus* DSM 738 and compare it to other publications, a table of general characteristics for each one of the replicons was constructed (**Table 4.4**). The first characteristic studied was the replicons sizes. The chromosome, with 6,748 Mb, is 263 bp larger than the chromosome sequence published by Hwang *et al.*, 2019, although this difference could be due to the lack of telomeres in the latter sequence. In contrast, both industrial strains F613-1 and F1D-5 present larger chromosomes than DSM 738. Our sequence of pSCL4 is also slightly larger than the Hwang sequence (203 bp of difference) and significantly larger than the pSCL4 in F613-1 and F1D-5. Furthermore, all the replicons present a linear topology and a GC content of 70 to 72%, characteristic common from *Streptomyces* chromosomes and plasmids.

Table 4.4. List of general characteristics of the genome of *S. clavuligerus* DSM 738.

	Chrom	pSCL4	pSCL3	pSCL2	pSCL1
Size (bp)	6,748,854	1,795,698	455,364	149,779	11,696
Topology	linear	linear	linear	linear	linear
GC content (%)	72.7	71.9	70.8	70.1	71.9
Coding sequences	5,825	1,728	528	175	12
rRNA operons	18	0	0	0	0
tRNA genes	65	0	1	0	0
Putative biosynthetic gene clusters	25	17	1	0	0
Terminal inverted repeats length (bp)	293	194	1,597	413	917

Moreover, the chromosome and plasmids of *S. clavuligerus* DSM 738 were uploaded to RAST in order to obtain the annotations of each replicon. This resulted in identification of a total number of 8,268 coding sequences. The chromosome carries 70% of the coding sequences, pSCL4 the 21%, and pSCL3, pSCL2 and pSCL1 carry the 6, 2 and 0.15% of the coding sequences respectively. In addition, 1,997 annotations subsystems (collections of functionally related protein families) were identified that were distributed by their biological process as indicated in **Table 4.5**. The chromosome carries 96% of these annotation subsystems, and pSCL4 the 4%. In addition, all the rRNA and tRNA necessary for primary metabolism were found on the chromosome, as well as a putative tRNA gene on pSCL3.

Furthermore, in order to compare these annotations to other *Streptomyces* species, the annotated genome of *S. clavuligerus* was uploaded into OrthoVenn2 and ortholog gene clusters were identified between *S. clavuligerus* and all the available *Streptomyces* species in OrthoVenn2: *S. albus*, *S. avermitilis*, *S. bingchengensis*, *S. cattleya*, *S. coelicolor*, *S. fulvissimus*, *S. globisporus*, *S. griseus*, *S. lividans*, *S. rapamycinicus*, *S. scabiei*. The OrthoVenn2 analysis resulted in 2,168 ortholog gene clusters shared between all the organisms (**Figure 4.17**). These conserved regions represent the core genome and, composed of 2,205 genes, it is the 40% of the total genome of *S. clavuligerus*. In addition, in order to identify if the plasmids contribute to the core genome, these genes were analysed and it was confirmed that the megaplasmid pSCL4 presents 28 of these genes, while pSCL3 carries 3 core genes, representing 1.3% and 0.14% of the core genome respectively. These 31 genes were individually characterised, and it was observed that they code either hypothetical proteins, proteins involved in secondary metabolism or they were genes with homologs on the chromosome. These observations suggest that all the core genes involved on primary metabolism are located on the chromosome.

Additionally, in order to obtain graphical representation of the annotated sequences, the chromosome and plasmids sequences were uploaded to CGView Server. The circular graphs of each replicon were constructed, indicating the coding sequences located on the forward and reverse strands, RNA genes and GC skew (**Figure 4.18**)

Table 4.5. List of RAST annotation subsystems identified on each *S. clavuligerus* DSM 738 replicon.

Annotations Subsystems	Chrom	pSCL4	pSCL3	pSCL2	pSCL1
Amino acids and derivatives	377	15	3	0	0
Carbohydrates	293	9	0	0	0
Protein metabolism	214	2	3	0	0
Cofactors, vitamins, prosthetic groups, pigments	187	0	0	0	0
Fatty acids, lipids and isoprenoids	160	0	0	0	0
Respiration	109	1	0	0	0
Nucleosides and nucleotides	96	3	0	0	0
DNA metabolism	87	0	4	0	0
Membrane transport	43	7	0	0	0
Stress response	43	0	0	0	0
RNA metabolism	40	2	0	0	0
Virulence, disease and defence	38	3	0	0	0
Phosphorus metabolism	34	0	0	0	0
Iron acquisition and metabolism	34	0	0	0	0
Cell wall and capsule	31	15	0	0	0
Regulation and cell signalling	24	2	0	0	0
Metabolism of aromatic compounds	17	4	0	0	0
Secondary metabolism	15	0	0	0	0
Potassium metabolism	12	0	0	0	0
Dormancy and sporulation	11	0	0	0	0
Nitrogen metabolism	10	0	0	0	0
Sulphur metabolism	8	8	0	0	0
Phages, prophages, transposable elements	3	0	0	0	0
Miscellaneous	27	3	0	0	0

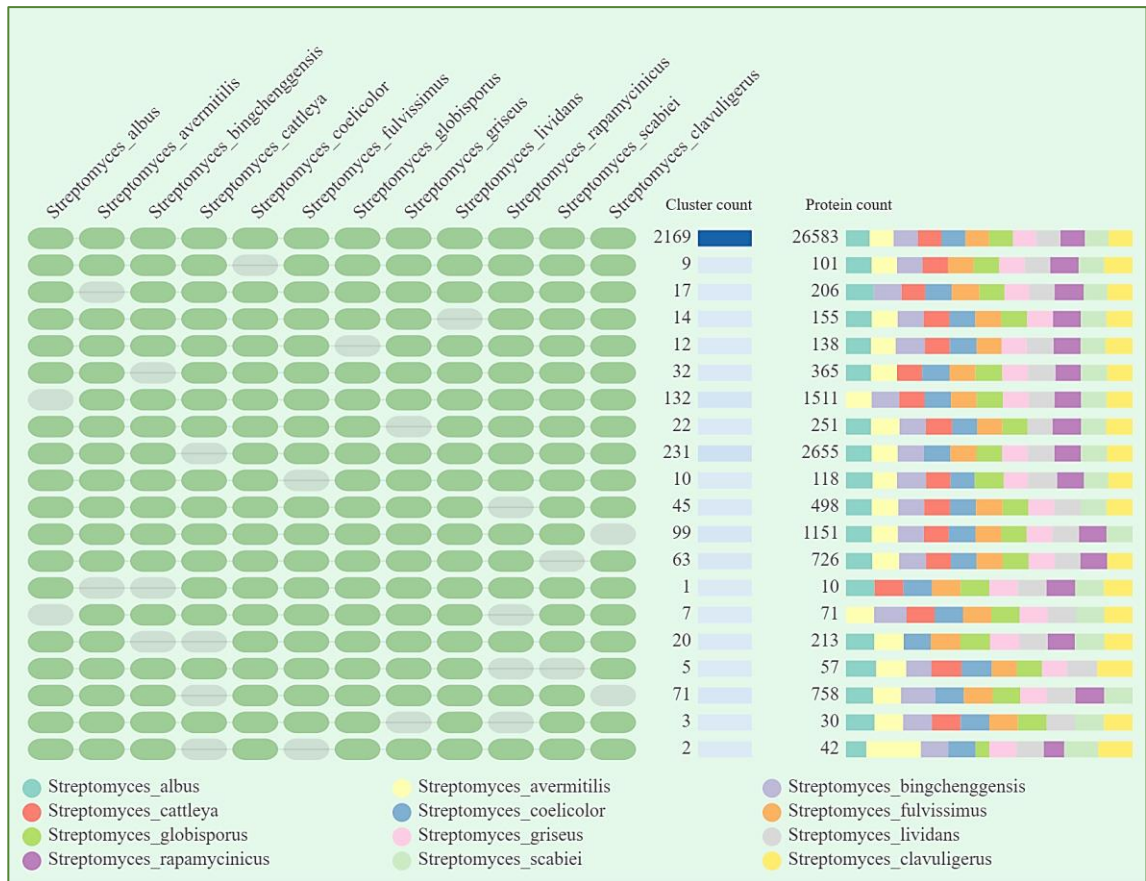


Figure 4.17. OrthoVenn2 diagram of ortholog gene clusters shared between *S. clavuligerus* and eleven different species of *Streptomyces*.

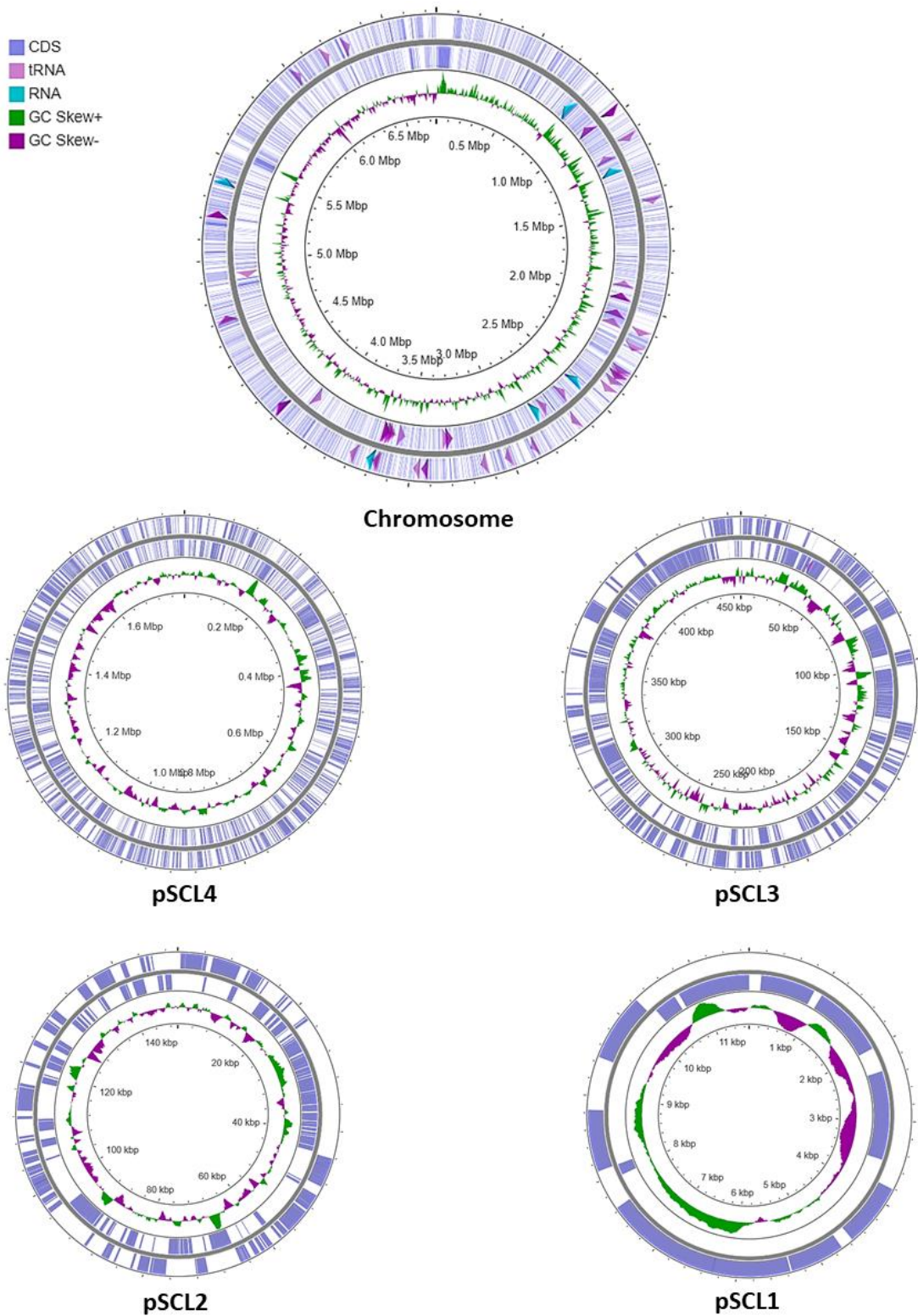


Figure 4.18. Schematic representation of the genome of *S. clavuligerus* DSM 738. Chromosome and plasmids sequences presented as circles (dark grey). The coding sequences (CDS) of the forward and reverse strains are presented in blue at each side of the sequence. The concentric circle presents the GC skew in green (positive skew) and purple (negative skew). Not to scale.

Finally, we characterised the length of the terminal inverted repeats of each replicon. To obtain this, the first 10 kb of each end of the sequences were aligned to the reverse complementary sequence of the last 10kb. The first and last 239 bp of the chromosome sequence exhibit a 93.27% similarity. At the same time, the first and last 194 bp of pSCL4 show a sequence identity of 96.39%. Although a second section of the pSCL4 ends shows similarity beyond these 194 bp, it is not included on the TIR as it is not contiguous, and the nucleotide identity is low (62%). In the case of pSCL3, the high sequence similarity of 99.50% was found between the first and last 1,597 bp of the sequence. Moreover, the first and last 413 bp of the pSCL2 sequence have a sequence identity of 99.03%. Lastly, the alignment of the ends of the pSCL1 sequence shows a of 99.24% similarity between the first and last 917 bp of the sequence. These results are illustrated on the **Figure 4.19**, exhibiting the alignment of the first and last 2,000 bp of each sequence. In summary, we report the TIR length of the *S. clavuligerus* replicons: chromosome 239 bp, pSCL4 194 bp, pSCL3, 1,597 bp, pSCL2 413 bp and pSCL1 917 bp. Considering *Streptomyces* TIR have been reported to range from a few bp (14 bp in the case of *S. hygrosopicus*) to over 1 Mb (1.06 Mb in *S. coelicolor*), *S. clavuligerus* replicons carry relatively short TIRs.

4.7 The genome of *S. clavuligerus* carries 43 putative biosynthetic gene clusters, 17 of them are located on the megaplasmid.

In order to investigate the gene clusters responsible for the biosynthesis of secondary or specialised metabolites on the genome of *S. clavuligerus* DSM 738, the sequences of the chromosome and plasmids were processed using antiSMASH 5.0. A total number of 43 biosynthetic gene cluster were identified in the genome of this strain, located on the chromosome, megaplasmid pSCL4 and pSCL3.

Twenty-five biosynthetic gene clusters were predicted for the chromosome of *S. clavuligerus* (**Table 4.6**). The position of these clusters on the sequence is illustrated on **Figure 4.20**, which shows how more than half of the clusters are located on the right arm of the chromosome, and only five of the predicted clusters lie on the core region. These predicted biosynthetic gene clusters include five polyketide-like (PKS) clusters, such as cluster 23, identified as the naringenin cluster; cluster 1 that has 72% similarity to a cluster found in *S. varsoviensis*; and cluster 18 whose polyketide motif shares similarity to a spore pigment cluster in *S. avermitilis*.

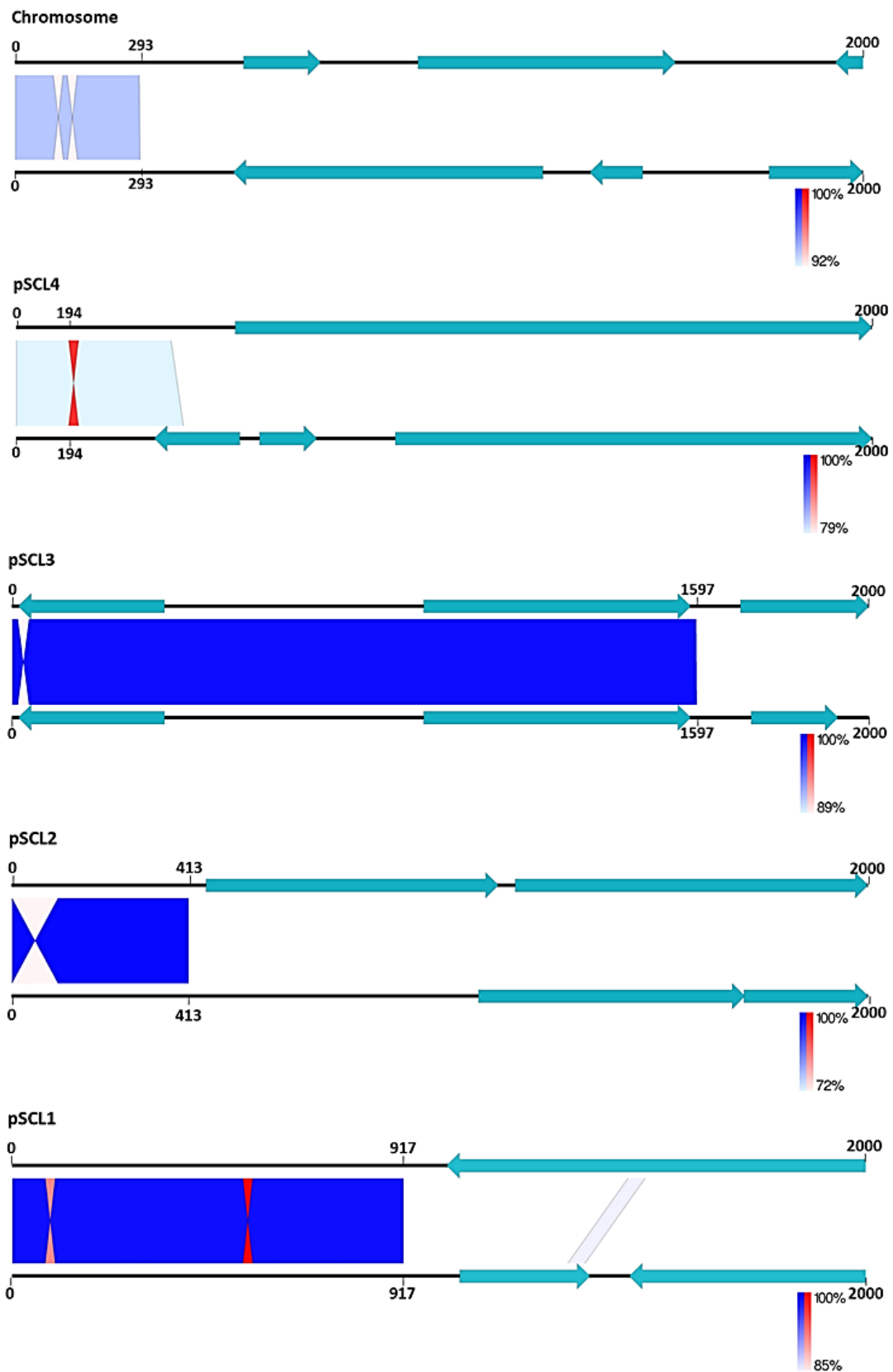


Figure 4.19. Easyfig alignments of the first 2,000 bp and the last 2,000 bp (reverse complement) of each *S. clavuligerus* replicon. Nucleotide identity is presented in dark blue and red (inverted sequence) which represent the terminal inverted repeats of each sequence.

Table 4.6. List of putative biosynthetic gene clusters of the *S. clavuligerus* DSM 738 chromosome predicted by antiSMASH 5.0. TPKS: Type-X Polyketide synthase, NRPS: Non-ribosomal peptide synthetase.

Cluster	Type	Position in sequence		Most similar cluster	Similarity
		From	To		
1	T1PKS	539	94,685	JBIR-100	72%
2	Terpene	222,299	243,465	Geosmin	100%
3	T1PKS, butyrolactone	599,820	682,423	4-hexadecanoyl-3-hydroxy-2-(hydromethyl)-2H-furan-5-one	54%
4	Ectoine	1,314,476	1,324,901	Ectoine	100%
5	Siderophore	2,305,381	2,315,785	Desferrioxamin B	100%
6	NRPS	2,672,295	2,719,118		
7	Butyrolactone	2,725,599	2,736,561	Lactonamycin	3%
8	Lanthipeptide	2,889,407	2,913,738		
9	β -lactam	3,427,743	3,446,127	Alanylclavam/2-hydroxymethylclavam	75%
10	Melanin	4,506,239	4,516,673	Melanin	100%
11	NRPS, β -lactam	4,835,754	4,886,296	Cephamycin C	89%
12	Nucleoside	4,944,602	4,965,279	Tunicamycin B1	85%
13	Lanthipeptide	5,069,285	5,093,158		
14	T1PKS, NRPS	5,149,875	5,203,445	Kanamycin	1%
15	Siderophore	5,446,264	5,455,958		
16	NRPS	5,506,790	5,582,877	A-201A	15%
17	Bacteriocin	5,695,620	5,706,993		
18	T2PKS, Lanthipeptide	5,728,686	5,822,210	Spore pigment	83%
19	NRPS	6,019,077	6,062,698		
20	Terpene	6,143,358	6,169,923	Hopene	69%
21	NRPS	6,177,505	6,226,737	Holomycin	100%
22	NRPS	6,239,865	6,293,694	Nucleocidin	43%
23	T3PKS	6,429,141	6,470,196	Naringenin	100%
24	Terpene, NRPS, T1PKS	6,572,923	6,670,695	Heat-stable antifungal factor	75%
25	Siderophore	6,722,894	6,736,462		



Figure 4.20. Graphical representation of the *S. clavuligerus* DSM 738 chromosome indicating the 25 putative biosynthetic gene clusters and their location on the sequence. Obtained from antiSMASH 5.0.

Moreover, five clusters were identified as Non-ribosomal peptide-like (NRPS) clusters, which includes cluster 21, which was identified as the holomycin cluster. Two terpene-like clusters were predicted, the geosmin (cluster 2) and hopene (cluster 20) clusters, which are common in *Streptomyces* species genomes. In addition, cluster 24 carries terpene, polyketide and non-ribosomal peptide-like motifs, and exhibits 75% similarity to a heat-stable antifungal factor previously found in *Lysobacter enzymogenes*. Additionally, other clusters common to streptomycetes genomes were found on *S. clavuligerus*, like the ectoine (cluster 4), desferrioxamin B (cluster 5) and melanin (cluster 10) clusters. Another interesting cluster is the number 12, which is predicted as a nucleoside-coding cluster with 85% similarity to the tunicamycin B1 cluster in *Streptomyces chartreusis*. Finally, two beta-lactam clusters were identified; the cluster 9 was identified as an alanylclavam and cluster 11 as the cephamycin C cluster. However, it was noticed that cluster 11 is, in fact, the clavulanic acid-cephamycin C supercluster. The blast of this supercluster against the individual cephamycin and clavulanic acid cluster is illustrated in **Figure 4.21**.

Seventeen biosynthetic gene clusters were identified on the sequence of the megaplasmid pSCL4 (**Table 4.7**). These clusters, unlike the ones on the chromosome, are dispersed across the entire plasmid sequence (**Figure 4.22**). The first two clusters on pSCL4 were identified as the terpenes (+)-T-muurolo (cluster 1) and (-)- δ -cadinene (cluster 2). In addition, seven more clusters carry terpene-like motifs, such as cluster 6 that has 100% similarity to the lanthipeptide venezuelin cluster from *S. venezuelae*. Furthermore, three of those clusters (7, 9 and 16) also carry NRPS-like motifs but show low similarity to previously described specialised metabolite clusters. Two more clusters exhibit NRPS motifs (15 and 17) and three clusters also show Type-1PKS-like sequences (9, 14 and 15), but were not identified as other known metabolite clusters. Nevertheless, cluster 11 shows 82% similarity to the alkaloid staurosporine; and cluster 10, that was identified as a beta-lactam cluster, and has 100% similarity to an alanylclavam cluster. Interestingly, this beta-lactam cluster carries four genes identical to ones located on the clavulanic acid cluster on the chromosome.

Only one biosynthetic gene cluster was identified on the sequence of pSCL3 (**Table 4.8**), which is located at the left arm of the replicon (**Figure 4.23**). This cluster is composed by a terpene and a nucleoside motif. Three of the genes (18%) are identical to genes located on the pseudouridimycin biosynthetic gene cluster. Furthermore, no biosynthetic gene clusters were predicted from the pSC2 and pSCL1 sequences.

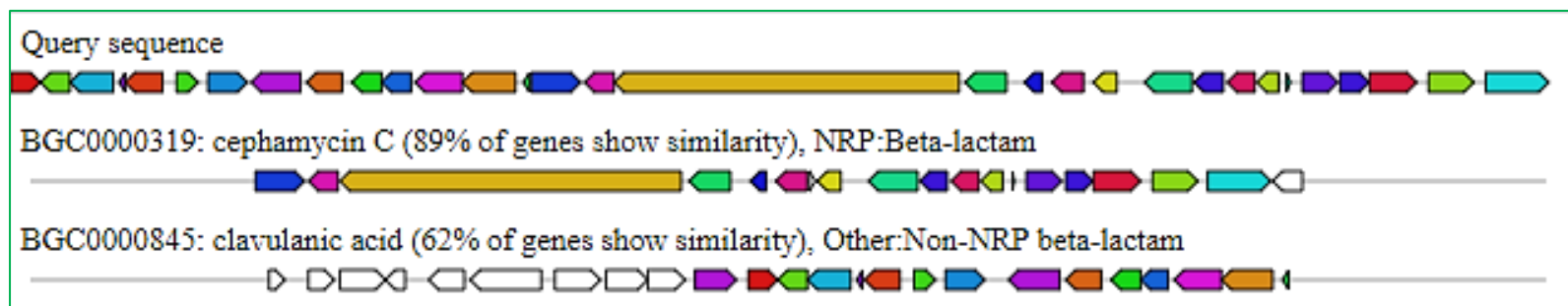


Figure 4.21. Blast of cluster 11 identified on the *S. clavuligerus* DSM 738 chromosome (query sequence) against the known clusters of cephamycin C and clavulanic acid. Obtained from antiSMASH 5.0.

Table 4.7. List of putative biosynthetic gene clusters of the *S. clavuligerus* DSM 738 megaplasmid pSCL4 predicted by antiSMASH 5.0. T1PKS: Type-1 Polyketide synthase, NRPS: Non-ribosomal peptide synthetase.

Cluster	Type	Position in sequence		Most similar cluster	Similarity
		From	To		
1	Terpene	77,916	96,721	(+)-T-muurolol	100%
2	NRPS, terpene	256,145	334,825	(-)- δ -cadinene	100%
3	Lasso peptide	390,959	413,425		
4	Terpene	473,406	492,468	Cyslabdan	18%
5	Indole, NRPS-like, terpene	545,312	599,353	Abyssomicin	9%
6	Terpene, lanthipeptide	644,157	670,992	Venezuelin	100%
7	Terpene	800,625	821,214		
8	Butyrolactone	861,688	872,620		
9	Amglyccycl, terpene, T1PKS, NRPS	1,026,159	1,096,055	Daptomycin	7%
10	β -lactam	1,142,939	1,164,522	Alanylclavam/ 2-hydroxymethylclavam/ 2-formyloxymethylclavam	100%
11	Indole	1,199,889	1,223,506	Staurosporine	82%
12	Bacteriocin	1,229,798	1,241,792		
13	Terpene	1,286,367	1,328,530	Primycin	5%
14	Melanin, T1PKS	1,334,288	1,386,273	Neocarzinostatin	15%
15	NRPS-like, T1PKS, phosphoglycolipid	1,392,562	1,557,852	Maduropetin	30%
16	Terpene, NRPS	1,594,575	1,656,337	Rapamycin	7%
17	NRPS-like	1,668,990	1,713,063	Indigoidine	40%



Figure 4.22. Graphical representation of the *S. clavuligerus* DSM 738 pSCL4 indicating the 17 putative biosynthetic gene clusters and their location on the sequence. Obtained from antiSMASH 5.0.

Table 4.8. List of putative biosynthetic gene clusters of the *S. clavuligerus* DSM 738 plasmid pSCL3 predicted by antiSMASH 5.0.

Cluster	Type	Position in sequence		Most similar cluster	Similarity
		From	To		
1	Terpene, nucleoside	62,948	103,867	Pseudouridimycin	18%



Figure 4.23. Graphical representation of the *S. clavuligerus* DSM 738 pSCL3 indicating the putative biosynthetic gene cluster and its location on the sequence. Obtained from antiSMASH 5.0.

Overall, these results show how the chromosome carries 58% of the biosynthetic gene clusters, including the cephamycin-clavulanic acid supercluster, and the clusters encoding holomycin, naringenin, tunicamycin-like metabolite and a beta-lactam molecule. At the same time, pSCL4, with 40% of the biosynthetic gene clusters, carries genes related to the synthesis of interesting compounds such as venezuelin, staurosporine and another beta-lactam metabolite.

4.8 Whole genome sequence comparison suggest *S. clavuligerus* DSM 738 carries the same chromosome and megaplasmid as the ATCC 27064 Strain.

In order to study the genome sequence synteny between *S. clavuligerus* DSM 738 strain and the other published strains with closed sequences, alignments and dot plots were constructed against the type strain ATCC 27064 (Hwang *et al.*, 2019) and the industrial strains F613-1 and F1D-5 (Cao *et al.*, 2016).

Comparison to the ATCC strain is illustrated on **Figure 4.24**. A Mauve alignment was constructed that shows full consensus of the chromosome and megaplasmid pSCL4 between both strains (**Figure 4.24A**). This is again confirmed on the dot plots constructed individually for each replicon (**Figure 4.24B**). At a nucleotide level, the major difference between the chromosome sequences is the lack of the telomeric sequences on the ATCC strain, more specifically, this strain is missing the first 126 bp and the last 129 bp from the sequence. Beyond the telomeres, only nine inconsistencies are observed between both chromosome sequences, seven of them are single nucleotide substitutions, one is an insertion of 1 nucleotide on the ATCC sequence, and the last variation is the deletion of 6 nucleotides on the ATCC chromosome. All these inconsistencies were checked on our BAM files and it was confirmed that these differences of the chromosome published by Hwang *et al.* do not match our sequencing data and so, they could be due to sequencing errors or polymorphisms between strains. In the case of pSCL4, the ATCC sequence (published as pCLA1 plasmid) is also missing both telomeres, 145 bp and 130 bp at the left and right arm respectively. In addition, 282 nucleotides mismatches were found between sequences, of which 19 were gaps or insertions.

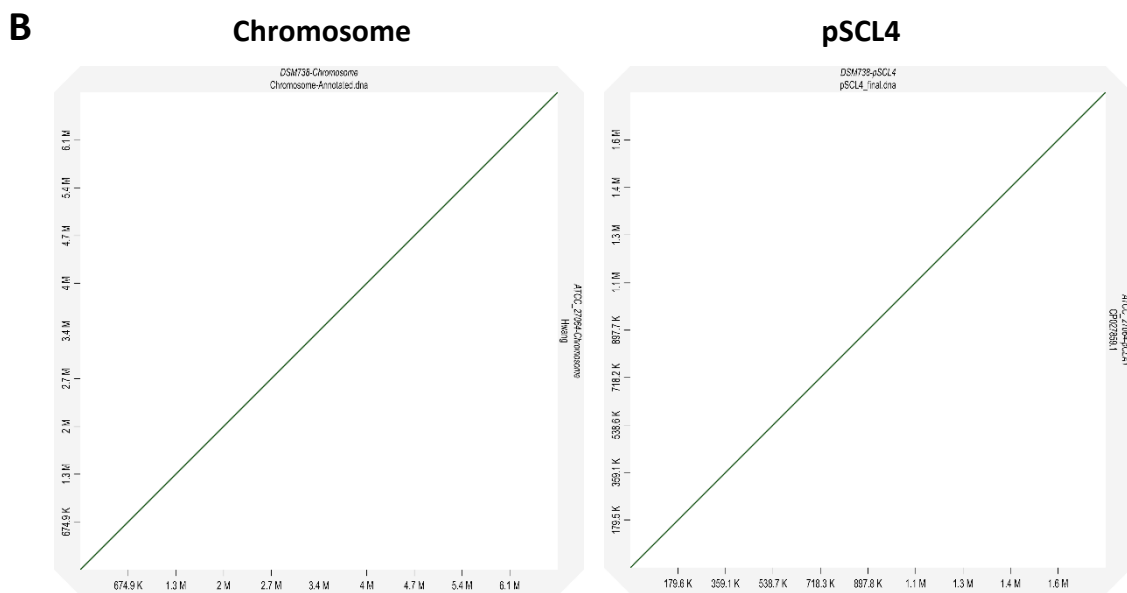


Figure 4.24. Comparison of the genomes of *S. clavuligerus* DSM 738 and ATCC 27064 (Hwang *et al.*, 2019) strains. **A:** Mauve alignment of the genome sequences of DSM 738 and ATCC 27064 strains. **B:** D-GENIES dot plots of chromosome and pSCL4 sequences alignments.

The chromosome and megaplasmid sequences of *S. clavuligerus* DSM 738 were also compared to the industrial strains F613-1 and F1D-5. Unlike the ATCC strain, these strains do not show full consensus to the DSM 738 sequences (**Figure 4.25**). In the case of the chromosome sequences, the major difference is shown on both the alignment and dot plots, which is an insertion of 130 kb on the industrial strains at the right arm. Many other inconsistencies are visualised on the alignments, some of them are shared between the industrial strains which suggest these two strains might belong to the same evolutionary lineage. When comparing the megaplasmid sequences, it is observable that the F613-1 sequence is missing 744 kb and 389 kb at the left and right arm respectively, in reference to DSM 738. The pSCL4 sequence of the F1D-5 strain shows consensus to a combination of material from the DSM 738 and F613-1 sequences, which suggest an incorrect assembly of the F1D-5 megaplasmid sequence.

In summary, these results show that *S. clavuligerus* DSM 738 and ATCC 27064 chromosome sequences are almost identical, while the pSCL4 sequences are highly similar. In contrast the industrial strain F613-1 and F1D-5 show lower similarity to the DSM 738 strain.

4.9 The genetic architecture of *S. clavuligerus* suggests chromosomal gene complementation by plasmids.

In order to study the phylogeny of *S. clavuligerus* DSM 738 in reference to other *Streptomyces* species, a multi-locus sequence analysis tree was constructed and contrasted to the genetic architecture of each genome.

In order to construct the tree, the following 14 *Streptomyces* species were chosen to be compared to *S. clavuligerus*: *S. ambofaciens* ATCC 23877, *S. avermitilis* MA-4680, *S. bingchenggensis* BCW-1, *S. coelicolor* A3(2), *S. collinus* Tu 365, *S. glaucescens* GLA.O, *S. hygrosopicus subsp. jinggangensis* 5008, *S. lavendulae subsp. lavendulae* CCM 239, *S. leeuwenhoekii* C34, *S. lividans* TK24, *S. malaysiensis* DSM 4137, *S. parvulus* 2297, *S. rochei* 7434AN4 and *S. scabiei* 87.22. These strains were selected as they all exhibit complete genome sequences and carry archetypal telomeres. The chromosome sequences of these strains were processed by autoMLST, which compares the sequences of housekeeping genes producing a phylogenetic tree.

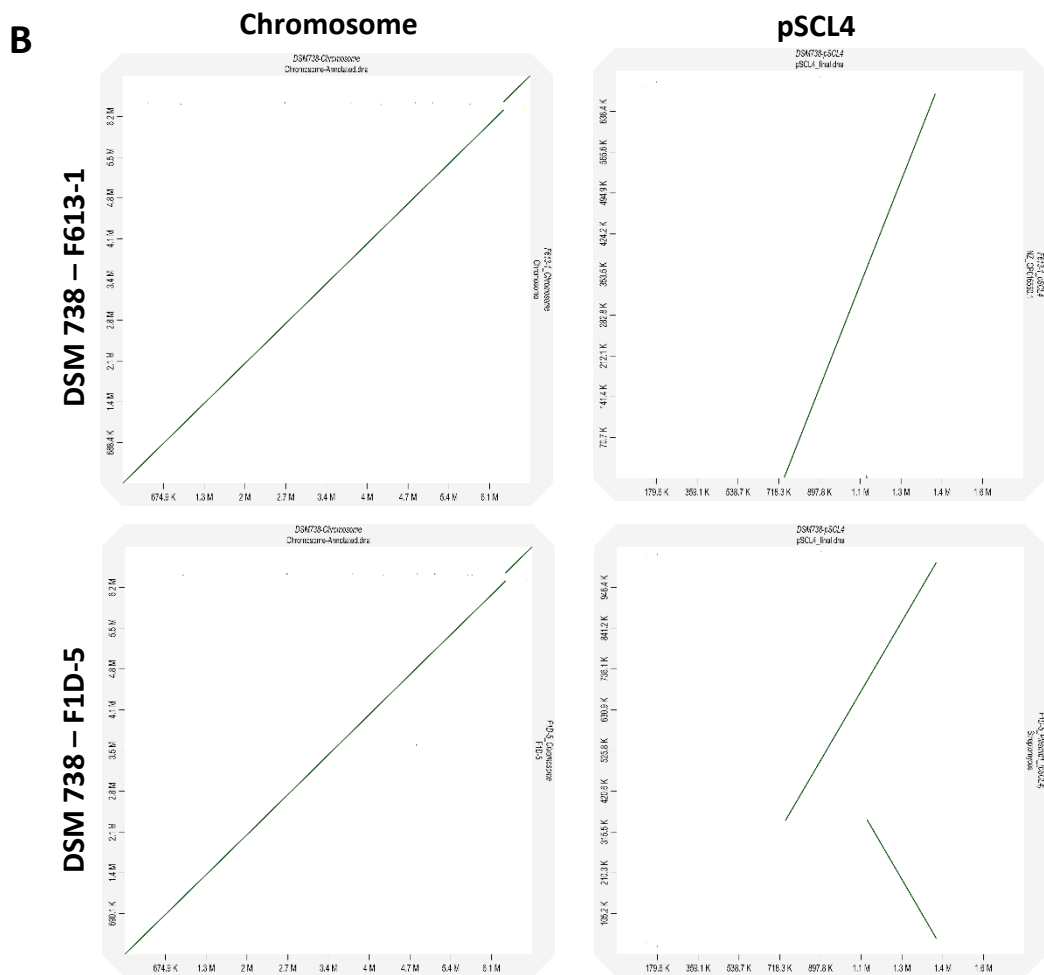
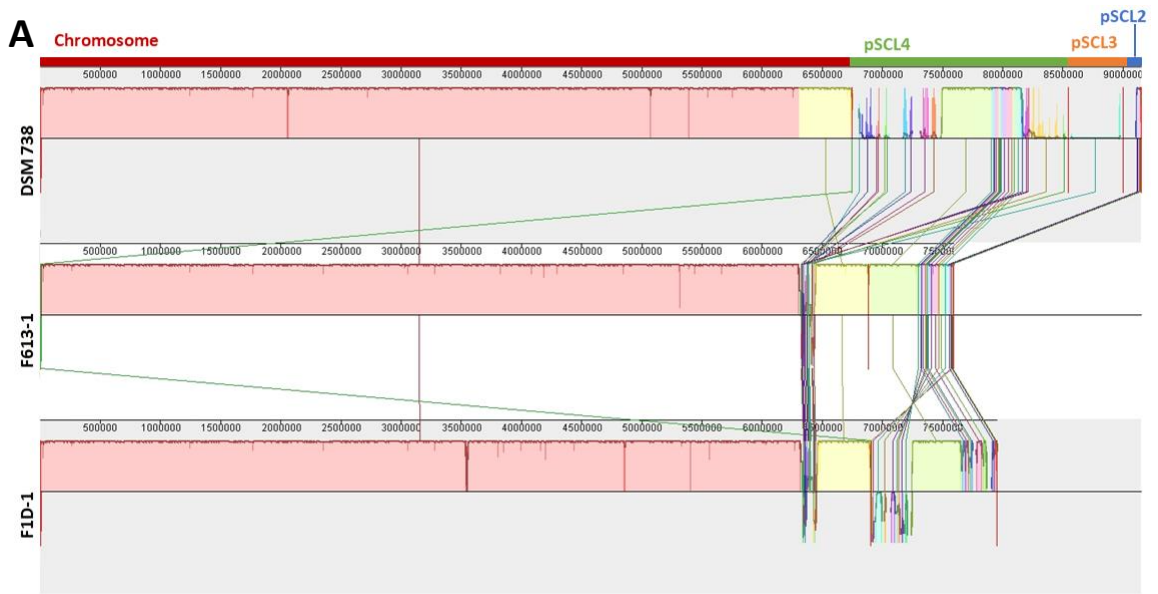


Figure 4.25. Comparison of the genomes of *S. clavuligerus* DSM 738 and the industrial strains F613-1 and F1D-5 strains (Cao *et al.*, 2016). **A:** Mauve alignment of the genome sequences of the three strains. **B:** D-GENIES dot plots of chromosome and pSCL4 sequences alignments.

The obtained tree shows the 15 strains distributed in three main clades (**Figure 4.26**). The chromosome of *S. clavuligerus* is clustered together with *S. lavendulae*, *S. malaysiensis* and *S. bingchenggensis*. Furthermore, in order to compare the genetic architecture of all these species, genome diagrams were created for each strain containing the chromosome and plasmids exhibiting key genes (**Figure 4.26**). Only linear plasmids carrying archetypal telomeres were included, so plasmids such as *S. coelicolor* SCP1 with non-archetypal telomeres were excluded. Seven genes were shown on these diagrams: the housekeeping genes *dnaA* and *recA* (Jakimowicz *et al.*, 1998; Muth *et al.*, 1997), the genes coding the partitioning proteins ParA and ParB (Donczew *et al.*, 2016), the genes coding the terminal proteins Tap and Tpg, and the *ttrA* gene that codes a helicase-like protein thought to be involved in conjugal transfer of linear chromosomes and plasmids (Huang *et al.*, 2003). All the chromosomes exhibited the presence of the genes *dnaA*, *recA*, *parA* and *parB*, showing a similar pattern of the position and orientation of *dnaA* and partitioning genes *parA* and *parB*, while the *recA* is located either upstream (4 strains) or downstream (11 strains) of *dnaA*. The terminal protein-coding genes *tap* and *tpg* are found in all the chromosomes except from *S. clavuligerus*, *S. lavendulae*, *S. bingchenggensis*, *S. rochei* and *S. parvulus*. In addition, all the chromosome sequences present two copies of *ttrA* at the sequence ends, except from *S. lavendulae*, *S. malaysiensis* and *S. glaucescens* that carry one copy, and *S. clavuligerus*, *S. bingchenggensis* and *S. parvulus* chromosomes that carry no copies of *ttrA*. In the case of the plasmids, all of them carry copies of *parA* and *parB* except from *S. clavuligerus* pSCL3 and pSCL1, where only homologs of *parA* were found. Similarly, all the plasmids carried copies of *tap* and *tpg*, except from *S. clavuligerus* pSCL1 that has no copies of *tap* and *tpg* and *S. lividans* SLP2 that shows two copies of *tpg*. In addition, all the plasmids except from pSCL1 carry at least one copy of *ttrA*.

Overall, these results suggest there is little correlation between phylogeny and genetic architecture, as organisms phylogenetically similar show very different genetic architecture. Furthermore, with these results, it is noticeable that some strains lacking important genes on the chromosome, still carry them on plasmids which is the case of *S. clavuligerus*, *S. rochei* and *S. parvulus*, which suggest complementation of these genes by plasmids on these strains. Moreover, the lack of key genes in *S. clavuligerus* chromosome and some of the plasmids suggest complementation between plasmids as well.

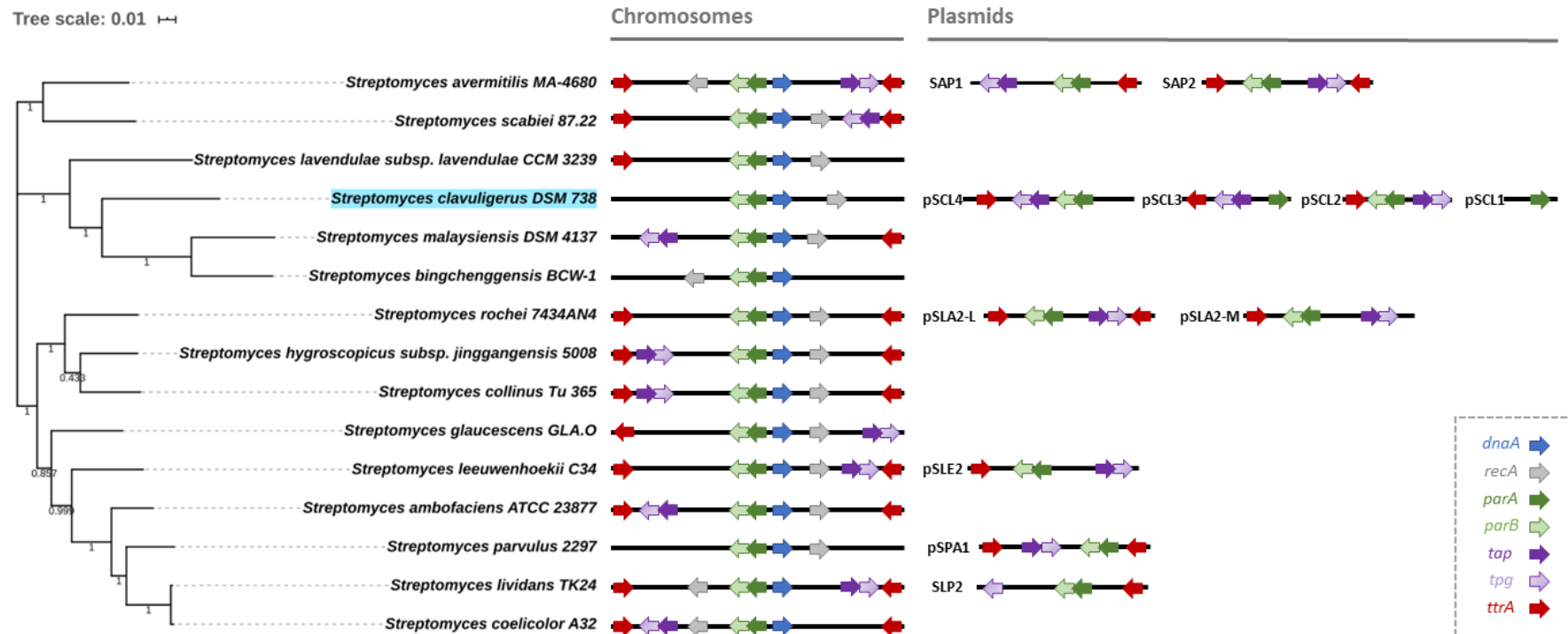


Figure 4.26. AutoMLST tree of 15 different *Streptomyces* strains (left), and their genetic architecture of the chromosome and plasmid sequences. Not to scale.

4.10 Summary

With these results, we report the achievement of the complete genome sequence of *S. clavuligerus* DSM 738. We confirm that combination of Illumina and PacBio sequencing technologies obtains good quality *de novo* assemblies of *Streptomyces* chromosomes and linear plasmids. However, automated complete assembly can still be improved for streptomycetes genomes, as a manual component is still necessary. This is proven particularly essential for terminal inverted repeats, given that common assembly tools fail to differentiate between these large sequence repeats at each end of *Streptomyces* replicons. This was observable on all our replicon assemblies, as one of the ends of the assembled sequence was very close to the end while the other one was missing >100 bp. This is particularly significant on replicons with larger TIRs such as pSCL3, that was lacking 1.5 kb at one end, while replicons with shorter TIRs such as pSCL4 was missing up to 133 bp. This is an important point to take into account when performing *de novo* assembly of streptomycetes genomes, considering TIR length can reach values higher than 1Mb (Weaver *et al.*, 2004) and much material could be ignored.

Our results are consistent with the hypothesis that PacBio sequencing yields higher quality assemblies on actinobacterial genomes (Shin *et al.*, 2013), given that only by utilising PacBio reads, it was possible to obtain a contig with almost the entirety of the *S. clavuligerus* chromosome. In contrast, it was observed on our data that Illumina reads got closer to the sequence ends and obtained a more homogenous read depth than PacBio reads. This evidenced the necessity of combining both technologies for complete genome sequencing of *Streptomyces* strains. However, it was also evidenced the limitation of Illumina sequencing to obtain the last bp of the sequences. This could be due to unavailability of the ends to link to sequence adapters during library preparation, since DNA fragmentation and tagging is done by transposomes, which leaves the ends untagged. This emphasises the importance of utilising physical methods to obtain the sequences ends.

The self-ligation method proved successful for purifying the ends of *S. clavuligerus* chromosome and plasmids. This could be extrapolated to other actinobacterial linear sequences and be considered the simplest most effective method for telomere identification. However, we have identified a few limitations of this methodology. Firstly, it is necessary to have sequence data close to the telomeres, in order to be able to design primers and amplify the unknown material. In addition, restriction sites could be present on the unknown

sequence that would hinder amplification of the telomeres; this was the case we encountered when purifying the ends of other *Streptomyces* species, nevertheless, this allowed extension of sequence and new primers were designed that amplified the telomeres.

Our results confirm the presence of four different types of archetypal telomeres on *S. clavuligerus*, considering the chromosome and megaplasmid carry the same telomere. These data correlate to our Southern blot analysis, where the chromosomal telomere probe hybridised to many bands of different intensities. Having now all the telomeric sequences of the organism, we can identify strong bands corresponding to both ends of the chromosome, and to both ends of the megaplasmid. Fainter bands could represent partial hybridisation of the probe to other telomeres. The fact that chromosome and megaplasmid share the same telomere could be linked to the theory proposed by Medema *et al.*, 2010, that the megaplasmid originated from recombination events of chromosomal material. Moreover, a similar case has been reported for other streptomycetes, such as *S. rochei*, whose chromosome carries the same telomeres as the plasmids pSLA2-L and pSLA2-M (Nindita *et al.*, 2015). In addition, our results indicate that the telomeres of pSCL2 are significantly more similar to phylogenetically distant *Streptomyces* species, which suggest this plasmid could have originated from conjugal transfer from a different species. In general, our observations confirm that *S. clavuligerus* carries a high diversity of telomeres, never seen on the same actinobacterial organism before.

Furthermore, with this work we propose that read coverage depth can be used to estimate the plasmid copy number in reference to the chromosome coverage. Calculation of mean coverage depth values resulted in a copy number of >2 for pSCL2, while a value of <1 for pSCL4 and pSCL3 and copy number of 1 for pSCL1 in reference to the chromosome. These values are relatively low, considering a copy number of 7 per chromosome was estimated for *S. coelicolor* A3(2) plasmid SCP1 (Yamasaki *et al.*, 2003), or 60 for plasmid pSLA2-S in *S. rochei* (Hirochika and Sakaguchi, 1982). However, it is important to mention that these values were calculated by electrophoresis analyses and these two plasmids are smaller than pSCL3 and considerably smaller than pSCL4.

Our sequence comparison analyses confirm that the chromosome sequences of DSM 738 and ATCC 27064 strain are almost identical, and that the megaplasmid sequences of both strains are highly similar. In addition, even though no sequences for pSCL3, pSCL2 and pSCL1 were included on the ATCC 27064 sequence publication by Hwang *et al.*, 2019, evidence of

the presence of these plasmids in the ATCC strain has been published previously (Song et al., 2010; Huang et al., 1998; Wu and Roy, 1993). Since the DSM 738 sequence was obtained independently from the ATCC 27064, we hypothesise that that these *S. clavuligerus* strains are potentially the same strain, and the ATCC genome sequence is simply incomplete.

Genome annotation allowed identification of 8,268 coding sequences, which is the highest number of genes for the *S. clavuligerus* type strain observed to date. Ortholog genes identification within different species of *Streptomyces* resulted in a core genome that represents 40% of the *S. clavuligerus* genome. This correlates with studies that state the core genomes of streptomycetes being the 33%-45% of the whole genome (Zhou et al., 2012). We also identified that all the core genes involved in primary metabolism are encoded on the chromosome, fact that is consistent with the observations made by Medema et al., 2010.

AntiSMASH analyses identified a total number of 43 putative biosynthetic gene clusters on *S. clavuligerus* DSM 738. 25 of these were identified on the chromosome. As expected, these included the cephamycin-clavulanic acid supercluster, as well as other clusters of metabolites previously proven to be produced by *S. clavuligerus*, such as holomycin, tunicamycin and naringenin (Li and Walsh, 2010; Chen et al., 2010; Álvarez-Álvarez et al., 2015). 17 putative biosynthetic gene clusters were predicted from the megaplasmid sequence, these include a β -lactam cluster that was identified by Southern blotting before knowing of the existence of the megaplasmid, and was referred to as a clavulanic acid paralogous cluster, as it shares early genes involved in the biosynthesis of clavulanic acid and so, it probably encodes the production of another β -lactam antibiotic (Tahlan et al., 2004). In addition, several unknown biosynthetic gene clusters were identified as unique to *S. clavuligerus*, that shows a variety of motifs such as NRPS, PKS, lanthipeptides and terpenes, that exposes the metabolic potential of this organism. When comparing these results to the transcriptomics data published by Hwang et al., 2019, we can identify all our predicted clusters on their work, while they split some of the clusters predicted by antiSMASH into individual clusters summing up to 58 clusters, 45 of which were confirmed to be expressed at early exponential growth phase. This transcription levels could be extrapolated to our organism until proven differently.

Characterisation of key genes allowed us to confirm the theory proposed by Medema et al., 2010, that stated a lack of *tap* and *tpg* genes on the *S. clavuligerus* chromosome, while being present on the megaplasmid pSCL4. Moreover, we confirm the presence of *tap* and *tpg* genes

on the plasmids pSCL3 and pSCL2 as well. Since the chromosome must rely on the terminal proteins coded by at least one of these plasmids, we hypothesise that chromosome and megaplasmid utilise the same terminal proteins for telomere replication, given they carry almost identical telomeres. Chromosomal telomeres maintained by plasmid proteins have been reported before in *Streptomyces* species, like the case of *S. rochei* (Nindita *et al.*, 2015). Similarly, the telomeres of pSCL1 must be maintained by the proteins encoded on the other plasmids, like the case of pSLA2-S in *S. rochei*.

Furthermore, we report the absence of the *ttrA* gene from the *S. clavuligerus* chromosome, while copies were found on pSCL4, pSCL3 and pSCL2. The TtrA helicase was shown to be important for chromosome transfer during conjugation in *S. lividans* (Huang *et al.*, 2003), however its specific mechanism is still uncharacterised and complementation of chromosomal TtrA activity by plasmids is yet to be proved.

5. THE TERMINAL PROTEINS OF *S. CLAVULIGERUS*

Terminal proteins (TPs) were named for their role of capping the ends of linear DNA molecules by covalently attaching to the termini. They were first identified in *Bacillus subtilis* phage ϕ 29, where they are necessary for recruiting the DNA polymerase and subsequent full-length phage DNA synthesis by strand displacement (Salas *et al.*, 1978; Salas, 1991). This protein-primed DNA replication by terminal proteins was also observed in adenoviruses (van der Vliet, 1995). In the case of *Streptomyces*, terminal proteins were first identified in the linear plasmids pSLA1 and pSLA2, as it was observed that plasmid DNA exhibited no electrophoretic mobility when omitting the proteinase treatment during plasmid purification (Hirochika and Sakaguchi, 1982). Similarly, terminal proteins attached to the ends of chromosome sequences were first confirmed in *S. lividans* by PFGE of proteinase-untreated DNA (Lin *et al.*, 1993). However, unlike phage ϕ 29 and adenoviruses, DNA replication in *Streptomyces* has an internal origin and extends newly synthesised strands in a bidirectional manner, which implies that the terminal proteins have a different biological role in streptomycetes (Chang and Cohen, 1994; Musialowski *et al.*, 1994). It was then proposed that terminal proteins in *Streptomyces* were involved in replication of telomeres by patching the single stranded 3' end overhang (Qin and Cohen, 1998).

Characterisation of the terminal protein Tpg in *S. rochei* and *S. lividans* was first done by Bao and Cohen, 2001, where they identified a helix-turn-helix motif similar to the DNA-binding thumb of the human immunodeficiency virus reverse transcriptase, which could be responsible for protein-primed DNA synthesis. In addition, Tpg encoded on the *S. coelicolor* chromosome was sequenced and an amphiphilic β -sheets domain was identified, along with the DNA-binding motif, which implied a protein-protein interaction (Yang *et al.*, 2002).

Immediately upstream of *tpg*, another conserved gene was identified, *tap*, whose translation product includes another helix-turn-helix motif. The Tap protein was characterised by Bao and Cohen, 2003, and they reported specific binding of Tap to single-stranded telomeric sequences by electrophoretic mobility shift assay, and more specifically binding to the second and third palindrome of the *S. rochei* pSLA2-S telomere. In addition, they confirmed a Tap-Tpg interaction by immunoprecipitation and yeast two-hybrid analysis. These experiments suggested the role of Tap in interacting with telomeres and recruiting Tpg for subsequent end patching.

The *tap-tpg* operon is usually located toward the end of chromosome and plasmid sequences and its conserved products denominate archetypal terminal proteins. Non-archetypal terminal proteins have been identified, such as the TPs that cap the non-archetypal telomeres of the *S. coelicolor* plasmid SCP1, named Tac and Tpc (Huang, Hsiu H. Tsai, *et al.*, 2007).

In the case of *S. clavuligerus*, the terminal proteins have never been studied, and they have received little mention in the literature, only the fact that the chromosome lacks the *tap* and *tpg* genes (Medema *et al.*, 2010). In this chapter we focus on the characterisation of the terminal proteins of *S. clavuligerus* DSM 738, knock out of the *tap-tpg* operon from pSCL4 and study the interactions between the different sets of proteins.

5.1 Characterisation of three sets of terminal proteins in the genome of *S. clavuligerus*.

Annotation analyses of the complete genome sequence of *S. clavuligerus* DSM 738 allowed identification of three sets of *tap-tpg* operons, located on pSCL4, pSCL3 and pSCL2. These were confirmed by RAST blast of the Tap and Tpg amino acid sequences from the *S. coelicolor* chromosome against the whole *S. clavuligerus* genome (**Figure 5.1A**). The *tap-tpg* operon is located on the central part of pSCL4, while in the case of pSCL3 and pSCL2, it is positioned towards the end of the replicons (**Figure 5.1B**). Comparison to other *Streptomyces* chromosomes and plasmids revealed that pSCL4 is the only case of a centrally located *tap* and *tpg* in complete chromosome and plasmid sequences.

In order to characterise the amino acid sequences of these three TP sets in *S. clavuligerus*, alignments were constructed using the *S. coelicolor* chromosome sequences as reference. The alignment of the Tap amino acid sequences is illustrated on **Figure 5.2**. This alignment exhibits many regions of high conservation between the four sequences, this is particularly visible among the first 86 aa, and the last 574 aa in reference to the *S. coelicolor* Tap sequence. The Tap sequence from pSCL2 (2Tap) shows a 74% identity to the *S. coelicolor* sequence, while the Taps from pSCL4 (4Tap) and pSCL3 (3Tap) exhibit 48 and 44% identity to the reference respectively. Furthermore, the shared identity among the Tap sequences of the three plasmids is between 42% and 50%. The helix-turn-helix motif observed in archetypal Tap sequences was identified on our alignment, this region presents a residue

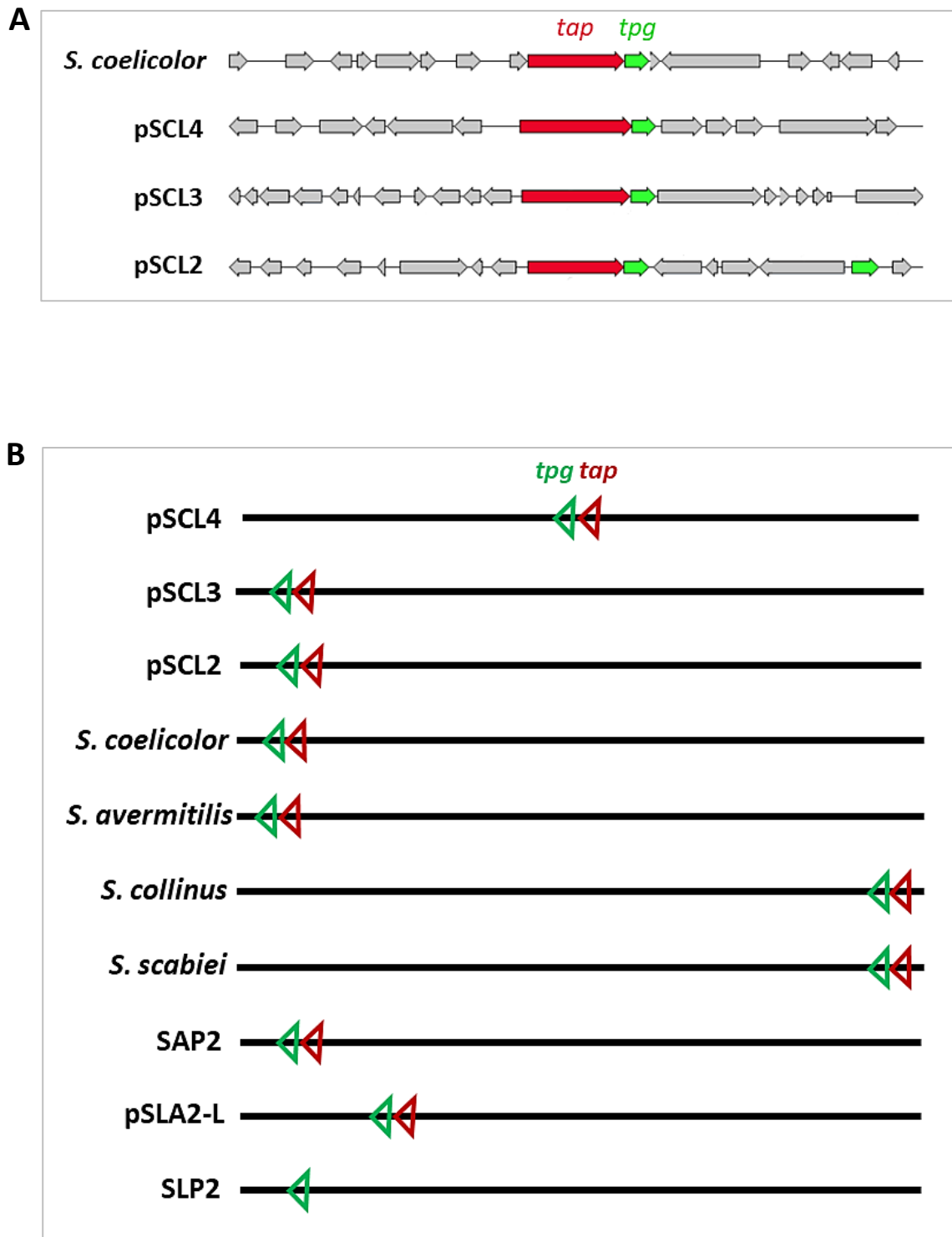


Figure 5.1. The three *tap-tpg* operons in *S. clavuligerus*. **A:** *S. coelicolor* chromosome *tap* and *tpg* homologs found on pSCL4, pSCL3 and pSCL2. Graphs obtained from RAST. **B:** Graphical representation of the location of the *tap-tpg* operon on *Streptomyces* chromosomes and plasmids. Not to scale.

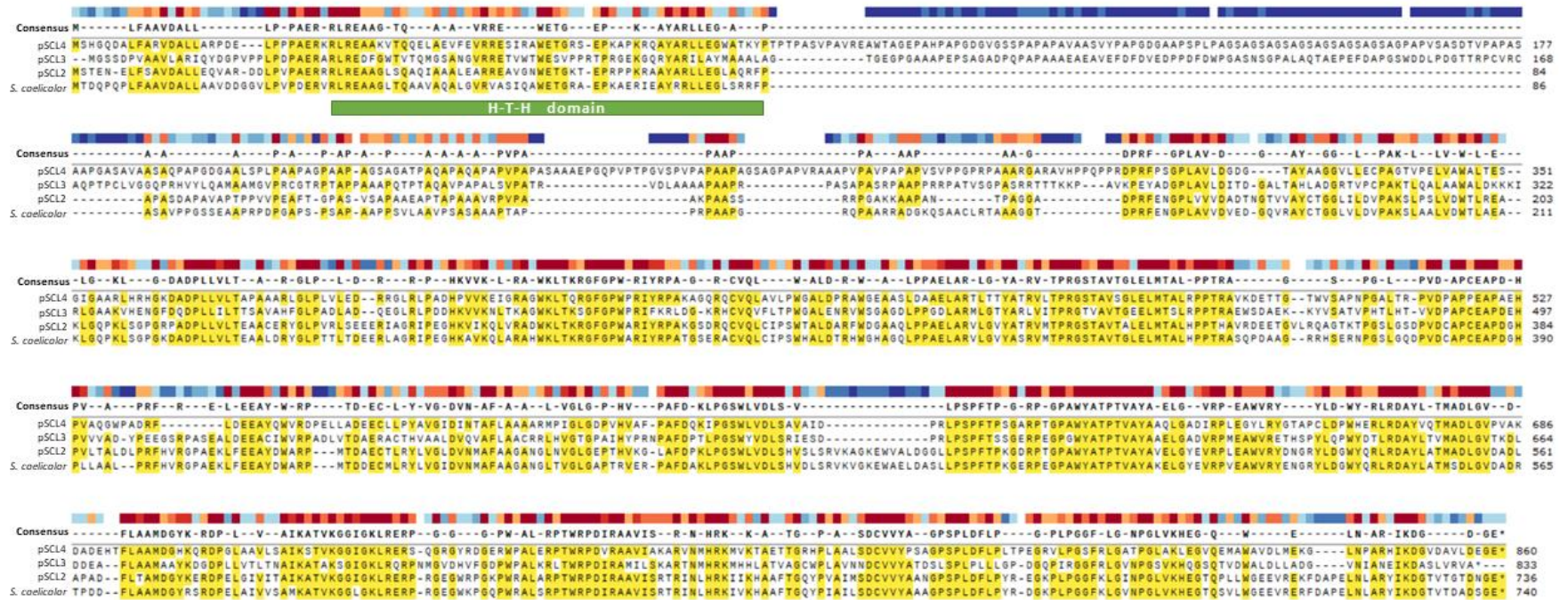


Figure 5.2. Alignment of Tap amino acid sequences from pSCL4, pSCL3, pSCL2 and *S. coelicolor* chromosome. Consensus highlighted in yellow and conservation indicated in coloured blocks (red: high, blue: low). The helix-turn-helix domain is indicated by a green box.

conservation of 21% among the four sequences, while this value increases to 41% when looking at the conservation between pSCL4, pSCL2 and *S. coelicolor* Tap sequences alone. In addition, the Tap amino acid sequences were analysed through secondary structure prediction servers (InterPro and ExPASy) and the same helix-turn-helix motif was identified on the pSCL4, pSCL2 and *S. coelicolor* Tap proteins. In contrast, no helix-turn-helix domain was predicted for the pSCL3 Tap sequence.

Another alignment was constructed using the four Tpg amino acid sequences (**Figure 5.3**). This alignment shows a high level of conservation between the four sequences. Similar to the Tap proteins, the pSCL2 Tpg (2Tpg) shows higher identity to the *S. coelicolor* sequence (72%), than the pSCL4 (4Tpg) and pSCL3 (3Tpg) proteins, which present a shared identity of 50% and 43% to *S. coelicolor* respectively. In addition, Tpgs of the three plasmids share an identity lower than 50% between each other. The helix-turn-helix (H-T-H) motif was identified on the alignment, based on literature, including the HIV reverse transcriptase motif. This H-T-H domain shows conserved regions among the four sequences, however, the 3Tpg carries 6 extra residues in the central section of the motif. The amphiphilic β -sheet is also identified on our alignment, where high conservation is observable among the four sequences. Furthermore, motif prediction servers identified the helix-turn-helix domain in the 2Tpg sequence, which extended the motif 16 residues further. In contrast, no domains were predicted for the 3Tpg, 4Tpg and *S. coelicolor* Tpg protein sequences.

Moreover, in order to compare the amino acid sequences of the Tap and Tpg proteins from *S. clavuligerus* plasmids to other *Streptomyces* archetypal terminal proteins, we constructed a maximum likelihood phylogenetic tree per protein using protein sequences from 16 streptomycete chromosomes and 13 plasmids. The obtained trees are presented on **Figure 5.4**. The Tap proteins tree is divided in three main branches. The three *S. clavuligerus* Tap proteins are located on each of these main branches. The 3Tap is clustered together with the protein of *S. lavendulae* plasmid, while 4Tap is located on a clade carrying Taps from *S. cineauber*, *S. dengpaensis* and *S. anulatus*. On the other hand, 2Tap shows higher similarity to proteins from *Streptomyces sp.* FR1, *S. coelicolor* and *S. lividans*. In the case of the Tpg proteins tree, it is also divided in three main clades, however, the 3Tpg and 4Tpg proteins are clustered on the same branch this time, while the 4Tpg protein is still more similar to Tpgs from *S. cineauber*, *S. dengpaensis* and *S. anulatus*. Similarly to the Tap proteins tree, 2Tpg i

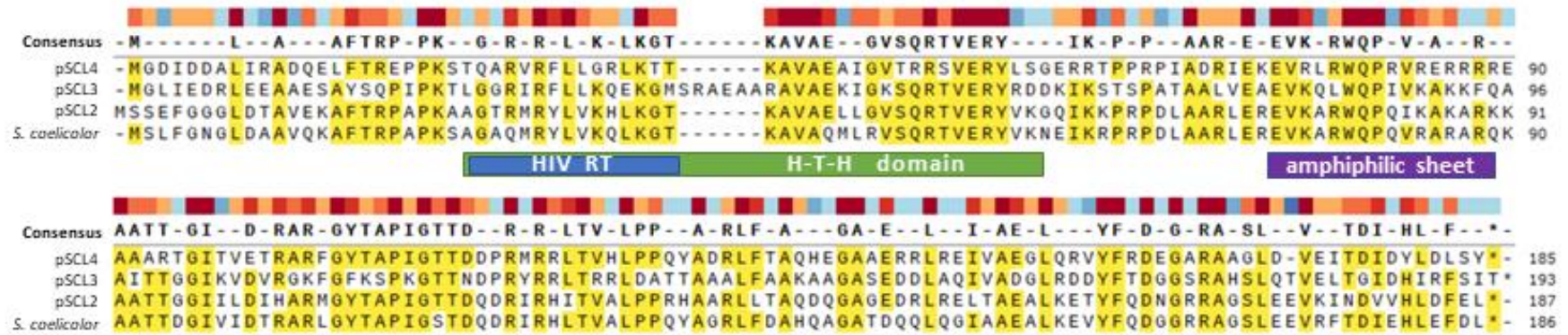


Figure 5.3. Alignment of Tpg amino acid sequences from pSCL4, pSCL3, pSCL2 and *S. coelicolor* chromosome. Consensus highlighted in yellow and conservation indicated in coloured blocks (red: high, blue: low). The helix-turn-helix domain is indicated by a green box, including the human immunodeficiency virus reverse transcriptase motif in blue. The amphiphilic β -sheet is indicated by a purple box.

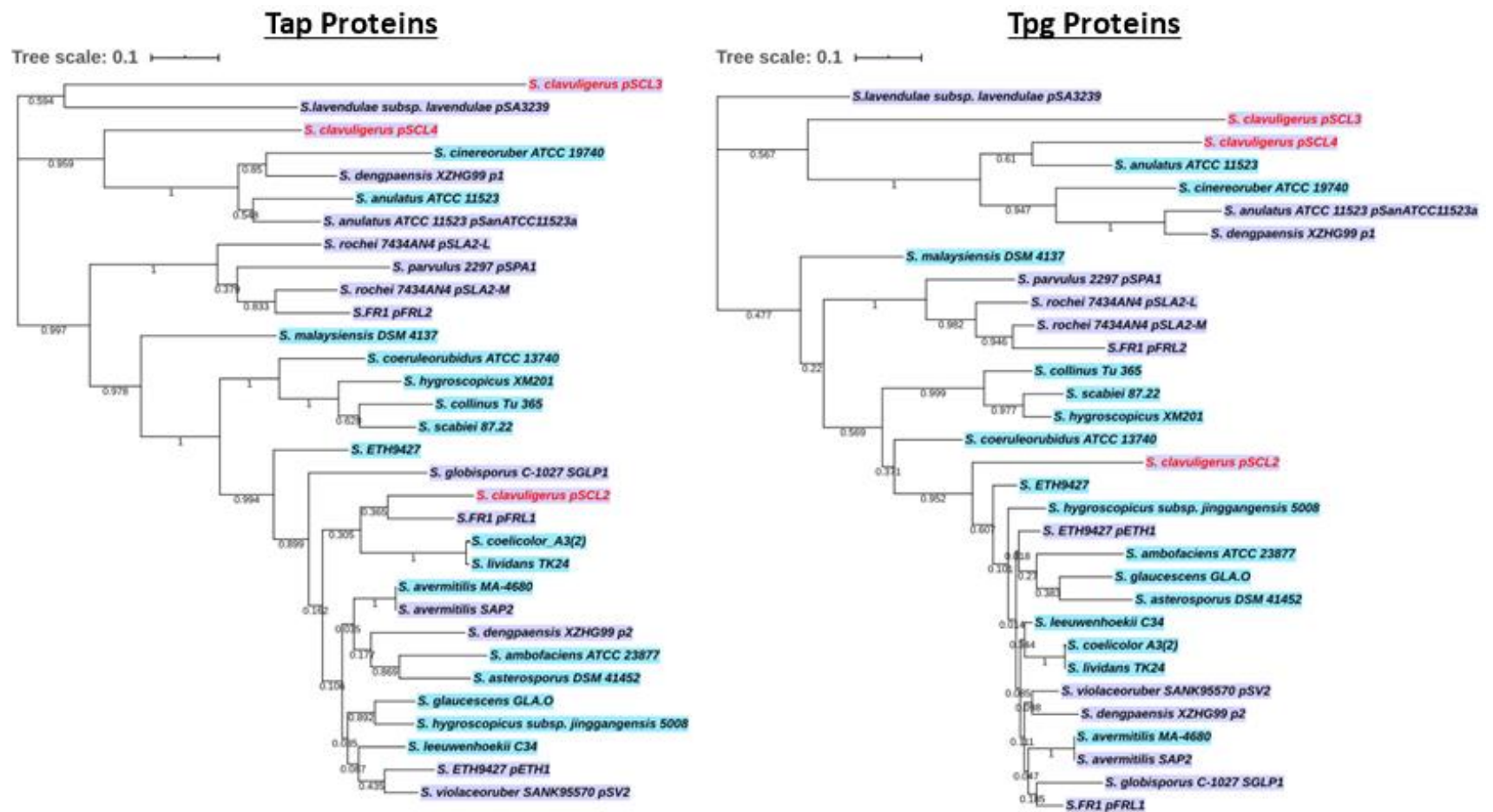


Figure 5.4. Maximum likelihood phylogenetic trees of archetypal Tap and Tpg amino acid sequences of different *Streptomyces* species. Chromosomal proteins are highlighted in blue and plasmid proteins in purple. *S. clavuligerus* proteins are indicated by red font. Tree constructed on MEGA with 1000 bootstraps and edited on iTOL.

more similar to proteins of organisms phylogenetically distant to *S. clavuligerus*, such as *S. coelicolor*, *S. avermitilis* or *S. hygroscopicus*. Moreover, both the Tap and Tpg phylogenetic trees exhibit a similar clustering pattern.

In conclusion, the high conservation of the three sets of Tap and Tpg in *S. clavuligerus* to those encoded on the *S. coelicolor* chromosome suggests that these plasmid proteins are required for end-patching of archetypal telomeres. However, the maximum likelihood trees exhibit substantial phylogenetic distance between the proteins of the three plasmids.

5.2 A copy of pSCL4 *tap-tpg* operon was incorporated on *S. clavuligerus* chromosome.

In order to study the role of the Tap in Tpg sets in *S. clavuligerus*, we intend to perform analyses including *tap-tpg* knockouts and plasmid curing. In order to provide the chromosome with a copy of these genes prior to knockout experiments, we made pSCL4 *tap-tpg* duplicated strains of *S. clavuligerus*. For this, we constructed the integrative plasmid pLIS22 which carries the *tap-tpg* operon from pSCL4. This plasmid is based on the pMS82 vector, which carries a hygromycin-resistance gene, the bacteriophage ϕ BT1 integrase gene and an *attP* site that recognises the *attB* site on *Streptomyces* chromosomes. The plasmid was constructed by restriction cloning and introduced into *S. clavuligerus* strains by intergeneric conjugation.

Firstly, the plasmid was constructed *in silico* using SnapGene (**Figure 5.5A**). The forward primer PTP-F-HindIII was designed to start annealing to the 304 bp upstream of *4tap*, and a *HindIII* restriction site was included. The reverse primer PTP-R-NsiI was designed to start annealing to the 51 bp downstream of *4tpg*, and a *NsiI* restriction site was included on the sequence. The pSCL4 *tap-tpg* operon (3.5 kb) was amplified and cloned into pMS82 using the restriction enzymes *HindIII* and *NsiI*.

Wet lab construction of the plasmid was initiated by PCR amplification of the operon from genomic DNA of *S. clavuligerus* DSM 738 using the previously mentioned primers. The amplified product was confirmed by visualisation of a band of 3.5 kb by gel electrophoresis analysis (**Figure A4** in Appendix). The amplified and purified genes were digested with *HindIII* and *NsiI* restriction enzymes alongside the vector pMS82, and the products were ligated. *E. coli* competent cells were transformed with ligation products and hygromycin-resistant

transformants were analysed by colony PCR. Four transformants exhibited bands at 3.5 kb on gels (**Figure A5** in Appendix), which were then grown for plasmid isolation. In order to confirm the correct construction of pLIS22, the purified plasmids were digested with *Bam*HI, *Hind*III, *Kpn*I and *Eco*RI and the products were separated by electrophoresis. The gel shows bands at the expected sizes and therefore confirms the construction of pLIS22 (**Figure 5.5B**).

In order to introduce pLIS22 into *S. clavuligerus*, the plasmid was first used to transform the *E. coli* ET12567/pUZ8002 conjugative strain, and hygromycin-resistant colonies were picked and grown for conjugation. In addition, conjugative *E. coli* was transformed with the empty vector pMS82 to use as a control. *E. coli* strains containing either pLIS22 or pMS82 were conjugated into the *S. clavuligerus* strains DSM 738, DSM 41826 and Sc2. After applying selection for hygromycin and nalidixic acid resistance, transconjugants appeared, which were then streaked one more time on plates containing hygromycin and nalidixic acid. Both pLIS22 and pMS82 transconjugants exhibited growth on these antibiotics, which suggested a successful conjugation (**Figure A6** in Appendix). Transconjugant strains were named SCLA22 (*S. clavuligerus* DSM 738 plus pLIS22), SCLA23 (*S. clavuligerus* DSM 41826 plus pLIS22), and SCLA24 (Sc2 plus pLIS22).

Furthermore, in order to confirm the presence of an additional copy of the pSCL4 *tap*-*tpg* operon on the genome, we performed a Southern blot analysis of the *tap*-*tpg*-duplicated strains against the parental strains. For this, purified genomic DNA of DSM 738, DSM 41826, Sc2, SCLA22, SCLA23 and SCLA24 was digested with the restriction enzyme *Sac*I. This enzyme was chosen because it does not cut within the *tap*-*tpg* operon, and so, one digested fragment contains the whole operon, thus one band represents one *tap*-*tpg* copy. The digested DNA was separated by electrophoresis and the gel was blotted and hybridised to a probe that binds to pSCL4 *tap*-*tpg*. The Southern blot shows one band at around 4.5 kb for the parental strains, which represents the copy on pSCL4 (**Figure 5.6A**). The *tap*-*tpg*-duplicated strains present the same band at 4.5 kb, plus an additional band at around 10 - 15 kb that represents the duplicated copy in the chromosome (**Figure 5.6B**).

With these results we can confirm the incorporation of pSCL4 *tap* and *tpg* on the chromosome of the strains of *S. clavuligerus* DSM 738, DSM 41826 and Sc2.

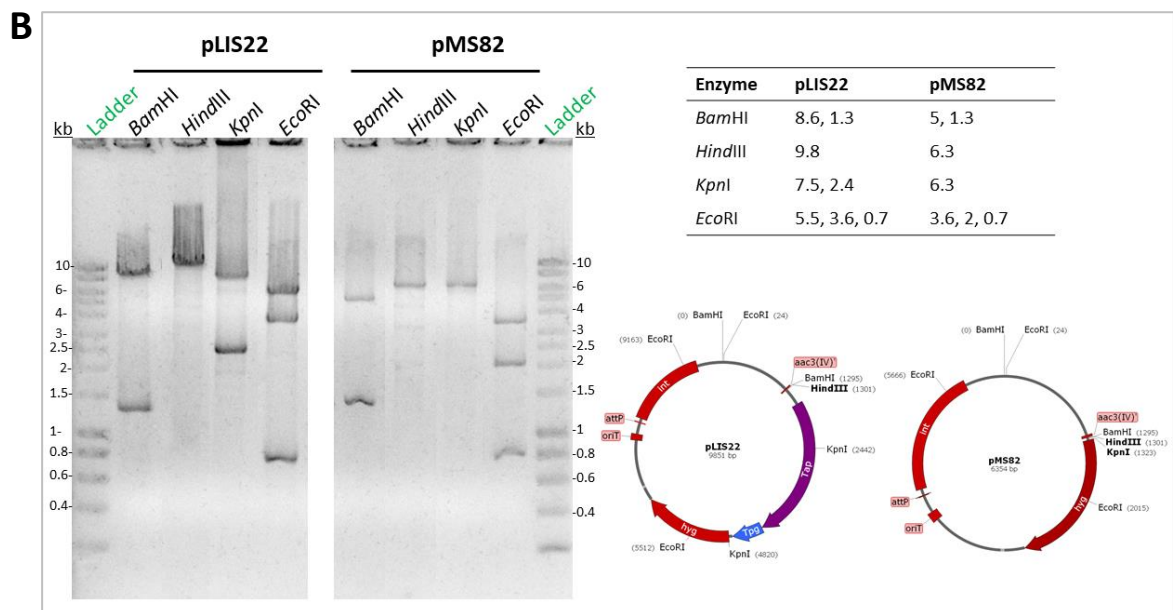
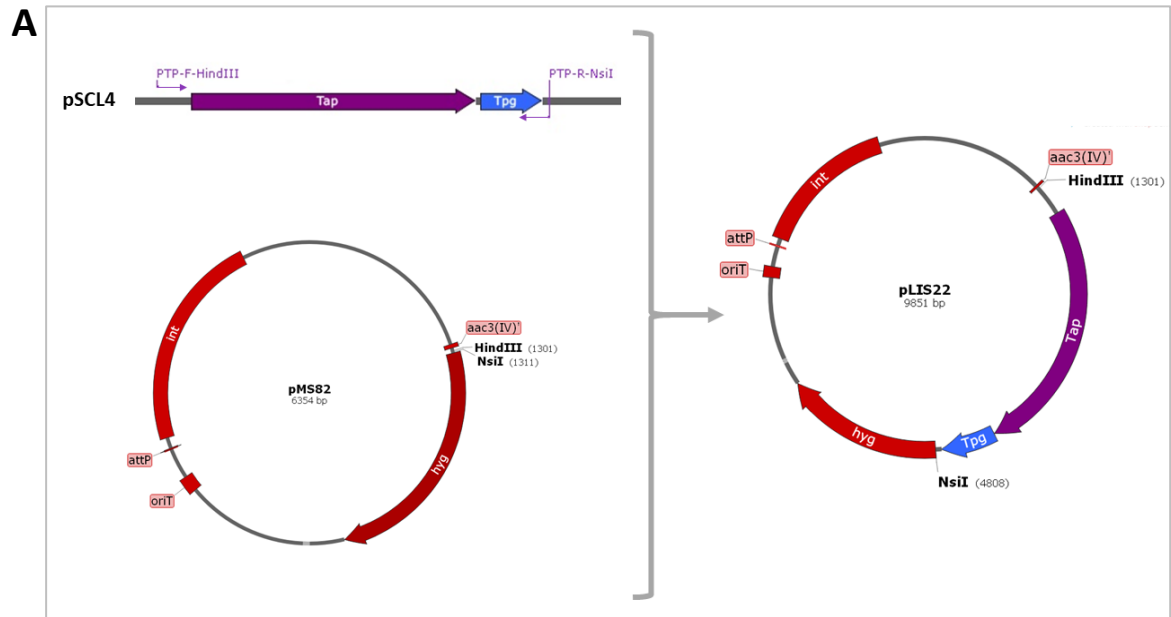


Figure 5.5. Construction of integrative plasmid pLIS22 **A:** *In silico* cloning of the *tap-tpg* operon into pMS82 **B:** Gel electrophoresis of the digest products of pLIS22 and pMS82 with *Bam*HI, *Hind*III, *Kpn*I and *Eco*RI. Including a table with the expected product sizes and the plasmid maps indicating the restriction sites.

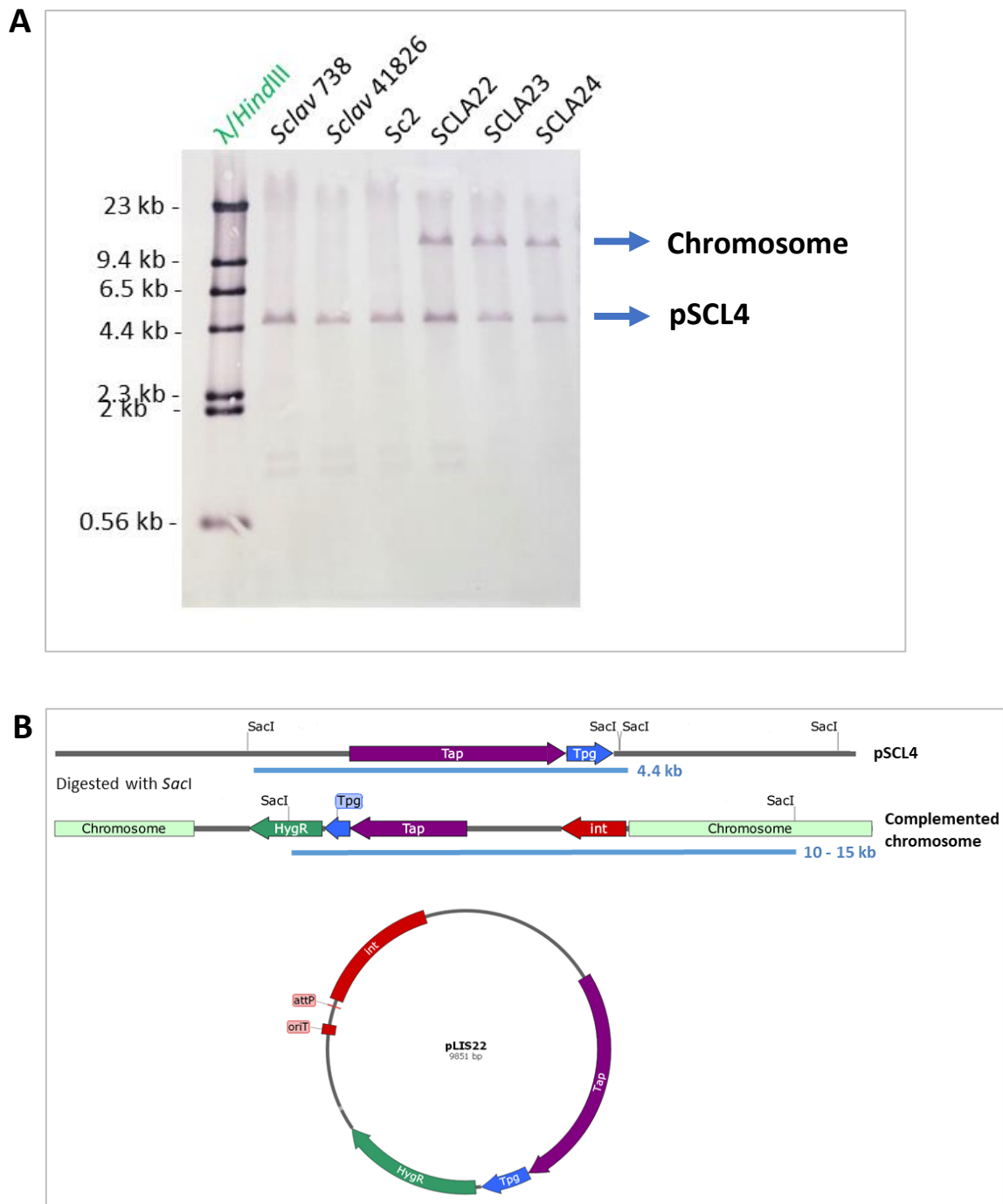


Figure 5.6. Southern blot analyses confirm an extra copy of *tap-tpg* on duplicated strains. **A:** Southern blot performed with pSCL4 *tap-tpg* as probe on digested DNA of the three *tap-tpg*-duplicated strains and their parental strains. Lambda DNA digested with *Hind*III as DNA size standard. **B:** Maps of the regions of pSCL4 and *tap-tpg*-duplicated chromosome (estimated) containing the *tap-tpg* operon, indicating *Sac*I restriction sites, and the size of expected bands.

5.3 Homologous recombination did not allow elimination of *tap* and *tpg* from pSCL4.

The fact that the chromosome lacks its own copies of *tap* and *tpg*, led to the hypothesis that it utilises the proteins encoded on the plasmids. Our theory was that the chromosome relies on the terminal proteins provided by the megaplasmid pSCL4. In order to confirm this theory and further study the role of these proteins on the megaplasmid, we intended to eliminate the *tap-tpg* genes from pSCL4 by allelic replacement (homologous recombination of the operon's flanking regions and substitute the genes by an antibiotic-resistance gene). The plan for knocking out *tap* and *tpg* is illustrated on **Figure 5.7**. A knockout plasmid called pLIS405, which carries the pWHM3 vector backbone (thiostrepton-resistant) and the 3kb flanking regions at each side of *tap-tpg* separated by the apramycin-resistance gene, was utilised for the recombination. For total elimination of *tap* and *tpg*, two recombination events are necessary. The first cross-over would incorporate the pLIS405 plasmid on pSCL4, providing resistance to both apramycin and thiostrepton. The second cross-over would eliminate *tap*, *tpg* and the thiostrepton-resistance gene, thus colonies should only be resistant to apramycin.

In silico construction of the knockout plasmid pLIS405 was performed using SnapGene (**Figure 5.8**). Firstly, primers were designed in order to amplify the flanking regions of the *tap-tpg* operon in pSCL4. The downstream flanking region (referred here as SP-Left) of 2,993 bp of size is amplified using the primers SP-L-SphIXbaI and SP-L-BamHI, which include the *SphI* and *XbaI*, and the *BamHI* restriction sites respectively. The upstream flanking region (referred here as SP-Right) of 3,189 bp of size is amplified using the primers SP-R-BamHI and SP-R-EcoRIXbaI, which include the *BamHI*, and the *EcoRI* and *XbaI* restriction sites respectively. The flanking regions were cloned individually into pGEM-T Easy vector. SP-Right was subcloned into pBGS19 constructing the plasmid pLIS401 using *BamHI* and *EcoRI*. SP-Left was then cloned onto pLIS401 using *BamHI* and *SphI*, forming the new plasmid pLIS402. The apramycin-resistance gene was amplified from pIJ773 and cloned between SP-Right and SP-Left using *BamHI* constructing pLIS404. The plasmid pLIS405 would then result from the cloning of the knockout cassette to pWHM3. Due to the lack of the DNA sequence of pWHM3, we estimated the map of pLIS405 utilising published maps of pWHM3.

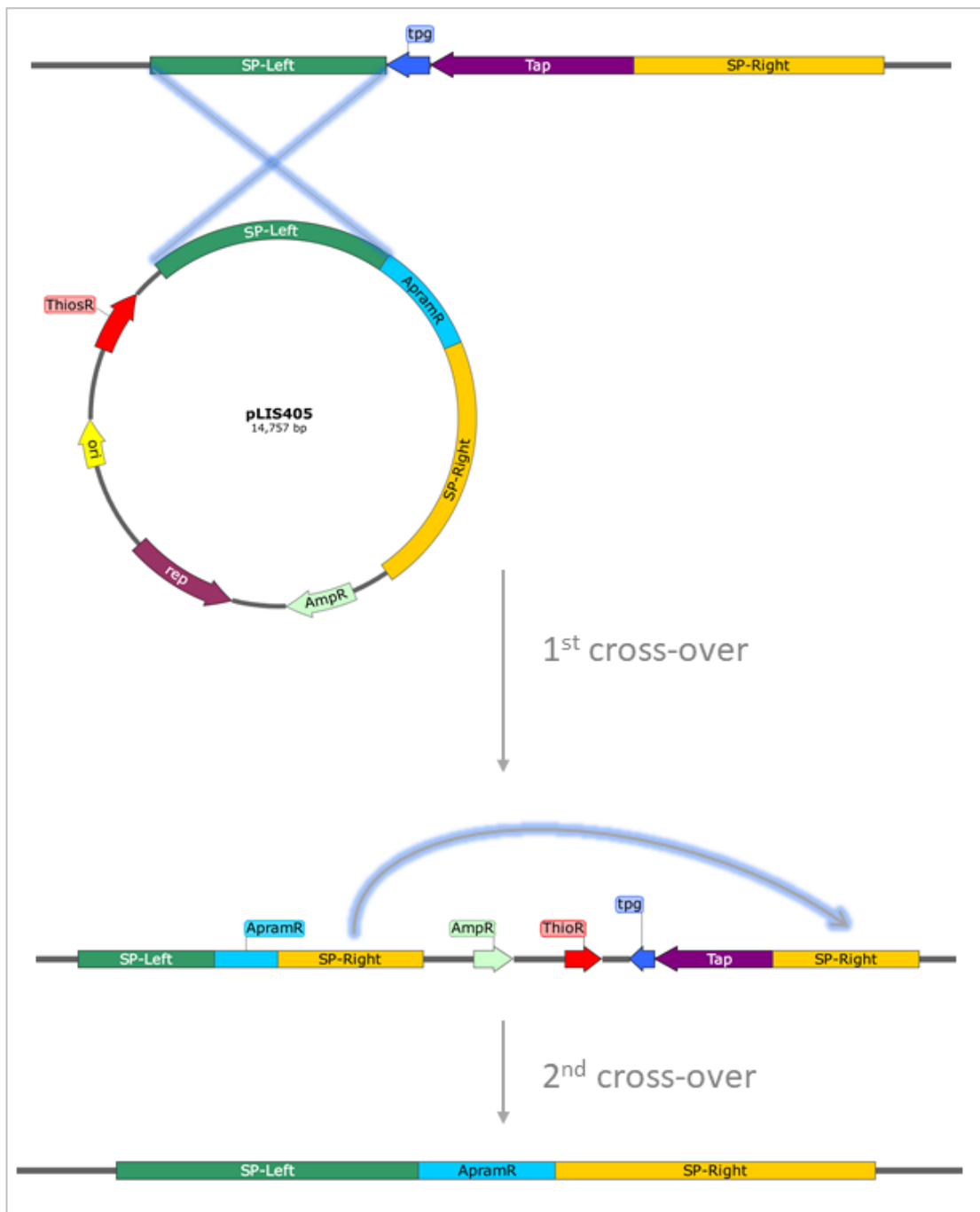


Figure 5.7. Schematic representation of the plan for eliminating the *tap* and *tpg* genes from pSCL4 by homologous recombination of the operon's flanking regions and substitution by apramycin-resistance gene. Two recombination events (presented in blue) against the plasmid pLIS405 are necessary.

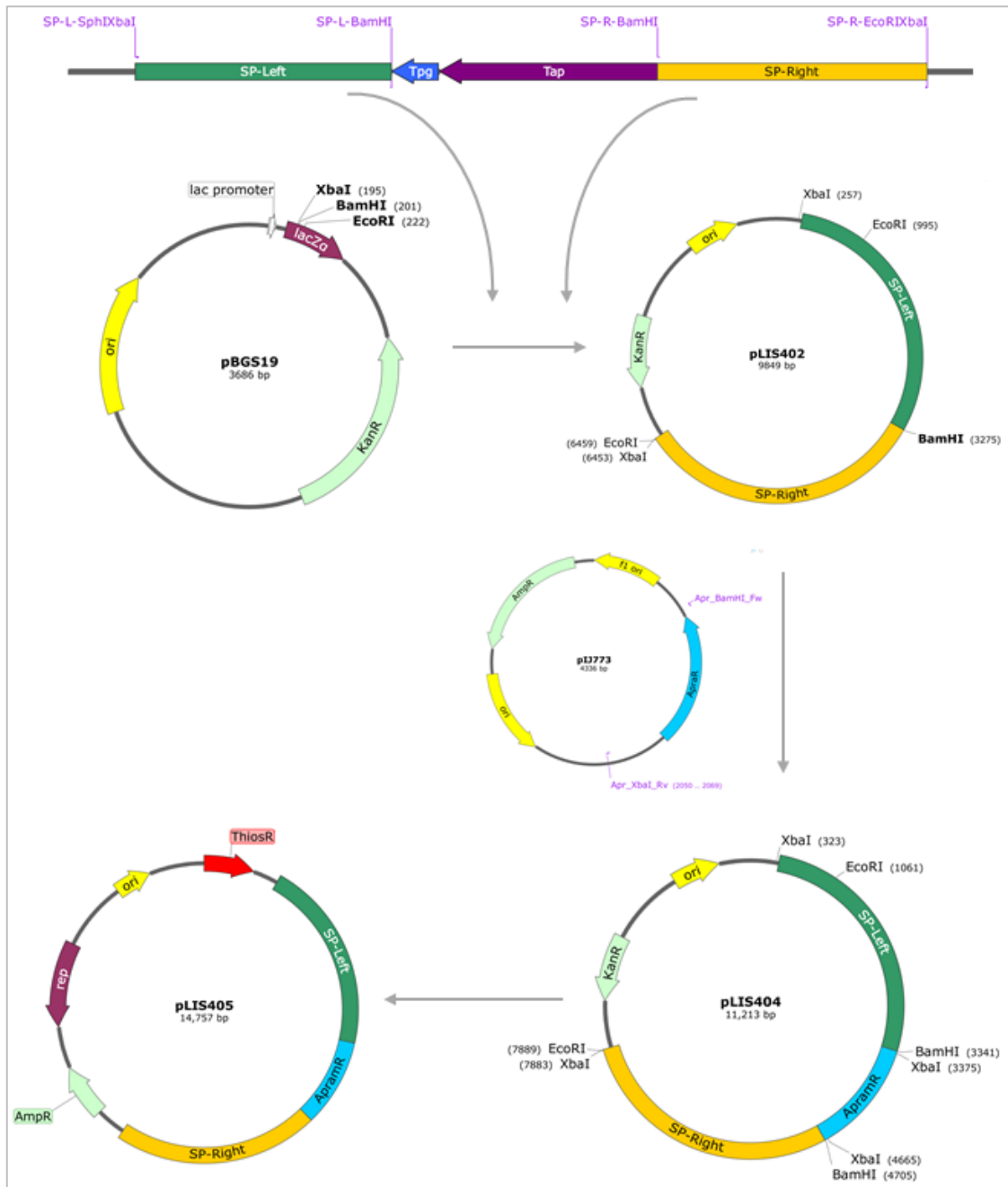


Figure 5.8. Schematic representation of the *in silico* cloning to construct pLIS405. The pSCL4 *tap-tpg* flanking regions are cloned onto pBGS19. The apramycin-resistance gene is amplified from pIJ773 and cloned between the flanking regions. The knockout cassette is finally subcloned onto pWHM3.

Wet lab construction of pLIS405 was initiated by PCR amplification of SP-Right and SP-Left, which was confirmed by visualisation of bands at around 3 kb on agarose gels (**Figure A7** in Appendix). Purified amplified products were ligated to linearised pGEM-T Easy vector individually. Ligation products were used to transform *E. coli* competent cells and ampicillin-resistant white colonies were picked and grown for plasmid purification. Confirmation of correct cloning was done by digestion of the pGEM-SP-Left or SP-Right plasmids with restriction enzymes (**Figure A8** in Appendix). In order to construct the plasmid pLIS401, pGEM-SP-Right and pBGS19 were digested independently with *Bam*HI and *Eco*RI by sequential digestion with a clean-up step in between. Following purification, the digested products were ligated and used to transform competent *E. coli* cells. Kanamycin-resistant colonies were selected, and the presence of the insert was confirmed by colony PCR. Correct construction of the plasmid was confirmed by restriction digest with *Bam*HI and *Sal*I (**Figure A9** in Appendix). Exactly the same procedure was followed to clone SP-Left into pLIS401 utilising this time the enzymes *Bam*HI and *Sph*I, resulting in the plasmid pLIS402. Furthermore, the apramycin resistance gene was amplified from DNA of pIJ773 using the [Apr_BamHI_Fw](#) and [Apr_BamHI_Rv](#) primers, and then cloned into pGEM-T Easy. This gene was then subcloned into pLIS402 using *Bam*HI, resulting in the plasmid pLIS404, which construction was confirmed by restriction digestion (**Figure A10** in Appendix). Finally, the knockout cassette was subcloned into pWHM3, for this pLIS404 was digested with *Pvu*II and *Hind*III, while pWHM3 with *Eco*ICRI and *Hind*III. Following ligation and transformation, we obtained the knockout plasmid pLIS405. The insertion of the cassette in pWHM3 was confirmed by restriction digest with *Hind*III, *Bam*HI, *Eco*RI and *Sal*I (**Figure 5.9**).

In order to introduce the knockout plasmid into *S. clavuligerus*, the plasmid was first used to transform competent *E. coli* ET12567/pUZ8002. Thiostrepton and apramycin resistant colonies were selected and grown to conjugate the wild type strain *S. clavuligerus* DSM 738 and the *tap-tpg*-duplicated strain SCLA22. After applying selection with nalidixic acid and apramycin, a total number of 7 transconjugants were obtained, 6 from conjugation of the wild type (called SCLA405-1 to SCLA405-6), and 1 colony from the conjugation of the *tap-tpg*-duplicated strain (called SCLA22-405). These 7 strains were then streaked in the presence of apramycin and thiostrepton separately, and they all exhibited growth on plates containing both antibiotics, which implied that one cross-over had occurred integrating the whole pLIS405 plasmid on the megaplasmid (**Figure 5.10A**).

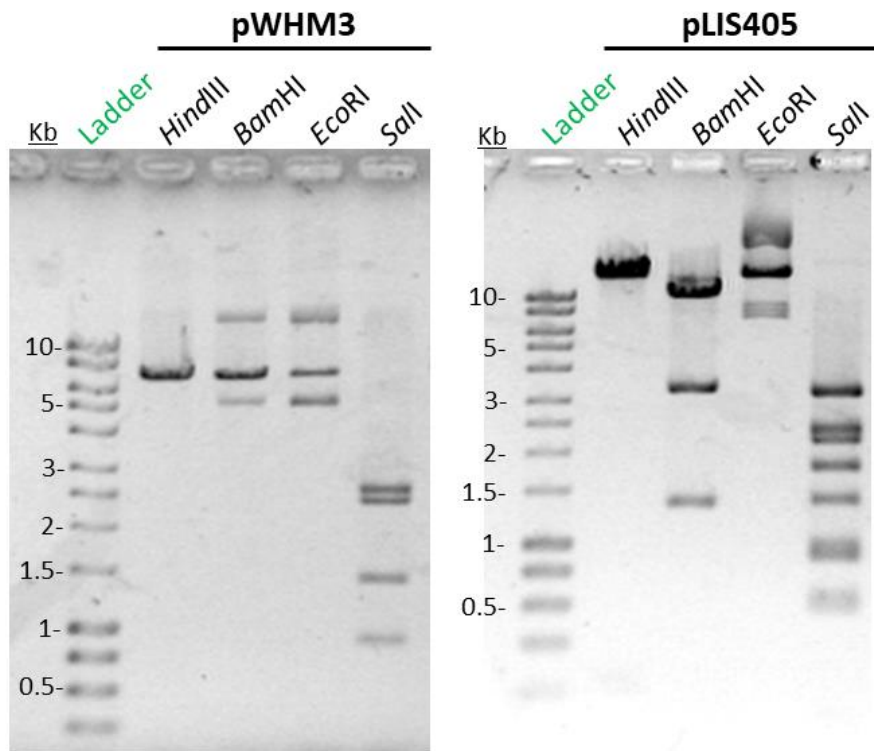


Figure 5.9. Confirmation of cloning of the knockout cassette into pWHM3 resulting in pLIS405. Gel electrophoresis of restriction digest products of pLIS405 and empty vector pWHM3 digested with *HindIII*, *BamHI*, *EcoRI* and *Sall*.

In order to allow a second cross-over to occur and eliminate *tap* and *tpg* from pSCL4, the transconjugant strain were grown without antibiotic and spores were harvested. These spores were diluted 1:50, 1:2500 and 1:25000 and were plated on agar without antibiotics.

For the purpose of screening for colonies where a second recombination event happened, single colonies were picked and patched on plates containing apramycin, thiostrepton and no antibiotics separately, so each colony was tested against the three conditions (**Figure 10.B**). In addition, the wild type strain *S. clavuligerus* DSM 738 was used as negative control, and a *S. clavuligerus* strain conjugated with pIJ6902 was used as positive control as it is resistant to both apramycin and thiostrepton. Between 900 to 1000 colonies were tested for a second cross-over and only two strains were identified as knockout candidates. These two strains, named SCLA405-4.27 and SCLA405-3.72, presented resistance to apramycin but sensitivity to thiostrepton and they originate from conjugation of the wild type with pLIS405 (**Figure 5.11A**). No thiostrepton-sensitive colonies were obtained from conjugation of the *tap-tpg*-duplicated strain (SCLA22) with pLIS405.

In order to confirm if these two strains had the *tap-tpg* genes eliminated from their genome, we extracted genomic DNA from SCLA405-4.27 and SCLA405-3.72 and performed a PCR amplification using primers that bind to the *tap* gene in pSCL4. Both candidate strains present the same band as the wild type (**Figure 5.11B**). This suggest that both colonies still contain this gene and therefore the second crossover did not happen.

Overall, these results suggest that it was not possible to knockout the *tap* and *tpg* genes from pSCL4 using allelic replacement with 3kb of homologous regions.

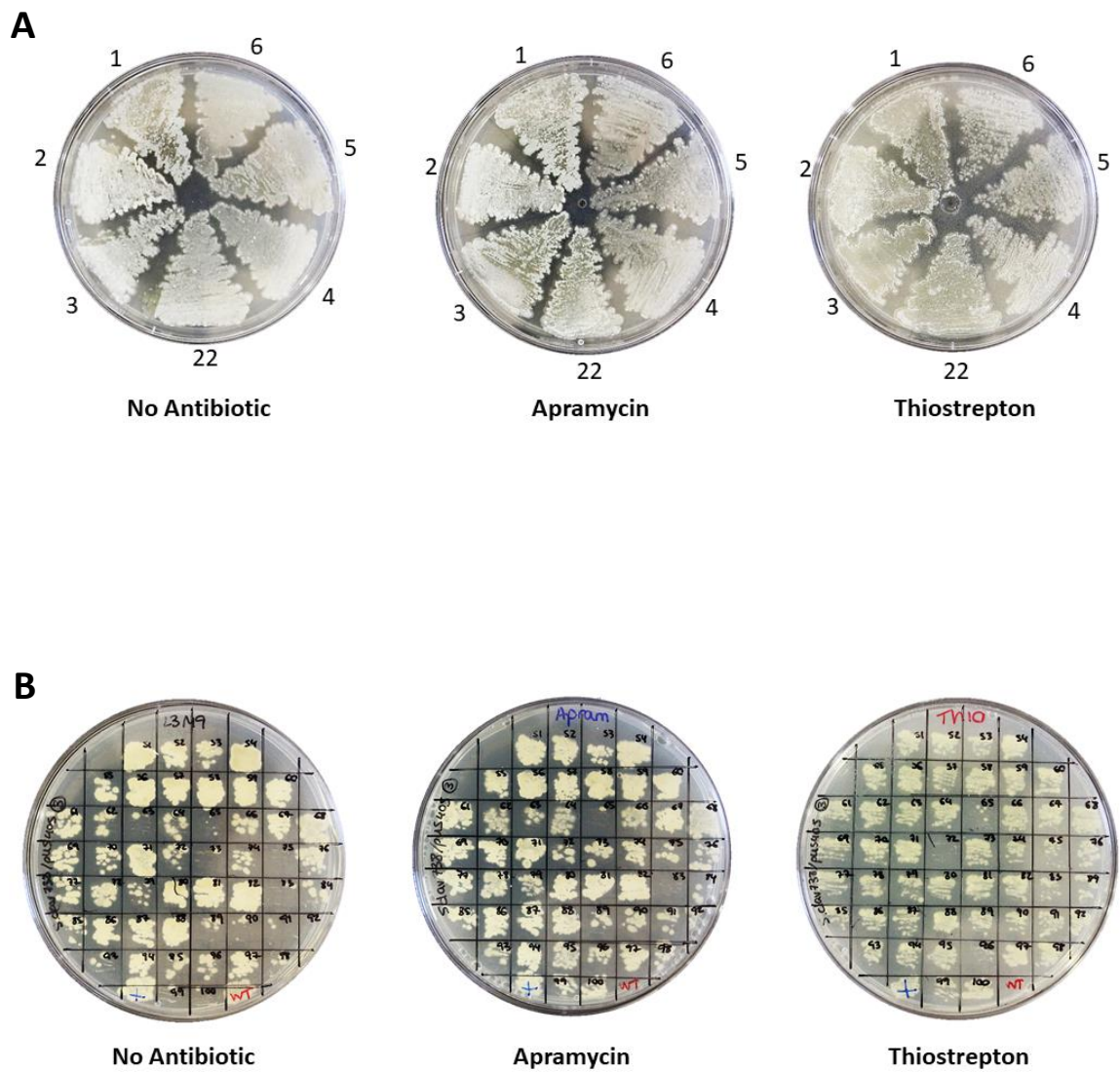


Figure 5.10. Resistance to apramycin and thiostrepton confirms recombination of pLIS405 into *S. clavuligerus* megaplasmid. **A:** 7 transconjugants were obtained that were resistant to both antibiotics, 6 from the wild type (1 to 6) and 1 from the *tap-tpg*-duplicated strain (22). **B:** Example of 50 colonies patched on the three conditions in order to search for knockout strains.

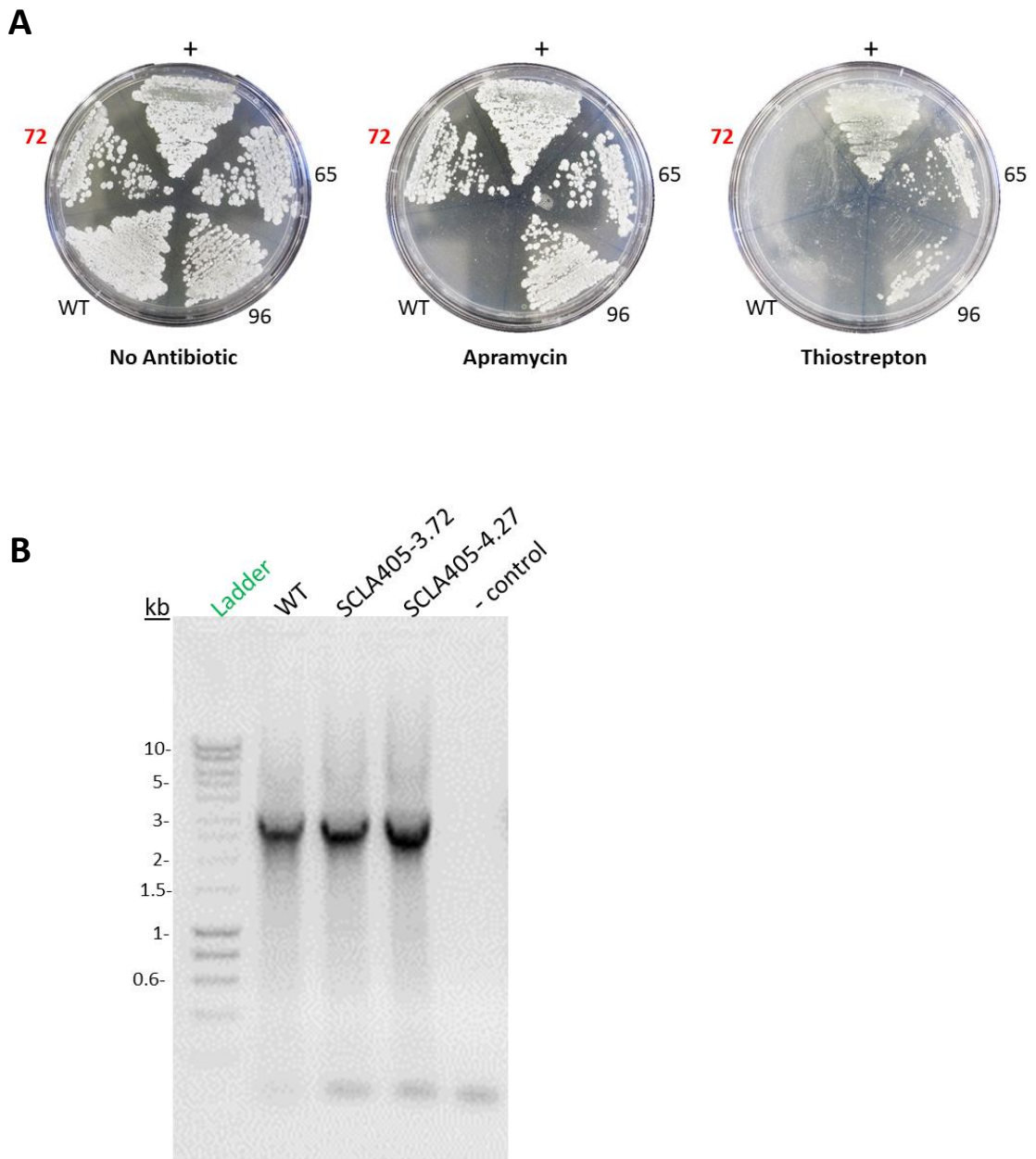


Figure 5.11. Two colonies exhibited the expected double-crossover phenotype (Apra^R, Thio^S) but were excluded from being *tap*-*tpg* knockouts as the *tap* gene was amplified from their genomic DNA. **A:** Plates of the three conditions containing candidate strain SCLA405-3.72 (**72**), WT (DSM 738), positive control (pIJ6903 conjugated), and two one-crossover controls (65 and 96). **B:** Electrophoresis gel of products from amplification of the *tap* gene from DNA of strains WT, SCLA405-3.72 and SCLA405-4.27.

5.4 BACTH experiments confirm interaction of the Tap protein encoded on pSCL4 with itself.

For the purpose of studying the interaction between the Tap and Tpg proteins encoded on the megaplasmid pSCL4, bacterial adenylate cyclase two-hybrid (BACTH) analyses were performed. In order to do so, it was necessary to clone the *4tap* and *4tpg* genes onto the BACTH vectors pKT25 and pKNT25 to translationally fuse the gene to the T25 unit of the adenylate cyclase (AC) enzyme at both orientations, as well as onto the vectors pUT18 and pUT18C, to translationally fuse the other gene to the T18 subunit of the AC at both orientations. Interaction would be then assessed by co-transforming *E. coli* with the two plasmids carrying the T25 and T18 subunits and observing restoration of the adenylate cyclase activity by inducing β -galactosidase degradation of X-Gal, which is observable by the apparition of blue colonies. In contrast, the lack of interaction would yield white colonies. In order to study all the possible interactions between these two proteins, we decided to clone both genes onto the four BACTH vectors, which would result in the construction of the eight plasmids presented on **Table 5.1**.

Table 5.1. List of the eight BACTH plasmids to be constructed in order to study all the possible interactions between the Tap and Tpg proteins encoded on *S. clavuligerus* plasmid pSCL4.

Construct	Genotype	Expression
pLIS611	4Tap + pKT25	N— T25 — 4 Tap — C
pLIS621	4Tap + pKNT25	N— 4 Tap — T25 — C
pLIS631	4Tap + pUT18	N— 4 Tap — T18 — C
pLIS641	4Tap + pUT18C	N— T18 — 4 Tap — C
pLIS711	4Tpg + pUT18	N— 4 Tpg — T18 — C
pLIS721	4Tpg + pUT18C	N— T18 — 4 Tpg — C
pLIS731	4Tpg + pKT25	N— T25 — 4 Tpg — C
pLIS741	4Tpg + pKNT25	N— 4 Tpg — T25 — C

Firstly, *in silico* cloning of the eight plasmids was performed in SnapGene. The restriction enzymes *Xba*I and *Bam*HI were chosen to be used in the cloning as they have restriction sites present on the multiple cloning site on the four vectors and they do not cut within *tap* and *tpg*. The forward primers were designed so it would maintain the gene in frame with the T25

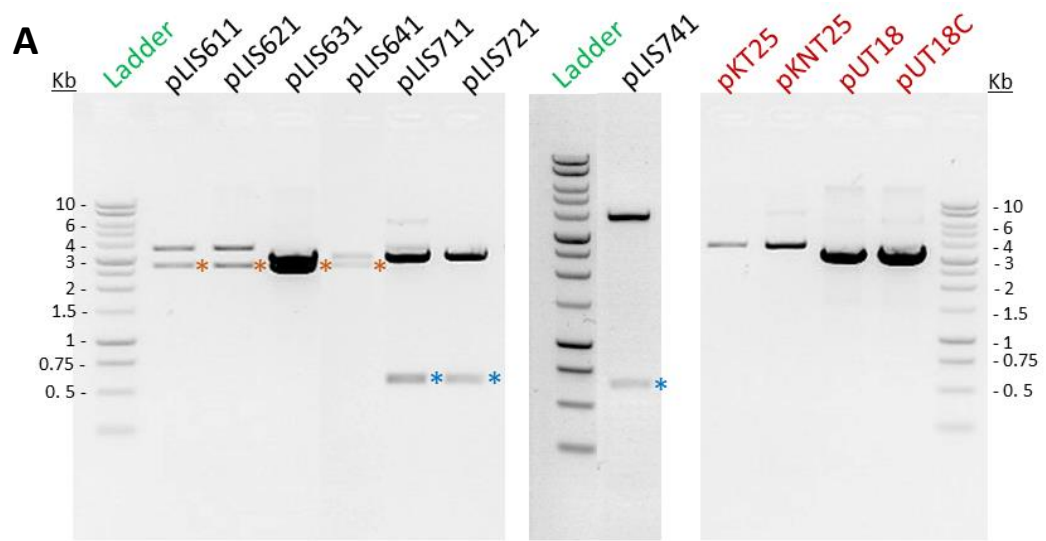
or T18 in the case of pKT25 and pUT18C vectors, or in frame with the AC subunit promoter in the case of pKNT25 and pUT18 vectors. In addition, the *Xba*I restriction site was added to these primers and they were named: BACTH-p4-Tap-Fw and BACTH-p4-Tpg-Fw. In the case of the reverse primers, they were designed so the natural stop codon of *tap* and *tpg* would be eliminated, and so, only the AC subunit stop codon would be expressed. The *Bam*HI restriction site was added to the beginning of the primers and they were named: BACTH-p4-Tap-Rv and BACTH-p4-Tpg-Rv. The *in silico* construction of the eight plasmids was performed by amplification of the *tap* and *tpg* genes and insertion into the four vectors using *Xba*I and *Bam*HI (Figure 5.12).

Wet lab construction of the plasmids was done following the steps of the *in silico* cloning. The *tap* and *tpg* genes were amplified by PCR using the mentioned primers and Q5 High-fidelity polymerase, which was confirmed by gel electrophoresis (Figure A11 in Appendix). PCR products and BACTH vectors were digested with *Xba*I and *Bam*HI individually and digested products were separated by electrophoresis and purified. The products were ligated and used to transform competent *E. coli* cells. Several colonies were selected, and the plasmids were isolated. Screening for the correct construct was performed by restriction enzyme digest with *Xba*I and *Bam*HI. Confirmation of correct cloning is presented on Figure 5.13; unfortunately, the *tpg* gene could not be cloned into pKT25 vector (pLIS731); nevertheless, the seven constructed plasmids are sufficient for performing the interaction assays.

In order to study the interaction between Tap and Tpg from pSCL4, we co-transformed *E. coli* BTH101 competent cells with the plasmids pLIS611 (T25-Tap) and pLIS711 (Tpg-T18). Transformants were streaked on the presence of X-Gal and white colonies were observed, which suggests no protein-protein interaction. In addition, to confirm if the orientation of the proteins in reference to the AC subunit is relevant, we also tested the plasmids pLIS621 (Tap-T25) and pLIS721 (T18-Tpg), and following all the possible combinations of the four plasmids, only white colonies were observed in the presence of X-Gal. Furthermore, the opposite combinations of protein-AC subunit were tested (Tap plus T18 and Tpg plus T25), but still only white colonies were visualised. *E. coli* transformed with pKT25-ZIP and pUT18C-ZIP was used as a positive interaction control and *E. coli* transformed with the empty vectors as negative control, which yielded blue and white colonies respectively. These results suggest



Figure 5.12. Schematic representation of the *in silico* cloning of the BACTH plasmids containing the pSCL4 *tap* gene. The gene was cloned into pKT25, pKNT25, pUT18 and pUT18C using XbaI and BamHI, resulting in the plasmids pLIS611, pLIS621, pLIS631 and pLIS641 respectively.



B

Plasmid	XbaI + BamHI (kb fragment)
pLIS611	3.44, 2.6
pLIS621	3.44, 2.6
pLIS631	3.02, 2.6
pLIS641	3.01, 2.6
pLIS711	3.02, 0.56
pLIS721	3.01, 0.56
pLIS741	3.46, 0.56
pKT25	3.44
pKNT25	3.46
pUT18	3.02
pUT18C	3.01

Figure 5.13. Confirmation of construction of BACTH pSCL4 plasmids **A:** Gel electrophoresis of the seven constructs against the parental vectors (red) digested with *XbaI* and *BamHI*. The *tap* genes are highlighted by an orange asterisk, and the *tpg* genes by a blue asterisk. **B:** List of expected band sizes following digestion with *XbaI* and *BamHI*.

that the pSCL4 Tap and Tpg do not interact in these conditions regardless of the protein orientation. Moreover, in order to study if the Tap and Tpg proteins from pSCL4 interact with themselves, we transformed *E. coli* cells with the following combinations: pLIS611 (T25-Tap) and pLIS641 (T18-Tap), and pLIS711 (Tpg-T18) and pLIS741 (Tpg-T25). Bright blue colonies were obtained from the *E. coli* transformed with Tap plasmids, which suggest strong interaction of the Tap protein with itself. On the other hand, white colonies were observed for the Tpg-Tpg co-transformation and so, no interaction of Tpg with itself is confirmed.

To summarise, Tap and Tpg proteins from pSCL4 do not interact with each other when expressed in *E. coli*. In contrast, Tap shows strong interaction with itself, while no interaction was detected between Tpg proteins. Examples of the co-transformed *E. coli* colonies are presented on **Table 5.2**.

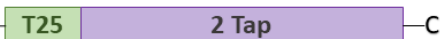

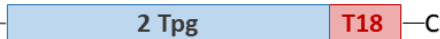
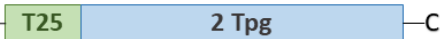
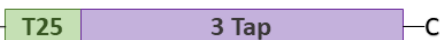

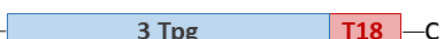
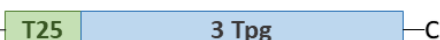
Table 5.2. List of pSCL4 BACTH plasmids combinations tested for studying the protein-protein interactions: 4Tap-4Tpg, 4Tap-4Tap and 4Tpg-4Tpg. Leucine zipper plasmids are used as positive control (green), and empty vectors as negative control (red).

Plasmid combination	Expression	Phenotype
pLIS611 pLIS711	T25-Tap Tpg-T18	
pLIS611 pLIS721	T25-Tap T18-Tpg	
pLIS621 pLIS711	Tap-T25 Tpg-T18	
pLIS621 pLIS721	Tap-T25 T18-Tpg	
pLIS611 pLIS641	T25-Tap T18-Tap	
pLIS711 pLIS741	Tpg-T18 Tpg-T25	
pKT25-ZIP pUT18C-ZIP	T25-Leucine zipper T18-Leucine zipper	
pKT25 pUT18	T25 T18	

5.5 Interaction of the Tpg protein with itself was confirmed by BACTH experiments using the pSCL2 and pSCL3 proteins.

In order to study protein-protein interactions of the terminal proteins encoded on pSCL3 and pSCL2, BACTH analyses were performed similarly to the pSCL4 proteins interaction assays. For this, we proposed to clone the *tap* and *tpg* genes from pSCL3 and pSCL2 onto the BACTH vectors pKT25 and pUT18. The BACTH plasmids constructed to study interactions between terminal proteins of these two plasmids are presented on **Table 5.3**.

Table 5.3. List of the eight BACTH plasmids to be constructed in order to study interactions between the Tap and Tpg proteins encoded on *S. clavuligerus* plasmids pSCL2 and pSCL3.

Construct	Genotype	Expression
pLIS612	2Tap + pKT25	N—  —C
pLIS632	2Tap + pUT18	N—  —C
pLIS712	2Tpg + pUT18	N—  —C
pLIS732	2Tpg + pKT25	N—  —C
pLIS613	3Tap + pKT25	N—  —C
pLIS633	3Tap + pUT18	N—  —C
pLIS713	3Tpg + pUT18	N—  —C
pLIS733	3Tpg + pKT25	N—  —C

In silico construction of the plasmids was done in a similar manner to the pSCL4 plasmids. In the case of the terminal proteins from pSCL2, we also chose the restriction enzymes *Xba*I and *Bam*HI for cloning *2tpg*, but in the case of *2tap*, it was observed that *Bam*HI cuts within the gene and so, *Eco*RI was chosen instead. However, the T25 stop codon on pKT25 lies between the *Xba*I and *Eco*RI sites and cloning of the *2tap* gene would eliminate this codon. It is then necessary to maintain the original stop codon of the *2tap* gene when cloning it into pKT25, while the *2tap* stop codon should be eliminated for cloning into pUT18, in order to allow translational fusion to T18. Taking this into account, the forward primers [BACTH-p2-Tap-Fw](#) and [BACTH-p2-Tpg-Fw](#) were designed including the *Xba*I site so the genes are in frame with the AC subunit promoter. The two reverse primers designed for cloning *2tap* are [BACTH-p2-Tap-Rv](#) and [BACTH-p2-TapEND-Rv](#) for insertion into pUT18C and pKT25 respectively, that

include the *EcoRI* site. The reverse primer for the *2tpg* gene is **BACTH-p2-Tpg-Rv**, that include the *BamHI* site. The four plasmids were constructed using SnapGene using these primers and restriction enzymes.

During *in silico* construction of the pSCL3 BACTH plasmids, it was observed that both the *3tap* and *3tpg* genes are cut with *BamHI*, and so the enzymes *XbaI* and *EcoRI* were chosen to clone both genes. Similarly to the case of the *2tap*, it is necessary to maintain the original stop codon to clone into pKT25, while disrupting the codon for cloning into pUT18, this time for both *3tap* and *3tpg*. The following forward primers were design in order to maintain the genes within the AC promoter: **BACTH-p3-Tap-Fw** and **BACTH-p3-Tpg-Fw**, that include the *XbaI* site. And the reverse primers: **BACTH-p3-Tap-Rv** and **BACTH-p3-TapEND-Rv** for *3tap*, and **BACTH-p3-Tpg-Rv** and **BACTH-p3-TpgEND-Rv** for *3tpg*, all of them containing the *EcoRI* site. The four plasmids were constructed using SnapGene with these conditions.

The wet lab construction of the four pSCL2 and four pSCL3 BACTH plasmids was performed in the same way as the pSCL4 BACTH plasmids. The final products were confirmed by restriction enzyme digest (**Figure 5.14**).

In order to study the interaction between the terminal proteins 2Tap and 2Tpg, we co-transformed *E. coli* BTH101 competent cells with the plasmids pLIS612 (T25-Tap) and pLIS712 (Tpg-T18), as well as, pLIS632 (Tap-T18) and pLIS732 (T25-Tpg). In both cases, the *E. coli* colonies exhibited a white colour, suggesting no protein-protein interaction in these conditions (**Table 5.4**). Furthermore, the plasmids pLIS612 (T25-Tap) and pLIS632 (Tap-T18) were used to transform *E. coli* cells in order to study 2Tap-2Tap interaction, which yielded white cells. In contrast, when transforming cells with pLIS712 (Tpg-T18) and pLIS732 (T25-Tpg) to study interaction of 2Tpg with itself, we obtained blue colonies, as well as some white colonies. In summary, these experiments confirmed no interaction between 2Tap and 2Tpg in *E. coli*, while 2Tpg exhibits interaction with itself.

Moreover, the same plasmid combinations were used to study the interactions of the pSCL3 terminal proteins. Similarly to pSCL2 proteins, no interaction was visualised between the 3Tap and 3Tpg proteins encoded on pSCL3 when expressed in *E. coli* (**Table 5.5**). White colonies were obtained for *E. coli* transformed with 3Tap and 3Tap plasmids, while both white and blue colonies were observed for 3Tpg-3Tpg colonies, which confirms at least partial interaction of 3Tpg with itself for the pSCL3 proteins.

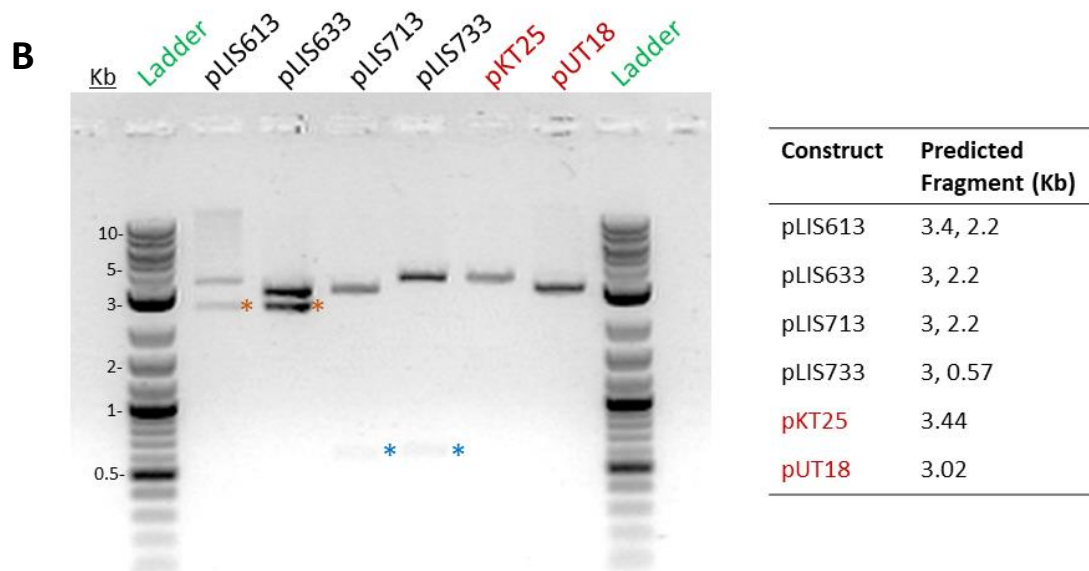
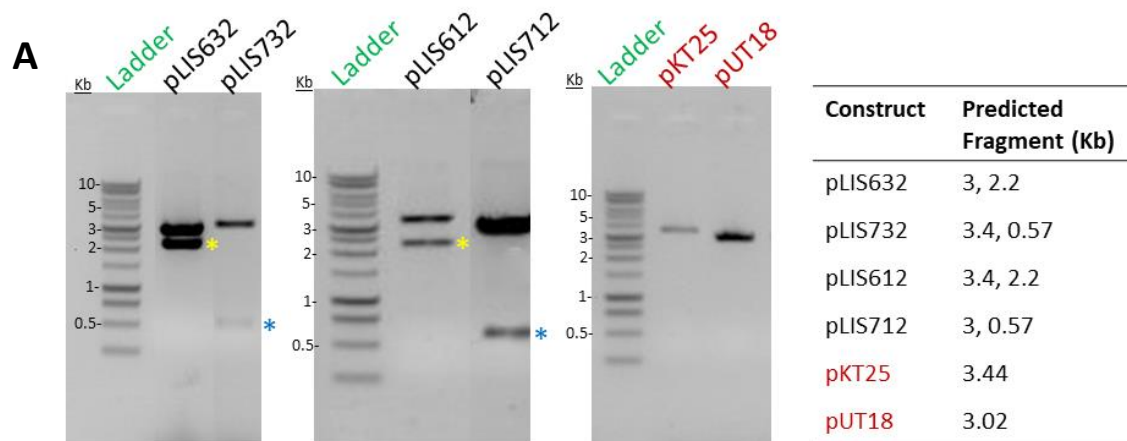


Figure 5.14. A: Confirmation of construction of BACTH pSCL2 plasmids. Gel electrophoresis of the four constructs against the parental vectors (red) digested with *Xba*I and *Bam*HI or *Eco*RI. The *2tap* genes are highlighted by a yellow asterisk, and the *2tpg* genes by a blue asterisk. The table indicates the expected band sizes. **B:** Confirmation of construction of BACTH pSCL3 plasmids. Gel electrophoresis of the four constructs against the parental vectors (red) digested with *Xba*I and *Bam*HI or *Eco*RI. The *3tap* genes are highlighted by an orange asterisk, and the *3tpg* genes by a blue asterisk. The table indicates the expected band sizes.

Table 5.4. List of pSCL2 BACTH plasmids combinations tested for studying the protein-protein interactions: 2Tap-2Tpg, 2Tap-2Tap and 2Tpg-2Tpg. Leucine zipper plasmids are used as positive control (green), and empty vectors as negative control (red).










Plasmid combination	Expression	Phenotype
pLIS612 pLIS712	T25-Tap Tpg-T18	
pLIS632 pLIS732	Tap-T18 T25-Tpg	
pLIS612 pLIS632	T25-Tap Tap-T18	
pLIS712 pLIS732	Tpg-T18 T25-Tpg	
pKT25-ZIP pUT18C-ZIP	T25-Leucine zipper T18-Leucine zipper	
pKT25 pUT18	T25 T18	

Table 5.5. List of pSCL3 BACTH plasmids combinations tested for studying the protein-protein interactions: 3Tap-3Tpg, 3Tap-3Tap and 3Tpg-3Tpg. Leucine zipper plasmids are used as positive control (green), and empty vectors as negative control (red).

Plasmid combination	Expression	Phenotype
pLIS613 pLIS713	T25- Tap Tpg -T18	
pLIS633 pLIS733	Tap -T18 T25- Tpg	
pLIS613 pLIS633	T25- Tap Tap -T18	
pLIS713 pLIS733	Tpg -T18 T25- Tpg	
pKT25-ZIP pUT18C-ZIP	T25-Leucine zipper T18-Leucine zipper	
pKT25 pUT18	T25 T18	

5.6 The terminal proteins from the three plasmids pSCL4, pSCL3 and pSCL2 do not interact with each other.

In order to investigate if the terminal proteins from different plasmids can interact with each other, we performed some BACTH analyses using the plasmids constructed as described on sections 5.4 and 5.5.

Firstly, in order to study possible interaction between the Tap and Tpg proteins of pSCL4 and pSCL2, we co-transformed *E. coli* with plasmids combinations to study 4Tap-2Tpg, 2Tap-4Tpg, 4Tap-2Tap and 4Tpg-2Tpg interactions. All the obtained colonies exhibited a white colour, and so, no protein-protein interaction was confirmed for any of the combinations of terminal proteins from pSCL4 and pSCL2 (**Table 5.6**).

Similarly, plasmids carrying the *tap* and *tpg* from pSCL4 and pSCL3 were used to transform *E. coli* and study the interactions of the four proteins. The pSCL4 proteins do not present interaction to the pSCL3 proteins either, as we only obtained white transformants (**Table 5.7**).

Lastly, the same transformation combinations were carried out to study interactions between the terminal proteins of pSCL2 and pSCL3. Once again, only white *E. coli* colonies were observed, thus no protein-protein interaction was confirmed (**Table 5.8**)

In summary, BACTH experiments suggest no protein-protein interactions between the Tap and Tpg from pSCL4, pSCL3 and pSCL2.

Table 5.6. List of pSCL4 and pSCL2 BACTH plasmids combinations tested for studying the protein-protein interactions: 4Tap-2Tpg, 2Tap-4Tpg, 4Tap-2Tap and 4Tpg-2Tpg. Leucine zipper plasmids are used as positive control (green), and empty vectors as negative control (red).






Plasmid combination	Expression	Phenotype
pLIS611 pLIS712	T25-4Tap 2Tpg-T18	
pLIS711 pLIS612	4Tpg-T18 T25-2Tap	
pLIS611 pLIS632	T25-4Tap 2Tap-T18	
pLIS711 pLIS732	4Tpg-T18 T25-2Tpg	
pKT25-ZIP pUT18C-ZIP	T25-Leucine zipper T18-Leucine zipper	
pKT25 pUT18	T25 T18	

Table 5.7. List of pSCL4 and pSCL3 BACTH plasmids combinations tested for studying the protein-protein interactions: 4Tap-3Tpg, 3Tap-4Tpg, 4Tap-3Tap and 4Tpg-3Tpg. Leucine zipper plasmids are used as positive control (green), and empty vectors as negative control (red).













Plasmid combination	Expression	Phenotype
pLIS611 pLIS713	T25-4Tap 3Tpg-T18	
pLIS711 pLIS613	4Tpg-T18 T25-3Tap	
pLIS611 pLIS633	T25-4Tap 3Tap-T18	
pLIS711 pLIS733	4Tpg-T18 T25-3Tpg	
pKT25-ZIP pUT18C-ZIP	T25-Leucine zipper T18-Leucine zipper	
pKT25 pUT18	T25 T18	

Table 5.8. List of pSCL2 and pSCL3 BACTH plasmids combinations tested for studying the protein-protein interactions: 2Tap-3Tpg, 3Tap-2Tpg, 2Tap-3Tap and 2Tpg-3Tpg. Leucine zipper plasmids are used as positive control (green), and empty vectors as negative control (red).

Plasmid combination	Expression	Phenotype
pLIS612 pLIS713	T25-2Tap 3Tpg-T18	
pLIS712 pLIS613	2Tpg-T18 T25-3Tap	
pLIS612 pLIS633	T25-2Tap 3Tap-T18	
pLIS712 pLIS733	2Tpg-T18 T25-3Tpg	
pKT25-ZIP pUT18C-ZIP	T25-Leucine zipper T18-Leucine zipper	
pKT25 pUT18	T25 T18	

5.7 Summary.

In this chapter we report the identification and characterisation of three set of terminal proteins in the genome of *S. clavuligerus*. While the *tap* and *tpg* genes from pSCL4 were identified in the work published by Medema *et al.*, 2010, no previous findings regarding the *tap* and *tpg* from pSCL3 and pSCL2 have been described. Considering no *tap-tpg* genes are present on both the chromosome and the smallest plasmid pSCL1, and they still display a linear topology, these replicons must rely on the terminal proteins encoded on the other plasmids to maintain the telomeric sequences. This is not unknown, as it has been described how in other *Streptomyces* species, such as *S. rochei*, the chromosome utilises the terminal proteins encoded on plasmids (Nindita *et al.*, 2015).

Moreover, it was observed that the *4tap-4tpg* operon is located on the central region of pSCL4, a fact that is unique to this replicon, as these genes are usually located towards the termini of *Streptomyces* chromosomes and plasmids. This observation contributes to Medema's theory about the origin of the megaplasmid, which states that the plasmid comes from recombination events of the chromosome with a smaller plasmid (Medema *et al.*, 2010). The *tap-tpg* genes could have once been present at the end of the chromosome or plasmid and acquired the central position upon recombination between the molecules.

Comparison of the Tap and Tpg amino acid sequences to other *Streptomyces* terminal proteins allowed us to visualise a high conservation, which confirms that the three sets of proteins are archetypal. However, phylogenetic studies of these sequences suggested that the proteins encoded on pSCL4 and pSCL3 are among the rarest archetypal terminal proteins of *Streptomyces* species. In contrast, the terminal proteins from pSCL2 are more similar to those from streptomycetes phylogenetically distant to *S. clavuligerus*, hence it could be suggested that pSCL2 originated from conjugal transfer of an organism closer to species like *S. coelicolor* and *S. avermitilis*. In addition, we predicted the helix-turn-helix motifs on the Tap and Tpg amino acid sequences of the three plasmids, as well as the amphiphilic β -sheet on the Tpg proteins.

As it was confirmed by Southern blotting, we succeeded in introducing a copy of the pSCL4 *tap-tpg* operon on the *S. clavuligerus* chromosome. This integration was achieved by utilisation of the bacteriophage ϕ BT1 *att*-site-specific recombination mechanism described by Gregory and Smith, 2003. *attB* sites are found on the *S. clavuligerus* chromosome and integration of plasmids with the *attP*-integrase system have been used on previous studies






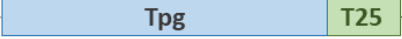
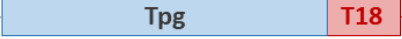
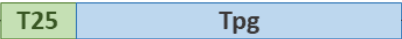
on this microorganism (Combes et al. 2002; Trepanier et al. 2002). However, these studies utilised plasmids containing ϕ C31 *attP-int* locus, which integrate at different *attB* sites than the ϕ BT1-based plasmids (Gregory and Smith, 2003). With this work we confirm that the ϕ BT1 *att*-site-specific mechanism works successfully for the integration of elements on the chromosome of *S. clavuligerus*, even with such large inserts as the *tap-tpg* operon (3.5 kb).

Once a *tap-tpg*-chromosome strain was obtained, we intended to knock out the genes from pSCL4 by allelic replacement using 3 kb of homologous regions. Unfortunately, we confirmed that it was not possible to eliminate these genes by this methodology. While many single-crossover transconjugants (Apra and Thio resistant) were obtained, only two mutants had lost their resistance to thiostrepton. Nevertheless, the presence of *4tap* on these mutants was confirmed by PCR, which confirmed no second recombination event happened and the loss of thiostrepton resistance could probably be explained by apparition of random mutations on the resistance gene. The lack of knockout mutants might indicate that these *4tap* and *4tpg* genes are essential for *S. clavuligerus*. However, the genes could not be eliminated from the *tap-tpg*-duplicated strain either, which suggests that either the copies on the chromosome are not functional, or the loss of *tap* and *tpg* from pSCL4 has detrimental effects in addition to their terminal function in end-patching. It would be interesting to see if this is the same effects when trying to silence the *4tap-4tpg* genes, instead of eliminating the whole genes, by new technology such as CRISPR-BEST or CRISPR/dCas9 systems; as well as studying the effects of disrupting the other *tap* and *tpg* genes on pSCL2 and pSCL3 (Gjaltema and Schulz, 2018; Tong et al., 2019).

As was evidenced in our BACTH experiments on the Tap and Tpg proteins of pSCL4, pSCL3 and pSCL2, these two proteins do not interact when expressed in *E. coli*. This was confirmed regardless of the position of the AC subunit in reference to the terminal protein. Previous studies confirmed *in vivo* interaction between the Tap and Tpg of *S. lividans* by co-immunoprecipitation using cell extract (Bao and Cohen, 2003). It could be then hypothesised that the Tap protein must bind to telomeric DNA first in order to interact with Tpg, and so this interaction is not possible in *E. coli*. The *in vitro* studies of Yang et al., 2015 suggested that without the attachment of Tap to the telomeric Palindromes II and III, the synthesis efficiency of the 13 nucleotides-primer for end patching decreased. Following our hypothesis, this impaired end-patching could be explained by a poor Tap-Tpg interaction in the absence of specific telomeric DNA.

In contrast, significant interaction was observed between Tap proteins of pSCL4. This interaction was not visible on the Tap proteins of pSCL3 and pSCL2, however, this could be due to the orientation of the Tap protein in relation to the AC subunit. The 4Tap protein tested positive for interaction when both AC subunits were positioned at the N-terminus of the protein, while no interaction was observed when one of the pSCL3 and pSCL2 proteins carried the subunit at the C terminus. If this is extrapolated to the three sets of proteins, it can be assumed that the AC subunits at the C-termini are blocking interaction of the Tap proteins (**Table 5.9**). Similarly, the 4Tpg proteins showed no interaction when both the subunits were located at the C-terminus; while the pSCL3 and pSCL2 Tpg proteins, which one of the plasmids was cloned so one of the subunit would be located at the N-terminus, and the other plasmids at the C-terminus, exhibited partial interaction. The AC subunit could be blocking the interaction site at the C-terminus as well. These results suggest that both Tap and Tpg proteins interact with themselves and the interaction sites are located towards the C-termini of the proteins. This could explain the high level of conservation we observed on the amino acid sequences of the second half of the Tap and Tpg proteins, and correlates with the fact that the proteins N-termini are rather involved in interaction with DNA, as suggested by identification of helix-turn-helix motifs.

Table 5.9. The combinations of Tap-Tap and Tpg-Tpg plasmids tested by BACTH analysis and if interaction was observed.

Combination		Interaction
N—	 Tap —C	✓
N—	 Tap —C	
N—	 Tap —C	✗
N—	 Tap —C	
N—	 Tpg —C	✗
N—	 Tpg —C	
N—	 Tpg —C	✓
N—	 Tpg —C	

Interaction between same terminal proteins was previously confirmed by chemical cross-linking of *Streptomyces* mycelium (Tsai *et al.*, 2011). Although this interaction was explained

to be due to Tpg-Tpg association, considering our results, Tap-Tap interactions could also be responsible for this. In this publication they evidenced interaction of terminal proteins located at the ends of same molecule, which would provide linear chromosomes and plasmids a circular conformation. In addition, it was confirmed the interaction of terminal proteins between chromosome and plasmid even when carrying different terminal proteins. Unfortunately, the terminal proteins from pSCL4, pSCL3 and pSCL2 do not show any kind of interaction between each other according to our BACTH experiments, suggesting a high specificity for each replicon's terminal proteins to interact. However, considering the chromosome must utilise one of the *tap-tpg* sets from pSCL4, pSCL3 or pSCL2, it cannot be excluded the idea of chromosomal telomeres associating with the telomeres of the providing plasmid.

6. ESSENTIALITY OF THE MEGAPLASMID pSCL4 IN *S. CLAVULIGERUS*

Streptomyces often carry extra-chromosomal elements which present either a linear or circular topology. These plasmids frequently carry transposable elements that allow mobilisation of biosynthetic genes into chromosomes enabling acquisition of specialised metabolism pathways, which evidences the important role that plasmids have had in evolution and adaptation of streptomycetes. In some cases, plasmids even carry the entire clusters responsible for the biosynthesis of antibiotic compounds. These are plasmids larger than 100 kb and referred as giant linear plasmids.

While small linear plasmids were easily identified by conventional gel electrophoresis in organisms such as *S. rimosus* (pSRM) and even *S. clavuligerus* (pSCL1) (Chardon-Loriaux *et al.*, 1986; Keen *et al.*, 1988), it was not until the development of PFGE that larger plasmids were observed. The studies of Kinashi *et al.*, for example, revealed the presence of linear plasmids in six *Streptomyces* species with sizes 17 to 590 kb, that carry genes for the biosynthesis of antibiotics (Kinashi and Shimaji, 1987; Kinashi *et al.*, 1988).

The most studied of the giant linear plasmids is SCP1 from *S. coelicolor* A3(2). This was first identified by Hopwood *et al.*, 1969 as a fertility factor, whose replication was independent of the chromosome. It was also confirmed that SCP1 could integrate into different regions of the chromosome and be transferred by conjugation to different *Streptomyces* species (Vivian and Hopwood, 1975; Vivian, 1971). SCP1 carries the genes necessary for the biosynthesis of methylenomycin (Kirby *et al.*, 1975), and also important regulatory genes for antibiotic production, as well as and extracytoplasmic function sigma factors genes, and even genes important for cell division cell development like the *wbl* genes (Takano, 2006; Bentley *et al.*, 2002; Chater and Chandra, 2006). However, it was confirmed that the plasmid could be lost either spontaneously or after UV exposure (Vivian and Hopwood, 1970).

Other examples of giant linear plasmids are the pSLA2-L and pSLA2-M plasmids of *S. rochei*. These plasmids carry many secondary metabolite clusters, such as the ones encoding the antibiotics lankacidin and lankamycin (Mochizuki *et al.*, 2003). In addition, it was evidenced that these plasmids are essential for providing the telomere replication machinery for the chromosome, as plasmid loss resulted in chromosomal circularisation. This circularisation of the chromosome was also evidenced to be avoided by complementing the chromosome with a copy of the plasmid *tap-tpg* genes previously to curing the plasmid (Nindita *et al.*, 2015).

S. clavuligerus presents three giant linear plasmids (pSCL2, pSCL3 and pSCL4), and is similar to *S. rochei* in the way that the plasmids are responsible for expressing Tap and Tpg in the organism. Extrapolating the findings for *S. rochei*, it could be hypothesised that plasmid loss in *S. clavuligerus* might induce circularisation of the chromosome as well; in particular the loss of pSCL4, as the megaplasmid and chromosome carry the same telomeres, similarly to *S. rochei*.

Spontaneous loss of the megaplasmid pSCL4 from *S. clavuligerus* was previously reported on the studies by Charusanti *et al.*, 2012, when they co-cultured *S. clavuligerus* ATCC 27064 and methicillin-resistant *Staphylococcus aureus* N315. The pSCL4-free strain was also reported to have an increased production of the antitumoral metabolite holomycin. In addition, in a publication by Álvarez-Álvarez *et al.*, 2014, the authors cured pSCL4 by deletion of the plasmid *parA* and *parB* genes in a *S. clavuligerus* strain that lacks pSCL3. This yielded non-sporulating and slow growing strains, which suggested that the plasmid might be involved in mechanisms for the optimal growth of *S. clavuligerus*. Furthermore, they reported a slight decrease of clavulanic acid production and an increase in holomycin production when pSCL4 was missing, as well as the loss of the chromosomal right end.

These studies suggested that, while the megaplasmid could be eliminated, it has important consequences for *S. clavuligerus*. With this chapter we aimed to test this by studying the effect of the loss of the megaplasmid in *S. clavuligerus* DSM 738 physiology, metabolism, and chromosomal topology.

6.1 Elimination of the megaplasmid pSCL4 from *S. clavuligerus*.

In order to obtain pSCL4-free mutants of *S. clavuligerus*, we utilised CRISPR/Cas9 technology. The CRISPR/Cas9 plasmid pGE240 was constructed and kindly provided by Dr. Gómez-Escribano (John Innes Centre, Norwich). This plasmid carries the CRISPR/Cas9 cassette from pCRISPomyces2 (Cobb *et al.*, 2015), guide-RNA spacer to target *parB* in pSCL4, and the apramycin-resistance gene. The spacer has the following sequence: CCCCAGCAAATACTGGAAT, obtained from the middle section of *parB* in pSCL4. The spacer was cloned into pCRISPomyces2 and the CRISPR/Cas9-spacer functionality was subcloned into pIJ86, resulting in pGE240.

Expression of the Cas9-guide-RNA in *S. clavuligerus* should result in the Cas9 nuclease incorporating a double-stranded break in pSCL4 *parB*. This would induce loss of nucleotides at the point of the break and, even if re-ligation occurs, the gene would stay disrupted, which would prevent the correct expression of the partitioning machinery necessary for the segregation of pSCL4. Taking this into account, we decided to use pGE240 to eliminate pSCL4 from the *S. clavuligerus* wild type (DSM 738) and the strain with an additional copy of pSCL4 *tap-tpg* on the chromosome (SCLA22) (**Figure 6.1**). Both these strains contained all four linear plasmids.

In order to introduce pGE240 into *S. clavuligerus*, DNA of pGE240 was first used to transform *E. coli* ET12567/pUZ8002 conjugative strain, and apramycin-resistant colonies were picked and used for conjugation. This strain was then conjugated with spores of *S. clavuligerus* DSM 78 and SCLA22 and after 2 to 3 weeks of applying selection with nalidixic acid and apramycin, transconjugants appeared.

Transconjugants originating from the WT were named **SCLA-MX**, and the ones originating from SCLA22 were named **SCLA-McX**. The strains were grown in the absence of antibiotic and it was observed how some of the strains carried a partially bald phenotype (**Figure 6.2**). This is particularly characteristics for the strains M1, M3 and Mc3.

In order to assess the presence of the megaplasmid, we performed colony PCR analyses on the transconjugants. For the first PCR analysis, we used primers that bind to the *tpg* gene on pSCL4 (**Tpg-Fw** and **Tpg-Rv**), and genomic DNA from *S. clavuligerus* DSM 738 was used for a positive control. Products were separated by gel electrophoresis and it was confirmed that four strains lacked this gene: M1, M2, M3 and M5 (**Figure 6.3A**). In the case of the *tap-tpg*-duplicated mutants, all of them showed bands for *tpg*, which was to be expected as there is an extra copy of this gene on the chromosome. Furthermore, the same transconjugants were used for another colony PCR analysis, this time using primers that bind to the left arm end of pSCL4 (**SCL4-L-end-F** and **SCL4-L-end-R**). The same transconjugants originating from the WT (M1, M2, M3 and M5) lacked this region, as well as one strain from the *tap-tpg*-duplicated mutants (Mc3) (**Figure 6.3B**). These results confirm that the strains M1, M2, M3, M5 and Mc3 are missing both the *tpg* gene and the left arm region of the megaplasmid making them candidates for curing pSCL4.

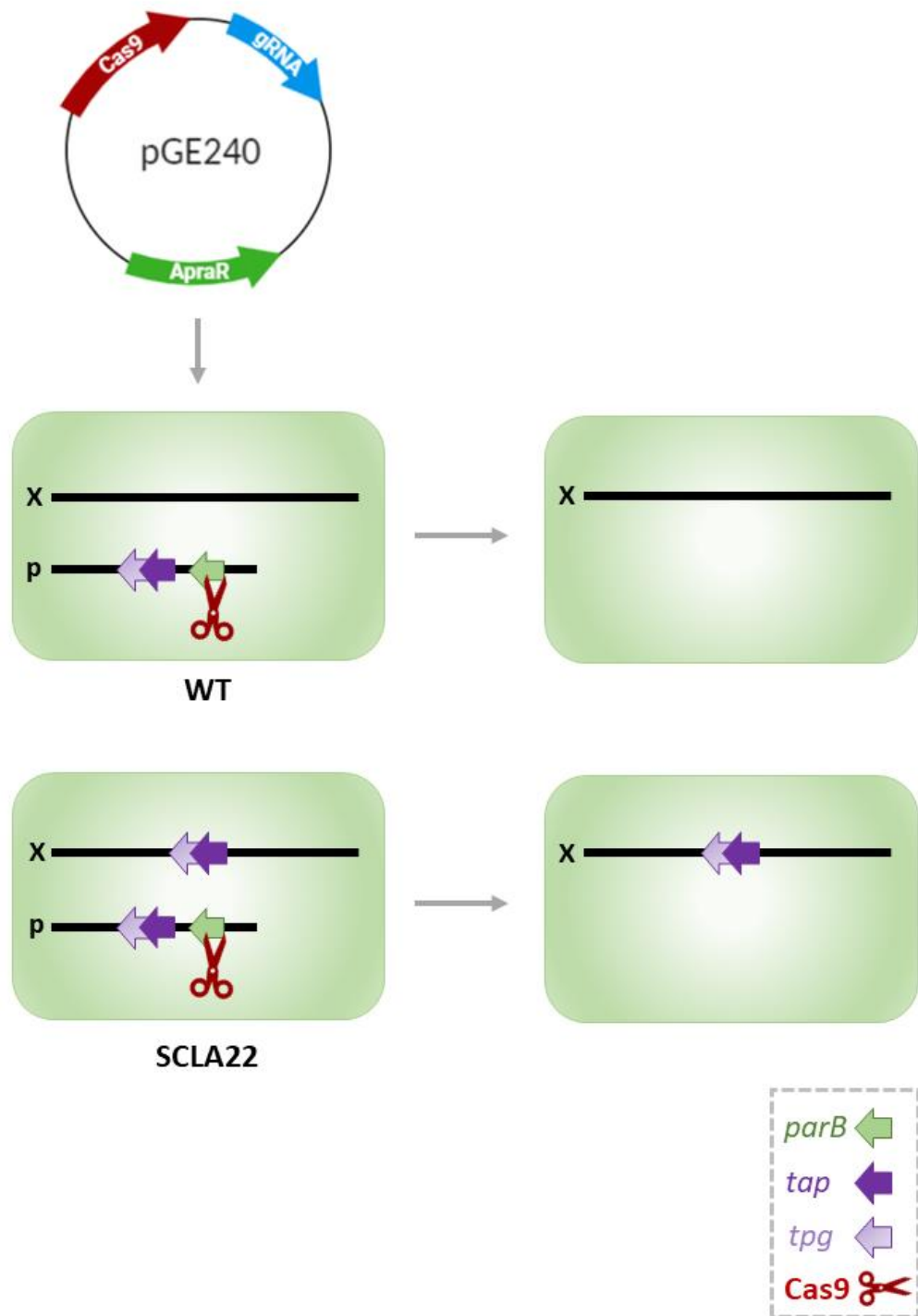


Figure 6.1. Schematic representation of the elimination of pSCL4 from *S. clavuligerus* WT (DSM 738) and the *tap-tpg*-duplicated strain SCLA22 using the CRISPR/Cas9 plasmid pGE240. pGE240 carries guide RNA (gRNA) that targets the Cas9 to cut the *parB* gene of pSCL4, which prevents further segregation of pSCL4. X: chromosome, p: pSCL4, ApraR: apramycin-resistance gene.

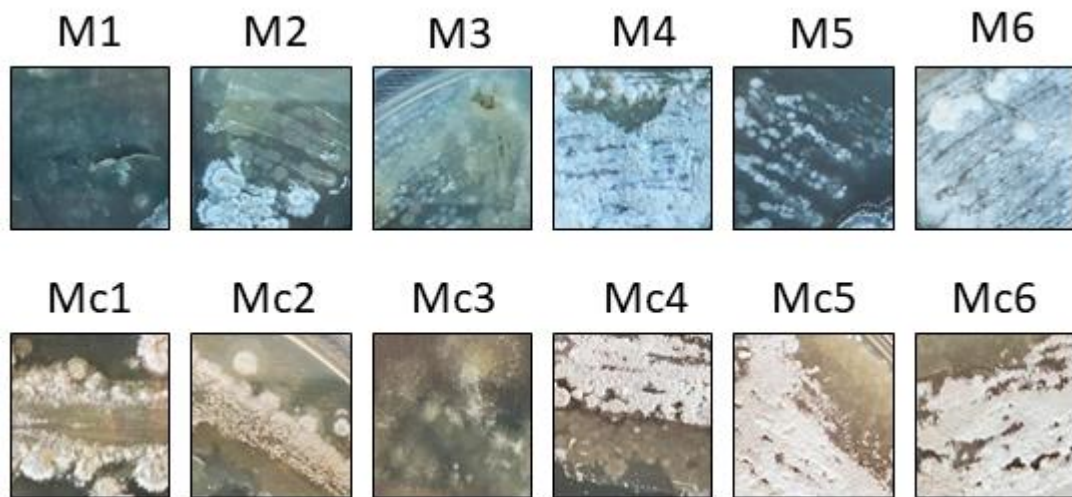


Figure 6.2. Examples of strains obtained from conjugation of *S. clavuligerus* DSM 738 (M1 to M6) and conjugation of SCLA22 (Mc1 to Mc6) with pGE240. The strains were grown on L3M9 media.

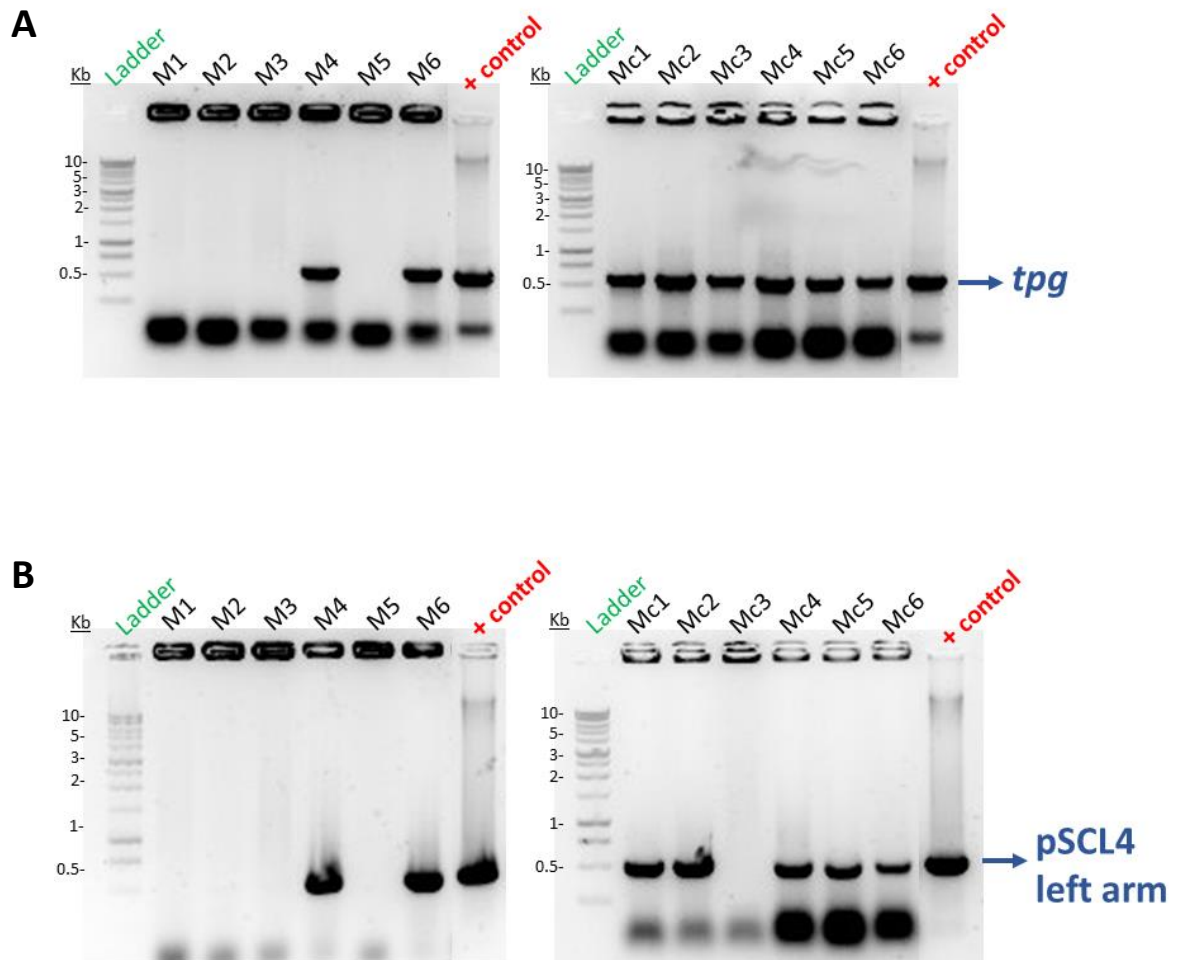


Figure 6.3. Gel electrophoresis of colony PCR results from 12 transconjugants. **A:** PCR performed using primers that bind to the *tpg* gene of pSCL4. **B:** PCR performed using primers that bind to the left arm end of pSCL4. *S. clavuligerus* DSM 738 genomic DNA was used for a positive control.

These five strains were used for further colony PCR analyses. The first PCR was performed using primers that bind the right arm end of pSCL4 (*SCL4-R-end-F* and *SCL4-R-end-R*). None of the mutant strains showed bands for this region (**Figure 6.4A**). The second PCR was carried out with primers that anneal to the chromosomal telomere (*CaoChr-L-end-F* and *CaoChr-L-end-R*). Similarly, none of the mutants showed bands for the telomere (**Figure 6.4B**). In addition, another PCR was performed using primers that amplify the *ftsZ* gene (*FtsZ-Fw* and *FtsZ-Rv*), which is located on the chromosome of *S. clavuligerus*, in order to use as a chromosome positive control. All the tested strains showed correct bands for *ftsZ* (**Figure 6.4C**). These results confirm that all five strains are *S. clavuligerus* mutants that have at least partially lost the megaplasmid, as both left and right arm ends are missing, as well as the *tpg* gene located on the central region of the pSCL4. In addition, all the mutants have lost the chromosome telomeres, which suggests that the chromosome might have acquired a circular form.

Finally, in order to confirm the loss of the megaplasmid and the genomic architecture of the mutants, a PFGE was performed. Samples of the five mutants, as well as *S. clavuligerus* DSM 738 as the wild type (WT), were prepared and loaded undigested on the PFGE gel, and electrophoresis was performed at 4 V/cm for 33 hours using HEPES buffers. The gel showed bands at 1,800, 440 and 150 kb for the WT sample, which correspond to the three giant linear plasmids pSCL4, pSCL3 and pSCL2 respectively (**Figure 6.5**). In contrast, the 1,800 kb band was missing in all the mutant strains. In addition, the band corresponding to pSCL3 was also missing on strains M5 and Mc3. Moreover, a larger band for pSCL3 was observed on the M2 strain. These results suggest that all the mutants have lost pSCL4, and in the case of M5 and Mc3 also pSCL3. Furthermore, M2 strain carries a larger pSCL3, which could be due to a translocation event with pSCL4 material.

Overall, we can confirm that elimination of the megaplasmid is possible by targeting the plasmids *parB* gene with CRISPR/Cas9 technology. Elimination of the megaplasmid has an effect on the chromosome sequence, as well as on the pSCL3 plasmid in some cases.

We decided to continue characterising the mutants, and for that we chose one strain per different phenotype observed: **M1** missing pSCL4, **M2** missing pSCL4 and carrying a larger pSCL3 and **Mc3** missing both pSCL4 and pSCL3.

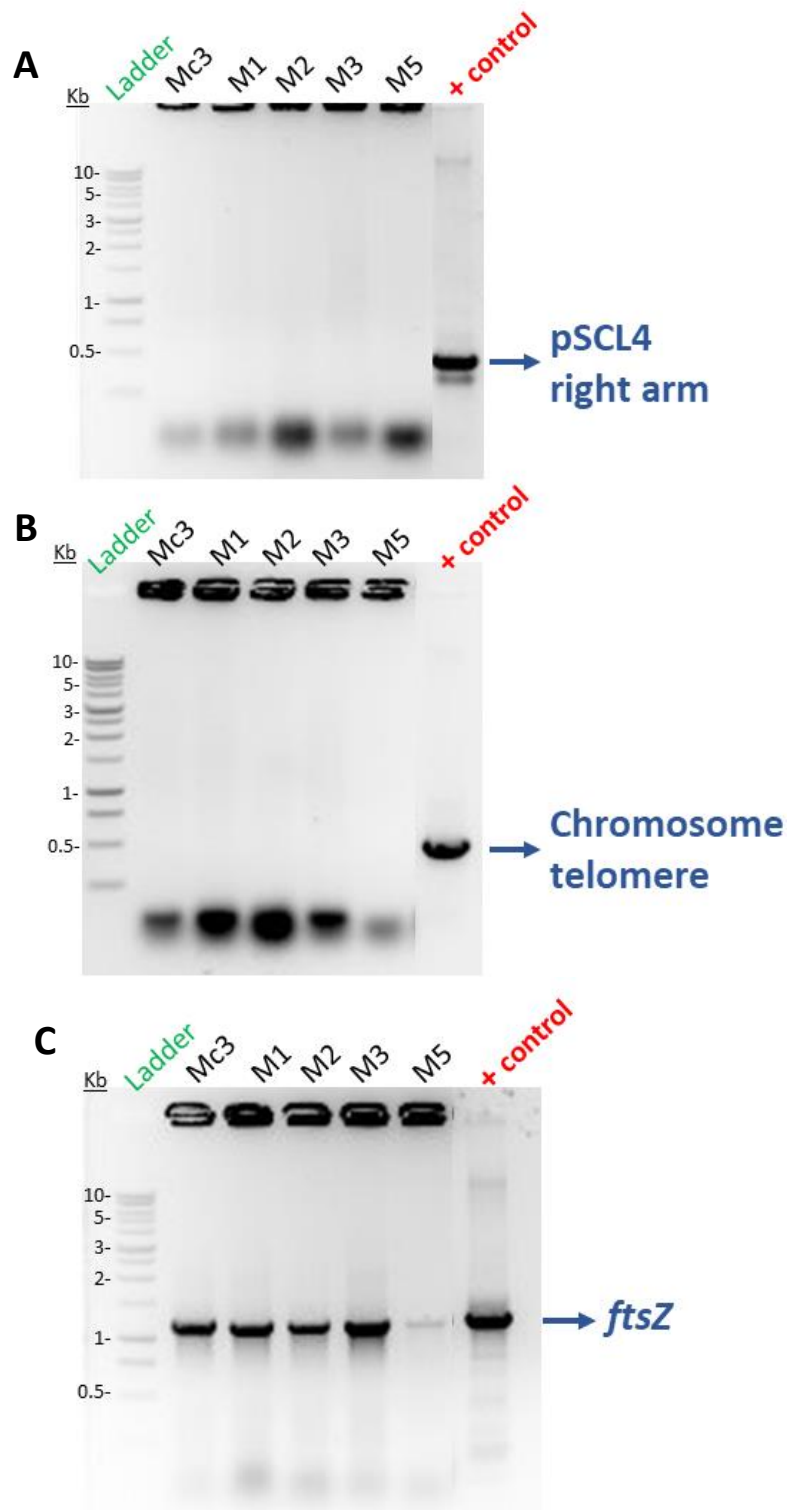


Figure 6.4. Gel electrophoresis of colony PCR results from the five pSCL4-knockout candidates. **A:** PCR performed with primers that amplify the right arm end of pSCL4. **B:** PCR using primers that anneal to the chromosome telomere. **C:** PCR performed to amplify the *ftsZ* gene. *S. clavuligerus* DSM 738 genomic DNA was used for a positive control.

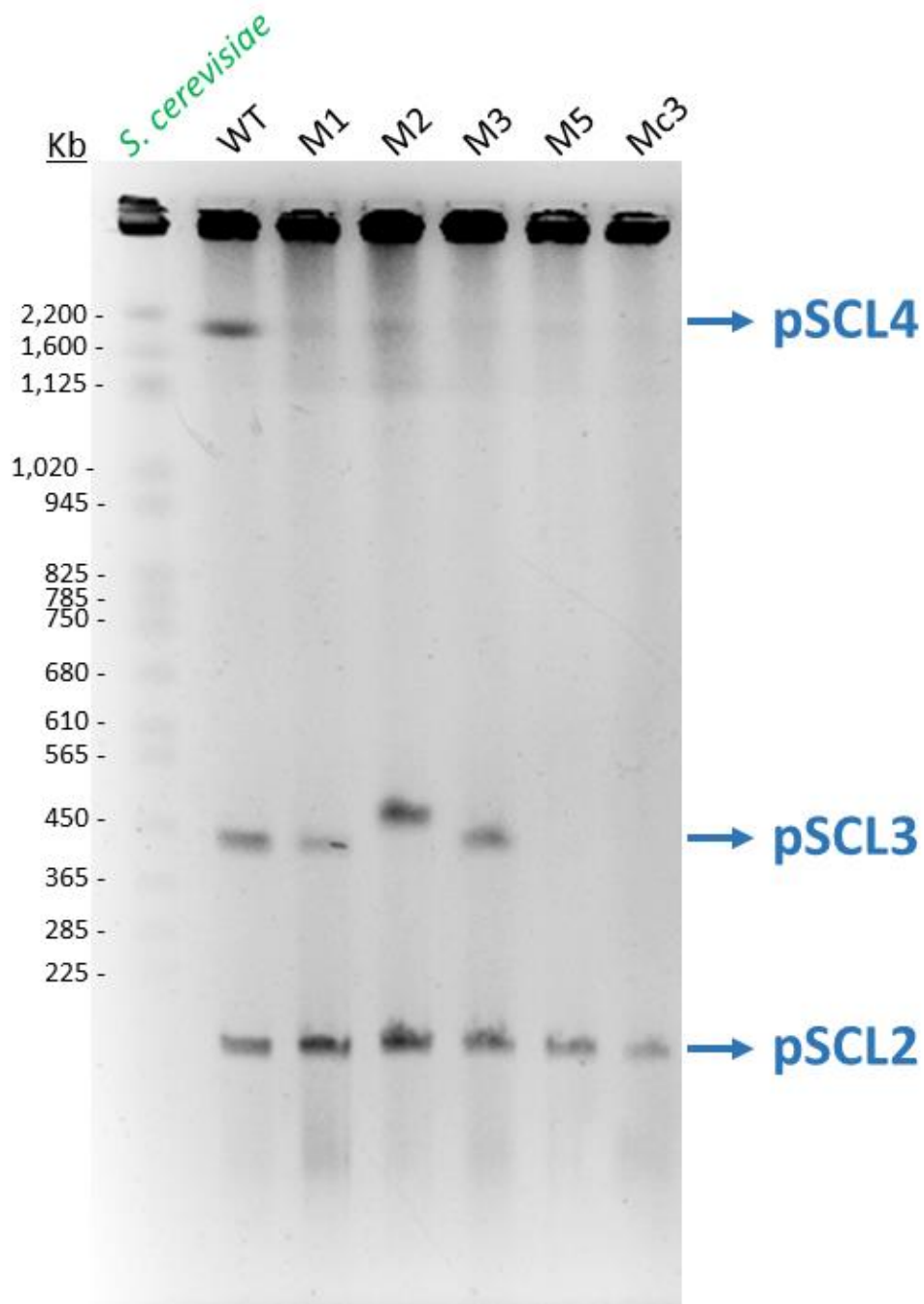


Figure 6.5. PFGE of undigested samples of the mutant strains M1, M2, M3, M5 and Mc3, as well as the WT strain (*S. clavuligerus* DSM 738). Electrophoresis performed using HEPES buffer at 4 V/cm for 33 hours. *Saccharomyces cerevisiae* YNN295 was used as DNA size standard (Bio-Rad).

6.2 Whole genome sequencing of pSCL4-free strains confirms loss of plasmid and chromosome ends.

In order to study the whole genome sequence of the three mutant strains M1, M2 and Mc3, and confirm total loss of pSCL4, we used the services of MicrobesNG (Birmingham) to obtain Illumina sequencing data for the three strains. The Illumina reads were aligned to the WT assembly (*S. clavuligerus* DSM 738 sequences obtained as described in Chapter 4). The alignments were viewed using Integrated Genome Browser (IGB) and compared to the WT reads alignment to study the coverage.

The coverage map of the mutant strain M1 Illumina reads aligned to the WT assembly exhibited sequencing data for most of the chromosome sequence (**Figure 6.6**). As expected from the PCR analyses of the chromosome telomere, no reads from the M1 strain were obtained that aligned to the chromosome ends. Almost no reads were observed to align to the pSCL4 sequence. The few reads that were observed for pSCL4 were confirmed to align to regions with homology to other parts of the genome (chromosome and pSCL2). In addition, we observed M1 Illumina reads that aligned to the entire sequences of pSCL3, pSCL2 and pSCL1. These results confirm that the entirety of pSCL4 is missing from the M1 strain, as well as the chromosome ends, while pSCL3, pSCL2 and pSCL1 remain intact.

The Illumina reads coverage map for strain M2 is presented on **Figure 6.7**. This figure shows reads aligning to the chromosome sequence, except for the ends, which correlates with the colony PCR results described in the previous section. In this case, the missing ends were extended further by a much lower number of reads as compared to the rest of the chromosome coverage, however, no reads were observed for the chromosome telomeres. Similarly to the M1 map, very few reads aligned to the pSCL4 sequence. A new small section of reads that aligned towards the left arm of the megaplasmid was found on the M2 map that was not present on the M1 figure. In addition, we obtained reads for all the sequence of pSCL3, with the exception of a small internal region at the left arm of the plasmid. Furthermore, we observed reads for all the pSCL2 and pSCL1 sequences. These results suggest that the M2 strain lacks the megaplasmid pSCL4 and the chromosome ends, while maintaining pSCL3, pSCL2 and pSCL1. Additionally, the presence of reads for a region of pSCL4 suggests that a translocation event from pSCL4 to another replicon might have occurred during curing the megaplasmid.

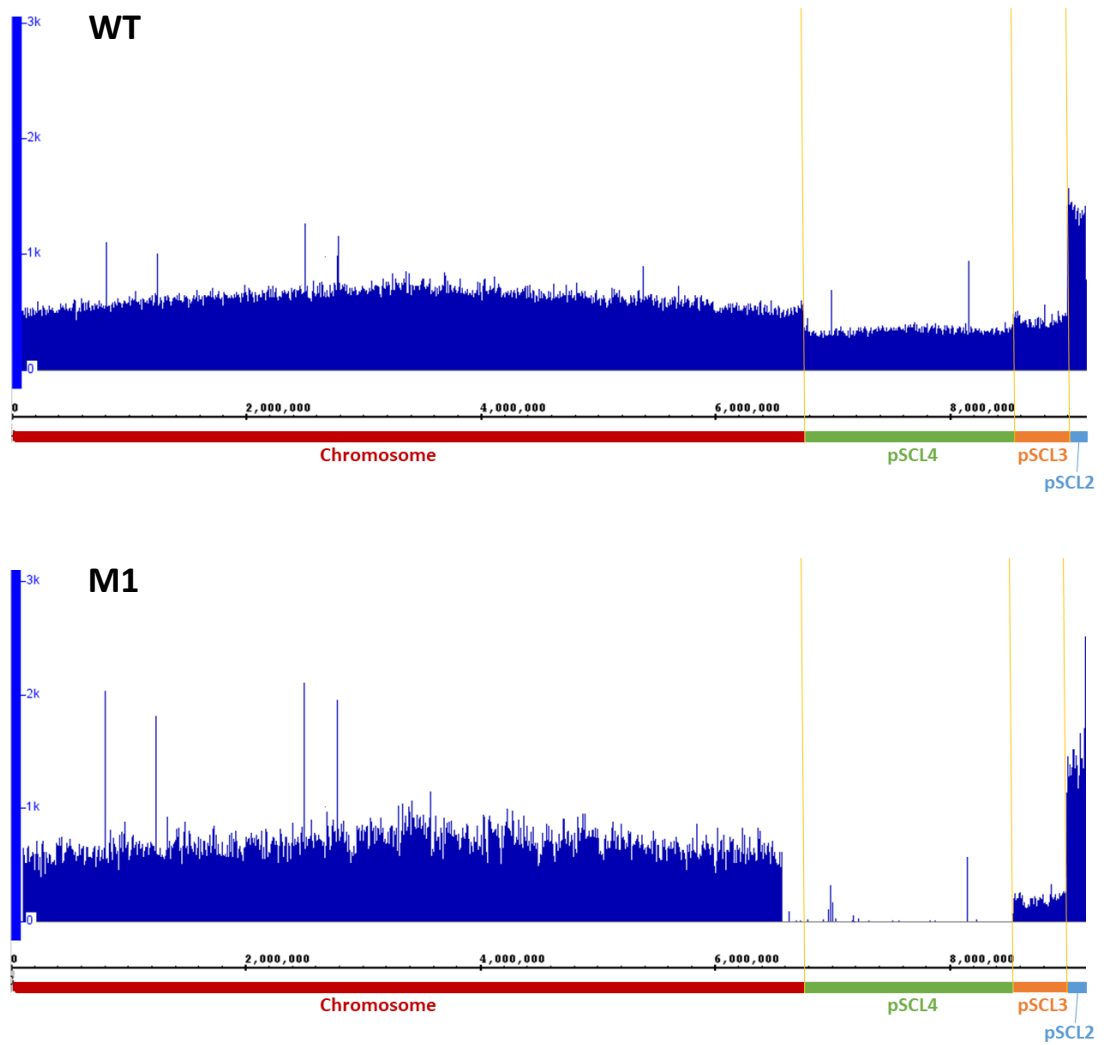


Figure 6.6. Illumina reads of the WT (*S. clavuligerus* DSM 738) and mutant strain M1 mapped on the WT assembly. View from IGB.

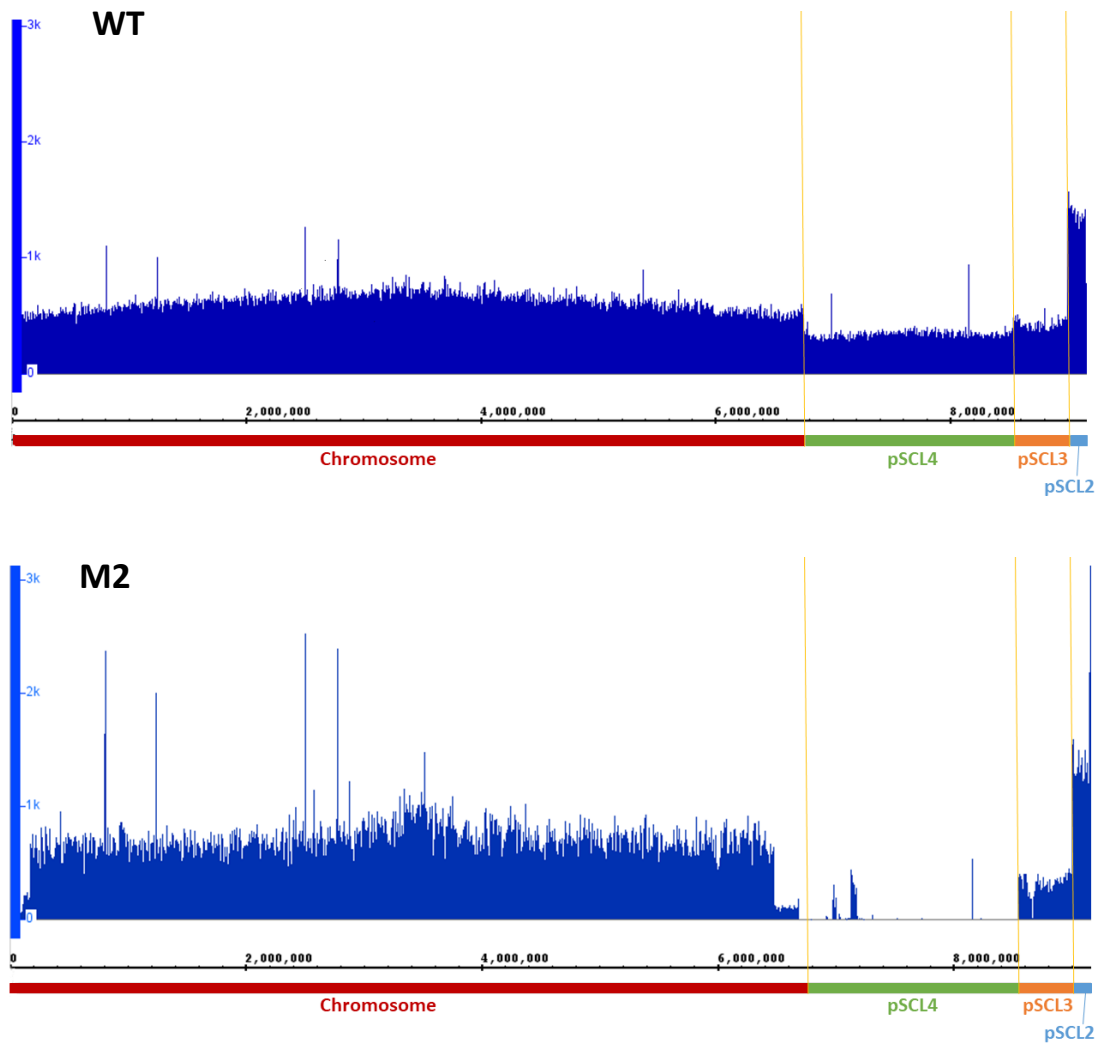


Figure 6.7. Illumina reads of the WT (*S. clavuligerus* DSM 738) and mutant strain M2 mapped on the WT assembly. View from IGB.

In the case of the mutant strain Mc3, Illumina reads aligned to most of the chromosome sequences, with the exception of the sequence ends, in a similar way to the other mutant strain M1 (**Figure 6.8**). In the same way, no reads aligned to sequences exclusive of pSCL4. Only reads that aligned to the *tap-tpg* operon were present this time, as this strain contains an additional copy of pSCL4 *tap* and *tpg* on the chromosome. In addition, very few reads were observed that aligned to pSCL3, and the regions with higher coverage contain genes homologous to those located on the chromosome. In contrast, we obtained reads that covered the whole sequences of pSCL2 and pSCL1. These results confirm the loss of pSCL4 in mutant strain Mc3, as well as the loss of pSCL3 and chromosome ends, while pSCL2 and pSCL1 remain present and conserve their whole sequence.

In order to further assess the loss of the chromosome ends on the mutant strains, we constructed **Figure 6.9**. In this figure, the Illumina reads from the WT, and mutant strains M1, M2 and Mc3 are mapped onto the first and last 300 kb of the WT chromosome sequence. As expected, the WT reads show a full coverage of the chromosome ends. The strain M1 exhibits loss of the first 108 and last 177 kb of the chromosome. In the case of strain M2, reads exhibiting an expected chromosomal coverage depth suggest this strain is missing the first 178 and last 270 kb from the chromosome sequence. However, reads with much lower coverage depth are observable beyond these points, which suggest that either these regions were maintained on a separated replicon, or that the sequenced DNA came from a population of strains with different genotypes. Lastly, the Illumina reads obtained for Mc3 show how this strain is missing the first 55 and last 155 kb of the chromosome sequence. Overall, these results confirm that the three mutant strains are missing different lengths of the chromosomal ends.

Furthermore, read coverage maps and visualisation of BAM files allowed the observation of a difference in coverage for the smallest plasmid pSCL1 of the mutant strains from the WT sequencing data. In order to quantify the difference in read coverage for this and the rest of the plasmids, mean coverage values were obtained individually for each replicon of the three mutants. Additionally, considering mean coverage as an estimation of replicon copy number, the mean coverage values for the chromosome were normalised to 1, and the replicon:chromosome ratios were calculated for the rest of the plasmids. The results are presented on **Figure 6.10**. As it was expected the calculated ratios for the megaplasmid pSCL4 were close to 0 on the three mutant strains (M1: 0.02, M2: 0.03 and Mc3: 0.02).

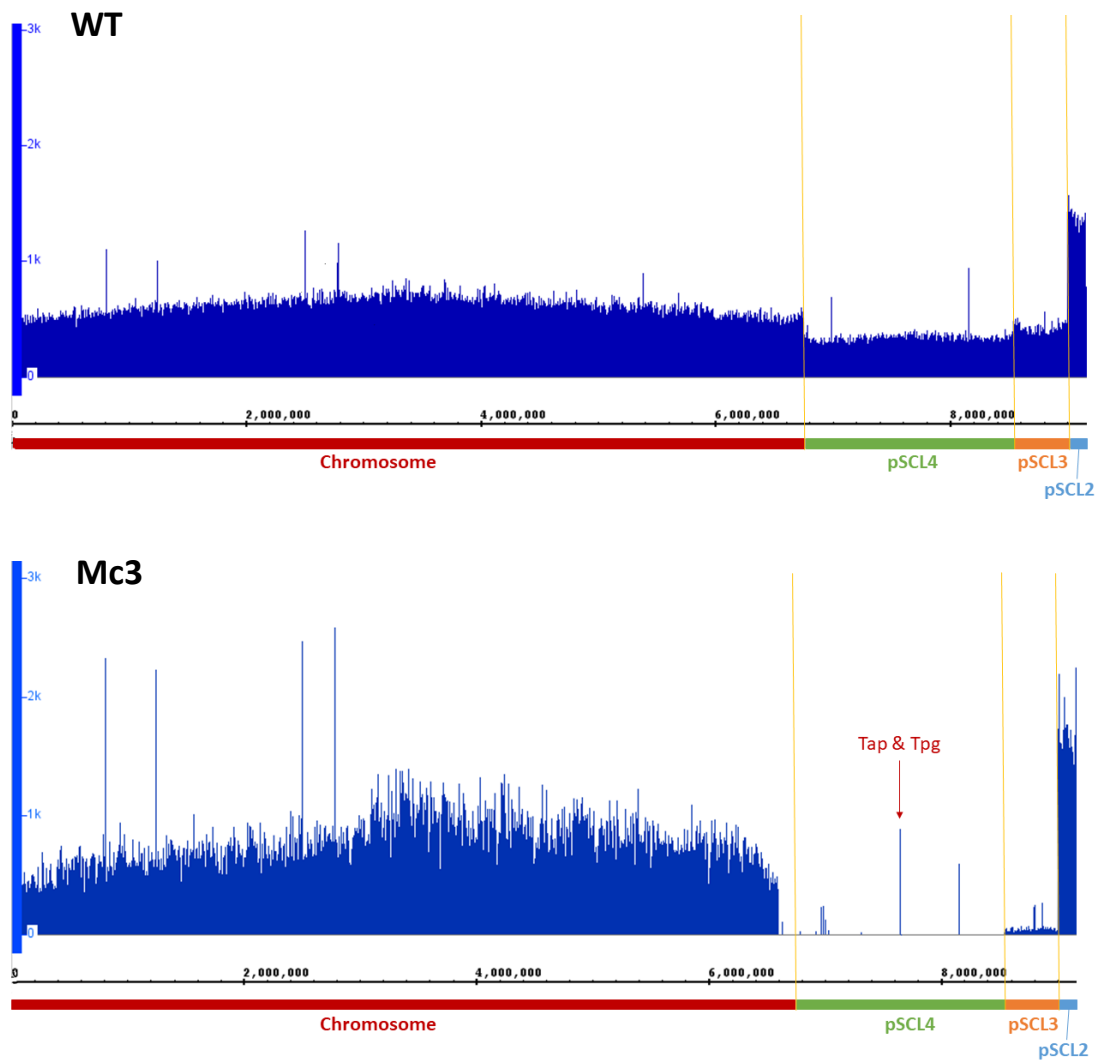


Figure 6.8. Illumina reads of the WT (*S. clavuligerus* DSM 738) and mutant strain Mc3 mapped on the WT assembly. View from IGB.

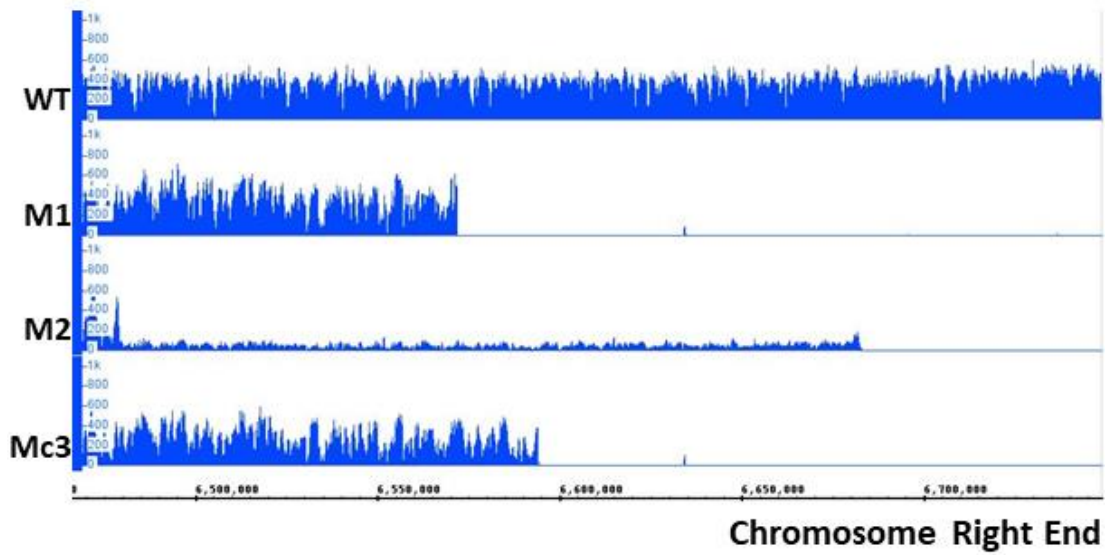
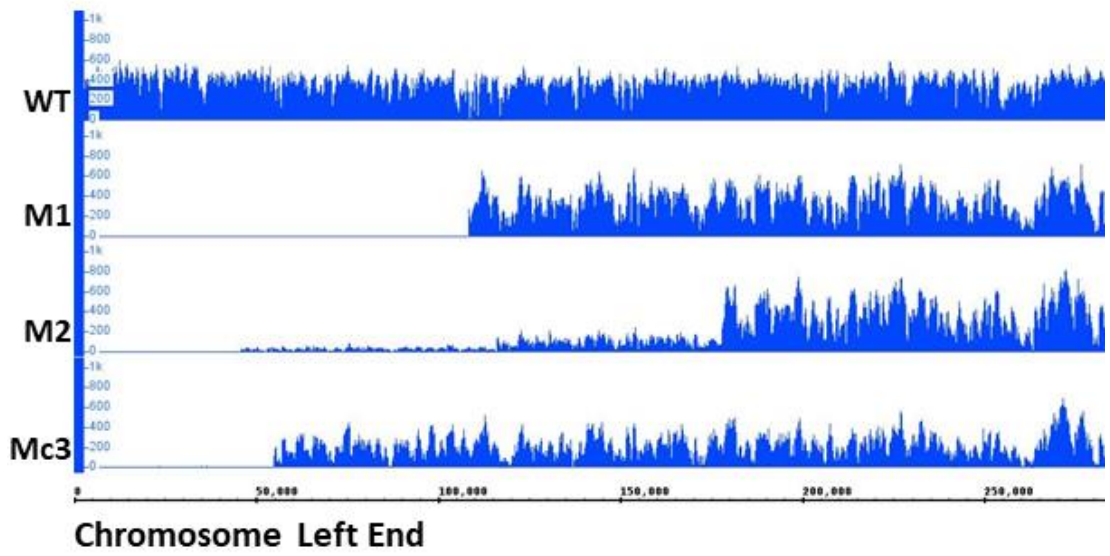


Figure 6.9. Illumina reads of the WT (*S. clavuligerus* DSM 738) and mutant strains M1, M2 and Mc3 mapped on the WT chromosome sequence ends: first 300 kb as left end and last 300 kb as right end. View from IGB.

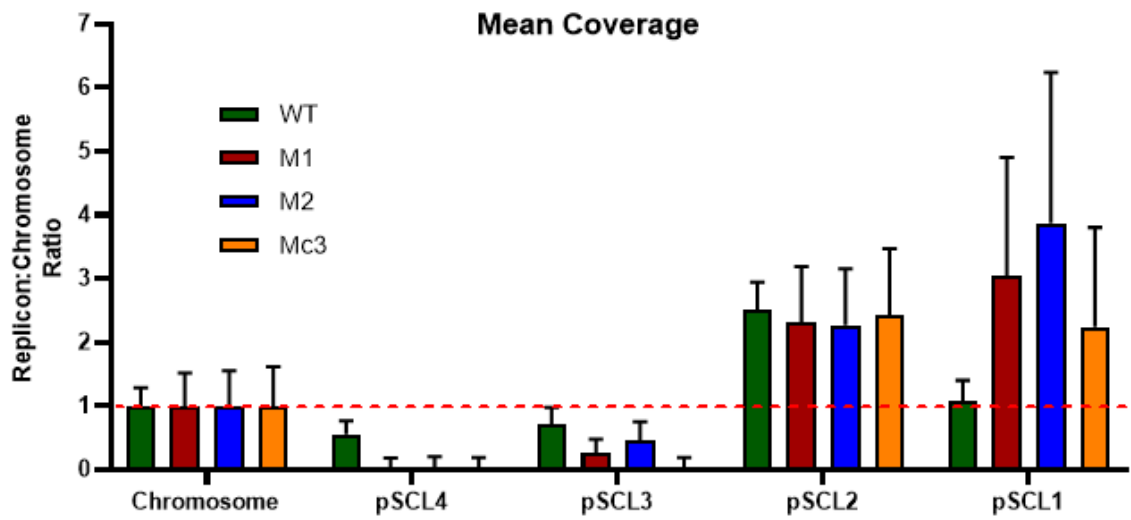


Figure 6.10. Illumina sequencing reads mean coverage for each replicon of the WT strain (*S. clavuligerus* DSM 738) and mutant strains M1, M2 and Mc3, presented as replicon:chromosome ratio.

In the case of pSCL3, the calculated ratios were reduced from 0.7 on the WT to 0.27, 0.46 and 0.04 on M1, M2 and Mc3 respectively. In contrast, the values for pSCL2 on the mutant strains (M1:2.31, M2:2.27 and Mc3:2.43) do not show a substantial variation from the WT (2.5). Interestingly, the calculated ratio for the pSCL1 plasmid has considerably increased on the mutants. Strain M1 shows a value of 3.04, M2 3.87, and Mc3 2.24, in contrast to the WT ratio of 1.07. These results suggest that pSCL1 copy number might have increased 2-, 3- and even 4-fold on the pSCL4-free strains. However, the large variation in coverage depth for pSCL1 implies this could also be due to a flawed sequencing of this replicon.

6.3 Translocation of 50 kb region from pSCL4 to pSCL3 on strain M2.

As mentioned on the previous section, visualisation of Illumina reads coverage maps allowed identification of a region on the pSCL4 sequence where reads obtained from sequencing of the M2 strain aligned. This characteristic was exclusive to the M2 strain. Moreover, the PFGE results presented on section 6.1 show how the M2 sample exhibits a larger band for the pSCL3 plasmid, which suggested the acquisition of <100 kb of material. Considering this, we hypothesise that the 50 kb region from pSCL4 was translocated to pSCL3 during curing pSCL4. In addition, it is observable that the pSCL4-50kb region exhibits a similar read coverage depth to that of pSCL3, as illustrated in **Figure 6.11A**. Also, the fact that the pSCL3 map of M2 reads shows a missing region of around 13 kb on pSCL3 for this strain, suggests that this could be the insertion point which again correlates with our theory.

In order to find evidence of this transposition event, we screened the sequencing data of M2 assembled by Unicycler. Unfortunately, no contig was found to extend to both pSCL4 and pSCL3 sequences. Furthermore, we decided to study these regions at the Illumina read level. It was then observed that reads aligning to the beginning of the pSCL4-50kb region had their mate read at the end of the missing 13kb region of pSCL3. Similarly, reads at the end of the pSCL4-50kb region had their mate read located right before the 13kb missing region of pSCL3. The location of these paired reads is illustrated on **Figure 6.11B**. These results confirm that indeed the 50 kb region from pSCL4 was translocated to that specific region of pSCL3 on the M2 mutant strain.

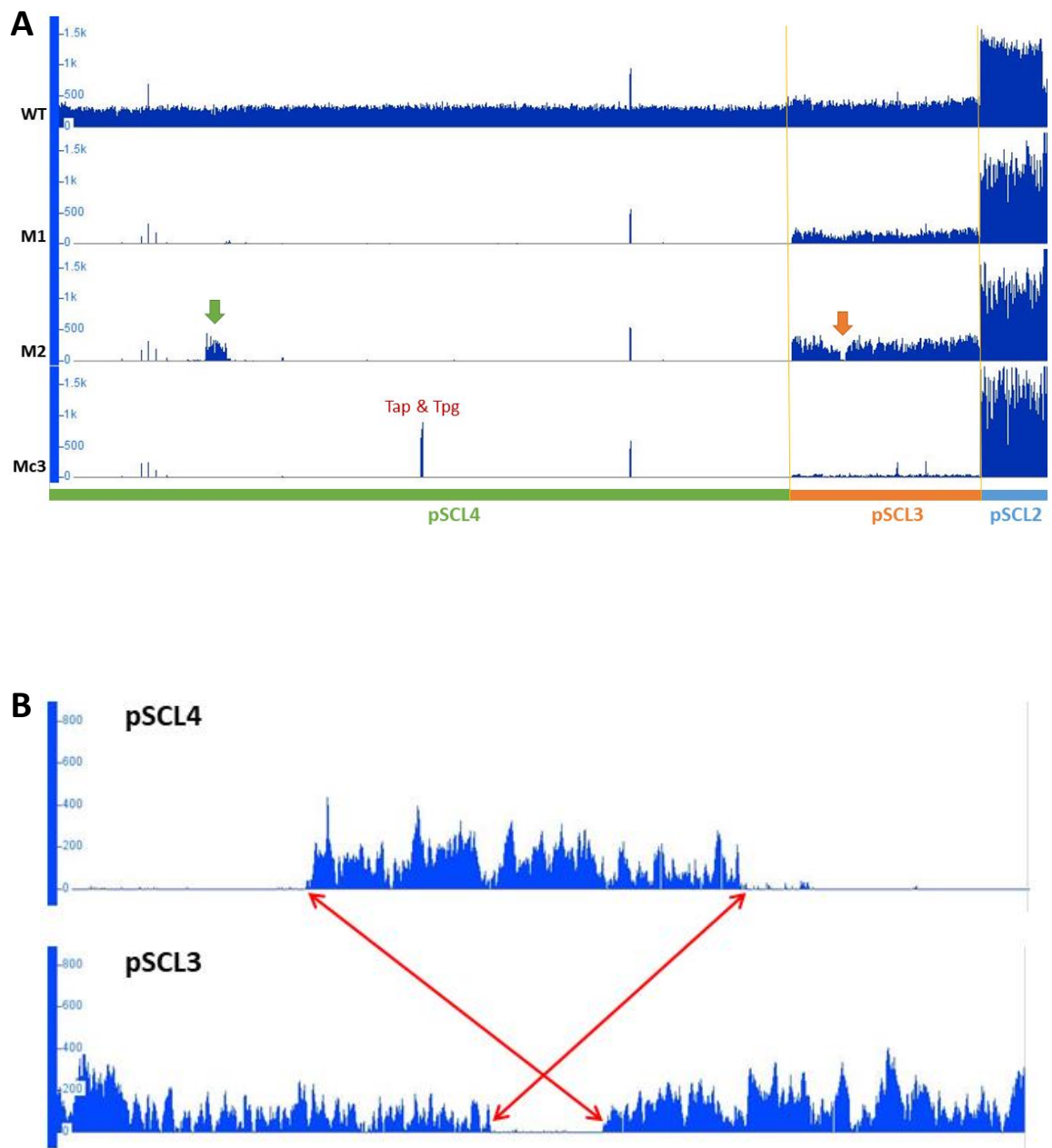


Figure 6.11. Transposition of 50 kb region from pSCL4 to pSCL3 on the M2 strain. **A:** Illumina reads of the WT (*S. clavuligerus* DSM 738) and mutant strains M1, M2 and Mc3 mapped on the WT pSCL4, pSCL2 and pSCL2 sequences. View from IGB. The green arrow indicates the translocated region from pSCL4, and the orange arrow indicates the insertion point in pSCL3. **B:** Reads of the strain M2 mapping to regions of the WT pSCL4 and pSCL3 sequences of the translocated region and insertion point. The red arrows indicate the location of the paired reads

Furthermore, in order to study the translocated genes from pSCL4 to pSCL3, we obtained the annotations of both the translocated pSCL4-50kb region and missing pSCL3-13kb region as presented on **Figure 6.12**. It was confirmed that the translocated region contains genes encoding 50 hypothetical proteins, transcriptional regulator, putative transposase, acetyltransferase, two DNA-binding proteins, aminodeoxychorismate lyase and a small Ras-like GTPase. We also observed that this region includes the biosynthetic gene cluster predicted by antiSMASH as a novel lasso peptide (Cluster 2 in Table 4.7). We were particularly interested on the putative transposase feature, as it could be responsible of the translocation. Protein blast analyses suggested that this protein belongs to the IS360 family of transposases. However, no insertion sites for IS360-type transposition were found on the pSCL3 sequence, or IS360 inverted repeats on pSCL4. Moreover, we observed a 15 kb region of homology between the sequences of pSCL4 and pSCL3 at the translocation points as indicated on **Figure 6.12**. Around 6 kb of these 15 kb of homology are missing from the M2 sequence of pSCL3 and were potentially replaced by the homologous material from pSCL4. These results suggest that the translocation could have occurred by homologous recombination between the plasmids.

6.4 Identification of *attB* site on the *S. clavuligerus* chromosome for ϕ BT1 integration.

On Chapter 5, we described the successful integration of ϕ BT1-based plasmids such as pLIS22 and pMS82, which is mediated by recognition of *attB* site on the chromosome. However, the *attB* site for ϕ BT1-integration on *S. clavuligerus* is unknown. The mutant strain Mc3 originates from the *tap-tpg*-duplicated strain SCLA22, which carries pLIS22 on the chromosome, and so, in order to locate this integration point, we screened the sequencing data obtained for Mc3. Sequences matching to the pLIS22 plasmid were found on two contigs obtained from assembling the Mc3 reads using Unicycler, these are contigs 363 and 136. These two contigs extend each side of pLIS22 to the chromosome sequence, which allowed construction of the map containing the plasmid integrated on the chromosome (**Figure 6.13A**).

In addition, characterisation of the integration point allowed us to identify the *attB* site on *S. clavuligerus*. The site is composed of the following sequence of 9 nucleotides: GGTGACCCA,

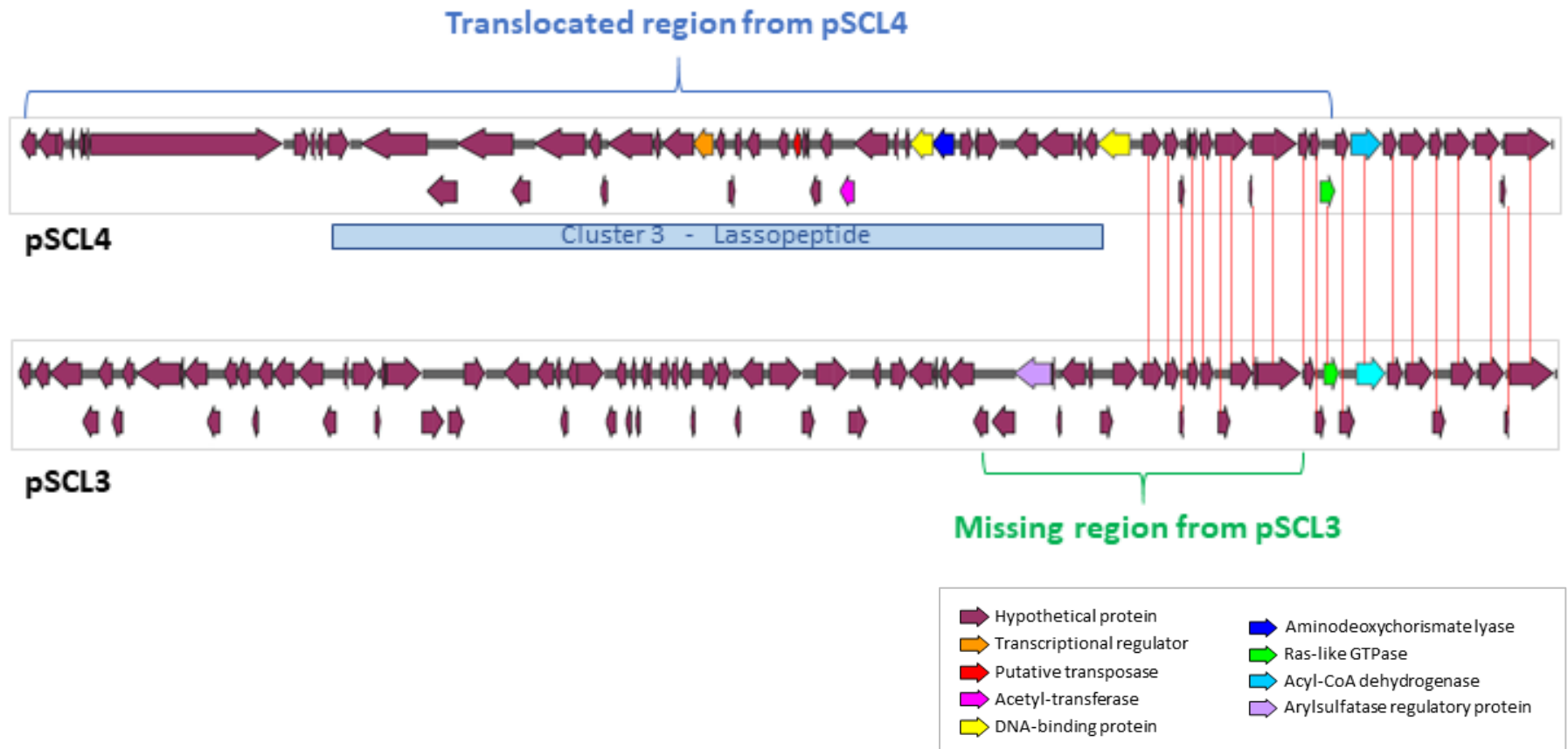


Figure 6.12. Annotated regions of the WT sequences of pSCL4 (377,996 – 437,482 bp) and pSCL3 (108,590 – 167,590 bp, inverted) where the translocation event happened. The homologous features are linked by red lines.

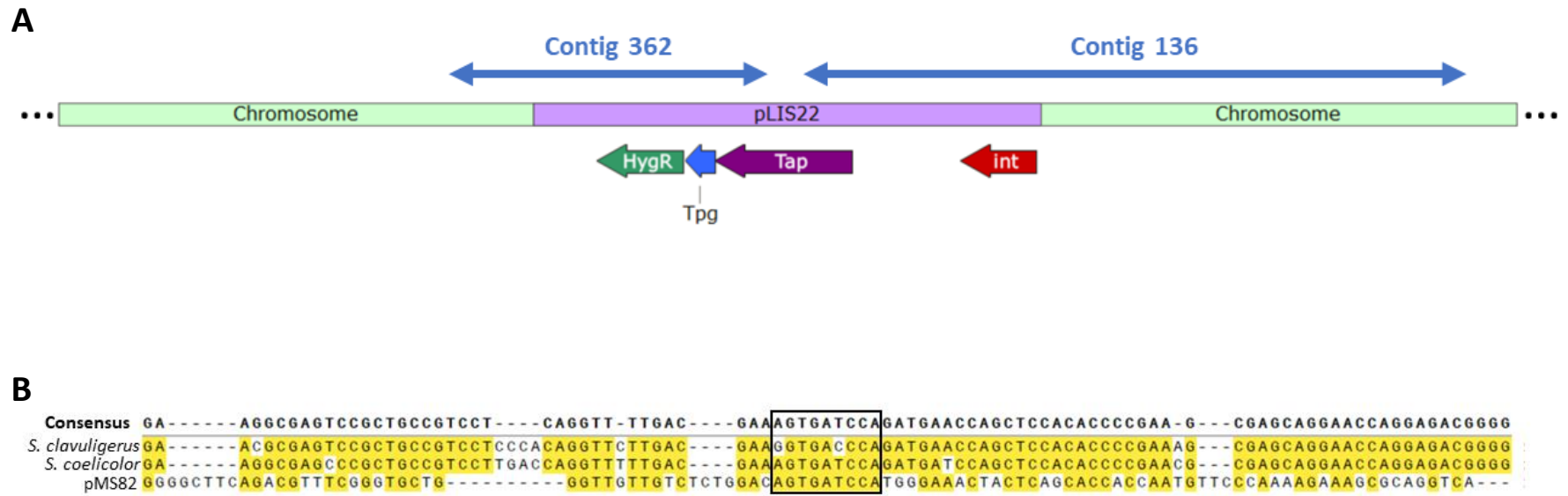


Figure 6.13. Identification of *attB* sites on *S. clavuligerus* chromosome for ϕ BT1-integration. **A:** Map of the integration of construct pLIS22 on the chromosome obtained by mapping of the Mc3 contigs 363 and 136. **B:** Alignment of the integration points on the *S. clavuligerus* and *S. coelicolor* chromosome and the pMS82. The *attB/P* sites are indicated by the black box.

and the plasmid integrated between the first and second guanine nucleotides. This sequence was also identified by aligning the integration point to the *S. coelicolor* and pMS82 sequences carrying the *attB* and *attP* site respectively that have the following sequence: AGTGATCCA (**Figure 6.13B**). Furthermore, the high similarity of these regions between *S. clavuligerus* and *S. coelicolor* evidences the conservation of the ϕ BT1-integration site between different *Streptomyces* species.

In summary, with these results we report identification of the *attB* sites on the *S. clavuligerus* chromosome for integration of ϕ BT1-based plasmids.

6.5 Strains M1 and M2 carry circular chromosomes.

Considering the three mutant strains M1, M2 and Mc3 are missing the chromosome ends, we decided to study the possibility of the circularisation of the chromosome induced by the loss of the megaplasmid. Firstly, we examined the first and last sets of Illumina sequencing reads that mapped to the WT chromosome sequence. In the case of the M1 strain, we observed that some of the first reads (left arm) had their paired mate read located towards the right end, as illustrated on **Figure 6.14A**, which suggest circularisation of the chromosome at that point. When investigating the data for the M2 strain, we found a similar event where reads located on the left arm end have their mate read towards the right arm end. However, we observed two specific locations where this occurred on the M2 data. The first location links reads at the beginning of a low read coverage region, and 282 kb upstream of the end of reads (presented in yellow in **Figure 6.14A**). The second location links reads at the start and end of regions with a similar read coverage depth to the rest of the chromosome sequence (presented in green in **Figure 6.14A**). This suggests that the sequenced data could have belonged to mixed populations of *S. clavuligerus* mutants carrying chromosomes that circularised at different points. Given the difference in read coverage, we consider the one exhibiting the highest coverage depth as more abundant, and so, the one we will regard as circularisation point for the sequence of strain M2. In contrast to M1 and M2, no linkage between reads at both arms was found for the Mc3 strain. These results show that the raw Illumina sequencing data suggest that chromosomes of M1 and M2 are circular. The predicted junction and size are illustrated on **Figure 6.14B**.

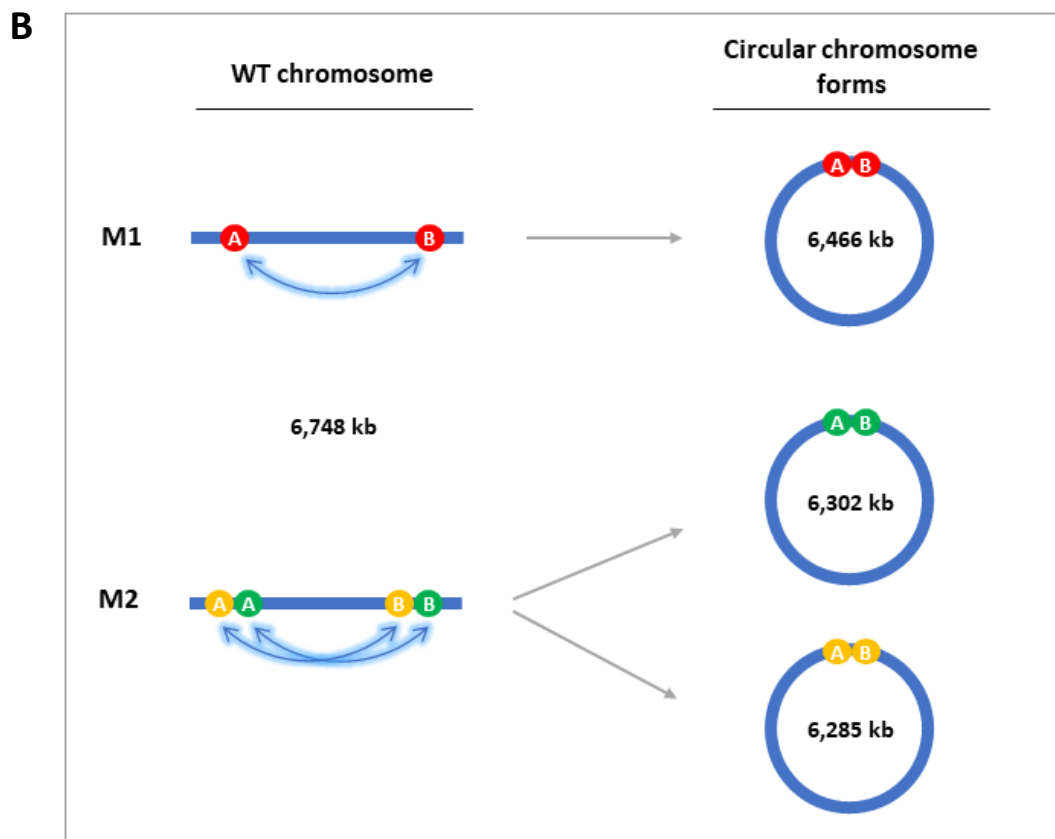
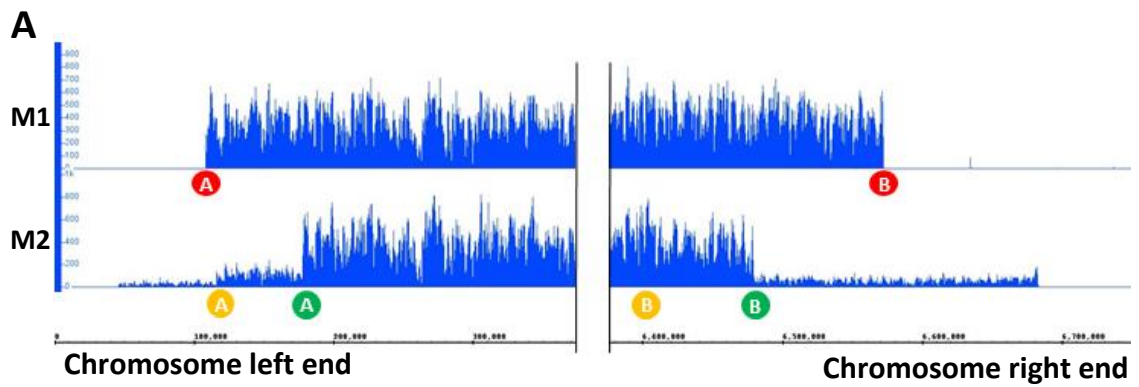


Figure 6.14. Chromosome circularisation on strains M1 and M2 **A:** Illumina reads from M1 and M2 mapped to the WT chromosome ends. The letters represent the position of paired reads: reads at A have their mate reads at B. M2 data shows two points of circularisation (green and yellow). **B:** Schematic representation of predicted chromosomal circularisation occurred at those specific points. Not at scale.

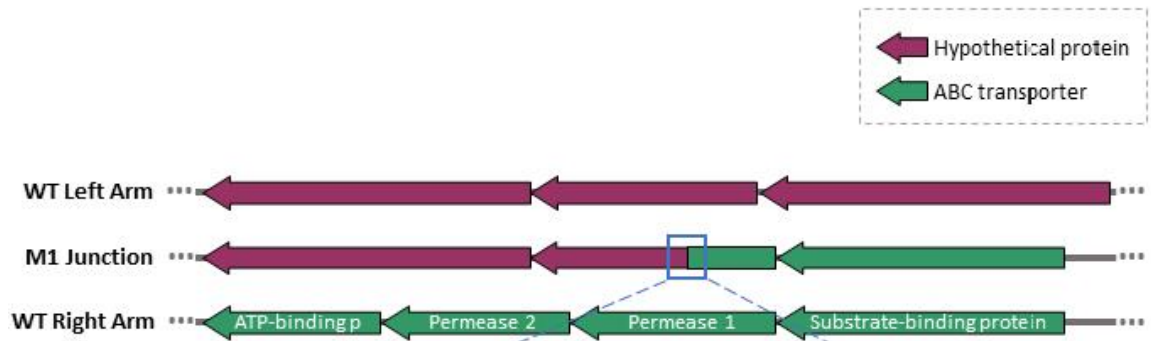
Furthermore, we searched for chromosomal circularisation evidence on the assembled genome data of the three strains. In the case of M1, we found a contig (contig 489) that extended to both the left and right arm ends of the chromosome. A circular version of the *S. clavuligerus* chromosome was constructed using this contig 489 to fuse both arms. This was used to compare to the original junction points as presented on **Figure 6.15A**. The fusion occurred at a gene encoding a hypothetical protein from the left arm and a gene encoding an ABC transporter permease from the right arm of the chromosome. In addition, in order to study the junction from a nucleotide level, we constructed an alignment with the original left and right arm sequences and the contig 489 sequence. The alignment showed an overlapping region of 6 nucleotides (GTACGC) at the fusion point (**Figure 6.15B**). Given the short overlap, we excluded the idea of circularisation by homologous recombination.

For the case of the M2 strain, we identified contig 446 that aligned to both the left at right arm of the chromosome. Similarly to the M1 sequence, we constructed a circular chromosome form using this contig and compared to the original sequences at the junction point. As presented on **Figure 6.16A**, the fusion happened between a gene encoding a putative cytochrome P450 hydroxylase at the left arm, and a gene encoding the B chain of a potassium-transporting ATPase at the right arm. An alignment was also constructed for this strain at the junction. In contrast, no homology was identified at the fusion point for this strain (**Figure 6.16B**). In the case of the Mc3 strain, no contigs were identified to extend the left and right arm of the chromosome.

Moreover, in order to obtain physical evidence of the chromosome circularisation, we performed a PCR analysis to amplify the circularisation junction. Firstly, primers were designed so they would anneal towards the ends of the predicted sequences of M1, M2 and Mc3 following divergent orientations (**Figure 6.17A**). The primers were named: **M1-junction-F/R**, **M2-junction-F/R** and **Mc3-junction-F/R** (Table 2.5 in Chapter 2). PCR amplification was then performed from genomic DNA of the three mutant strains using the described primers and the PCR products were separated by gel electrophoresis. The gel showed bands for the circularisation junction of M1 and M2 with the expected sizes of 1,1 kb and 0.88 respectively (**Figure 6.17B**). In contrast, no bands were observed for junction of the chromosome arms on the Mc3 strain.

To summarise, with these results we confirm that the M1 and M2 strains carry circular chromosomes. In contrast no evidence was found for chromosomal circularisation in Mc3.

A



B

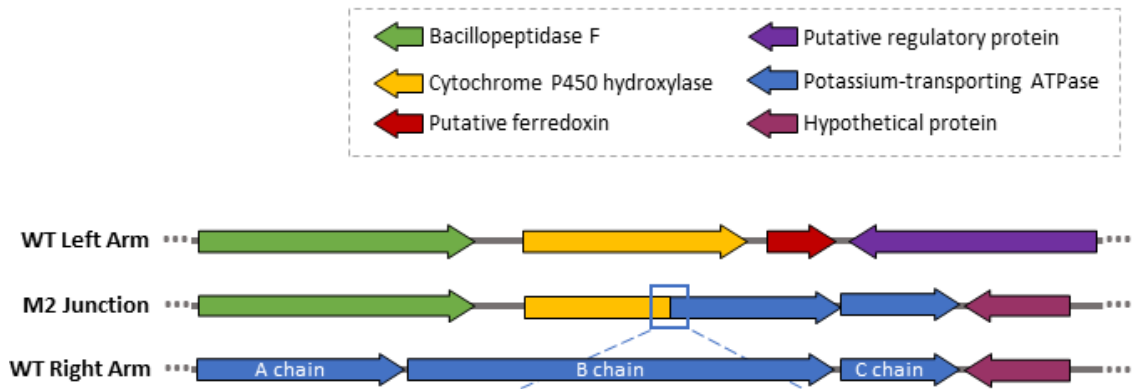
Consensus	TGCCGCGCCGGGCCCCCGCGGGTGCCGTCCGGCGGAAGGACGGGTGCCGTGGCGCGTGGTGG
Left Arm	TGCCGCGCCGGGCCCCCGCGGGTGCCGTCCGGCGGAAGGACGGGTGCCGTGGCGCGTGGTGG
Contig 489	TGCCGCGCCGGGCCCCCGCGGGTGCCGTCCGGCGGAAGGACGGGTGCCGTGGCGCGTGGTGG
Right Arm	GCCCCCGCGGGGCAAGGACGCCAGCTCAGGGAGAAGAACCAGATGGCGAGCAGCCCCAG

Consensus	TCTGCGGACGCACGGCGGAGGACGGGGGTACGGGGCACCGGAGGTCGCGCGGTGGAACCG
Left Arm	TCTGCGGACGCACGGCGGAGGACGGGGGTACGGGGCACCGGAGGTCGCGCGGTGGAACCG
Contig 489	TCTGCGGACGCACGGCGGAGGACGGGGGTACGGGGCACCGGAGGTCGCGCGGTGGAACCG
Right Arm	CCAGAAAGACGGGGCGGGCTTCGAGGGCGTACGGGGCACCGGAGGTCGCGCGGTGGAACCG

Consensus	CCCGCCCCGGCGGGCGGGCGAGGACGCCGAGCGCGGTGCCGAGGACGACGGCGAGGAG
Left Arm	CGGGGGAGTGCCTGCGAGCCGTGCGACGGGGCGGGACCCGGTGGCACGACGGGAGGC
Contig 489	CCCGCCCCGGCGGGCGGGCGAGGACGCCGAGCGCGGTGCCGAGGACGACGGCGAGGAG
Right Arm	CCCGCCCCGGCGGGCGGGCGAGGACGCCGAGCGCGGTGCCGAGGACGACGGCGAGGAG

Figure 6.15. Circularisation junction of M1 chromosome **A:** Graphic representation of 5 kb region where chromosomal circularisation occurred in strain M1, and the same regions on the WT chromosome. **B:** Alignment of 180 bp region of circularisation junction in M1 (Contig 489) and the original sequences from WT chromosomal left and right arms. The black box indicates the fusion point with 5 bp of overlap. Consensus is highlighted in yellow.

A



B

```

Consensus  ACACGGACC6GA6CC6GTTCCAG6GGGTACACC6CC6C6CTCAG6GG6GTCGAC6TGACCT
Left Arm   ACACGGACC6GA6CC6GTTCCAG6GGGTACACC6CC6C6CTCAG6GG6GTCGAC6TGACCT
Contig 446 ACACGGACC6GA6CC6GTTCCAG6GGGTACACC6CC6C6CTCAG6GG6GTCGAC6TGACCT
Right Arm  CGACATCGACGGCCGCAGGGTCCGCAAGGGCGCGGGCGCGGGCCGTCATCGGCTGGGTGGG

Consensus  TGGAGGAGCAGATCGCCGCCATGGCGGACCGCGCGCGCGCGTGGACGCCGTCGCGGCCGA
Left Arm   TGGAGGAGCAGATCGCCGCCATGGCGGACCGCTCTCCTCGTACATCCAAGAACTGATCACCG
Contig 446 TGGAGGAGCAGATCGCCGCCATGGCGGACCGCGCGCGCGCGTGGACGCCGTCGCGGCCGA
Right Arm  CGCGCGCGGGCGGCGGGTGGACGAGGCGGTGCGCGCGCGCGTGGACGCCGTCGCGGCCGA

Consensus  GGGCGGCACCCCGCTGGCGGTGGCCGTCGAGGACGGCGAAGGCGCCCGGGTGCTGGGGGT
Left Arm   CCAAGCGGAAGCGCCACCGACGACCTGCTCAGCGACCTCGTCAACGACGGCGGGTTCA
Contig 446 GGGCGGCACCCCGCTGGCGGTGGCCGTCGAGGACGGCGAAGGCGCCCGGGTGCTGGGGGT
Right Arm  GGGCGGCACCCCGCTGGCGGTGGCCGTCGAGGACGGCGAAGGCGCCCGGGTGCTGGGGGT
  
```

Figure 6.16. Circularisation junction of M2 chromosome **A:** Graphic representation of 5 kb region where chromosomal circularisation occurred in strain M2, and the same regions on the WT chromosome. **B:** Alignment of 180 bp region of circularisation junction in M2 (Contig 446) and the original sequences from WT chromosomal left and right arms. The black line indicates the fusion point. Consensus is highlighted in yellow.

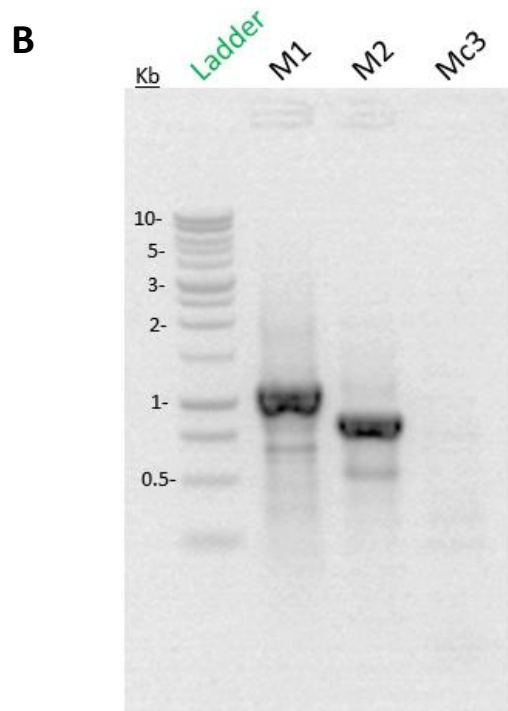
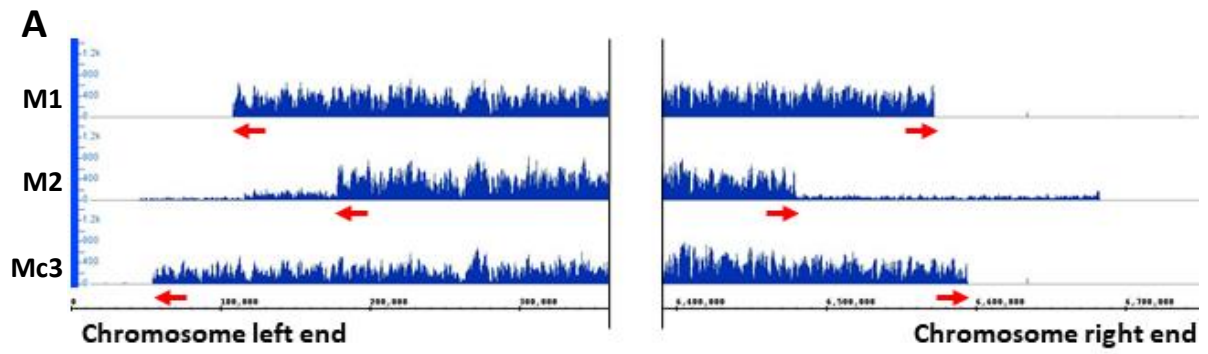


Figure 6.17. PCR analyses confirm chromosomal circularisation in M1 and M2 **A:** Illumina reads from M1, M2 and Mc3 mapped to the WT chromosome ends. The red arrows indicate the location and orientation of the six primers used. **B:** Gel electrophoresis of the PCR results from amplification of circularisation junction of M1, M2 and Mc3.

6.6 Identification of the missing annotations on the pSCL4-free strains.

Considering the loss of pSCL4, chromosome ends, and in the case of strain Mc3, pSCL3, many coding sequences are missing from these mutant strains which can have important consequences for the physiology of *S. clavuligerus*. In order to assess the missing genes on the three strains, we first constructed the estimated genome sequences of M1, M2 and Mc3 by excluding the regions that have no read coverage such as the megaplasmid and chromosome ends. These sequences were then annotated using RAST, and the annotation subsystem were obtained for each strain. These were compared to the WT annotations at both the chromosome level or whole genome, and the number of missing subsystems were obtained and presented on **Table 6.1**. As observed on the table, the majority of missing subsystems are due to the loss of the megaplasmid and few of them are missing from the chromosome.

Furthermore, we established a list of the annotations located on the missing chromosomal regions of the three strains in order to identify relevant chromosomal genes that these strains are lacking. The strain M1 lacks a total number of 187 genes from the original chromosomal ends, 92 of them were annotated as hypothetical proteins. M2 is missing 338 of the WT chromosome genes, 173 of them are hypothetical proteins. Mc3 lacks 133 chromosomal genes, of them 54 are hypothetical proteins. The complete list missing genes with a known function is shown on **Table A1** in Appendix. We identified four transcriptional regulator genes belonging to different families: an AfsR regulator, a MarR regulator, an AsnC regulator and a LysR-type regulator, which are missing in the three pSCL4-free strains. In addition, M2 is also missing an AraC regulator, another LysR regulator and two AcrR-type regulators. Furthermore, M1 and M2 also lack a transcriptional regulator homologous to a SARP regulator in *Streptomyces sparsogenes*, as confirmed by NCBI BLAST. These types of transcriptional regulators are often associated with the regulation of secondary metabolism and aerial mycelium development in streptomycetes. In addition, in order to confirm if these missing chromosomal genes in the pSCL4-free strains contain any of the genes from the core genome of *S. clavuligerus* that was established on Chapter 4, we created a BLAST database using the amino acid sequences belonging to the core genome, and subsequently performed BLAST searches using the annotations of the missing regions as query. These analyses identified the same two amino acid sequences on the missing regions of the three strains, which were identified as a hypothetical protein and an MFS (major facilitator superfamily) transporter by NCBI Protein BLAST. In addition, six more proteins were identified on the

missing regions from the M2 strain that includes proteins of the following families: metallophosphoesterase, electro transfer flavoprotein, NADPH:quinone oxidoreductase, acyl-Coa dehydrogenase and another MFS transporter.

These results served as an estimation of the differences between the annotations of the WT and the pSCL4-free strains and they will be contrasted to the physiological observations of these strains. In order to characterise the physiology of these mutant strains we decided to further investigate their growth, morphology and production of clavulanic acid in contrast to the WT.

Table 6.1. List of estimated RAST annotation subsystems missing on the mutant strains M1, M2 and Mc3 in reference to the WT (*S. clavuligerus* DSM 738). Chrom: number of missing subsystems from the chromosome exclusively. Total: number of missing subsystems from the whole genome.

Annotations Subsystems	M1		M2		Mc3	
	Chrom	Total	Chrom	Total	Chrom	Total
Amino acids and derivatives	10	53	15	55	10	59
Carbohydrates	9	37	14	42	9	39
Protein metabolism	0	23	0	23	0	25
Cofactors, vitamins, prosthetic groups, pigments	0	18	0	17	0	22
Fatty acids, lipids and isoprenoids	11	29	14	32	10	30
Respiration	0	5	2	7	0	5
Nucleosides and nucleotides	0	2	3	5	0	3
DNA metabolism	0	0	1	0	0	6
Membrane transport	0	23	1	24	0	24
Stress response	2	2	3	3	2	2
RNA metabolism	0	1	1	2	1	1
Virulence, disease and defence	0	8	1	9	0	8
Phosphorus metabolism	0	8	0	8	0	8
Iron acquisition and metabolism	0	0	0	0	0	0
Cell wall and capsule	2	22	2	22	2	22
Regulation and cell signalling	0	4	0	4	0	4
Metabolism of aromatic compounds	2	6	2	6	2	6
Secondary metabolism	0	0	0	0	0	1
Potassium metabolism	0	0	1	1	0	0
Dormancy and sporulation	0	1	1	2	0	1
Nitrogen metabolism	0	0	0	0	0	0
Sulphur metabolism	0	12	0	12	0	12
Phages, prophages, transposable elements	0	2	0	2	0	2
Miscellaneous	0	4	1	5	0	4

6.7 Characterisation of the growth of plasmid-free strain in liquid cultures.

In order to study the influence of the megaplasmid on the growth of *S. clavuligerus*, we obtained growth curves for the pSCL4-free strains M1, M2 and Mc3, as well as the WT strain (*S. clavuligerus* DSM 738) in liquid cultures. For this, the four strains were pre-grown for 30 hours in TSB media. The samples were then standardised to the same OD₆₀₀ and inoculated in 24-well plates with TSB media. The strains were incubated for 72 hours, and OD₆₀₀ measurements were taken every 30 minutes. The Napierian logarithm of the OD₆₀₀ values were calculated and plotted against time to obtain the growth curves (**Figure 6.18A**). The graph shows that all the *S. clavuligerus* strains follow a typical growth curve profile, with an exponential or log phase during the first 20 hours, and a stationary phase the following hours. The mutant strains M1 and Mc3 presented a very similar profile to the WT throughout the three days of culture, and in particular, at the exponential phase where they presented almost the same OD₆₀₀ values. On the other hand, strain M2 exhibited a delayed start of exponential phase, reaching a linear correlation at around 10 hours from inoculation. However, M2 presented a similar stationary phase to the other strains.

Furthermore, the Specific Growth Rate (μ) was calculated to further quantify the fitness of the strains. For this, we selected the time points exhibiting a linear correlation during log phase for each strain and calculated the slope value (μ). This resulted in the following values: 0.31, 0.34, 0.41 and 0.33 for WT, M1, M2 and Mc3 respectively (**Figure 6.18B**). Statistical analyses confirmed that M2 had a significantly increased μ value in reference to the WT growth rate. In contrast, no significant difference was confirmed between the growth rate of M1, Mc3 and WT.

These results suggest that the loss of the megaplasmid does not have a negative effect on the vegetative growth of *S. clavuligerus*, and it can even increase the growth rate as it is the case of the M2 strain.

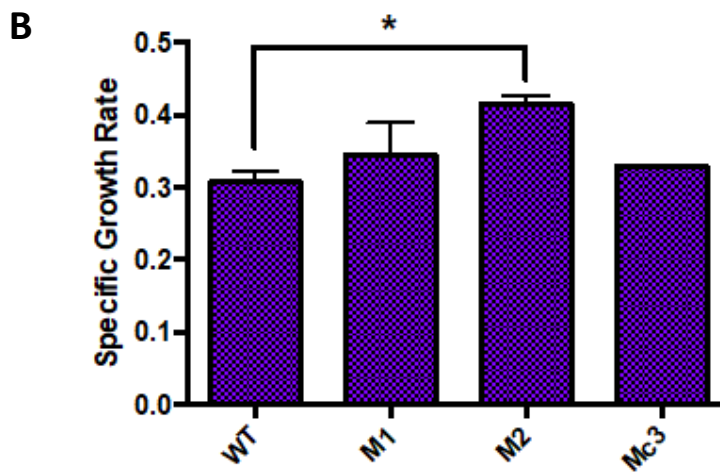
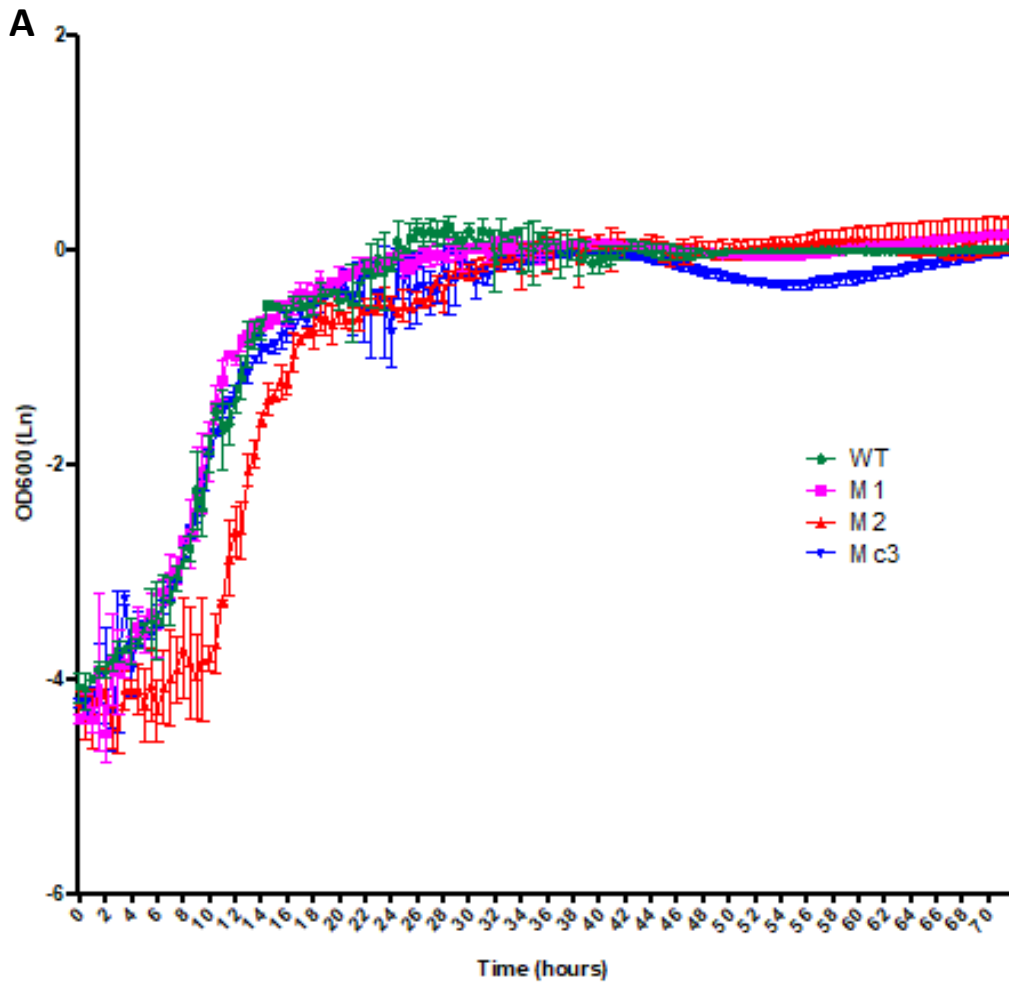


Figure 6.18. Growth of pSCL4-free strains in liquid cultures. **A:** Growth curves of *S. clavuligerus* strains WT (DSM 738), M1, M2 and Mc3 in TSB media. OD₆₀₀ was recorded for 72 hours every 30 minutes. **B:** Specific growth rate (μ) of each strain. μ was calculated using the time points at log phase that follow a linear correlation. Error bars represent the standard deviation. *: $p < 0.05$.

6.8 Microscopic characterisation of pSCL4-free strain confirms they present shorter, thinner, and more frequent branches.

While the growth curves data suggests that the loss of the megaplasmid does not compromise the growth of *S. clavuligerus* in liquid media, solid cultures of these mutant exhibited different morphologies (**Figure 6.19**). The strains M1 and Mc3 showed decreased aerial hyphae formation, while M2 presented abundant aerial hyphae when grown on *S. clavuligerus* sporulation media (L3M9) for 3 weeks. However, none of the mutants exhibited sporulation, in contrast with the WT that showed ample spore formation on this media. In order to further characterise the morphology of the pSCL4-free strains, we obtained microscopic images of the three mutant strains and the WT using an inverted widefield epifluorescence microscope. For this, the four strains were grown on glass coverslips inserted in L3M9 agar for 5 to 35 days. The cells were then fixated and stained using the Schwedock Staining procedure, which uses fluorescein isothiocyanate-wheat germ agglutinin (FITC-WGA) to dye the peptidoglycan, and propidium iodine to dye the nucleic acids. Phase contrast images were taken, as well as fluorescence images using a FITC emission filter that allows visualisation of stained peptidoglycan in green, and a 600 nm emission filter to visualise the nucleic acids in red. 10 images were taken of each sample grown for 5, 8, 15 and 35 days. The images channels were split, and the fluorescence green and red channels were merged in order to obtain the phase contrast and the fluorescence images separately.

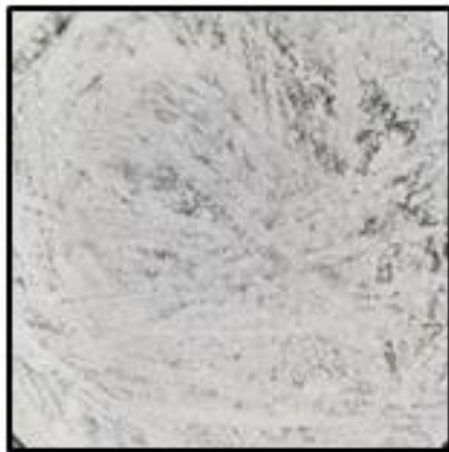
The images obtained from the WT sample grown for 5 days, exhibited long, hyphae that were often grouped in parallel and they branched into new hyphae of similar width and variable length. In contrast, the M1 sample presented generally short and curled branches that are considerably thinner than the main hyphae. In the case of the M2 sample, the hyphae presented short branches that occurred substantially more frequent than on the parent strain. Similarly to M1, the Mc3 sample presented thin and curled branches in contrast to the WT. Examples of images taken from samples growth for 5 days are illustrated on **Figure 6.20**. The fluorescence images allowed visualisation of the peptidoglycan wall in green, which was brighter on the younger branches, while on older hyphae the nucleoids presented in red were more visible.



WT



M1



M2



Mc3

Figure 6.19. The four strains of *S. clavuligerus*: WT (DSM 738), M1, M2 and Mc3 grown on L3M9 media. Images taken after 3 weeks of inoculation.

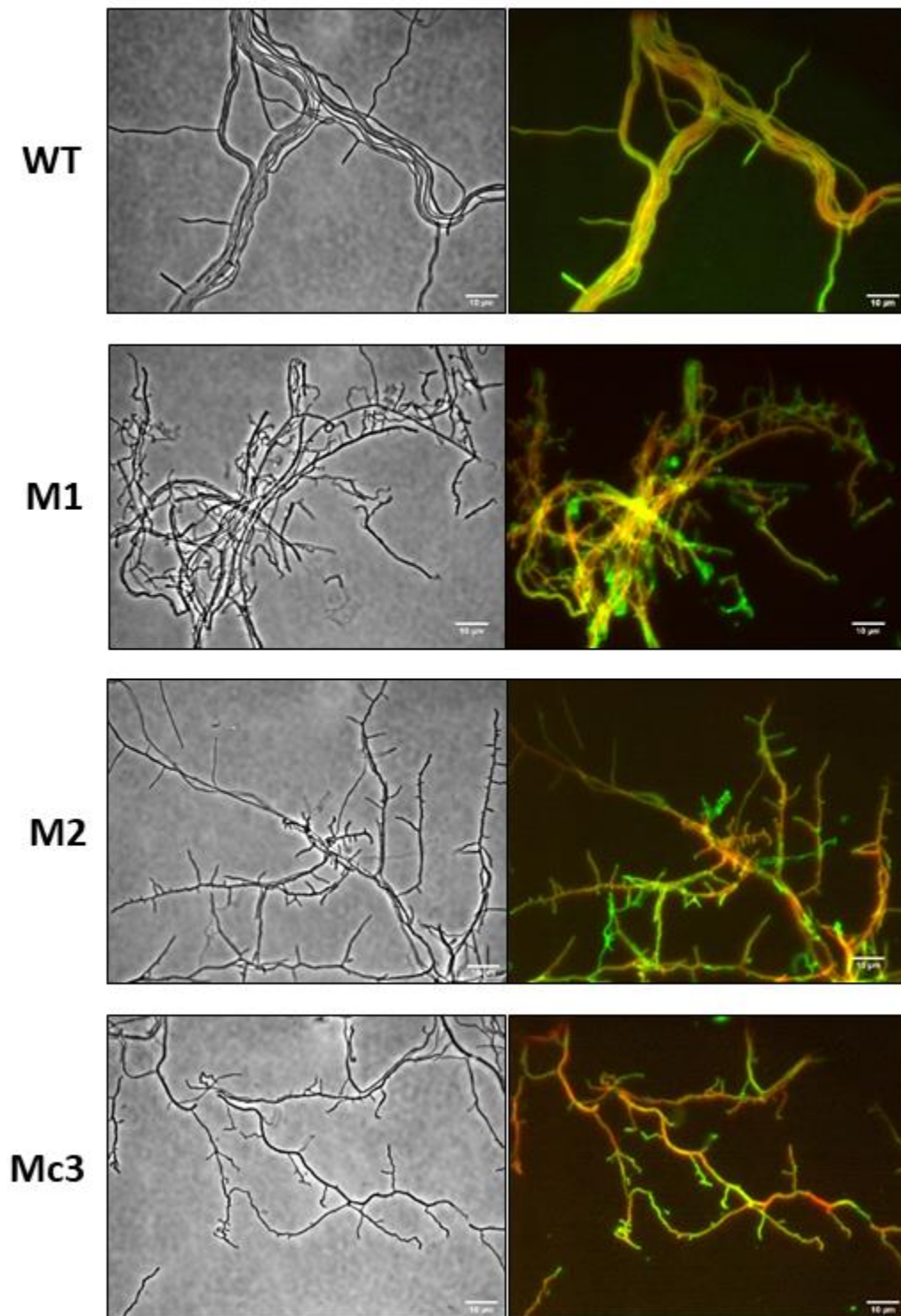


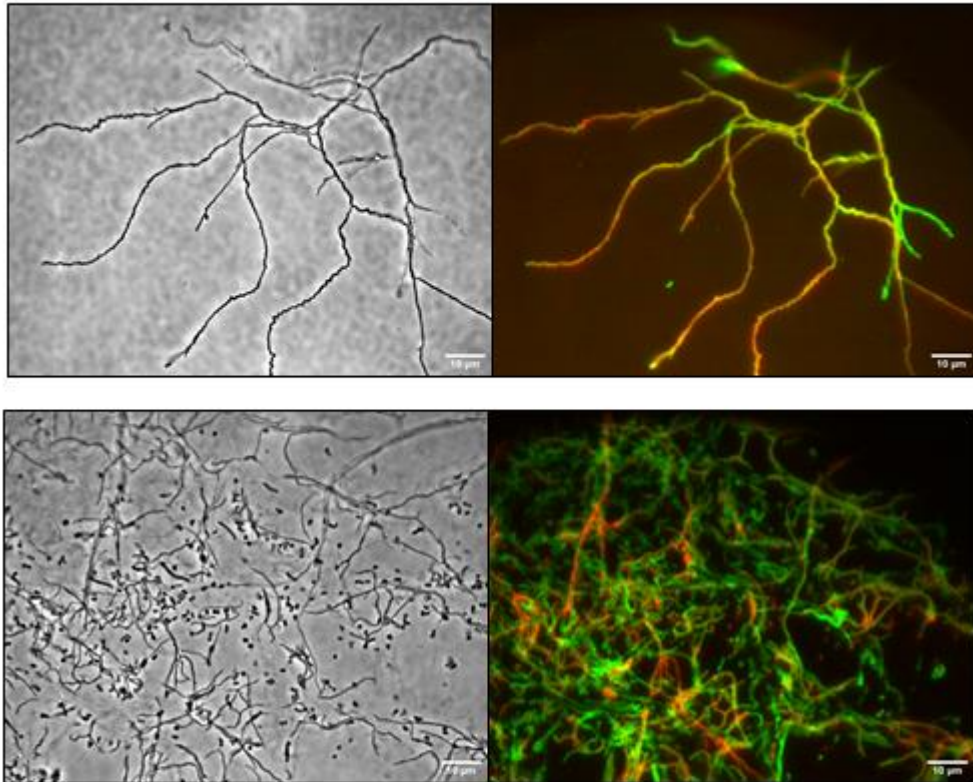
Figure 6.20. Examples of microscopy images of strains WT (*S. clavuligerus* DSM 738), M1, M2 and Mc3 grown on L3M9 agar for 5 days. Images obtained with an inverted widefield epi-fluorescence microscope. Phase contrast images are on the left. Fluorescence images are on the right; the peptidoglycan stained with fluorescein isothiocyanate-wheat germ agglutinin (FITC-WGA) is presented in green under a FITC emission filter; nucleic acids stained with propidium iodide and are presented in red under a 600 nm emission filter. Scale bar: 10μm.

The samples of WT mycelium grown for 8 and 15 days presented similar looking hyphae, with the addition of curled hyphae and some sporulation on the sample from day 15 (**Figure 6.21**) At day 35 the WT mycelium exhibited less integrity and abundant sporulation. In the case of the pSCL4-free strains, the samples grown for 8, 15 and 35 days maintained a similar pattern of short thin frequent branches to the samples from day 5. Moreover, no evidence of sporulation was found for the M1, M2 and Mc3 strains at any stage.

In general, visualisation of these microscopic images suggested a poor hyphal development of the pSCL4-free strains in comparison to the WT. Furthermore, in order to quantify these morphological differences of the mycelium of the four strains, we measured and compared the following aspects: branch length, hyphae width and inter-branch distance as indicated in **Figure 6.22**. In order to measure the length of the branches, we utilised the Simple Neurite Tracer plugin in the FIJI software on phase contrast images, which involved manual localisation of the start and end of the branch. The start was considered the branching point on the parent hypha, and the end was considered the tip of the branch. All the possible branches were measured on the 160 images (10 images x 4 samples x 4 time points). The width of the hyphae was measured using FIJI straight line selection on phase contrast images, which consisted of manually selecting the width of the hyphae by a straight line that was subsequently quantified. A minimum of 10 measurements were taken on each image, 160 images in total. Lastly, in order to quantify the branching frequency, the inter-branch distance was measured and compared between strains. For this, we utilised the Simple Neurite Tracer plugin (FIJI) again on phase contrast images by manually localising the branching points on the main hyphae. All the possible inter-branch distances were measured on the 160 images. Violin plots were constructed for each strain using the measured values of the four time points and statistical analyses were performed.

Measurements obtained from images taken at day 5 from inoculation confirmed the three pSCL4-free strains M1, M2 and Mc3 exhibited significantly shorter branches than the WT strain at this stage (**Figure 6.23**). Similarly, the width of the hyphae from these three mutant strains was significantly lower than the WT. In addition, the measurements of inter-branch distance showed that M1, M2 and Mc3 have significantly lower values, and therefore, higher branching frequency at day 5.

Day 15



Day 35

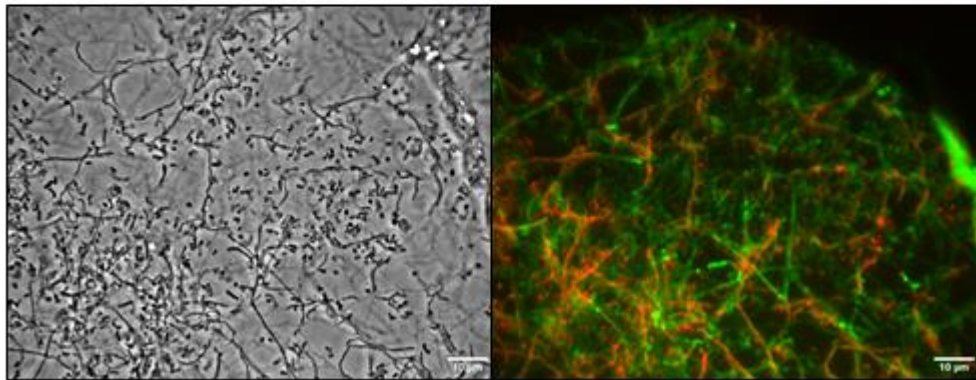
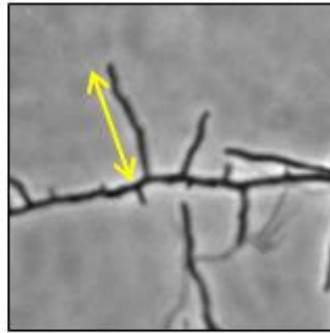
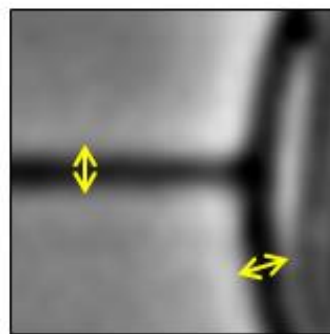


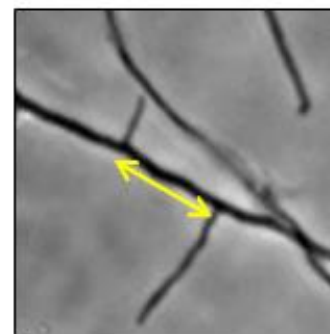
Figure 6.21. Examples of microscopy images of strains WT (*S. clavuligerus* DSM 738) grown on L3M9 agar for 15 and 35 days, exhibiting hyphae curling and sporulation. Images obtained with an inverted widefield epi-fluorescence microscope. Phase contrast images are on the left. Fluorescence images are on the right; the peptidoglycan stained with fluorescein isothiocyanate-wheat germ agglutinin (FITC-WGA) is presented in green under a FITC emission filter; nucleic acids stained with propidium iodide and are presented in red under a 600 nm emission filter. Scale bar: 10 μ m.



Branch length



Hyphae width



Inter-branch distance

Figure 6.22. Representation of how the three aspects (branch distance, hyphae width and inter-branch distance) were measured on phase contrast images of *S. clavuligerus* mycelia. The yellow arrow indicates the distance that was quantified.

Day 5

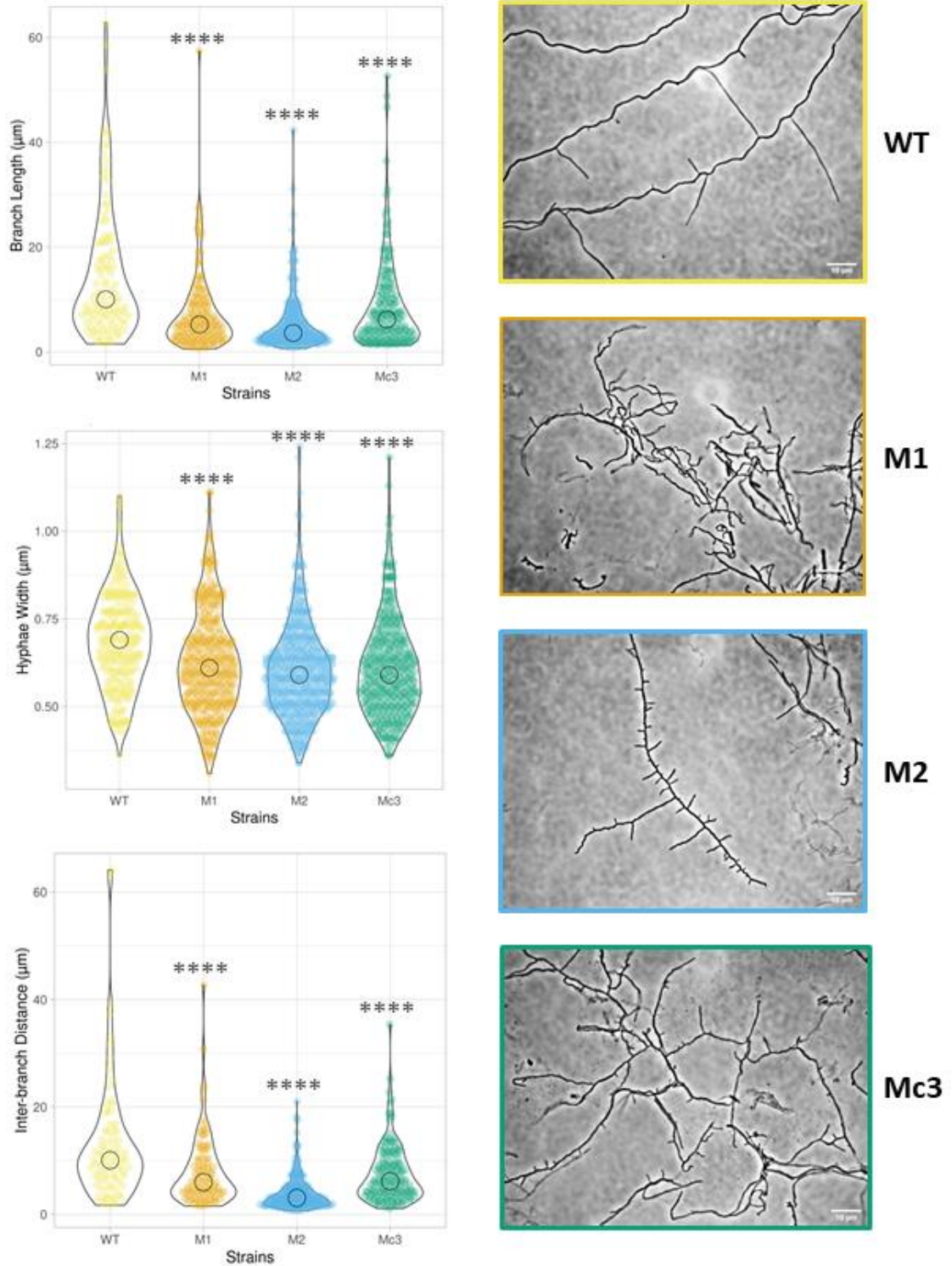


Figure 6.23. Violin plots representing the distribution of the branch length, hyphae width and inter-branch distance measurements for each strain on day 5 from inoculation. The circles indicate the median. ****: $p < 0.0001$, in reference to the WT. Examples of phase contrast images of the four strains at day 5.

The images taken of mycelium after 8 days from inoculation, showed a similar short and thin branches pattern for the three plasmid-free strains compared to the WT (**Figure 6.24**). The M1 and M2 strains also showed significant lower inter-branch distance than the WT, while no significance was found between the branching frequency of Mc3 and the WT at this point.

At day 15, it was observed that the significant lower values of branch length and hyphal width on the mutant strains were maintained with the exception of strain M2 that did not exhibit significant difference from the WT in hyphae width (**Figure 6.25**). In addition, M1 also showed an increased branching frequency while M2 and Mc3 exhibited similar values to those obtained from the WT images.

In contrast to the previous days, measurements from the images taken on day 35 from inoculation indicated that the strains M2 and Mc3 present a considerably larger distribution of different branch length measurements, which brought the mean branch length of these two samples to be significantly higher than the WT (**Figure 6.26**). The hyphal width was still generally lower in the three plasmid-free strains than in the WT samples at this point. No significant differences were observed for branching frequency for the WT, M1 and M2 samples while images from the Mc3 strain indicated a higher distance between branches. However, the difference in the branch length and inter-branch distance distribution pattern at this stage could be due to increased mycelial lysis and, in the case of the WT, sporulation.

In summary, observation and quantification of microscopic images confirm that the three pSCL4-free strains exhibit generally shorter and thinner mycelial branches as well as no sporulation. In addition, M1 and M2 present significantly more frequent branching than the WT.

Day 8

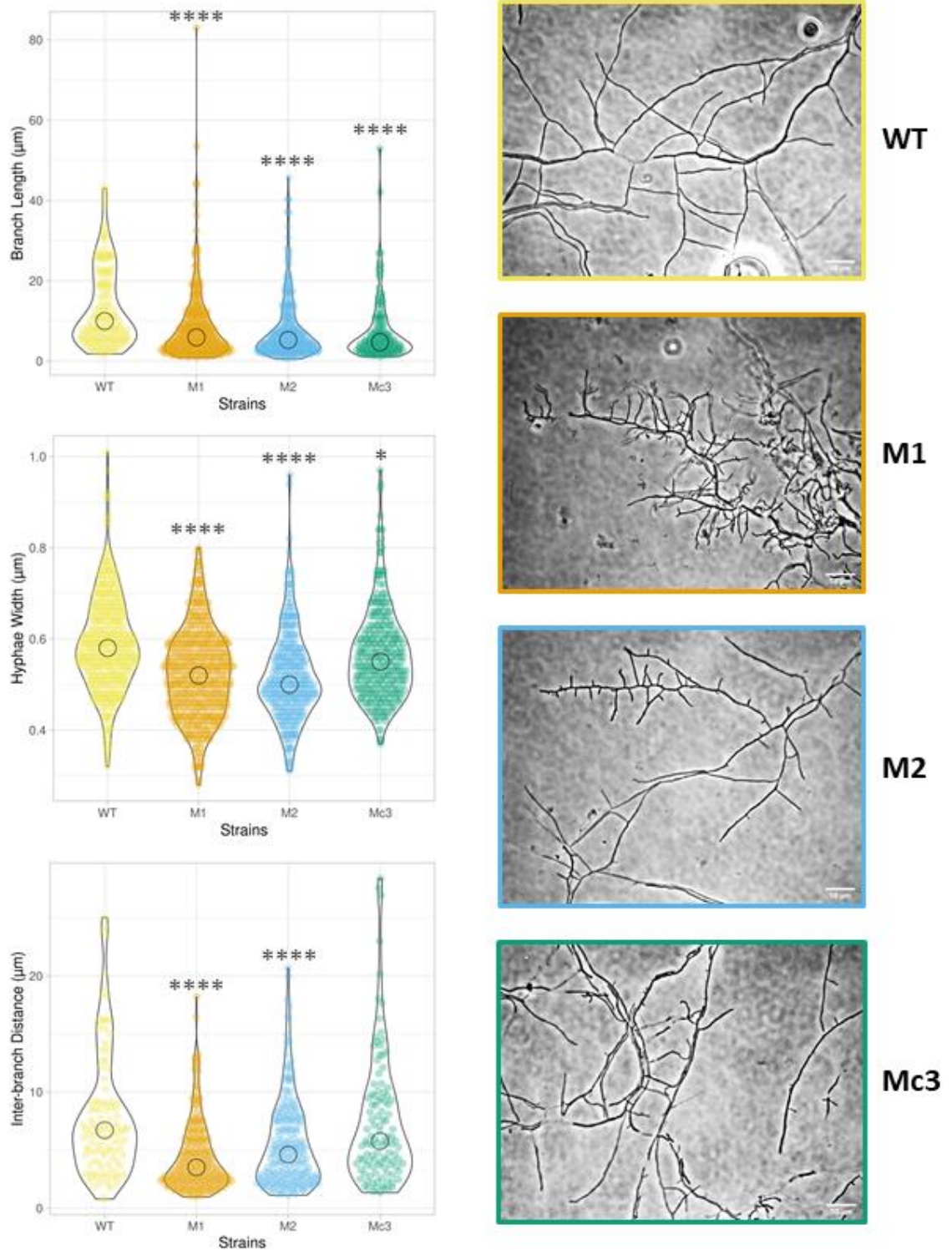


Figure 6.24. Violin plots representing the distribution of the branch length, hyphae width and inter-branch distance measurements for each strain on day 8 from inoculation. The circles indicate the median. ****: $p < 0.0001$, *: $p < 0.05$ in reference to the WT. Examples of phase contrast images of the four strains at day 8.

Day 15

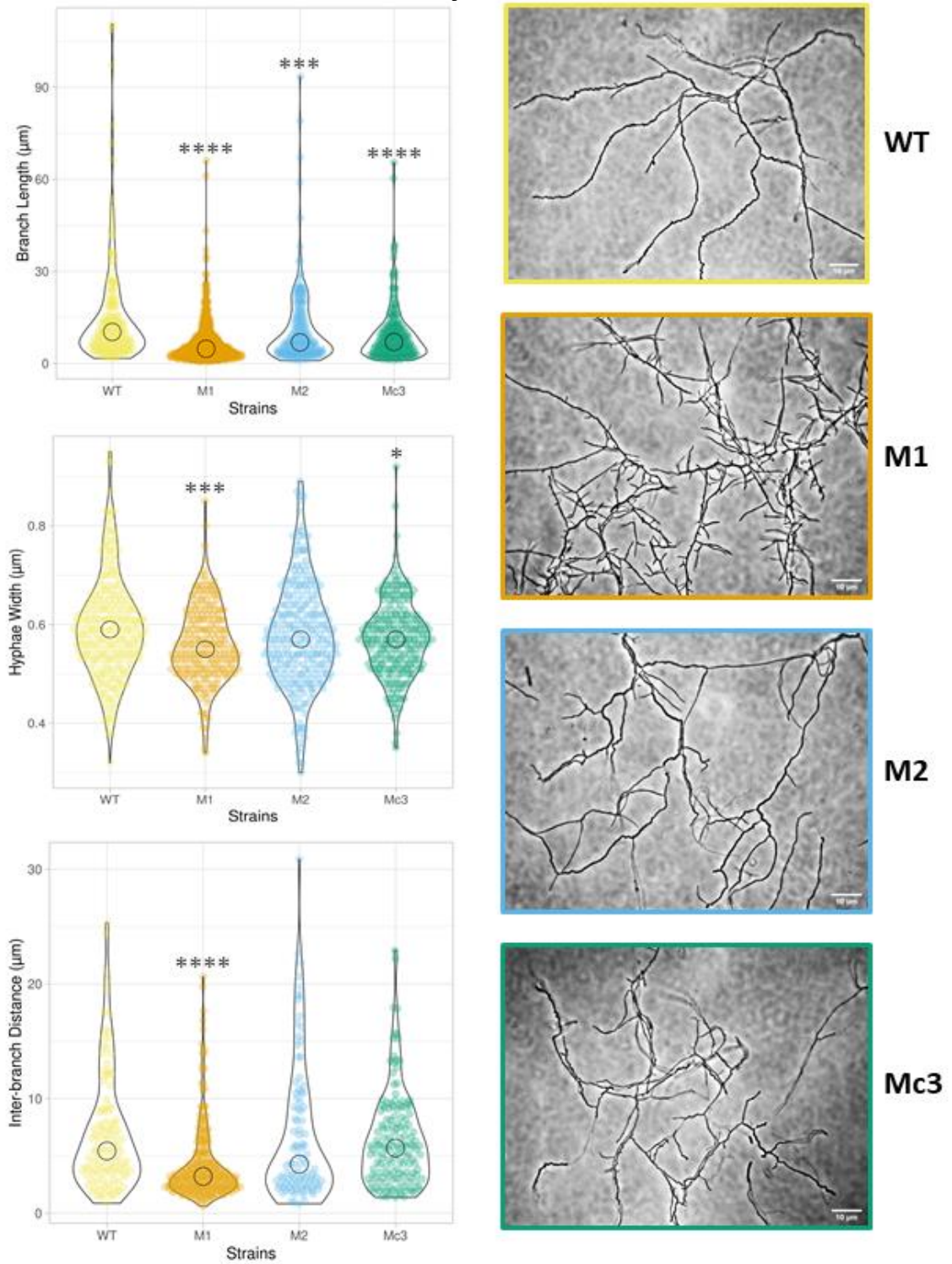


Figure 6.25. Violin plots representing the distribution of the branch length, hyphae width and inter-branch distance measurements for each strain on day 15 from inoculation. The circles indicate the median. ****: $p < 0.0001$, ***: $p < 0.001$, *: $p < 0.05$ in reference to the WT. Examples of phase contrast images of the four strains at day 15.

Day 35

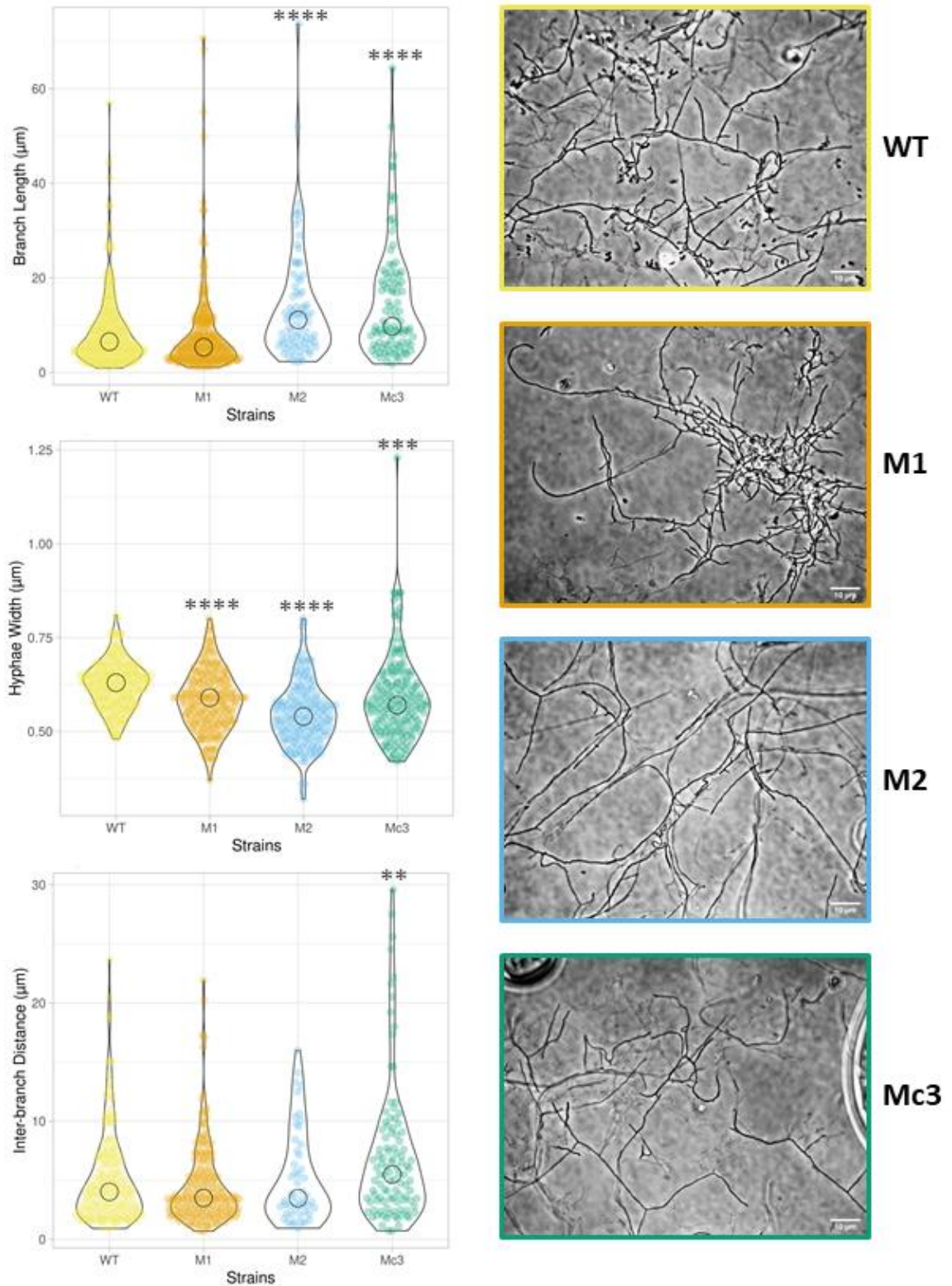


Figure 6.26. Violin plots representing the distribution of the branch length, hyphae width and inter-branch distance measurements for each strain on day 35 from inoculation. The circles indicate the median. ****: $p < 0.0001$, ***: $p < 0.001$, **: $p < 0.01$ in reference to the WT. Examples of phase contrast images of the four strains at day 35.

6.9 The pSCL4-free strains M1 and Mc3 exhibit decreased clavulanic acid production.

Our initial theory follows the idea that curing the megaplasmid would decrease the number of coding sequences by 21%, which would decrease the metabolic burden of the organism helping to improve the production of clavulanic acid. In order to prove this, we tested the production of clavulanic acid on our pSCL4-free strains M1, M2 and Mc3. For this, we grew the strains in liquid culture, and the supernatant was used for the clavulanic acid production assay. This assay is based on the reaction of clavulanic acid with imidazole that yields a derivative absorbing at 324 nm, which is detectable by a spectrophotometer.

Fermentations were initially performed by growing the WT, M1, M2 and Mc3 in TSB for 30 hours, and subculturing the strains at the same concentration of mycelium into fresh TSB for further 72 hours. The mycelium was then removed, and the supernatant used for the imidazole assay. Nevertheless, no clavulanic acid was detected from these fermentations for any strains.

The fermentations were then repeated, this time we used the GSK production broth for clavulanic acid production: CM5 medium. The strains were still pre-grown in TSB and subsequently inoculated at the same concentration in CM5, and grown for 72 hours. Following the imidazole assay, a concentration of clavulanic acid of 4.38 μg per ml of culture in the WT samples was detected as presented on the graph in **Figure 6.28**. In contrast, no clavulanic acid was detected in the cultures of the mutant strains. Nevertheless, the amount detected on the WT is still considerably low compared to the GSK early production strains, that exhibit at least 200 $\mu\text{g}/\text{ml}$ of CM5 broth (S. Kendrew, GSK). This low production could be due to a poor growth in this particular medium, however, since the media is composed of soy protein concentrate and rape seed oil, quantification of the mycelium by measurement of absorbance or cell dry weight was not possible.

In order to further investigate the performance of our strains in the GSK production medium, we cultured the four strains on a solid version of the CM5 medium. As illustrated on the images in **Figure 6.29A**, the WT strain showed abundant aerial mycelium while the three pSCL4-free strains exhibit a bald phenotype on this medium. In addition, to study the production of clavulanic acid in these cultures, we carried out some plug assays. For this, we excised plugs from the CM5 plates and placed them on a lawn of ampicillin-resistance *E. coli* (*E. coli*::pUC19) grown in the presence of carbenicillin. This assay is based on the idea that

the ampicillin resistance would be inhibited in the presence of clavulanic acid, allowing the carbenicillin in the media to prevent the growth of *E. coli*, which would be observable by the presence of inhibition halos around the CM5 plugs. A clear inhibition zone was observed around the WT plug, as well as around the plugs containing the M2 and Mc3, although both of smaller size than the WT (**Figure 6.29B**). In contrast, no inhibition was observed for the M1 strain. These results suggest a potential production of clavulanic acid of the WT strain in CM5 medium, and to a lesser level the pSCL4-free strains M2 and Mc3. However, this activity could be due to a combination of the action of clavulanic acid and cephamycin. In order to assess the action of cephamycin alone, we repeated the same plug assay with wild type *E. coli* grown in the absence of antibiotics. This resulted in inhibition zones of similar size than on the previous experiment (**Figure 6.29B**). These results suggest that the observed bioactivity is mainly due to the production of cephamycin, however, since the synthesis of cephamycin and clavulanic acid is co-regulated, we could extrapolate the results to the production of clavulanic acid.

Nevertheless, in order to quantify the amount of clavulanic produced by the four strains in CM5 agar, we extracted metabolites from agar plugs using double-distilled water. The extracts were then used to carry out an imidazole assay. A graph was constructed indicating the μg of clavulanic acid detected per g of agar, and statistical analyses were performed (**Figure 6.30**). This graph indicates a clavulanic acid concentration of 52 $\mu\text{g/g}$ of agar containing the WT strain. The pSCL4-free strains M1 and Mc3 presented significantly less clavulanic acid production, while M2 did not show any significant difference in clavulanic acid production in comparison to the WT.

To summarise, no clavulanic acid was detected on liquid fermentations of pSCL4-free strains. Additionally, a significantly decreased clavulanic acid production was observed on solid cultures of the M1 and Mc3 strains.

Clavulanic Acid Production

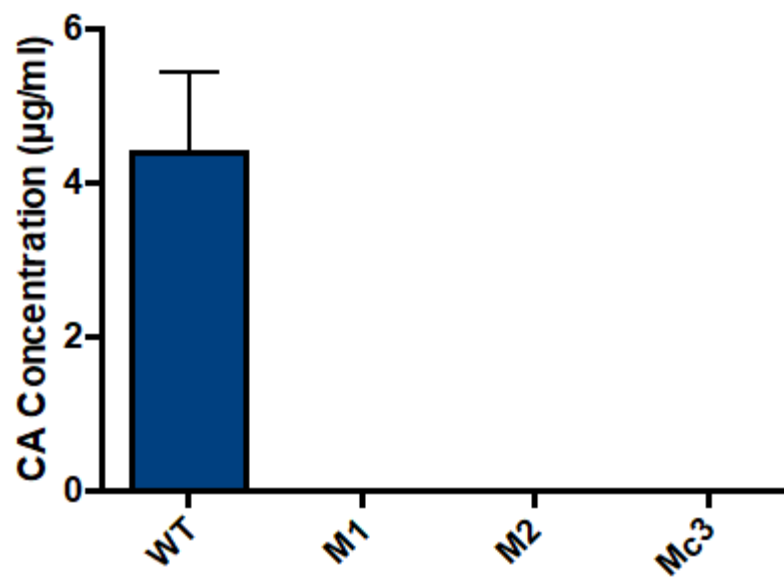


Figure 6.28. Clavulanic acid (CA) production from fermentation of the WT (*S. clavuligerus* DSM 738) and the pSCL4-free strains M1, M2 and Mc3 in CM5 liquid medium.

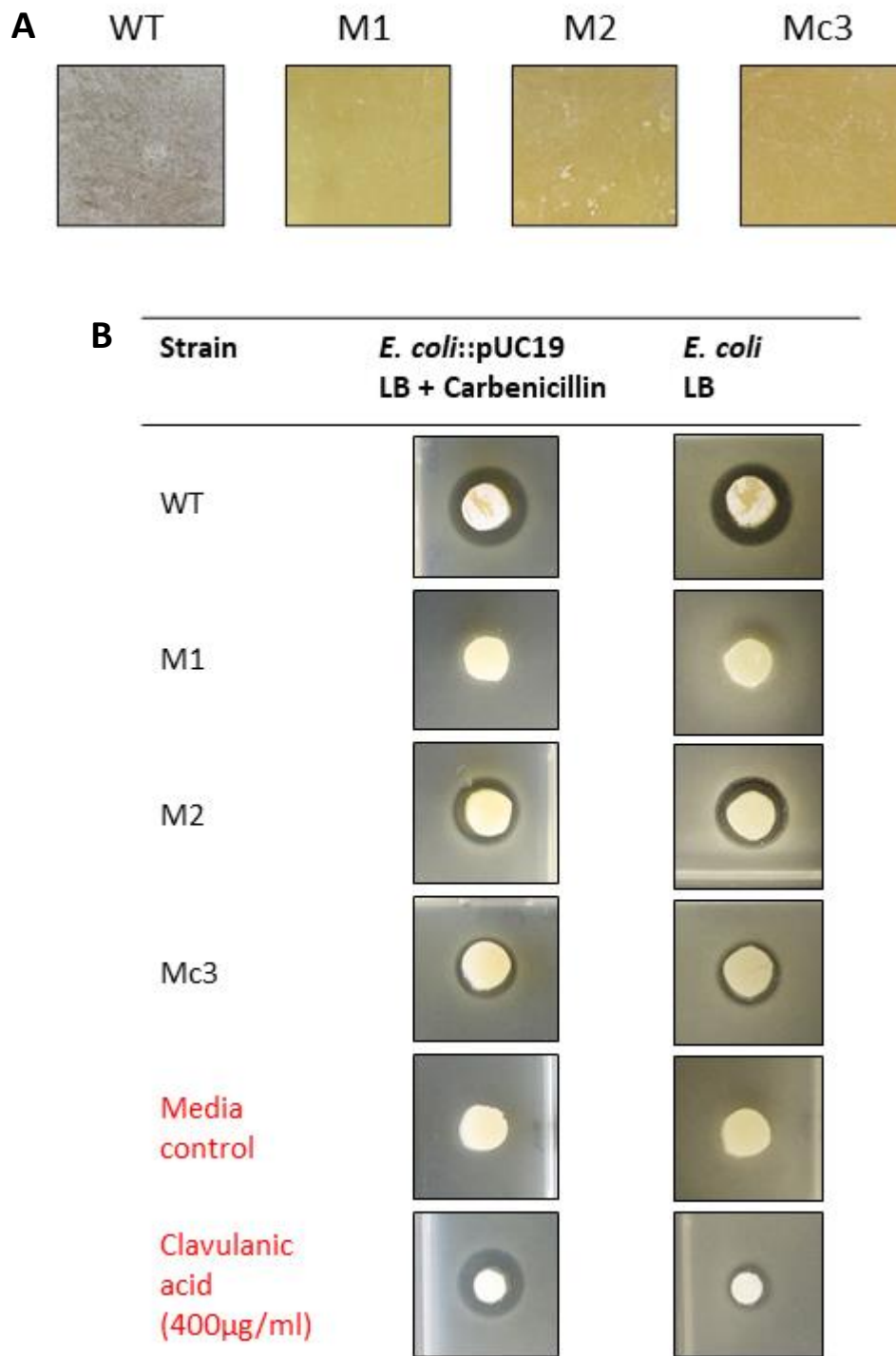


Figure 6.29. Bioactivity plug assay of *S. clavuligerus* strains WT, M1, M2 and Mc3 grown on CM5 agar. A: The four strains grown on CM5 agar. B: Images of the bioactivity assays. Plugs of *S. clavuligerus* cultures placed on *E. coli* lawns: ampicillin-resistance *E. coli* (::pUC19) grown in the presence of ampicillin, and *E. coli* (wild type) grown in the absence of antibiotics. The media control: CM5 agar. Clavulanic acid control: paper disc containing 10 µl of 400 µg/ml.

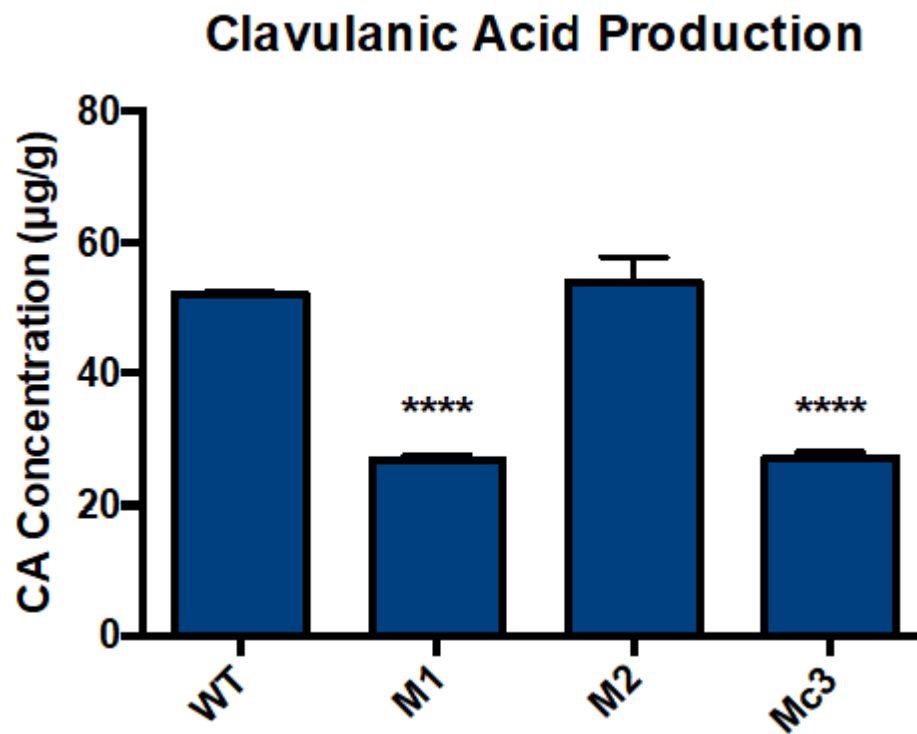


Figure 6.30. Clavulanic acid (CA) production from extraction of the WT (*S. clavuligerus* DSM 738) and the pSCL4-free strains M1, M2 and Mc3 grown on CM5 agar. CA concentration indicated in µg of CA per g of agar. ****: $p < 0.0001$ in reference to the WT.

6.10 Is the megaplasmid conjugative?

In order to study if the megaplasmid pSCL4 is conjugative and if it was possible to restore the WT phenotype in the mutants, we intended to conjugate pSCL4 back into one of our megaplasmid-free mutants. For this, we utilised the *S. clavuligerus* SCLA405 strain as plasmid donor and Mc3 as acceptor. SCLA405 comes from a crossover event of the WT pSCL4 with pLIS405, that tagged the megaplasmid with an apramycin-resistance gene at the *tap-tpg* operon. Mc3 contains a hygromycin-resistance gene at the *attB*/φBT1-integration site on the chromosome. Using these two strains for conjugation of pSCL4 from SCLA405 to Mc3 would allow identification of transconjugants by screening for strains that are resistant to both apramycin and hygromycin (**Figure 6.31A**). For this, both parent strains were grown together in the absence of antibiotics, and spores were harvested. The spores were then germinated and grown in the presence of both apramycin and hygromycin. This resulted in transconjugants resistant to both antibiotics, hence candidates for conjugation of pSCL4 (**Figure 6.31B**). This hygromycin and apramycin-resistant strain was called SCLA405-Mc3.

In order to confirm if this strain originated from the conjugation of pSCL4 into Mc3, we performed a PFGE. In the case of conjugation of pSCL4 into Mc3, the gel should have shown a band at 1,800 kb (pSCL4) but no band at 450 kb (pSCL3). However, both bands were observed on the sample of SCLA405-Mc3, exhibiting a similar profile to the WT, in contrast to Mc3 that showed only a band for pSCL2 (**Figure 6.32A**). Nevertheless, it could have been a case of pSCL4 and pSCL3 conjugation into Mc3. Furthermore, in order to investigate the topology of the chromosome and megaplasmid in SCLA405-Mc3, we extracted genomic DNA and performed PCR analyses. The primers *CaoChr-L-end-F* and *CaoChr-L-end-R* were used to amplify the chromosome telomeres from genomic DNA of WT, Mc3 and SCLA405-Mc3. The primers *CaoChr-L-end-F* and *SCL4-L-end-R2* were used to amplify the pSCL4 telomere. Additionally, the primers *FtsZ-Fw* and *FtsZ-Rv* were used to amplify the *ftsZ* gene as a positive control. The PCR products were then separated by gel electrophoresis. The image of the gel presented on **Figure 6.32B** indicates that SCLA405-Mc3 carries both the chromosome and megaplasmid telomeres and so, this strain maintains the WT ends of these two replicons and a linear topology. These results confirm that unlike Mc3, SCLA405-Mc3 carries the plasmid pSCL3 as well as the complete linear chromosome and pSCL4, which suggest that Mc3 was not the parent strain and so the megaplasmid was not conjugated. It is more likely that a recombination event occurred between the chromosomes of SCLA405 and Mc3, yielding a strain similar to SCLA405 with the addition of the hygromycin-resistance.

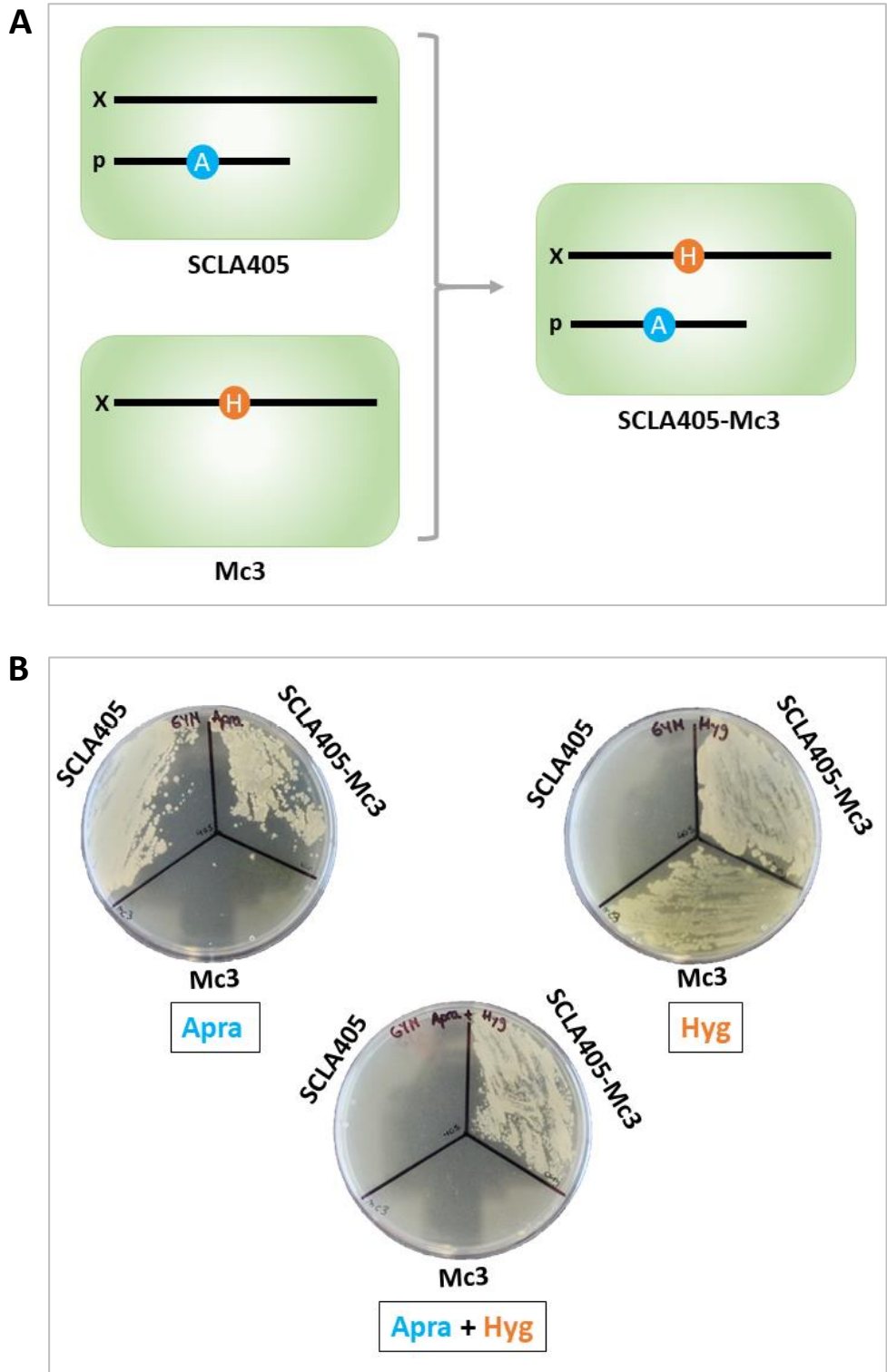


Figure 6.31. Plan for conjugating pSCL4 into Mc3. **A:** Schematic representation of the conjugation of pSCL4 from SCLA405 to Mc3. X: chromosome, p: pSCL4, A: apramycin-resistance gene. H: hygromycin-resistance gene. **B:** Plates of parent strains SCLA405 and Mc3 and daughter strains SCLA405-Mc3 grown on apramycin (Apra), hygromycin (Hyg) or both.

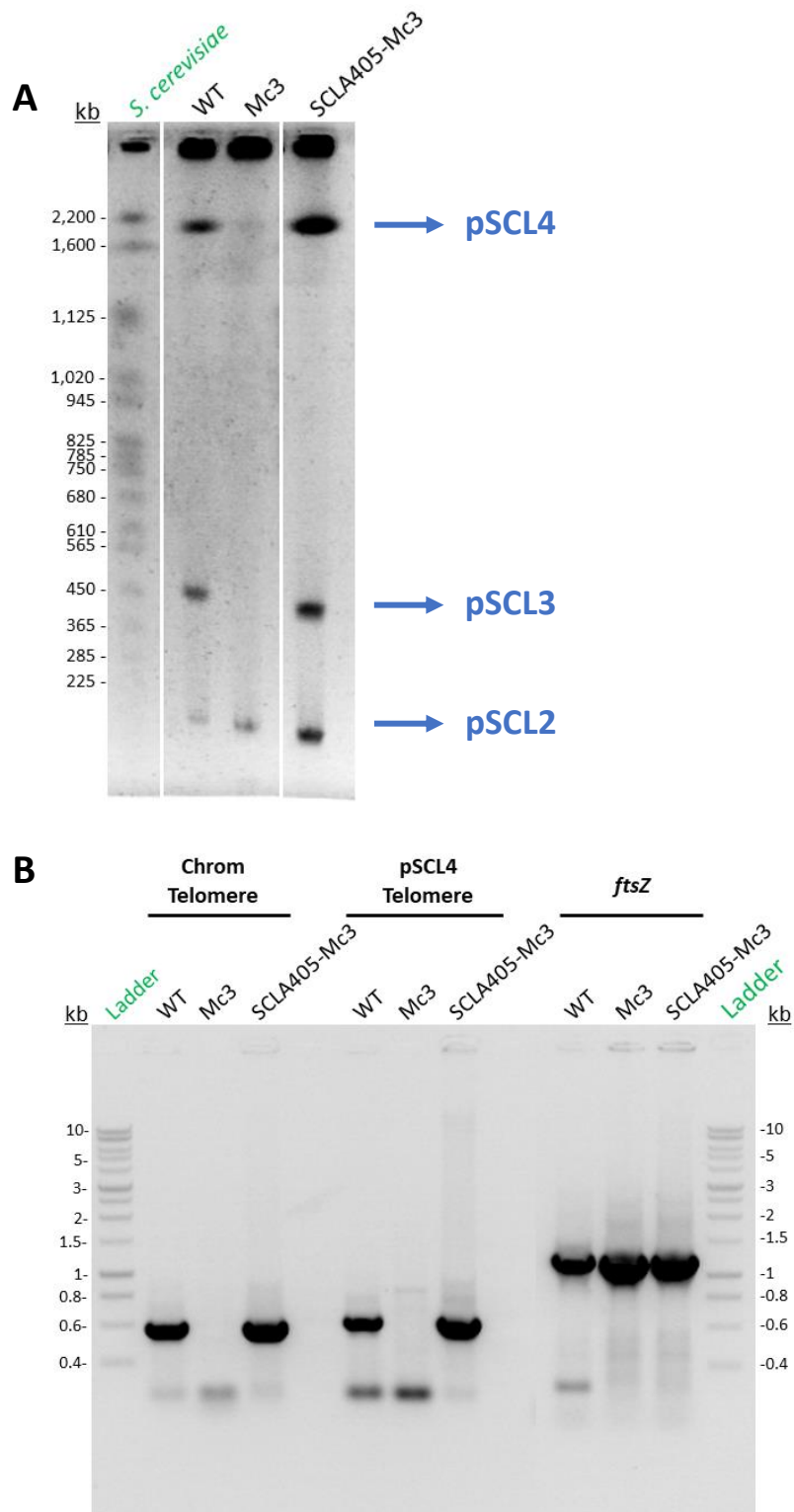


Figure 6.32. A: PFGE of undigested samples of WT, Mc3 and SCLA405-Mc3. Electrophoresis performed using HEPES buffer at 4 V/cm for 33 hours. *Saccharomyces cerevisiae* YNN295 was used as DNA size standard (Bio-Rad). **B:** Gel electrophoresis of PCR products from genomic DNA of WT, Mc3 and SCLA405-Mc3 using primers that anneal to the chromosome telomere, pSCL4 telomere and *ftsZ*.

gene. This would explain how this strain appeared resistant to both apramycin and hygromycin with still maintaining all the replicons complete.

6.11 Summary.

In this chapter we report the elimination of the megaplasmid pSCL4 from *S. clavuligerus*. This was achieved by cleaving the plasmid partitioning gene *parB* with CRISPR/Cas9 technology. This method has resulted a faster and simpler way for plasmid curing than traditional methods that involve recombination, protoplast formation, or many cloning steps. Additionally, PFGE was proven a good method for rapid screening of the cured candidates. Consequently, we confirm that targeting genes involved in plasmid segregation with CRISPR/Cas9, coupled with screening of mutant strains with PFGE, is a successful procedure to obtain plasmid-free strains of *S. clavuligerus*, a fact that could be extrapolated for curing giant linear plasmids in other *Streptomyces* species.

Three pSCL4-free strains: M1, M2 and Mc3, were chosen to be further studied. Their whole genome was sequenced by Illumina technology which confirmed the loss of the megaplasmid, as well as the loss of pSCL3 in the Mc3 strain, which agrees with the PFGE results. The loss of pSCL3 in Mc3 is unlikely to be connected to a CRISPR/Cas9 cleavage of this plasmid, as homology to the guide-RNA used is exclusive to pSCL4-*parB* and has low homology to the pSCL3 sequence. In addition, pSCL3 carries its own partitioning genes and so, a segregation machinery independent of pSCL4. Moreover, the fact that other pSCL4-free strains maintained pSCL3 suggest that the elimination of this plasmid in Mc3 might have occurred spontaneously. The studies of Álvarez-Álvarez *et al.*, 2014 were done in a pSCL3-free *S. clavuligerus* strain that is considered a wild type, which suggest that this plasmid is not essential and its loss was probably spontaneous. For the case of the other two smaller plasmids pSCL2 and pSCL1, sequencing data was obtained for the entirety of both replicons on the three mutant strains, suggesting they have not been altered, and so they are independent of pSCL4, which is again consistent with the pSCL4-free strain described by Álvarez-Álvarez *et al.*, 2014.

Sequencing data from these three pSCL4-free strains, allowed us to identify a translocation of a 50 kb region from pSCL4 to pSCL3 in M2, which is consistent to the larger pSCL3 observed by PFGE, making M2 the only strain that maintained some pSCL4 material after the curing.

Interestingly, this region contains a putative novel lasso peptide biosynthetic gene cluster that might be giving the M2 strain some specialised metabolism advantage that is missing in the other two strains. Furthermore, sequences with high homology were found on the inserted pSCL4 region and the insertion point in pSCL3 which suggests that this translocation was achieved by homologous recombination between the two plasmids. This fact, together with the spontaneous loss of pSCL3 in Mc3, emphasises how highly susceptible *S. clavuligerus* is to genomic rearrangement.

Furthermore, we identified the *attB* sites for ϕ BT1-based integration on the chromosome of *S. clavuligerus* by examining whole genome sequencing data from the Mc3 strain. This site is different to the *attB* site identified for ϕ C31-integration in *S. clavuligerus* described by Combes *et al.*, 2002, which is consistent with observations described by Gregory and Smith, 2003 in *S. coelicolor*. This finding will facilitate future experiments involving integration of vectors carrying the ϕ C31 *attP-int* locus into *S. clavuligerus*, allowing construction of an *in silico* version of the expected integration to complement the experimental results.

Our results confirmed a chromosome circularisation occurred in strains M1 and M2 concomitant to the loss of pSCL4. Characterisation of the fusion points between the left and right arms allowed observation of minimal or no overlapping sequences at the junctions, which implies that the circularisation occurred by non-homologous recombination of both arms. In addition, this fusion occurred at different locations within the chromosome on the two strains, which confirms that the junction point is unspecific. A similar circularisation event was reported in the studies by Nindita *et al.* on *S. coelicolor* and *S. rochei* (Nindita *et al.*, 2013; Nindita *et al.*, 2015), where they also described the loss of much material at both ends of the chromosome, similar to our strains where 100 to 300 kb were deleted. The authors explained their observations through circularisation by deletion of the ends occurring first and continued progressively until finding a fusion point that generated a stable circular chromosome that could be maintained through future generations. This circularisation occurs naturally as an adaptation to detrimental genetic modifications, and it is stably maintained, in contrast to cases of artificially constructed circular chromosomes in *S. lividans* and *S. ambofaciens* that yielded highly unstable strains (Lin and Chen, 1997; Volff *et al.*, 1997; Fischer *et al.*, 1997). This theory could be applied to our strains, suggesting that chromosome circularisation was induced by the loss of the chromosome ends, which, at the same time, was induced by loss of the telomere-replication machinery. This fact implies that

the chromosome utilises the Tap and Tpg proteins encoded on the megaplasmid, and its action could not be complemented by the other two sets of terminal proteins encoded on pSCL3 and pSCL2. This fact would also agree with the observations of Álvarez-Álvarez *et al.*, 2014, that described a loss of material at the chromosome right end when eliminating pSCL4. These facts lead us to the conclusion that loss of the megaplasmid deprived the chromosome of the proteins necessary for the replication and maintenance of the telomeres in the M1 and M2 strains, causing large deletions at the ends that were stopped by fusion of both arms creating a stable circular chromosome.

In the case of the Mc3 strain, no evidence for chromosome circularisation was found. This strain carries a copy of the *tap-tpg* operon from pSCL4 on the chromosome, and so, the telomeres should have been maintained unlike in M1 and M2. However, this strain is missing 210 kb of end material (55 + 155 kb), suggesting the chromosome is still missing a proper telomere-replication machinery. This could be due to the integrated *tap-tpg* not being functional or that other genes present exclusively on pSCL4 are also necessary for maintaining the chromosome ends. As it was observed on Chapter 4, there is another gene present on pSCL4, pSCL3 and pSCL2 that is missing on the chromosome, the *ttrA* gene. This gene is usually located at the very end of *Streptomyces* chromosome and plasmid sequences, and it encodes a helicase-like protein thought to be involved in conjugal transfer (Huang *et al.*, 2003). In the studies done on *S. rochei*, complementation of plasmid-*tap-tpg* on the chromosome was sufficient to maintain the telomeres in the absence of the plasmid (Nindita *et al.*, 2015); however, unlike in *S. clavuligerus*, the *S. rochei* chromosome carries two copies of *ttrA*. Although the studies of Bey *et al.*, 2000 confirmed that *S. lividans* strains with truncated *ttrA* genes still maintained linearity, it would be interesting to investigate the implications of TtrA in *S. clavuligerus*.

Characterisation of the growth of these pSCL4-free strains in rich liquid media confirmed the loss of the megaplasmid has no negative effect on the fitness of *S. clavuligerus*. And in the case of the M2 strain, a significant higher growth rate was observed; however this could be due to a secondary aspect of this strain since is not consistent with the other strains, implying the loss of pSCL4 is not directly related to an increased growth rate. In contrast, solid cultures of the three strains exhibited poor development and/or no sporulation on solid media. This was also confirmed in the microscopic images, where no spores were observed on the pSCL4-free strains, even after 35 days from inoculation. Moreover, further characterisation of the

mycelium morphology at a microscopic level lead us to the conclusion that the three pSCL4-free strains exhibit significant shorter and thinner branches. Additionally, the M1 and M2 strains showed a significant higher branching frequency than the wild type. Short and thin abortive branches have been previously linked to the absence of replisomes in these secondary hyphae (Wolánski *et al.*, 2011), however genes coding for proteins involved in associating replisomes to hyphae tips such as DivIVA, FilP or the Par proteins are intact in these pSCL4-free strains. Nevertheless, we identified a number of transcriptional regulators present on the missing chromosomal regions of the three mutant strains that are often associated with regulation of secondary metabolites biosynthesis and morphological development. Examples of these are the MarR-type and the AsnC-type regulators located towards the left end of the *S. clavuligerus* chromosome. Deletion of genes belonging to these families of regulators have previously resulted in the delay of aerial mycelium and sporulation of *Streptomyces roseosporus* and *S. coelicolor* cultured in solid media (Zhang *et al.*, 2015; Liu *et al.*, 2017). Another deleted gene in the three pSCL4-free strain is a AfsR-type transcriptional regulator, a family of regulators that have been proved to be involved in aerial mycelium formation in *S. griseus* (Umeyama *et al.*, 1999). Although many members of these three families of transcriptional regulators are still maintained in the chromosome of the three mutant strains, the loss of these genes might have repercussions on the morphological development and be responsible for the poor aerial mycelial formation and lack of sporulation exhibited on these strains and therefore, the action of these genes should be further studied. In contrast, the morphological differences of the mutant strains could also be due to the loss of pSCL4 alone. Interestingly, the pSCL4-free strains described on Álvarez-Álvarez *et al.*, 2014, were also reported as non-sporulating, which suggest the megaplasmid might have a role in regulation of chromosomal gene expression. Another option would be the effect of a circular chromosome might have on chromosome segregation or transcription. Chen *et al.*, 2010 reported strains of *S. avermitilis* carrying circular chromosomes exhibited a bald phenotype however the specific effect of chromosome circularisation on the development of mycelium is yet to be investigated.

Furthermore, the pSCL4-free mutants exhibited no clavulanic production in the GSK production medium. Culturing these strains on solid GSK production medium confirmed that pSCL4-free mutants develop poorly on this medium, which could explain the lack of clavulanic acid production in liquid fermentation. Extractions from solid cultures also confirmed a decreased clavulanic acid production in strains M1 and Mc3, while M2 extractions exhibited

similar levels to the WT. These results confirm that the loss of the megaplasmid does not have a positive effect on the production of clavulanic acid, revoking our initial theory. This agrees with the results by Álvarez-Álvarez *et al.*, 2014, who described a 20 to 30 % reduction in clavulanic acid in pSCL4-free strains. This could be due to direct loss of regulatory genes in pSCL4 or indirect deletion of chromosomal genes concomitant to chromosomal ends loss and circularisation. Among the identified deleted transcriptional regulators in the chromosome are the AfsR-type and LysR-type regulators. Members of these regulator families are known to be involved in activating clavulanic acid production and so their loss might also explain the decreased amounts of clavulanic acid in M1 and Mc3 cultures (Parajuli *et al.*, 2005; Pérez-Redondo *et al.*, 1998). In general, we can conclude that curing the megaplasmid is not a convenient resource for GSK to improve clavulanic acid production from *S. clavuligerus*, as it reduces its performance even in optimum production media.

Finally, we attempted to conjugate the megaplasmid back into the Mc3 strain, but this was unsuccessful. In contrast, we found evidence of inter-strain chromosomal recombination, suggesting conjugation of the chromosome mediated by a plasmid. Although there is evidence of conjugation of giant linear plasmid such as SLP2 from *S. lividans* and SCP1 from *S. coelicolor* (Chen *et al.*, 1993; Vivian and Hopwood, 1975), little is known about conjugation of larger replicons such as pSCL4 or chromosomes, and so, it would be something to investigate in future experiments.

In general, with these results we can confirm that *S. clavuligerus* can survive without pSCL4, making it dispensable. However, pSCL4 is essential to maintain an optimum fitness and the complete chromosome sequence. For this, eliminating pSCL4 does not represent an advantage, especially for the point of view of GSK, and a different approach must be taken for obtaining strains with improved production of clavulanic acid.

7. DISCUSSION

7.1 Combined physical and genetic map of the *S. clavuligerus* genome.

Before this work, different publications suggested that *S. clavuligerus* carries a complex genome architecture with the presence of several extra-chromosomal elements including giant linear plasmids. However, all the replicons were poorly defined and not always detected in this organism. In this work, we have obtained physical and computational evidence to establish the complete genome sequence of *S. clavuligerus* DSM 738.

In Chapter 3 we described the development of the optimised method for Pulsed-field Gel Electrophoresis of *S. clavuligerus* samples. This allowed visualisation of the three giant linear plasmids: pSCL4, pSCL3 and pSCL2 as well as large chromosomal fragments digested with restriction enzymes. This data was then used to corroborate the final assembly of the genome of *S. clavuligerus* described in Chapter 4. This assembly was obtained from combining whole genome sequencing and telomere purification, which resulted in the complete sequences of chromosome (6,750 kb), pSCL4 (1,796 kb), pSCL3 (455 kb), pSCL2 (150 kb) and pSCL1 (12 kb). Comparing these data to the PFGE observations, we can confirm that the sizes of the three giant linear plasmids correlate to the bands from undigested genomic DNA separated by PFGE. These results evidence the benefits of utilising PFGE as a method for confirming the presence and estimated size of giant linear plasmids in *Streptomyces*. Additionally, PFGE performed on samples digested with restriction enzymes allowed identification of bands that correlate to fragments obtained from *in silico* digest of the final assembly. This was used to construct a combined genetic and physical map of the genome of *S. clavuligerus* for two restriction enzymes as illustrated on **Figure 7.1**. This figure reflects that there are more *in silico* fragments than PFGE bands, which suggest that, either fragments of similar size are represented by the same PFGE band, or some of the expected bands are missing due to unspecific cleavage. Considering the PFGE band pattern visualised is consistent among all the electrophoresis analyses, we are inclined to opt for the former explanation, which stresses the importance of contrasting the PFGE and genetic data. Nevertheless, no unexpected bands were observed on PFGE, showing again good correlation between the physical and genetic data.

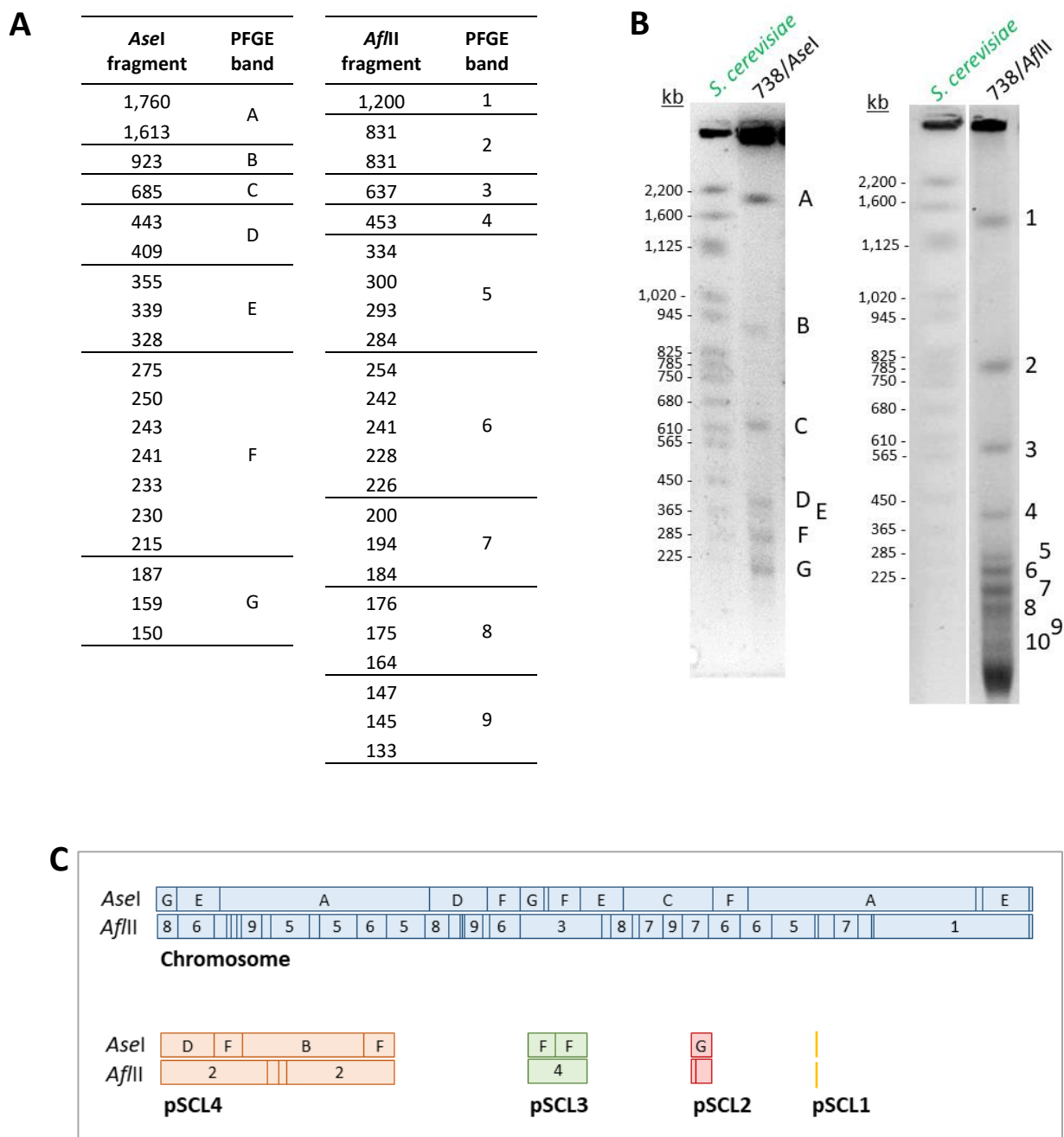


Figure 7.1. Combined genetic and physical map of the genome of *S. clavuligerus* DSM 738. **A:** List of *AseI* and *AfIII* *in silico* digest of the final *S. clavuligerus* genome assembly and the estimated PFGE band they represent. **B:** PFGE of digested genomic DNA with *AseI* and *AfIII* indicating the bands by letters or numbers. **C:** Constructed map of the genome of *S. clavuligerus* presenting the *AseI* and *AfIII* restriction sites and the PFGE fragments (at scale).

Unfortunately, nowadays the use of techniques such as PFGE have been minimised and generally replaced by whole genome sequencing. And, although whole genome sequencing techniques are now the fastest and simplest way to obtain good quality genome sequences, corroborating with physical evidence is important, especially in organism with large linear chromosomes and plasmids like streptomycetes. As we have evidenced with this work, whole genome sequencing alone is not always sufficient to obtain complete and closed sequences and PFGE remains a simple and economical procedure to confirm the presence and size of genomic elements, as well as specific genetic material as it can be coupled to Southern Hybridisation. In addition, as described on Chapter 6, PFGE is an optimum technique to screen for mutant strains that have undergone genomic modifications.

To summarise, with this work we want to stress the benefits of utilising a combination of computational analyses of sequence data and physical techniques like PFGE to investigate the genome of *S. clavuligerus*, and which could be extrapolated to other *Streptomyces* species.

7.2 Three sets of terminal proteins for five sets of telomeres.

As described in Chapter 4, we achieved purification and characterisation of the telomeres of the chromosome and plasmids of *S. clavuligerus*. It was established that the chromosome and megaplasmid carry almost identical telomeres while the other plasmids carried unique telomeric sequences that only resemble each other on the last 13 bp (Palindrome I). Nevertheless, all the described telomeres carry archetypal secondary structures of four loops that are key for the interaction with the terminal proteins. The diversity of these structures suggested that either more than one type of terminal proteins is necessary to maintain the different telomeres, or the same proteins act on different sequences in an unspecific manner.

From the sequencing data, we obtained evidence of the presence of three putative sets of Tap and Tpg proteins in *S. clavuligerus* plasmids pSCL4, pSCL3 and pSCL2, that were further characterised on Chapter 5. Although the actual interaction with the replicons telomeres is yet to be investigated, with our results we can estimate the mechanism of these three sets of terminal proteins. Considering the fact that *S. clavuligerus* carries five linear replicons, but only encodes for three types of terminal proteins, we can confirm that at least one of the sets must support more than one type of telomere during replication. While the telomeres

of pSCL4, pSCL3 and pSCL2 are probably maintained by the proteins encoded in each plasmid, the replication of the telomeres in the chromosome and pSCL1 is unclear. In order to explain telomere replication of these two molecules we offer the following theory. According to sequence similarity, which we illustrate in phylogenetic trees presented in **Figure 7.2**, the chromosome and megaplasmid carry almost the same telomere and so they are most likely supported by the same terminal proteins. In the case of pSCL1, its telomeres have a slightly higher similarity to the ones in pSCL2 than to the rest, which suggest that the pSCL1 telomeres are most likely maintained by the proteins encoded in pSCL2 rather than by the other two types. Additionally, regarding the secondary structures of the telomeres described in Chapter 4 (also presented in **Figure 7.2**), chromosome and pSCL4 present exactly the same folded form, supporting the idea that they interact with the same Tap and Tpg. In the case of pSCL1 and pSCL2, they present similar secondary structures, particularly the palindromes I, II and III. As confirmed by the work of Bao and Cohen, 2003, the second and third palindromes are key for the recruitment of Tap to the telomeres, which lead us to think that the pSCL1 and pSCL2 telomeres might be able to recruit the same protein. The telomeres of pSCL3, on the other hand, exhibit a unique secondary structure, with a distinct second palindrome that does not resemble any of other telomeres of *S. clavuligerus*. In conclusion, the sequence similarity and secondary structures of the telomeres suggest that the terminal proteins from pSCL4 act on the telomeres of the chromosome and pSCL4, while the proteins from pSCL2 interact with the telomeres of pSCL1 and pSCL2 and the telomeres of pSCL3 are supported by the proteins from the same replicon.

This theory agrees with the results observed in Chapter 6, following elimination of pSCL4, and also pSCL3 in the case of the Mc3 strain. The fact that the telomeres of the chromosome were eliminated in the absence of pSCL4 suggests that indeed the chromosome utilises the proteins from this plasmid, and the action could not be complemented by the other two terminal proteins of *S. clavuligerus*. Moreover, the telomeres of pSCL3, pSCL2 and pSCL1 remain stable in pSCL4-free strains, confirming the replication of these telomeres is independent of the megaplasmid. In addition, the sequencing data from the pSCL4- and pSCL3-free strain Mc3 indicates the conservation of the telomeres of pSCL2 and pSCL1, suggesting they are independent of both pSCL4 and pSCL3, which corroborates our theory that these two plasmids utilise the same terminal proteins.

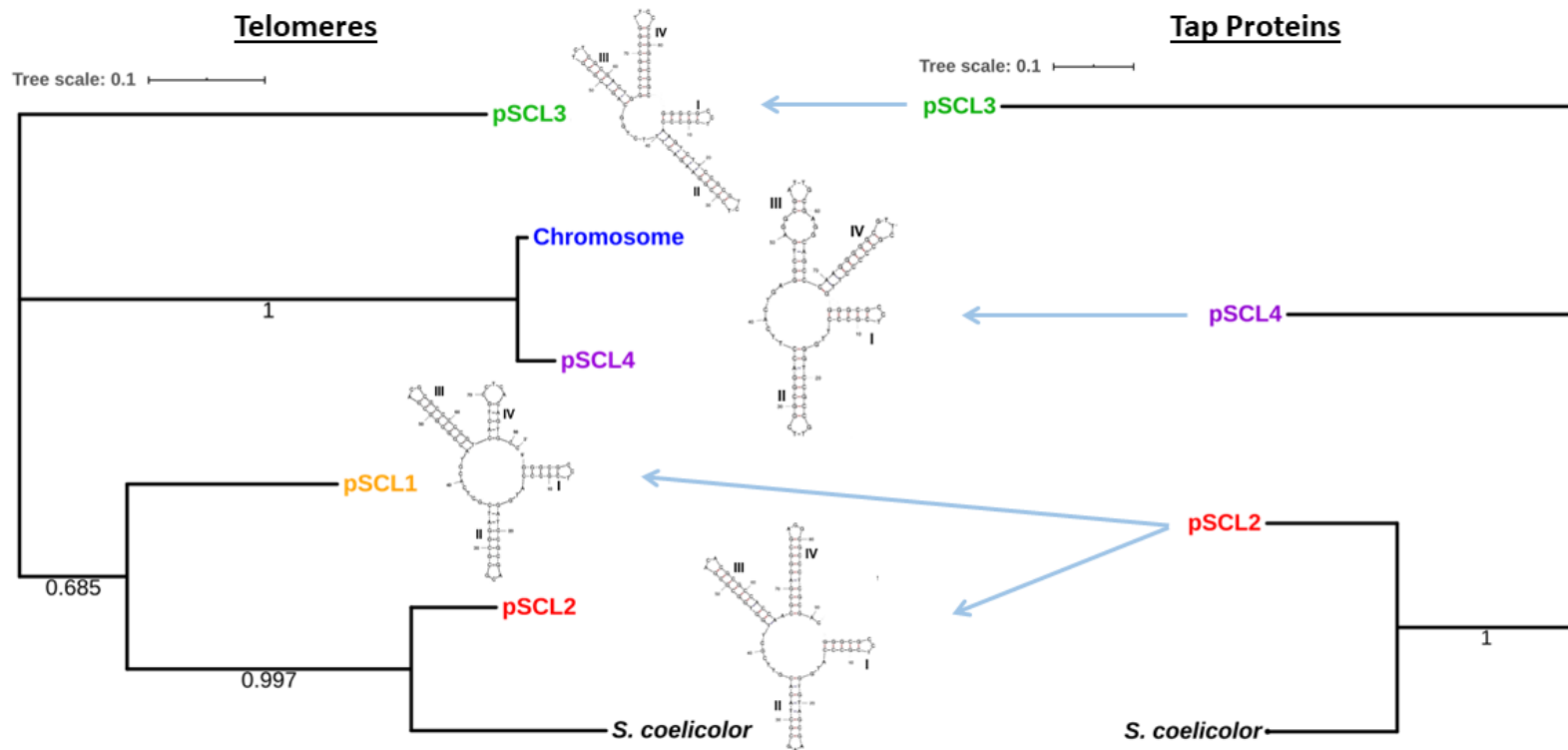


Figure 7.2. Maximum likelihood phylogenetic trees of the last 150 bp of the telomeres of *S. clavuligerus* chromosome and plasmids and *S. coelicolor* chromosome; and the amino acid sequences of the Tap proteins of pSCL4, pSCL3, pSCL2 and *S. coelicolor*. Trees constructed in MEGA with 1000 bootstrap replication and edited in iTOL. Secondary structures of the telomeres are shown next to the tree labels. As indicated by blue arrows, the Tap from pSCL3 supports the pSCL3 telomere; the Tap from pSCL4 supports the chromosome and pSCL4 telomere; and the Tap from pSCL2 supports the pSCL2 and pSCL1 telomeres.

Although the utilisation of pSCL4 terminal proteins by the chromosome is not unexpected, considering the high similarity of these two telomeres, the action of pSCL2 proteins on both the telomeres of pSCL2 and pSCL1 is rather unique. The sequence identity between the last 150 bp of these two replicons is only 64%, implying a promiscuous telomere-protein interaction of the pSCL2 terminal proteins. This also suggests that the secondary structure of telomeres is more important than the nucleotide sequence for interaction with the terminal protein. Although interaction of terminal proteins with more than one type of telomere has not been previously described, we hypothesise that it is not an uncommon event. For example, a similar case to *S. clavuligerus* is *S. rochei*. This organism contains two giant linear plasmids pSLA2-L and pSLA2-M that carry *tap* and *tpg* genes, while the chromosome and the smallest plasmid pSLA2-S carry none. This latter plasmid carries telomeres unique in the *S. rochei* organism, yet they must be replicated by the proteins encoded on the other plasmids. In addition, while the publication by Nindita *et al.*, 2015 exhibited a different secondary structure of the telomeres of pSLA2-S from the rest of the replicons, our observations following the Clover Leaves model (Yang *et al.*, 2017) confirm that this plasmid carries a very similar structure at the palindromes I II and III to the other telomeres. This agrees with our conclusions regarding the telomeres of pSCL1 in *S. clavuligerus* and the importance of the secondary structures for recognition of terminal proteins.

In addition, in Chapter 5 we described the observation of protein-protein interaction between the same Tap proteins, as well as the same Tpg proteins. This suggests the possibility of intramolecular interaction between the terminal proteins of each sequence end, giving linear chromosomes and plasmids a circular conformation, an event that was also exposed for *S. coelicolor* by the studies of Tsai *et al.*, 2011. This publication also demonstrated intermolecular interactions of proteins attached to chromosomal and plasmid telomeres. Having this into account we can hypothesise that *S. clavuligerus* chromosome and pSCL4 might be interacting at the terminal proteins since they utilise the same ones. Similarly, pSCL2 and pSCL1 could be interacting at the telomeres. This would be an interesting aspect to investigate in future *in vivo* experiments of *S. clavuligerus* by for example tagging Tap or Tpg with a fluorescent protein, or even the different telomeres by fluorescent *in situ* hybridisation.

7.3 pSCL4, a dispensable, yet essential plasmid of *S. clavuligerus*.

Since the publication of Medema *et al.*, 2010, we have known of the *S. clavuligerus* megaplasmid pSCL4, which still remains the largest linear plasmid identified. Considering the relatively small size of the *S. clavuligerus* chromosome (6.8 Mb) as opposed to other *Streptomyces* species (8 to 9 Mb), it is reasonable to think of the megaplasmid originated from material of a former larger chromosome. This was mentioned also by Medema *et al.*, 2010, and even proposed scenarios for the evolution of pSCL4, which included two recombination events between the chromosome and a smaller plasmid. Our findings described on Chapter 4 estate the high similarity between the telomeres of the chromosome and pSCL4, which agrees with Medema's theory of both replicons having a common origin. In addition, we observed that pSCL4 is the only case of a *Streptomyces* replicon with the *tap* and *tpg* genes located in the central region of the molecule and, following this theory, it could be hypothesised that these genes were once positioned at the termini of the chromosome.

Considering a putative large transfer of genetic material from the chromosome to the plasmid, and that pSCL4 carries 21% of the genome's coding sequences, it is surprising to find that all the genes involved in primary metabolism are still encoded on the chromosome. However, as confirmed by the work explained in Chapter 6, the megaplasmid is still necessary for providing the terminal proteins Tap and Tpg for the replication of the chromosome ends, and elimination of pSCL4 leads to the chromosome to adapt to a circular form. Nevertheless, the fact that this circularisation is caused by the absence of Tap and Tpg alone is yet to be proved, and the fact that our *tap-tpg*-duplicated pSCL4-free strain Mc3 did not maintain the chromosome ends suggests more elements from pSCL4 might be key for telomere replication. Additionally, another gene was found missing in the chromosome, while present in pSCL4, the helicase-like encoding gene *ttrA*. This gene is a common element found at the sequence ends of linear replicons in *Streptomyces*, and in some cases like in pSCL3, on the same operon as *tap* and *tpg*. This implies that TtrA might have some implications towards telomere replication and must be further investigated.

In contrast to the fact that pSCL4 appears to be dispensable for the primary metabolism of *S. clavuligerus*, our pSCL4-free strains showed poor development in solid media, as well as no sporulation. This was further evidenced by observation and characterisation of microscopic images of pSCL4-free mycelium that confirmed the presence of significantly thinner and shorter branches which links the megaplasmid with at least the regulation of some

developmental functions. Additionally, the strains confirmed to carry circular chromosomes exhibited a significantly higher branching frequency, which could also associate this distinct morphology of frequent abortive branches to a change in chromosome segregation. In terms of secondary metabolism, while the M1 and Mc3 showed lower production of clavulanic acid, M2 exhibited similar levels to the WT. It is important to notice that this strain is the only one that maintains some pSCL4 material, as a 50 kb region was translocated into pSCL3, while carrying the shortest chromosome of the three strains. This could be a direct effect on production of clavulanic acid, or an indirect effect due to a better general fitness of this strain, as M2 shows an increased growth rate, the presence of this pSCL4 material appears to be beneficial for *S. clavuligerus*. Therefore, the effect of pSCL4 on the regulation of chromosomal genes must be further investigated to help answer the question of its role in this microorganism. This will also allow GSK to know the essential parts of pSCL4 to maintain and help improve *S. clavuligerus* fermentations.

Nevertheless, the different morphology and antibiotic production could also be explained by the effects that pSCL4-curing had in the chromosome. As we observed, a number of transcriptional regulators were deleted from the chromosome ends that could have key effects in *S. clavuligerus*. But also, the fact that the chromosome acquired a circular form might have influenced the expression of developmental genes. The construction of an artificial circular *S. clavuligerus* chromosome could help understand the effects of such conformation in this organism, as well as confirm the repercussion of pSCL4 in the chromosome beyond providing the telomere-replication machinery.

In conclusion, *S. clavuligerus* survives without pSCL4 making it dispensable, however, pSCL4-free strains exhibit deteriorated aspects in primary and secondary metabolism suggesting it is still essential for obtaining an optimum physiology. Whether this is a direct effect from the loss of essential unknown elements in the plasmid or an indirect effect from the deletion of chromosome ends and circularisation is still to be clarified.

8. CONCLUSIONS & FUTURE WORK

8.1 Conclusions.

Within this project we proposed to study the genome architecture of *S. clavuligerus*. To do so, we have established an optimised method for pulsed-field gel electrophoresis of *S. clavuligerus* strains. This procedure includes utilisation of HEPES-based buffers after confirming that *S. clavuligerus* is susceptible to Tris-dependent DNA degradation during electrophoresis. PFGE analyses allowed visualisation of the giant linear plasmids pSCL2, pSCL3, and, for the first time, the megaplasmid pSCL4. Additionally, PFGE of digested samples of *S. clavuligerus* was used to study chromosomal fragments and contrast them to the computational sequencing data to construct a physical map of the genome of this microorganism. Having an optimised method to study the genome architecture of *S. clavuligerus*, as well as the established genome map of the wild type strain, will benefit research done in this bacterium, particularly for GSK, in order to screen mutant strains in a rapid and economical manner.

Moreover, we have obtained the complete genome sequence of *S. clavuligerus* DSM 738. This was achieved by combination of next generation sequencing technologies Illumina and PacBio. Bioinformatic processing as well as manual assembly were necessary to obtain a closed assembly of the chromosome and the four plasmids of this organism. However, both Illumina and PacBio sequencing failed to obtain the telomeres of the five replicons and so further physical methodology was applied to obtain the sequence ends. The telomeres were purified using the self-ligation method and the sequences were characterised. It was confirmed that all the replicons carry archetypal telomeres, with a predicted folded structure of four loops following the Clover Leaves model. Interestingly the chromosome and megaplasmid carry the same telomeric sequences, while the rest of the replicons exhibit different telomeres, having only the first palindrome in common between the four sets of telomeres. The complete genome was finalised by the addition of the telomeres and the sequences were compared to other published *S. clavuligerus* genomes. It was observed that *S. clavuligerus* strains DSM 738 and ATCC 27064 strain carry almost identical chromosome sequences, implying they are probably the same strain. The final sequences were annotated yielding 8,268 coding sequences, 21% of them encoded on the megaplasmid. Nevertheless, all the core genes involved in primary metabolism were found on the chromosome. 43 putative biosynthetic gene clusters were predicted from the sequence of *S. clavuligerus*, 25

of them encoded on the chromosome, including the cephamycin-clavulanic acid supercluster. Furthermore, we identified three genes missing from the *S. clavuligerus* chromosome which must be complemented by the plasmids: *tap*, *tpg* and *ttrA*.

Given the terminal protein functionality is provided solely by the plasmids and the diversity of telomeric sequences in *S. clavuligerus*, we proposed to characterise the Tap and Tpg proteins in this organism. Three sets of proteins were identified, located on pSCL4, pSCL3 and pSCL2, and their amino acid sequence revealed conserved helix-turn-helix motifs. Similarly to the telomere sequences, these proteins exhibited low sequence identity with each other. In order to provide the chromosome its own *tap* and *tpg* genes, we incorporated a copy of the pSCL4 *tap-tpg* operon using ϕ BT1 *att*-site-specific integration on the chromosome. Following this, we confirmed it was not possible to eliminate these genes from pSCL4 using a double homologous recombination event as only strains that underwent a single crossover were obtained. Moreover, bacterial two-hybrid experiments confirmed Tap-Tap and Tpg-Tpg interactions that support the idea of linear chromosome and plasmids acquiring circular conformations mediated by association of proteins at each telomere. In contrast Tap-Tpg interactions were not observed in these kinds of experiments.

Finally, in order to study the essentiality of the megaplasmid, we obtained pSCL4-free strains by targeting the plasmid's *parB* gene with CRISPR/Cas9 technology. Whole genome sequencing of these revealed that at least two of the three characterised mutant strains carry circular chromosomes, confirming that the chromosome relies exclusively on the terminal proteins from pSCL4 to maintain the telomeres. This circularisation was predicted to have occurred at unspecific locations by non-homologous recombination of both arms. In addition, one of the strains is also missing pSCL3, while maintaining complete pSCL2 and pSCL1, which suggested that pSCL1 telomeres are supported by the proteins encoded on pSCL2. Furthermore, pSCL4-free strains exhibited a poor development in solid media and no sporulation, which was also confirmed by microscopic characterisation, revealing a phenotype of significant short, thin and frequent branches in comparison to the wild type. Additionally, disproving the original theory that elimination of pSCL4 would improve the production of clavulanic acid, this was shown to be equal or even lower than on the wild type. These results confirm that pSCL4 might have a role in regulation of important chromosomal genes and so, it is essential for the optimum fitness of *S. clavuligerus*.

8.2 Future work.

Future experiments would include utilising different techniques to disrupt the *tap-tpg* operon instead of a knockout, such as CRISPR latest technologies: CRISPR-BEST to introduce targeted mutations on these genes (Tong *et al.*, 2019), or CRISPR/dCas9 to temporally inhibit the transcription of these genes (Gjaltema and Schulz, 2018). This will confirm if the circularisation of the chromosome is due solely to the loss of these two genes or if other elements in pSCL4 are involved in the maintenance of chromosomal telomeres. In addition, in case of these technologies being successful in genetic modification of *S. clavuligerus* will represent an important breakthrough for future investigations of interesting genes, as we have evidenced in this work the difficulty for gene deletion by homologous recombination in this organism. An example of a gene to be studied would be *ttrA*, as it is missing from the chromosome and its role is unclear.

Another interesting experiment would be to overexpress the different Tap and Tpg proteins from *S. clavuligerus* in *E. coli* for subsequent purification. The purified proteins could be used for electrophoretic mobility shift assay (EMSA) to study the specific interactions of the three sets of terminal proteins with the five telomeres. Furthermore, obtaining the crystal structures of Tap and Tpg would help characterise the mechanism of action of the terminal proteins.

In order to study more specifically the effect of the loss of pSCL4 on gene regulation of *S. clavuligerus* chromosome, RNA sequencing of the pSCL4-mutants would establish their transcriptome and compare it to the wild type. It would be particularly interesting to study the expression of genes involved in mycelium development and sporulation, given the observations in this project. The expression level of genes involved in clavulanic acid production would also confirm the implication of pSCL4 which could be crucial for GSK to improve their industrial strains.

Furthermore, since we have confirmed the successful elimination of pSCL4 by targeting portioning genes with CRISPR/Cas9, this could be used to cure the other three plasmids and study their significance in the physiology of *S. clavuligerus*.

9. REFERENCES

- Abraham, E. P. and Chain, E. (1940) 'An enzyme from bacteria able to destroy penicillin. 1940.', *Reviews of infectious diseases*, pp. 677–678.
- Álvarez-Álvarez, R., Rodríguez-García, A., Martínez-Burgo, Y., Robles-Reglero, V., Santamarta, I., Pérez-Redondo, R., Martín, J. F., and Liras, P. (2014) 'A 1.8-Mb-reduced *Streptomyces clavuligerus* genome: Relevance for secondary metabolism and differentiation', *Applied Microbiology and Biotechnology*, 98(5), pp. 2183–2195.
- Álvarez-Álvarez, R., Botas, A., Albillos, S. M., Rumbero, A., Martín, J. F., and Liras, P. (2015) 'Molecular genetics of naringenin biosynthesis, a typical plant secondary metabolite produced by *Streptomyces clavuligerus*', *Microbial Cell Factories*, 14(1), p.178.
- Ambler, R. P. (1980) 'The structure of beta-lactamases.', *Phil. Trans. R. Soc. Lond.*, B289(1036), p. 321.
- Bachmann, B. O., Li, R. and Townsend, C. a. (1998) 'Beta-Lactam synthetase: A new biosynthetic enzyme', *Proceedings of the National Academy of Sciences*, 95(16), pp. 9082–9086.
- Baggaley, K. H., Brown, A. G., and Schofield, C. J. (1997) 'Chemistry and biosynthesis of clavulanic acid and other clavams', *Nat. Prod. Rep.*, 14, pp. 309–333.
- Baldwin, J. E., Lloyd M. D., Wha-Son B., Schofield C. J., Elson S. W., Baggaley K. H. and Nicholson N. H. (1993) 'A substrate analogue study on clavaminic acid synthase: possible clues to the biosynthetic origin of proclavamic acid', *Journal of the Chemical Society, Chemical Communications*, 6, pp. 500–502.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012) 'SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing', *Journal of Computational Biology*, 19(5), pp. 455-477.
- Bao, K. and Cohen, S. N. (2001) 'Terminal proteins essential for the replication of linear plasmids and chromosomes in *Streptomyces*', *Genes and Development*, 15(12), pp. 1518–1527.
- Bao, K. and Cohen, S. N. (2003) 'Recruitment of terminal protein to the ends of *Streptomyces* linear plasmids and chromosomes by a novel telomere-binding protein essential for linear DNA replication', *Genes and Development*, 17(6), pp. 774–785.
- Barka, E. A., Vatsa, P., Sanchez, L., Gaveau-Vaillant, N., Jacquard, C., Meier-Kolthoff, J. P., Klenk, H. P., Clément, C., Ouhdouch, Y., and van Wezel, G. P. (2016) 'Taxonomy, Physiology, and Natural Products of Actinobacteria', *Microbiology and Molecular Biology Reviews*, 80(1), pp. 1–43.
- Bentley, S.D., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Harper, D., Bateman, A., Brown, S., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Larke, L., Murphy, L., Oliver, K., Rabinowitsch, E., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Woodward, J., Barrell, B.G., Parkhill, J. and Hopwood, D. A. (2002) 'Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2)', *Nature*, 3(2), pp. 141–147.

- Bentley, S. D., Brown, S., Murphy, L. D., Harris, D. E., Quail, M. A., Parkhill, J., Barrell, B. G., McCormick, J. R., Santamaria, R. I., Losick, R., Yamasaki, M., Kinashi, H., Chen, C. W., Chandra, G., Jakimowicz, D., Kieser, H. M., Kieser, T., and Chater, K. F. (2004) 'SCP1, a 356 023 bp linear plasmid adapted to the ecology and developmental biology of its host, *Streptomyces coelicolor* A3(2)', *Molecular Microbiology*, 51(6), pp. 1615-1628.
- Bertani, G. (1951) 'Studies on lysogenesis I. The mode of phage liberation by lysogenic *Escherichia coli*.', *Journal of bacteriology*, 62(3), pp. 293–300.
- Bey, S. J., Tsou, M. F., Huang, C. H., Yang, C. C., and Chen, C. W. (2000) 'The homologous terminal sequence of the *Streptomyces lividans* chromosome and SLP2 plasmid', *Microbiology*, 146(4), pp.911-922.
- Beyazova, M. and Lechevalier, M. P. (1993) 'Taxonomic utility of restriction endonuclease fingerprinting of large DNA fragments from *Streptomyces* strains', *International Journal of Systematic Bacteriology*, 43(4), pp. 674-682.
- Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H., and Weber, T. (2019) 'antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline', *Nucleic acids research*, 47(1), pp. 81-87.
- Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114-2120.
- Boucher, H. W., Talbot, G. H., Bradley, J. S., Edwards, J. E., Gilbert, D., Rice, L. B., Scheld, M., Spellberg, B., and Bartlett, J. (2009) 'Bad bugs, no drugs: no ESCAPE! An update from the Infectious Diseases Society of America.', *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 48(1), pp. 1–12.
- Brown, A. G., Butterworth, D., Cole, M., Hanscomb, G., Hood, J. D., Reading, C., and Rolinson, G. N. (1976) 'Naturally-occurring β -lactamase inhibitors with antibacterial activity', *J Antibiot (Tokyo)*, 29(6), pp. 668–669.
- Bush, K. (1988) 'Beta-lactamase inhibitors from laboratory to clinic.', *Clinical microbiology reviews*, 1(1), pp. 109–23.
- Bush, L. M. and Johnson, C. C. (2000) 'Ureidopenicillins and beta-lactam/beta-lactamase inhibitor combinations.', *Infectious disease clinics of North America*, 14(2), pp. 409–33.
- Cantor, C. R., Gaal, A. and Smith, C. L. (1988) 'High-Resolution Separation and Accurate Size Determination in Pulsed-Field Gel Electrophoresis of DNA. 3. Effect of Electrical Field Shape', *Biochemistry*, 27(26), pp. 9204–9210.
- Cao, G., Zhong, C., Zong, G., Fu, J., Liu, Z., Zhang, G., and Qin, R. (2016) 'Complete Genome Sequence of *Streptomyces clavuligerus* F613-1, an Industrial Producer of Clavulanic Acid', *Genome Announcements*, 4(5), pp. 4–5.
- Chang, P. C. and Cohen, S. N. (1994) 'Bidirectional replication from an internal origin in a linear *Streptomyces* plasmid.', *Science (New York, N.Y.)*, 265(5174), pp. 952–4.
- Chardon-Loriaux, I., Charpentier, M. and Percheron, F. (1986) 'Isolation and characterization of a linear plasmid from *Streptomyces rimosus*', *FEMS Microbiology Letters*, 35(2-3), pp.151-155.
- Charusanti, P., Fong, N. L., Nagarajan, H., Pereira, A. R., Li, H. J., Abate, E. A., Su, Y., Gerwick,

W. H., and Palsson, B. O. (2012) 'Exploiting adaptive laboratory evolution of *Streptomyces clavuligerus* for antibiotic discovery and overproduction', *PLoS ONE*, 7(3), e33727.

Chater, K. F. (1993) 'Genetics of differentiation in *Streptomyces*', *Annual review of microbiology*, 47, pp. 685–713.

Chater, K. F. and Chandra, G. (2006) 'The evolution of development in *Streptomyces* analysed by genome comparisons', *FEMS Microbiology Reviews*, 30(5), pp. 651-672.

Chater, K. F. and Kinashi, H. (2007) '*Streptomyces* Linear Plasmids: Their Discovery, Functions, Interactions with Other Replicons, and Evolutionary Significance', in *Microbial Linear Plasmids, Microbiology Monographs, Springer*.

Chen, C. W., Yu, T. W., Lin, Y. S., Kieser, H. M., and Hopwood, D. A. (1993) 'The conjugative plasmid SLP2 of *Streptomyces lividans* is a 50 kb linear molecule', *Molecular Microbiology*, 7(6), pp.925-932.

Chen, C. W., Lin Y., Yang Y., Tsou M., Chang, H., Kieser H. M. and Hopwood D. A.(1994) 'The linear chromosomes of *Streptomyces*: structure and dynamics', *Actinomycetologica*, 8, pp. 103–112.

Chen, C. W., Huang, C. H., Lee, H. H., Tsai, H. H., and Kirby, R. (2002) 'Once the circle has been broken: Dynamics and evolution of *Streptomyces* chromosomes', *Trends in Genetics*, 18(10), pp. 522–529.

Chen, W., Qu, D., Zhai, L., Tao, M., Wang, Y., Lin, S., Price, N. P., and Deng, Z. (2010) 'Characterization of the tunicamycin gene cluster unveiling unique steps involved in its biosynthesis', *Protein and Cell*, 1(12), pp. 1093-1105.

Chen, W., He, F., Zhang, X., Chen, Z., Wen, Y., and Li, J. (2010) 'Chromosomal instability in *Streptomyces avermitilis*: Major deletion in the central region and stable circularized chromosome', *BMC Microbiology*, 10(1), p. 198.

Cobb, R. E., Wang, Y. and Zhao, H. (2015) 'High-Efficiency Multiplex Genome Editing of *Streptomyces* Species Using an Engineered CRISPR/Cas System', *ACS Synth. Biol.*, 4(6), pp. 2–7.

Combes, P., Till, R., Bee, S., and Smith, M. C. (2002) 'The *Streptomyces* genome contains multiple pseudo-attB sites for the ϕ C31-encoded site-specific recombination system', *Journal of Bacteriology*, 184(20), pp. 5746–5752.

Darling, A. C. E., Mau B., Blattner F. R. (2004) 'Mauve: Multiple alignment of conserved genomic sequence with rearrangements', *Genome Research*, 14(7), pp. 1394-1403.

Donczew, M., Mackiewicz P., Wróbel A., Flårdh K., Zakrzewska-Czerwinska J. and Jakimowicz D. (2016) 'ParA and ParB coordinate chromosome segregation with cell elongation and division during *Streptomyces* sporulation', *Open Biology*, 6(4), 150263.

Elander, R. P. (2003) 'Industrial production of beta-lactam antibiotics.', *Applied microbiology and biotechnology*, 61(5–6), pp. 385–392.

Evans, M. and Dyson, P. (1993) 'Pulsed-field gel electrophoresis of *Streptomyces lividans* DNA', *Trends in genetics : TIG*, 9(3), p. 72.

Fan, Y., Dai Y., Cheng Q., Zhang G., Zhang D., Fang P., Wu H., Bai L., Deng Z. and Qin Z. (2012) 'A self-ligation method for PCR-sequencing the telomeres of *Streptomyces* and

- Mycobacterium* linear replicons', *Journal of Microbiological Methods*. Elsevier B.V., 90(2), pp. 105–107.
- Fischer, G., Decaris, B. and Leblond, P. (1997) 'Occurrence of deletions, associated with genetic instability in *Streptomyces ambofaciens*, is independent of the linearity of the chromosomal DNA', *Journal of Bacteriology*, 179(14), pp. 4553–4558.
- Freese, N. H., Norris, D. C. and Loraine, A. E. (2016) 'Integrated genome browser: Visual analytics platform for genomics', *Bioinformatics*, 32(14), pp. 2089–2095.
- Fu, J., Qin, R., Zong, G., Zhong, C., Zhang, P., Kang, N., Qi, X., and Cao, G. (2019) 'The two-component system CepRS regulates the cephamycin C biosynthesis in *Streptomyces clavuligerus* F613-1', *AMB Express*, 9(1), p. 118.
- Gjaltema, R. and Schulz, E. (2018) 'CRISPR/dCas9 Switch Systems for Temporal Transcriptional Control.', *Methods Mol Biol*, 17767, pp. 167–185.
- Goffin, C. and Ghuysen, J.-M. (1998) 'Multimodular Penicillin-Binding Proteins: An Enigmatic Family of Orthologs and Paralogs', *Microbiol. Mol. Biol. Rev.*, 62(4), pp. 1079–1093.
- Gomez-Escribano, J. P., Castro, J. F., Razmilic, V., Chandra, G., Andrews, B., Asenjo, J. A., and Bibb, M. J. (2015) 'The *Streptomyces leeuwenhoekii* genome: De novo sequencing and assembly in single contigs of the chromosome, circular plasmid pSLE1 and linear plasmid pSLE2', *BMC Genomics*, 16(1), p. 485.
- Goshi, K., Uchida, T., Lezhava, A., Yamasaki, M., Hiratsu, K., Shinkawa, H., and Kinashi, H. (2002) 'Cloning and analysis of the telomere and terminal inverted repeat of the linear chromosome of *Streptomyces griseus*', *Journal of Bacteriology*, 184(12), pp. 3411–3415.
- Grant, J. R. and Stothard, P. (2008) 'The CGView Server: a comparative genomics tool for circular genomes', *Nucleic acids research*, 36 (Web Server issue), W181–W184.
- Gregory, M. and Smith, M. C. M. (2003) 'Integration Site for *Streptomyces* Phage ϕ BT1 and Development of Site-Specific Integrating Vectors', *Journal of Bacteriology*, 185(17), pp. 5320–5323.
- Gupta, R. S. (2011) 'Origin of diderm (Gram-negative) bacteria: Antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes', *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, pp. 171–182.
- Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013) 'QUAST: Quality assessment tool for genome assemblies', *Bioinformatics*, 29(8), pp. 1072–1075.
- Gust, B., Kieser, T. and Chater, K. (2002) 'PCR targeting system in *Streptomyces coelicolor* A3(2). http://streptomyces.org.uk/redirect/protocol_v1_4.pdf.
- Heidrich, C., Templin, M. F., Ursinus, A., Merdanovic, M., Berger, J., Schwarz, H., de Pedro, M. A., and Höltje, J. V. (2001) 'Involvement of N-acetylmuramyl-L-alanine amidases in cell separation and antibiotic-induced autolysis of *Escherichia coli*', *Molecular Microbiology*, 41(1), pp. 167–178.
- Higgins, C. E. and Kastner, R. E. (1971) '*Streptomyces clavuligerus* sp. nov., a B-Lactam Antibiotic Producer', *International Journal of Systematic Bacteriology*, 21(4), pp. 326–331.
- Hirochika, H. and Sakaguchi, K. (1982) 'Analysis of linear plasmids isolated from

- Streptomyces*: Association of protein with the ends of the plasmid DNA', *Plasmid*, 7(1), pp. 59–65.
- Hopwood, D. A. (2006) 'Soil to genomics: the *Streptomyces* chromosome.', *Annual review of genetics*, 40(1), pp. 1–23.
- Hsiao, N. H. and Kirby, R. (2008) 'Comparative genomics of *Streptomyces avermitilis*, *Streptomyces cattleya*, *Streptomyces maritimus* and *Kitasatospora aureofaciens* using a *Streptomyces coelicolor* microarray system', *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 93(1-2), pp. 1–25.
- Huang, C. H., Lin, Y. S., Yang, Y. L., Huang, S. W., and Chen, C. W. (1998) 'The telomeres of *Streptomyces* chromosomes contain conserved palindromic sequences with potential to form complex secondary structures', *Molecular Microbiology*, 28(5), pp. 905–916.
- Huang, C. H., Chen, C. Y., Tsai, H. H., Chen, C., Lin, Y. S., and Chen, C. W. (2003) 'Linear plasmid SLP2 of *Streptomyces lividans* is a composite replicon', *Molecular Microbiology*, 47(6), pp. 1563–1576.
- Huang, C. H., Tsai, H. H., Tsay, Y. G., Chien, Y. N., Wang, S. L., Cheng, M. Y., Ke, C. H., and Chen, C. W. (2007) 'The telomere system of the *Streptomyces* linear plasmid SCP1 represents a novel class', *Molecular Microbiology*, 63(6), pp. 1710–1718.
- Huang, J., Shi, J., Molle, V., Sohlberg, B., Weaver, D., Bibb, M. J., Karoonuthaisiri, N., Lih, C. J., Kao, C. M., Buttner, M. J., and Cohen, S. N. (2005) 'Cross-regulation among disparate antibiotic biosynthetic pathways of *Streptomyces coelicolor*', *Molecular Microbiology*, 58(5), pp. 1276–1287.
- Hung, T. V., Malla, S., Park, B. C., Liou, K., Lee, H. C., and Sohng, J. K. (2007) 'Enhancement of clavulanic acid by replicative and integrative expression of *ccaR* and *cas2* in *Streptomyces clavuligerus* NRRL3585', *Journal of Microbiology and Biotechnology*, 17(9), pp. 1538–1545.
- Hwang, S., Lee, N., Jeong, Y., Lee, Y., Kim, W., Cho, S., Palsson, B. O., and Cho, B. K. (2019) 'Primary transcriptome and translome analysis determines transcriptional and translational regulatory elements encoded in the *Streptomyces clavuligerus* genome', *Nucleic acids research*. Oxford University Press, 47(12), pp. 6114–6129.
- Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M., and Omura, S. (2003) 'Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*', *Nat. Biotechnol.*, 21(5), pp. 526–531.
- Jakimowicz, D., Majka, J., Messer, W., Speck, C., Fernandez, M., Martin, M. C., Sanchez, J., Schauwecker, F., Keller, U., Schrempf, H., and Zakrzewska-Czerwińska, J. (1998) 'Structural elements of the *Streptomyces* oriC region and their interactions with the DnaA protein', *Microbiology*, 144 (Pt 5), pp. 1281–1290.
- Jensen, S. E., Paradkar, A. S., Mosher, R. H., Anders, C., Beatty, P. H., Brumlik, M. J., Griffin, A., and Barton, B. (2004) 'Five additional genes are involved in clavulanic acid biosynthesis in *Streptomyces clavuligerus*', *Antimicrobial Agents and Chemotherapy*, 48(1), pp. 192–202.
- Jensen, S. E. and Paradkar, A. S. (1999) 'Biosynthesis and molecular genetics of clavulanic acid', *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 75(1–2), pp. 125–133.
- Kämpfer, P. (2006) 'The Prokaryotes: Volume 3: Archaea. Bacteria: Firmicutes,

- Actinomycetes', in Dworkin, M. et al. (eds). New York, NY: Springer New York, pp. 538–604.
- Karimova, G., Pidoux, J., Ullmann, A., and Ladant, D. (1998) 'A bacterial two-hybrid system based on a reconstituted signal transduction pathway', *Proceedings of the National Academy of Sciences of the United States of America*, 95(10), pp. 5752–5756.
- Keen, C. L., Mendelovitz, S., Cohen, G., Aharonowitz, Y., and Roy, K. L. (1988) 'Isolation and characterization of a linear DNA plasmid from *Streptomyces clavuligerus*', *MGG Molecular & General Genetics*, 212(1), pp. 172–176.
- Kieser, H. M., Kieser, T. and Hopwood, D. A. (1992) 'A combined genetic and physical map of the *Streptomyces coelicolor* A3(2) chromosome', *Journal of Bacteriology*, 174(17), pp. 5496–5507.
- Kieser, T., Hopwood D.A., Kieser T., Bibb M., Buttner M. and Chater K. (2000) 'Practical *Streptomyces* Genetics', *John Innes Centre Ltd.*, p. 529.
- Kim, H. S., Lee, Y. J., Lee, C. K., Choi, S. U., Yeo, S. H., Hwang, Y. I., Yu, T. S., Kinoshita, H., and Nihira, T. (2004) 'Cloning and characterization of a gene encoding the gamma-butyrolactone autoregulator receptor from *Streptomyces clavuligerus*.' *Archives of microbiology*, 182(1), pp. 44–50.
- Kinashi, H. and Shimaji, M. (1987) 'Detection of giant linear plasmids in antibiotic producing strains of streptomyces by the ofage technique', *Journal of Antibiotics*, 40(6), pp. 913–916.
- Kinashi, H., Shimaji, M. and Sakai, A. (1988) 'Giant linear plasmids in *Streptomyces* which code for antibiotic biosynthesis genes', *Nature*, 328(6129), pp. 454–456.
- Kirby, R. and Hopwood, D. A. (1977) 'Genetic Determination of Methylenomycin Synthesis by the SCP1 Plasmid of *Streptomyces coelicolor* A3(2)', *Microbiology*, 98(1), pp. 239–252.
- Kirby, R., Wright, L. F. and Hopwood, D. A. (1975) 'Plasmid-determined antibiotic synthesis and resistance in *Streptomyces coelicolor*', *Nature*, 254(5497), pp. 265–267.
- Kong, K. F., Schneper, L. and Mathee, K. (2010) 'Beta-lactam antibiotics: From antibiosis to resistance and bacteriology', *APMIS*, pp. 1–36.
- Koort, J. M., Lukinmaa, S., Rantala, M., Unkila, E., and Siitonen, A. (2002) 'Technical Improvement To Prevent DNA Degradation of Enteric Pathogens in Pulsed-Field Gel Electrophoresis Technical Improvement To Prevent DNA Degradation of Enteric Pathogens in Pulsed-Field Gel Electrophoresis', *Journal of clinical microbiology*, 40(9), pp. 3497–3499.
- Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018) 'MEGA X: Molecular evolutionary genetics analysis across computing platforms', *Molecular Biology and Evolution*, 35(6), pp. 1547–1549.
- Kumar, S., Stecher, G. and Tamura, K. (2016) 'MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets', *Molecular biology and evolution*, 33(7), pp. 1870–1874.
- Labeda, D. P., Goodfellow, M., Brown, R., Ward, A. C., Lanoot, B., Vannanneyt, M., Swings, J., Kim, S. B., Liu, Z., Chun, J., Tamura, T., Oguchi, A., Kikuchi, T., Kikuchi, H., Nishii, T., Tsuji, K., Yamaguchi, Y., Tase, A., Takahashi, M., Sakane, T., Suzuki K. I., and Hatano, K. (2012) 'Phylogenetic study of the species within the family *Streptomycetaceae*', *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 101(1), pp. 73–

104.

Leblond, P., Francou, F. X., Simonet, J. M., and Decaris, B. (1990) 'Pulsed-field gel electrophoresis analysis of the genome of *Streptomyces ambofaciens* strains', *FEMS Microbiology Letters*, 72(1–2), pp. 79–88.

Leblond, P., Redenbach, M. and Cullum, J. (1993) 'Physical map of the *Streptomyces lividans* 66 genome and comparison with that of the related strain *Streptomyces coelicolor* A3(2)', *Journal of Bacteriology*, 175(11), pp. 3422–3429.

Letunic, I. and Bork, P. (2019) 'Interactive Tree Of Life (iTOL) v4: recent updates and new developments', *Nucleic acids research*, 47(W1), W256–W259.

Levene, S. D. (1992) 'Pulsed-Field Gel Electrophoresis', *Methods in Molecular Biology*, 12, pp. 347–65.

Lezhava, A., Mizukami, T., Kajitani, T., Kameoka, D., Redenbach, M., Shinkawa, H., Nimi, O., and Kinashi, H. (1995) 'Physical map of the linear chromosome of *Streptomyces griseus*', *Journal of Bacteriology*, 177(22), pp. 6492–6498.

Li, B., and Walsh, C. T. (2010) 'Identification of the gene cluster for the dithiolopyrrolone antibiotic holomycin in *Streptomyces clavuligerus*.' , *Proceedings of the National Academy of Sciences of the United States of America*, 107(46), pp. 19731–5.

Lin, Y. S., Kieser, H. M., Hopwood, D. A., and Chen, C. W. (1993) 'The chromosomal DNA of *Streptomyces lividans* 66 is linear', *Molecular Microbiology*, 10(5), pp. 923–933.

Lin, Y. and Chen, C. W. (1997) 'Instability of artificially circularized chromosomes of *Streptomyces lividans* ', *Molecular Microbiology*, 26(4), pp. 709–719

Liras, P., Gomez-Escribano, J. P. and Santamarta, I. (2008) 'Regulatory mechanisms controlling antibiotic production in *Streptomyces clavuligerus*.' , *Journal of industrial microbiology & biotechnology*, 35(7), pp. 667–676.

Liras, P. and Rodríguez-García, A. (2000) 'Clavulanic acid, a β -lactamase inhibitor: biosynthesis and molecular genetics', *Applied Microbiology and Biotechnology*, 54(4), pp. 467–475.

Liu, G., Chater, K. F., Chandra, G., Niu, G., and Tan, H. (2013) 'Molecular regulation of antibiotic biosynthesis in *Streptomyces*', *Microbiology and molecular biology reviews : MMBR*, 77(1), pp. 112–143.

Liu, J., Li, J., Dong, H., Chen, Y., Wang, Y., Wu, H., Li, C., Weaver, D. T., Zhang, L., and Zhang, B. (2017) 'Characterization of an Lrp/AsnC family regulator SCO3361, controlling actinorhodin production and morphological development in *Streptomyces coelicolor*', *Applied Microbiology and Biotechnology*, 101(14), pp. 5773–5783.

MacNeil, D. J., Gewain, K. M., Ruby, C. L., Dezeny, G., Gibbons, P. H., and MacNeil, T. (1992) 'Analysis of *Streptomyces avermitilis* genes required for avermectin biosynthesis utilizing a novel integration vector.' , *Gene*, 111(1), pp. 61–68.

Majiduddin, F. K., Materon, I. C. and Palzkill, T. G. (2002) 'Molecular analysis of beta-lactamase structure and function.' , *International journal of medical microbiology : IJMM*, 292(2), pp. 127–37.

Manteca, A., Alvarez, R., Salazar, N., Yagüe, P., and Sanchez, J. (2008) 'Mycelium

differentiation and antibiotic production in submerged cultures of *Streptomyces coelicolor*', *Applied and Environmental Microbiology*, 74(12), pp. 3877–3886

Manteca, A., Fernandez, M. and Sanchez, J. (2006) 'Cytological and biochemical evidence for an early cell dismantling event in surface cultures of *Streptomyces antibioticus*', *Research in Microbiology*, 157(2), pp. 143–152.

McLeod, M. P., Warren, R. L., Hsiao, W. W., Araki, N., Myhre, M., Fernandes, C., Miyazawa, D., Wong, W., Lillquist, A. L., Wang, D., Dosanjh, M., Hara, H., Petrescu, A., Morin, R. D., Yang, G., Stott, J. M., Schein, J. E., Shin, H., Smailus, D., Siddiqui, A. S., Marra M.A., Jones S. J. M., Holt R., Brinkman F. S. L., Miyauchi K., Fukuda M., Davies J. E., Mohn W. W. and Eltis, L. D. (2006) 'The complete genome of *Rhodococcus* sp. RHA1 provides insights into a catabolic powerhouse', *Proceedings of the National Academy of Sciences*, 103(42), pp. 15582–15587.

Medema, M. H., Trefzer, A., Kovalchuk, A., van den Berg, M., Müller, U., Heijne, W., Wu, L., Alam, M. T., Ronning, C. M., Nierman, W. C., Bovenberg, R. A., Breitling, R., and Takano, E. (2010) 'The sequence of a 1.8-Mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways', *Genome Biology and Evolution*, 2(1), pp. 212–224.

Mochizuki, S., Hiratsu, K., Suwa, M., Ishii, T., Sugino, F., Yamada, K., and Kinashi, H. (2003) 'The large linear plasmid pSLA2-L of *Streptomyces rochei* has an unusually condensed gene organization for secondary metabolism', *Molecular Microbiology*, 48(6), pp. 1501–1510.

Mosher, R. H., Paradkar, A. S., Anders, C., Barton, B., and Jensen, S. E. (1999) 'Genes specific for the biosynthesis of clavam metabolites antipodal to clavulanic acid are clustered with the gene for clavamate synthase 1 in *Streptomyces clavuligerus*', *Antimicrobial Agents and Chemotherapy*, 43(5), pp. 1215–1224.

Musialowski, M. S., Flett, F., Scott, G. B., Hobbs, G., Smith, C. P., and Oliver, S. G. (1994) 'Functional evidence that the principal DNA replication origin of the *Streptomyces coelicolor* chromosome is close to the *dnaA-gyrB* region', *Journal of Bacteriology*, 176(16), pp. 5123–5125.

Muth, G., Frese, D., Kleber, A., and Wohlleben, W. (1997) 'Mutational analysis of the *Streptomyces lividans* *recA* gene suggests that only mutants with residual activity remain viable', *Molecular and General Genetics*, 255(4), pp. 420–428.

Netolitzky, D. J., Wu, X., Jensen, S. E., and Roy, K. L. (1995) 'Giant linear plasmids of beta-lactam antibiotic producing *Streptomyces*', *FEMS Microbiology Letters*, 131(1), pp. 27–34.

Neu, H. C. and Fu, K. P. (1978) 'Clavulanic acid, a novel inhibitor of β -lactamases', *Antimicrobial Agents and Chemotherapy*, 14(5), pp. 650–655

Nindita, Y., Nishikawa, T., Arakawa, K., Wang, G., Ochi, K., Qin, Z., and Kinashi, H. (2013) 'Chromosomal circularization of the model *Streptomyces* species, *Streptomyces coelicolor* A3(2)', *FEMS Microbiology Letters*, pp. 149–155.

Nindita, Y., Cao, Z., Yang, Y., Arakawa, K., Shiwa, Y., Yoshikawa, H., Tagami, M., Lezhava, A., and Kinashi, H. (2015) 'The *tap-tpg* gene pair on the linear plasmid functions to maintain a linear topology of the chromosome in *Streptomyces rochei*', *Molecular Microbiology*, 95(5), pp. 846–858.

- O'Neill, J. (2016) 'Tackling Drug-Resistant Infections Globally : Final Report and Recommendations the Review on', *London, United Kingdom: Review on Antimicrobial Resistance*, (May).
- Ohnishi, Y., Ishikawa, J., Hara, H., Suzuki, H., Ikenoya, M., Ikeda, H., Yamashita, A., Hattori, M., and Horinouchi, S. (2008) 'Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350', *Journal of Bacteriology*, 190(11), pp. 4050–4060.
- Okonechnikov, K., Conesa, A. and García-Alcalde, F. (2016) 'Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data', *Bioinformatics*, 32(2), pp. 292–294.
- Paradkar, A. (2013) 'Clavulanic acid production by *Streptomyces clavuligerus*: biogenesis, regulation and strain improvement.', *The Journal of antibiotics*, 66(7), pp. 411–420.
- Parajuli, N., Viet, H. T., Ishida, K., Tong, H. T., Lee, H. C., Liou, K., and Sohng, J. K. (2005) 'Identification and characterization of the *afsR* homologue regulatory gene from *Streptomyces peucetius* ATCC 27952', *Research in Microbiology*, 156(5-6), pp. 707–712.
- Pérez-Redondo, R., Rodríguez-García, A., Martín, J. F., and Liras, P. (1998) 'The *claR* gene of *Streptomyces clavuligerus*, encoding a LysR-type regulatory protein controlling clavulanic acid biosynthesis, is linked to the clavulanate-9-aldehyde reductase (*car*) gene', *Gene*, 211(2), pp. 311–321.
- Pérez-Redondo, R., Rodríguez-García, A., Martín, J. F., and Liras, P. (1999) 'Deletion of the *pyc* gene blocks clavulanic acid biosynthesis except in glycerol-containing medium: Evidence for two different genes in formation of the C3 unit', *Journal of Bacteriology*, 181(22), pp. 6922–6928.
- Piette, A., Derouaux, A., Gerkens, P., Noens, E. E., Mazzucchelli, G., Vion, S., Koerten, H. K., Titgemeyer, F., De Pauw, E., Leprince, P., van Wezel, G. P., Galleni, M., and Rigali, S. (2005) 'From dormant to germinating spores of *Streptomyces coelicolor* A3(2): New perspectives from the *crp* null mutant', *Journal of Proteome Research*, 4, pp. 1699–1708.
- Postma, M. and Goedhart, J. (2019) 'Plotsofdata—a web app for visualizing data together with their summaries', *PLoS Biology*, 17(3), e3000202.
- Qin, Z. and Cohen, S. N. (1998) 'Replication at the telomeres of the *Streptomyces* linear plasmid pSLA2', *Molecular Microbiology*, 28(5), pp. 893–903.
- Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., and Manke, T. (2016) 'deepTools2: a next generation web server for deep-sequencing data analysis', *Nucleic acids research*, 44(W1), W160–W165.
- Ray, T., Mills, A. and Dyson, P. (1995) 'Tris-dependent oxidative DNA strand scission during electrophoresis', *Electrophoresis*, 16(6), pp. 888–894.
- Ray, T., Weaden, J. and Dyson, P. (1992) 'Tris-dependent site-specific cleavage of *Streptomyces lividans* DNA', *FEMS Microbiology Letters*, 96(2–3), pp. 247–252.
- Reading, C. and Cole, M. (1977) 'Clavulanic acid: a beta-lactamase-inhibiting beta-lactam from *Streptomyces clavuligerus*', *Antimicrobial Agents and Chemotherapy*, 11(5), pp. 852–857.
- Romero, J., Liras, P. and Martin, J. F. (1986) 'Utilization of ornithine and arginine as specific

precursors of clavulanic acid', *Applied and Environmental Microbiology*, 52(4), pp. 892–897.

Salas, M., Mellado, R. P., and Viñuela, E. (1978) 'Characterization of a protein covalently linked to the 5' termini of the DNA of *Bacillus subtilis* phage $\phi 29$ ', *Journal of Molecular Biology*, 119(2), pp. 269–291.

Salas, M. (1991) 'Protein-Priming of DNA Replication', *Annual Review of Biochemistry*, 60, pp. 39–71.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual.*, New York.

Sanger, F. and Coulson, A. R. (1975) 'A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase', *Journal of Molecular Biology*, 94(3), pp. 441–448.

Santamarta, I., Pérez-Redondo, R., Lorenzana, L. M., Martín, J. F., and Liras, P. (2005) 'Different proteins bind to the butyrolactone receptor protein ARE sequence located upstream of the regulatory *ccaR* gene of *Streptomyces clavuligerus*', *Molecular Microbiology*, 56(3), pp. 824–835.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J. Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., and Cardona, A. (2012) 'Fiji: An open-source platform for biological-image analysis', *Nature Methods*, 9(7), pp. 676–682.

Schwartz, D. C. and Cantor, C. R. (1984) 'Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis', *Cell*, 37(1), pp. 67–75.

Schwedock, J., McCormick, J. R., Angert, E. R., Nodwell, J. R., and Losick, R. (1997) 'Assembly of the cell division protein FtsZ into ladder-like structures in the aerial hyphae of *Streptomyces coelicolor*', *Molecular Microbiology*, 25(5), pp. 847–858.

Shin, S. C., Ahn, D. H., Kim, S. J., Lee, H., Oh, T. J., Lee, J. E., and Park, H. (2013) 'Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes', *PLoS One*, 8(7), e68824.

Song, J. Y., Kim, E. S., Kim, D. W., Jensen, S. E., and Lee, K. J. (2009) 'A gene located downstream of the clavulanic acid gene cluster in *Streptomyces clavuligerus* ATCC 27064 encodes a putative response regulator that affects clavulanic acid production', *Journal of Industrial Microbiology and Biotechnology*, 36(2), pp. 301–311.

Song, J. Y., Jeong, H., Yu, D. S., Fischbach, M. A., Park, H. S., Kim, J. J., Seo, J. S., Jensen, S. E., Oh, T. K., Lee, K. J., and Kim, J. F. (2010) 'Draft genome sequence of *Streptomyces clavuligerus* NRRL 3585, a producer of diverse secondary metabolites', *Journal of Bacteriology*, 192(23), pp. 6317–6318.

Spratt, B. G., Hedge, P. J., te Heesen, S., Edelman, A., and Broome-Smith, J. K. (1986) 'Kanamycin-resistant vectors that are analogues of plasmids pUC8, pUC9, pEMBL8 and pEMBL9', *Gene*, 41(2-3), pp. 337–341.

Sullivan, M. J., Petty, N. K. and Beatson, S. A. (2011) 'Easyfig: A genome comparison visualizer', *Bioinformatics*, 27(7), pp. 1009–1010.

Tahlan, K., Park, H. U., Wong, A., Beatty, P. H., and Jensen, S. E. (2004) 'Two sets of

- paralogous genes encode the enzymes involved in the early stages of clavulanic acid and clavam metabolite biosynthesis in *Streptomyces clavuligerus*', *Antimicrobial Agents and Chemotherapy*, 48(3), pp. 930–939.
- Takano, E. (2006) 'γ-Butyrolactones: *Streptomyces* signalling molecules regulating antibiotic production and differentiation', *Current Opinion in Microbiology*, 9(3), pp. 287–294.
- Tong, Y., Whitford, C. M., Robertsen, H. L., Blin, K., Jørgensen, T. S., Klitgaard, A. K., Gren, T., Jiang, X., Weber, T., and Lee, S. Y. (2019) 'CRISPR-BEST: a highly efficient DSB-free base editor for filamentous actinomycetes', *Proceedings of the National Academy of Sciences of the United States of America*, 116(41), pp. 20366–20375.
- Tsai, H. H., Huang, C. H., Tessmer, I., Erie, D. A., and Chen, C. W. (2011) 'Linear *Streptomyces* plasmids form superhelical circles through interactions between their terminal proteins', *Nucleic Acids Research*, 39(6), pp. 2165–2174.
- Umeyama, T., Lee, P. C., Ueda, K., and Horinouchi, S. (1999) 'An AfsK/AfsR system involved in the response of aerial mycelium formation to glucose in *Streptomyces griseus*', *Microbiology*, 145 (Pt 9), pp. 2281–2292.
- Vara, J., Lewandowska-Skarbek, M., Wang, Y. G., Donadio, S., and Hutchinson, C. R. (1989) 'Cloning of genes governing the deoxysugar portion of the erythromycin biosynthesis pathway in *Saccharopolyspora erythraea* (*Streptomyces erythreus*).', *Journal of Bacteriology*, 171(11), pp. 5872–5881.
- 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.', *Microbiology and molecular biology reviews : MMBR*, 71(3), pp. 495–548.
- Vivian, A. (1971) 'Genetic control of fertility in *Streptomyces coelicolor* A3(2): plasmid involvement in the interconversion of UF and IF strains.', *J Gen Microbiol*, 69, pp. 353–364.
- Vivian, A. and Hopwood, D. (1970) 'Genetic control of fertility in *Streptomyces coelicolor* A3(2): the IF fertility type', *J Gen Microbiol*, 64, pp. 101–117.
- Vivian, A. and Hopwood, D. (1975) 'Genetic control of fertility in *Streptomyces coelicolor* A3(2): new kinds of donor strains.', *J Gen Microbiol*, 76, pp. 147–162.
- van der Vliet, P. C. (1995) 'Adenovirus DNA replication.', *Curr. Top. Microbiol. Immunol.*, 199, pp. 1–30.
- Volff, J. N., Viell, P. and Altenbuchner, J. (1997) 'Artificial circularization of the chromosome with concomitant deletion of its terminal inverted repeats enhances genetic instability and genome rearrangement in *Streptomyces lividans*', *Molecular and General Genetics*, 253(6), pp. 753–760.
- Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., and Earl, A. M. (2014) 'Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement', *PLoS One*, 9(11), e112963.
- Wang, L., Chen, S., Xu, T., Taghizadeh, K., Wishnok, J. S., Zhou, X., You, D., Deng, Z., and Dedon, P. C. (2007) 'Phosphorothioation of DNA in bacteria by *dnd* genes', *Nature Chemical Biology*, 3(11), pp. 709–710.

- Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009) 'Jalview Version 2-A multiple sequence alignment editor and analysis workbench', *Bioinformatics*, 25(9), pp. 1189–1191.
- Watve, M. G., Tickoo, R., Jog, M. M., and Bhole, B. D. (2001) 'How many antibiotics are produced by the genus *Streptomyces*?', *Archives of Microbiology*, pp. 386–390.
- Weaver, D., Karoonuthaisiri, N., Tsai, H. H., Huang, C. H., Ho, M. L., Gai, S., Patel, K. G., Huang, J., Cohen, S. N., Hopwood, D. A., Chen, C. W., and Kao, C. M. (2004) 'Genome plasticity in *Streptomyces*: Identification of 1 Mb TIRs in the *S. coelicolor* A3(2) chromosome', *Molecular Microbiology*, 51(6), pp. 1535–1550
- Weber, T., Charusanti, P., Musiol-Kroll, E. M., Jiang, X., Tong, Y., Kim, H. U., and Lee, S. Y. (2015) 'Metabolic engineering of antibiotic factories: New tools for antibiotic production in actinomycetes', *Trends in Biotechnology*, pp. 15–26.
- White, A. R., Kaye, C., Poupard, J., Pypstra, R., Woodnutt, G., and Wynne, B. (2004) 'Augmentin® (amoxicillin/clavulanate) in the treatment of community-acquired respiratory tract infection: a review of the continuing development of an innovative antimicrobial agent', *Journal of Antimicrobial Chemotherapy*, 53, pp. 3–20.
- Wolanski, M., Wali, R., Tilley, E., Jakimowicz, D., Zakrzewska-Czerwinska, J., and Herron, P. (2011) 'Replisome trafficking in growing vegetative hyphae of *Streptomyces coelicolor* A3(2)', *Journal of Bacteriology*, 193(5), pp. 1273–1275.
- Wu, T. K., Busby, R. W., Houston, T. A., McIlwaine, D. B., Egan, L. A., and Townsend, C. A. (1995) 'Identification, cloning, sequencing, and overexpression of the gene encoding proclavaminic amidino hydrolase and characterization of protein function in clavulanic acid biosynthesis', *Journal of Bacteriology*, 177(13), pp. 3714–3720.
- Wu, X. and Roy, K. L. (1993) 'Complete nucleotide sequence of a linear plasmid from *Streptomyces clavuligerus* and characterization of its RNA transcripts', *Journal of Bacteriology*, 175(1), pp. 37–52.
- Xu, L., Dong, Z., Fang, L., Luo, Y., Wei, Z., Guo, H., Zhang, G., Gu, Y. Q., Coleman-Derr, D., Xia, Q., and Wang, Y. (2019) 'OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species', *Nucleic acids research*, 47(W1), W52–W58.
- Yamasaki, M., Ikuto, Y., Ohira, A., Chater, K., and Kinashi, H. (2003) 'Limited regions of homology between linear and circular plasmids encoding methylenomycin biosynthesis in two independently isolated streptomycetes', *Microbiology*, 149(Pt 5), pp.1351–1356.
- Yang, C. C., Tseng, S. M., Pan, H. Y., Huang, C. H., and Chen, C. W. (2017) 'Telomere associated primase Tap repairs truncated telomeres of *Streptomyces*', *Nucleic Acids Research*, 45, pp. 785–794.
- Yang, C. C., Huang, C. H., Li, C. Y., Tsay, Y. G., Lee, S. C., and Chen, C. W. (2002) 'The terminal proteins of linear *Streptomyces* chromosomes and plasmids: a novel class of replication priming proteins', *Molecular Microbiology*, 43, pp. 297–305.
- Yang, C. C., Tseng, S. M. and Chen, C. W. (2015) 'Telomere-associated proteins add deoxynucleotides to terminal proteins during replication of the telomeres of linear chromosomes and plasmids in *Streptomyces*', *Nucleic Acids Research*, 43(13), pp. 6373–

6383.

Yanisch-Perron, C., Vieira, J. and Messing, J. (1985) 'Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mpl8 and pUC19 vectors', *Gene*, 33(1), pp. 103–119.

Zhang, Q., Chen, Q., Zhuang, S., Chen, Z., Wen, Y., and Li, J. (2015) 'A MarR family transcriptional regulator, DptR3, activates daptomycin biosynthesis and morphological differentiation in *Streptomyces roseosporus*', *Applied and Environmental Microbiology*, 81(11), pp. 3753–3765.

Zhou, X., Deng, Z., Firmin, J. L., Hopwood, D. A., and Kieser, T. (1988) 'Site-specific degradation of *Streptomyces lividans* DNA during electrophoresis in buffers contaminated with ferrous iron', *Nucleic Acids Research*, 16(10), pp. 4341–4352.

Zhou, Z., Gu, J., Li, Y. Q., and Wang, Y. (2012) 'Genome plasticity and systems evolution in *Streptomyces*.', *BMC bioinformatics*, 13(Suppl 10), S8.

Zuker, M. (2003) 'Mfold web server for nucleic acid folding and hybridization prediction', *Nucleic Acids Research*, 31(13), pp. 3406–3415.

10. APPENDIX

10.1 Custom scripts

The following custom macro script was written and used to edit microscopy images in FIJI in order to separate composite images, scale them, automated scale contrast, and merge fluorescence and phase contrast channels into one image for further downstream analyses:

```
open("image_location");

run("Set Scale...", "distance=1 known=0.09 unit=um");

run("Stack to Images");

selectWindow("Image-0001");

//run("Brightness/Contrast...");

run("Enhance Contrast", "saturated=0.35");

run("Apply LUT");

selectWindow("Image-0003");

run("Enhance Contrast", "saturated=0.35");

run("Apply LUT");

selectWindow("Image-0002");

run("Enhance Contrast", "saturated=0.35");

run("Apply LUT");

run("Merge Channels...", "c1=Image-0003 c2=Image-0002");

run("Scale Bar...", "width=10 height=8 font=28 color=White background=None
location=[Lower Right] bold overlay");

saveAs("Tiff", "saving_location");

close();

selectWindow("Image-0001");

run("Scale Bar...", "width=10 height=8 font=28 color=White background=None
location=[Lower Right] bold overlay");

saveAs("Tiff", "saving_location");

close();
```

10.2 Supplementary figures.

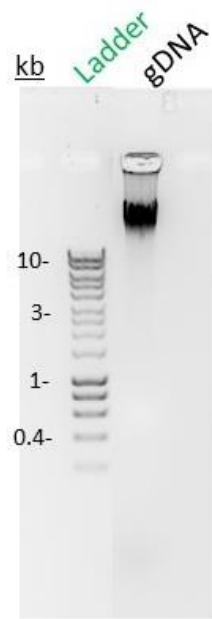


Figure A4. Gel electrophoresis of genomic DNA of *S. clavuligerus* DSM 738 used for Illumina sequencing and PacBio sequencing

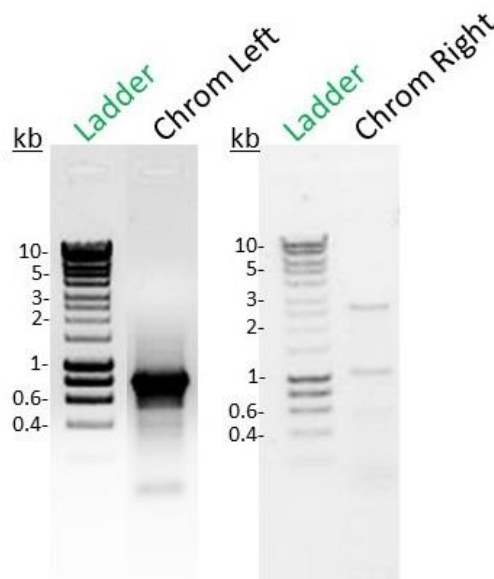


Figure A2. Gel electrophoresis of inverted-PCR amplification results of self-ligated chromosomal telomeres.

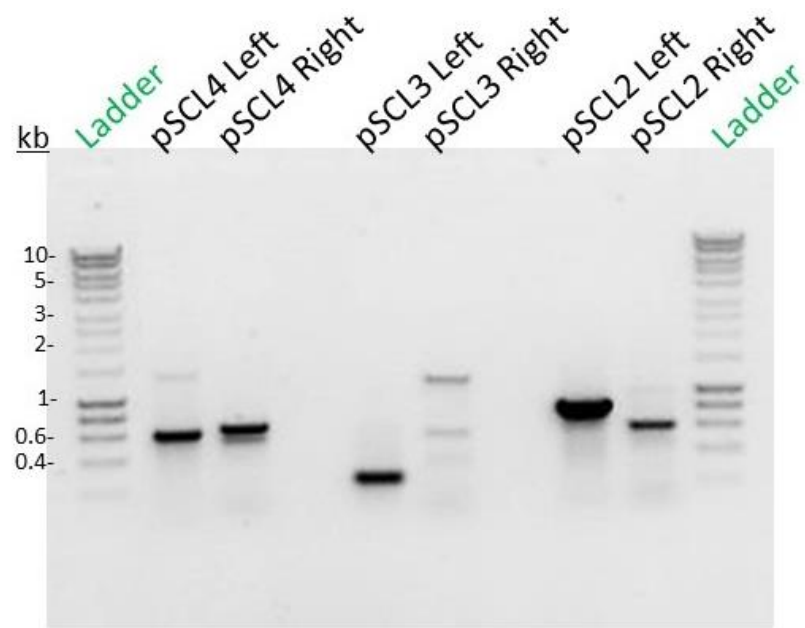


Figure A3. Gel electrophoresis of inverted-PCR amplification results of self-ligated pSCL4, pSCL3 and pSCL2 telomeres.

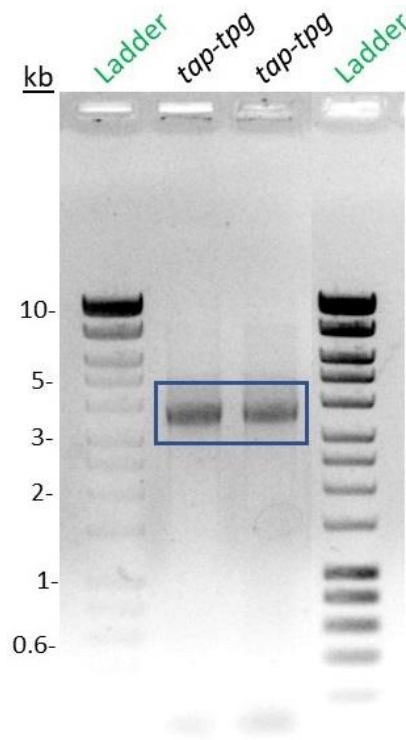


Figure A4. Gel electrophoresis of the *tap-tpg* genes amplified by PCR, with a size of 3.5 kb.

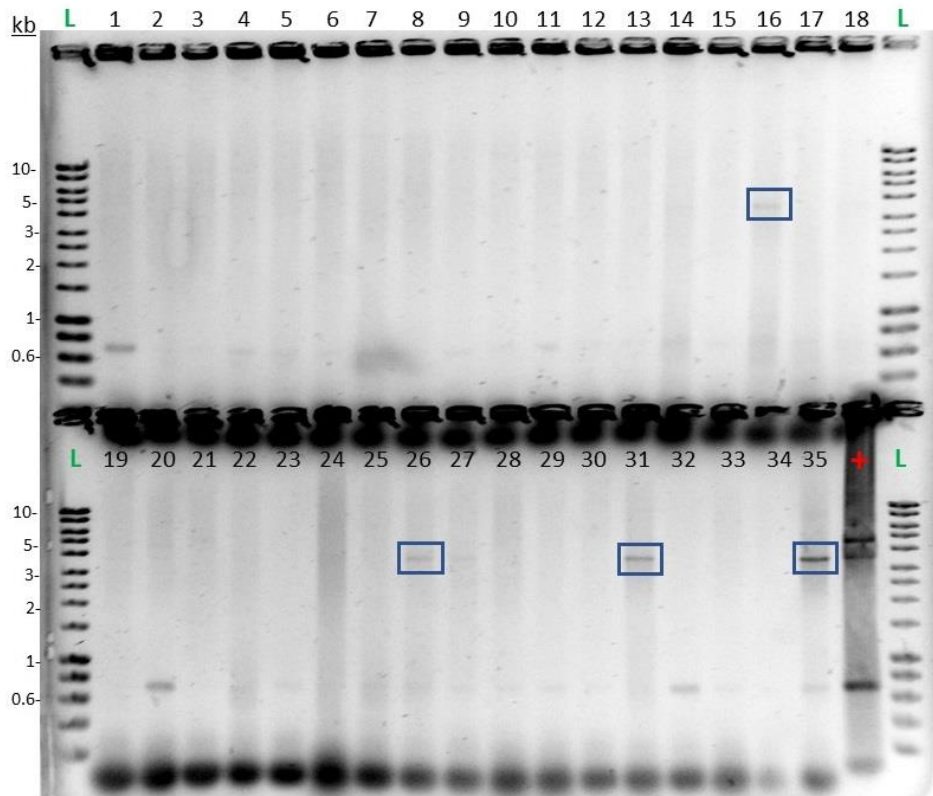


Figure A5. Gel electrophoresis of the products from colony PCR of 35 transformants. Four colonies (16, 26, 31 and 35) exhibited bands for the *tap-tpg* genes (band at 3.5 kb).



Figure A6. Confirmation of conjugation of pLIS22 and pMS82 into *S. clavuligerus*. pLIS22/pM82 transconjugants and WT (*S. clavuligerus* DSM 738) grown on GYM media with or without

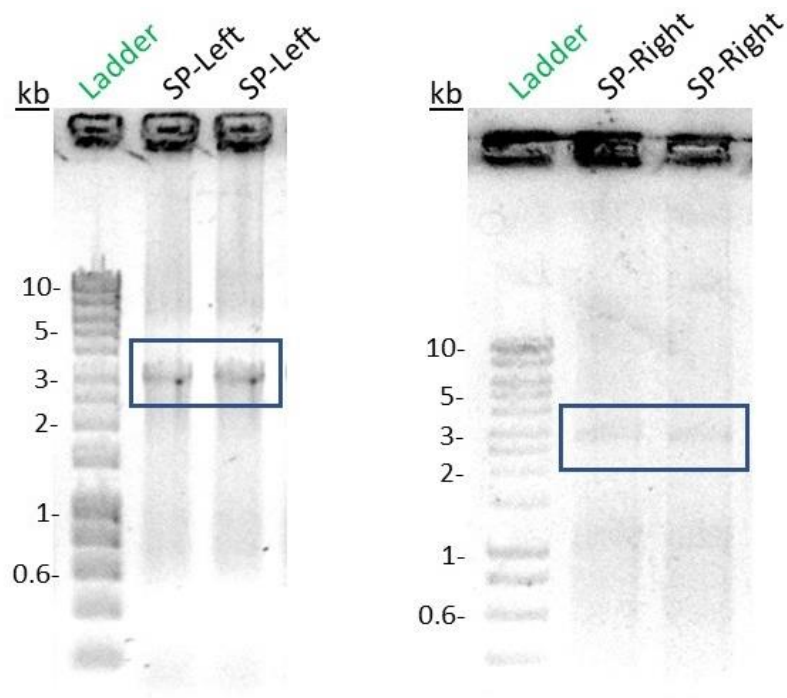


Figure A7. Gel electrophoresis of 3 kb flanking regions of the *tap-tpg* operon (SP-Left and SP-Right) amplified by PCR.

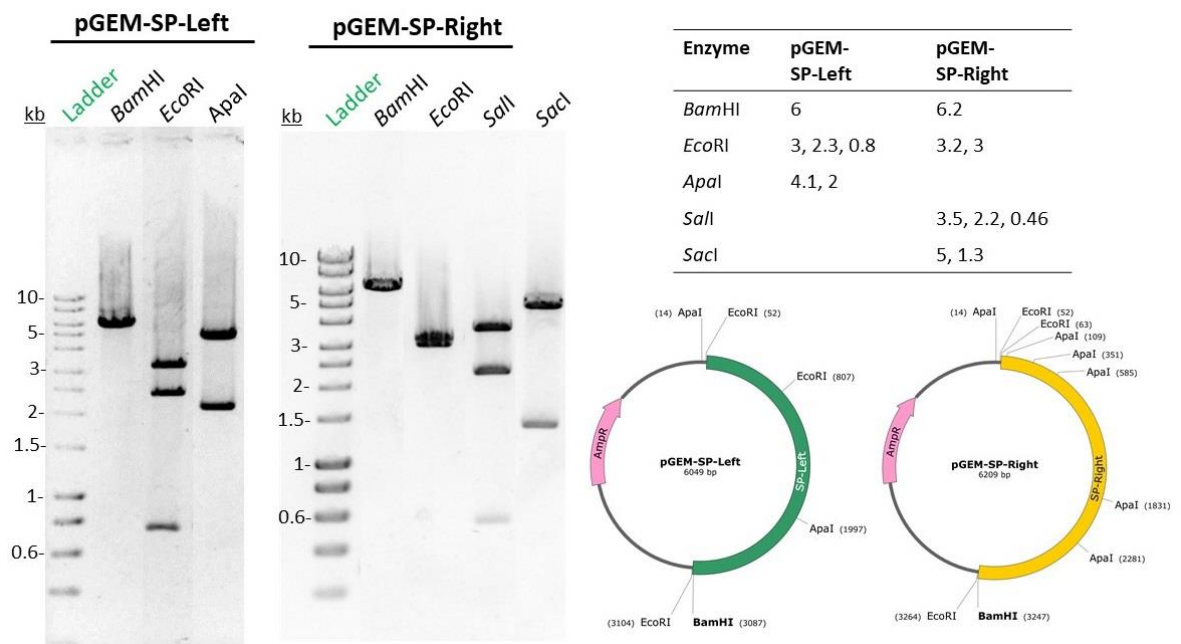


Figure A8. Confirmation of cloning of SP-Left and SP-Right into pGEM-T Easy. Gel electrophoresis of products from digest with *Bam*HI, *Eco*RI, *Apa*I, *Sal*I or *Sac*I. The table indicates the expected band sizes. Maps of the expected constructs with restriction enzyme sites.



Figure A9. Confirmation of construction of pLIS401 and pLIS402. Gel electrophoresis of products from digest with *Bam*HI, and *Sall*. The table indicates the expected band sizes. Maps of the expected constructs with restriction enzyme sites.

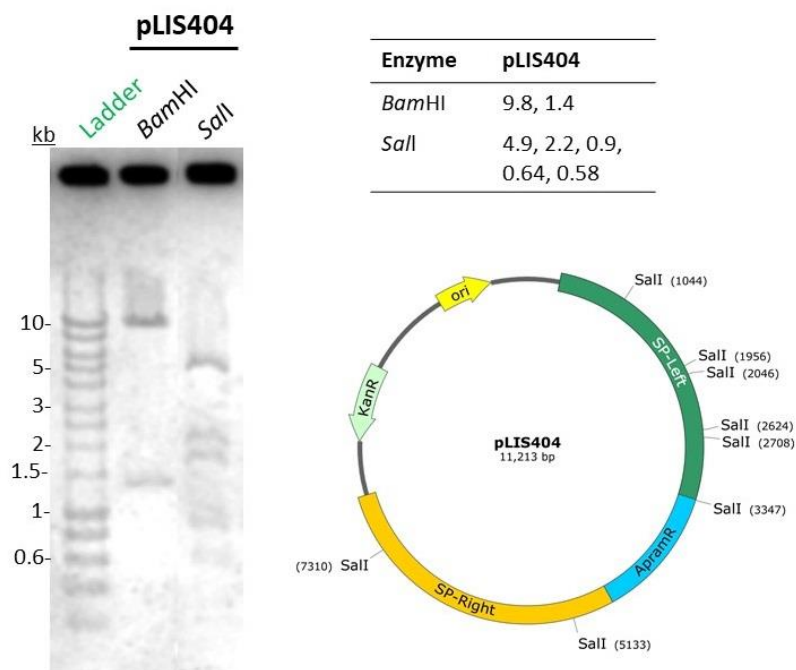


Figure A10. Confirmation of construction of pLIS404. Gel electrophoresis of products from digest with *Bam*HI, and *Sall*. The table indicates the expected band sizes. Maps of the expected construct with restriction enzyme sites.

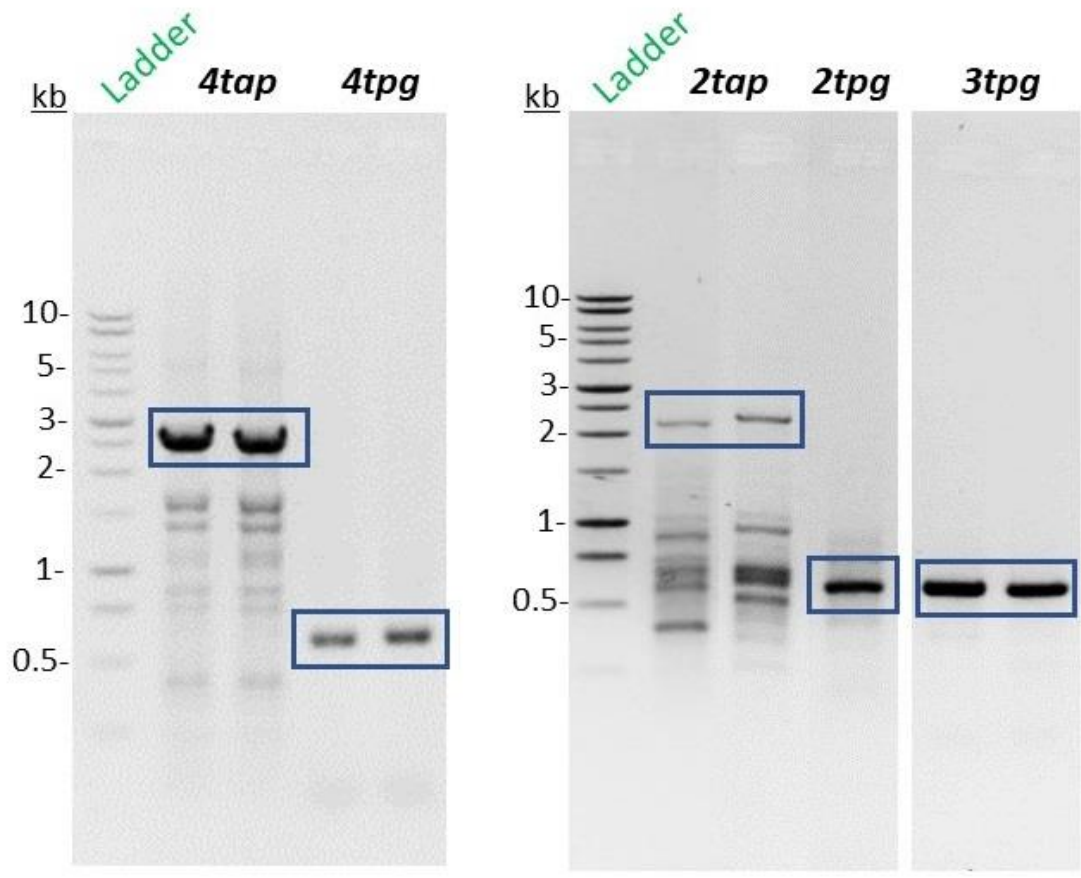


Figure A11. Gel electrophoresis of the *tap* and *tpg* genes from pSCL4, pSCL2 and pSCL3 amplified by PCR for the construction of BACTH plasmids.

10.3 Supplementary tables

Table A1. List of genes with known function deleted from the chromosome ends in the pSCL4-free strains M1, M2 and Mc3. Highlighted in orange are the genes belonging to the core genome. Highlighted in green are the discussed transcriptional regulators.

Gene function	Position	Missing in M1	Missing in M2	Missing in Mc3
Putative transposase	539-721	yes	yes	yes
Mobile element protein	956-1558	yes	yes	yes
O-methyltransferase family 3	9544-10206	yes	yes	yes
Polyketide synthase	10230-11318	yes	yes	yes
Acyl-CoA dehydrogenase	11315-12424	yes	yes	yes
Acyl carrier protein	12417-12725	yes	yes	yes
3-hydroxybutyryl-CoA dehydrogenase	12725-13609	yes	yes	yes
Modular polyketide synthase	13606-20031	yes	yes	yes
Modular polyketide synthase	20102-30892	yes	yes	yes
Modular polyketide synthase	30939-43265	yes	yes	yes
Modular polyketide synthase	43320-59720	yes	yes	yes
Modular polyketide synthase	59737-74754	yes	yes	
Malonyl CoA-acyl carrier protein transacylase	76147-77106	yes	yes	
Putative cytochrome P450 hydroxylase	77500-78768	yes	yes	
Putative membrane protein	79147-80256	yes	yes	
Phenylpropionate dioxygenase	83382-84449	yes	yes	
SAM-dependent methyltransferase	91403-92140	yes	yes	
Transcriptional regulator	93571-94338	yes	yes	
Phage tail sheath protein FI	96185-98449	yes	yes	
VgrG protein	103003-103707	yes	yes	
ATPase, AAA family	118766-120982		yes	
Efflux ABC transporter, permease/ATP-binding protein	121799-123700		yes	
Geranylgeranyl diphosphate synthase	125077-126183		yes	
Transcriptional regulator, AraC family	127468-128088		yes	
Excinuclease ABC subunit A paralog	140617-142905		yes	
Chitodextrinase	143224-145569		yes	
MBL-fold metallo-hydrolase superfamily	147951-148478		yes	
Metal-sensitive transcriptional repressor	149548-149829		yes	
Transcriptional regulator, AcrR family	151193-151777		yes	
Acetyltransferase, GNAT family	155176-155730		yes	

Putative phosphotransferase	155741-156538	yes
1,4-dihydroxy-2-naphthoate polyprenyltransferase	157744-158799	yes
Carbonic anhydrase, beta class	159517-160269	yes
Putative cytochrome P450 hydroxylase	160285-161805	yes
Alkaline serine exoprotease A precursor	162508-163746	yes
Aldo/keto reductase	164217-165203	yes
Gluconate 5-dehydrogenase	167454-168233	yes
Guanine deaminase	168441-169004	yes
Putative oxidoreductase	169170-169793	yes
Putative SimX4 homolog	170046-171017	yes
4'-phosphopantetheinyl transferase entD	171014-171781	yes
Secreted protein	174173-175525	yes
Putative regulatory protein	175672-177012	yes
Putative ferredoxin	177058-177276	yes
Putative cytochrome P450 hydroxylase	177346-177813	yes
Potassium-transporting ATPase B chain	6477878-6479980	yes
Potassium-transporting ATPase A chain	6479980-6481644	yes
Osmosensitive K ⁺ channel histidine kinase KdpD	6481993-6484530	yes
Sporulation control protein Spo0M	6485472-6486440	yes
Peptidase M48, Ste24p precursor	6486588-6487520	yes
Transcriptional regulator, Mecl family	6487528-6487899	yes
Lead, cadmium, zinc and mercury transporting ATPase	6489742-6492507	yes
Uncharacterized protein with VanW-like domain	6495159-6496952	yes
3-hydroxybutyrate dehydrogenase	6498294-6499091	yes
Uncharacterized MFS-type transporter	6499310-6500749	yes
Regulator of polyketide synthase expression	6500867-6502870	yes
Cysteine desulfurase	6502935-6503969	yes
Transcriptional regulator, LysR family	6507658-6508431	yes
Cytosine deaminase	6509066-6510256	yes

Dolichol-phosphate mannosyltransferase	6516217-6517452	yes
Glycosyl transferase	6519332-6520591	yes
Cytochrome P450	6530599-6532041	yes
Endo-1,4-beta-glucanase Z	6532188-6533147	yes
Ferredoxin	6533436-6533642	yes
Ferredoxin reductase	6533629-6534999	yes
Putative cytochrome P450 hydroxylase	6535025-6536257	yes
Hydrogenase assembly protein HoxX	6536304-6538013	yes
Homoserine O-acetyltransferase	6541212-6542363	yes
O-acetylhomoserine sulfhydrylase	6542372-6543730	yes
Antibiotic biosynthesis monooxygenase	6544441-6544779	yes
2-hydroxy-3-oxopropionate reductase	6544776-6545687	yes
Putative cytochrome P450 hydroxylase	6546688-6547944	yes
Acetyltransferase	6548024-6548533	yes
Thioesterase	6548565-6549356	yes
Electron transfer flavoprotein, alpha subunit	6549495-6550448	yes
Electron transfer flavoprotein, beta subunit	6550662-6551450	yes
Transcriptional regulator, AcrR family	6551629-6552315	yes
Transcriptional regulator, AcrR family	6553017-6553607	yes
Lactoylglutathione lyase	6553739-6554152	yes
ABC transporter, permease protein	6554300-6554980	yes
ABC transporter, permease protein	6554977-6555648	yes
ABC transporter, substrate-binding protein	6555648-6556760	yes
Acetyl xylan esterase	6558358-6559329	yes
ABC transporter, ATP-binding protein	6559651-6560814	yes
Endoglucanase celA precursor	6561006-6562508	yes
Putative prolyl aminopeptidase	6562649-6563950	yes
Nitroreductase family protein	6564889-6565590	yes
Putative oxidoreductase	6565797-6566798	yes
Acyl-CoA dehydrogenase, long-chain specific	6566869-6568053	yes

Acyl-CoA dehydrogenase	6568050-6569219		yes	
ABC transporter, substrate-binding protein	6569394-6571028		yes	
ABC transporter, permease protein 1	6571025-6572047	yes	yes	
ABC transporter, permease protein 2	6572040-6572999	yes	yes	
ABC transporter, ATP-binding protein	6572996-6573973	yes	yes	
ABC transporter, ATP-binding protein	6574006-6574668	yes	yes	
ABC transporter, ATP-binding protein	6574623-6574862	yes	yes	
Transmembrane protein	6576912-6577106	yes	yes	
Uncharacterized flavin-binding protein	6578119-6578119	yes	yes	
Oxidoreductase, short-chain dehydrogenase/reductase family	6582135-6582896	yes	yes	
Phosphoenolpyruvate-dihydroxyacetone phosphotransferase	6598966-6599676	yes	yes	yes
Phosphoenolpyruvate-dihydroxyacetone phosphotransferase	6599738-6600736	yes	yes	yes
Putative oxidoreductase	6600902-6601963	yes	yes	yes
Putative FAD-dependent oxidoreductase	6601968-6603713	yes	yes	yes
Putative FAD-dependent oxidoreductase	6603710-6605419	yes	yes	yes
Capsular polysaccharide biosynthesis fatty acid synthase WcbR	6605409-6615056	yes	yes	yes
Fatty acid hydroxylase family	6615368-6616291	yes	yes	yes
Phytoene dehydrogenase	6618625-6620247	yes	yes	yes
RsbU-domain-containing protein	6627781-6630582	yes	yes	yes
Chaperone protein ClpB	6630789-6633362	yes	yes	yes
Transcriptional regulator AfsR	6633442-6635367	yes	yes	yes
Dienelactone hydrolase	6636329-6637669	yes	yes	yes
Transcriptional regulator, MarR family	6637803-6638399	yes	yes	yes
Putative phosphotransferase	6640533-6641285	yes	yes	yes
Acetyltransferase	6642866-6643435	yes	yes	yes
Putative cytochrome P450 hydroxylase	6646612-6647835	yes	yes	yes
Polyketide synthase	6648827-6650695	yes	yes	yes
Putative cytochrome P450 hydroxylase	6652631-6653725	yes	yes	yes
Undecaprenyl-phosphate galactosephosphotransferase	6655089-6656567	yes	yes	yes
GDP-mannose 6-dehydrogenase	6656647-6657963	yes	yes	yes

Glycosyltransferase	6657953-6659224	yes	yes	yes
Putative membrane protein	6660614-6662188	yes	yes	yes
Asparagine synthetase	6662192-6664120	yes	yes	yes
Myo-inositol 2-dehydrogenase	6664154-6666343	yes	yes	yes
Putative membrane protein	6668653-6669951	yes	yes	yes
Glycosyl transferase, group 1 family protein	6669960-6671246	yes	yes	yes
SAM-dependent methyltransferase	6671243-6672478	yes	yes	yes
Glucose-1-phosphate cytidyltransferase	6672475-6673296	yes	yes	yes
UDP-glucose 4-epimerase	6673944-6674969	yes	yes	yes
MATE family protein	6676276-6677607	yes	yes	yes
Putative glycosyltransferase	6677604-6678560	yes	yes	yes
Polysaccharide pyruvyl transferase CsaB	6678557-6679810	yes	yes	yes
dTDP-4-dehydrorhamnose 3,5-epimerase	6679807-6680448	yes	yes	yes
Aminotransferase, Class III pyridoxal-phosphate dependent	6680456-6681781	yes	yes	yes
Putative membrane-associated phospholipid phosphatase	6681789-6682646	yes	yes	yes
Transcriptional regulator, AsnC family	6687606-6688049	yes	yes	yes
Pentalene oxygenase	6688130-6689491	yes	yes	yes
Terpene synthase, metal-binding	6689558-6690493	yes	yes	yes
Geranylgeranyl diphosphate synthase	6692434-6693522	yes	yes	yes
Aldehyde dehydrogenase	6694850-6696373	yes	yes	yes
Choline oxidase	6696559-6698133	yes	yes	yes
Urea carboxylase-related amino acid permease	6698294-6700009	yes	yes	yes
Betaine aldehyde dehydrogenase	6700006-6701475	yes	yes	yes
GAF domain-containing protein / Signal transduction response regulator	6701670-6702935	yes	yes	yes
N-acetylmuramoyl-L-alanine amidase	6703053-6703451	yes	yes	yes
SWF/SNF family helicase	6704539-6705969	yes	yes	yes
Helicase, SNF2/RAD54 family	6705966-6709088	yes	yes	yes
Integral membrane protein	6709315-6709494	yes	yes	yes
Acetyltransferase, GNAT family	6709534-6710208	yes	yes	yes
2-amino-3-ketobutyrate coenzyme A ligase	6712212-6713393	yes	yes	yes

L-threonine 3-dehydrogenase	6713552-6714580	yes	yes	yes
Transcriptional regulator, LysR family	6715767-6716930	yes	yes	yes
Putative regulatory protein	6717057-6719102	yes	yes	yes
Uncharacterized MFS-type transporter	6720097-6721308	yes	yes	yes
Putative membrane protein	6722540-6722989	yes	yes	yes
Putative integral membrane efflux protein	6724226-6725416	yes	yes	yes
Argininosuccinate lyase	6726704-6727897	yes	yes	yes
Diaminopimelate decarboxylase	6731459-6732751	yes	yes	yes
Glutathione synthase/Ribosomal protein S6 modification enzyme	6733814-6734749	yes	yes	yes
Competence-like protein	6734979-6735431	yes	yes	yes
Carbon starvation protein A	6735744-6737867	yes	yes	yes
Putative cytochrome P450	6738301-6739719	yes	yes	yes
Cysteine desulfurase	6740100-6741593	yes	yes	yes
Lipase	6741997-6743130	yes	yes	yes
Cobalt-zinc-cadmium resistance protein	6744006-6745136	yes	yes	yes
Histone acetyltransferase HPA2	6746154-6746609	yes	yes	yes
Salicylate esterase	6747142-6747873	yes	yes	yes