



**HAL**  
open science

# Towards combinatorial biosynthesis of pyrrolamide antibiotics in *Streptomyces*

Céline Aubry

► **To cite this version:**

Céline Aubry. Towards combinatorial biosynthesis of pyrrolamide antibiotics in *Streptomyces*. Biochemistry [q-bio.BM]. Université Paris Saclay (COMUE), 2019. English. NNT : 2019SACLS245 . tel-02955510

**HAL Id: tel-02955510**

**<https://theses.hal.science/tel-02955510>**

Submitted on 2 Oct 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Towards combinatorial biosynthesis of pyrrolamide antibiotics in *Streptomyces*

Thèse de doctorat de l'Université Paris-Saclay  
préparée à l'Université Paris-Sud

École doctorale n°577  
Structure et Dynamique des Systèmes Vivants (SDSV)  
Spécialité de doctorat : Sciences de la vie et de la Santé

Thèse présentée et soutenue à Orsay, le 30/09/19, par

**Céline AUBRY**

## Composition du Jury :

<b>Matthieu Jules</b> Professeur, Agroparistech (MICALIS)	Président du Jury
<b>Yanyan Li</b> Chargée de recherche, MNHN (MCAM)	Rapportrice
<b>Stéphane Cociancich</b> Chercheur, CIRAD (BGPI)	Rapporteur
<b>Annick Méjean</b> Professeure, Université Paris-Diderot (LIED)	Examinatrice
<b>Hasna Boubakri</b> Maitre de conférences, Université Claude Bernard Lyon I (Ecologie microbienne)	Examinatrice
<b>Sylvie Lautru</b> Chargée de recherche, CNRS (I2BC)	Directrice de thèse



# Acknowledgements

J'ai insisté pour rédiger l'ensemble de ma thèse en anglais. La logique voudrait donc que cette section soit écrite en anglais également, mais je n'ai pas pu m'y résoudre. Je vous présente mes excuses pour cette entorse linguistique.

Si le doctorat est l'occasion de creuser un projet principalement réalisé par le doctorant, ce n'est en aucun cas un travail individuel qui s'accomplit seul. De nombreuses personnes, par leur aide sur tous les aspects d'un projet doctoral, qu'ils soient scientifiques, administratifs, sociaux ou émotionnels, m'ont permis de vivre pleinement cette expérience. Ces personnes sont trop nombreuses pour être toutes citées ici, mais sachez que les moments partagés font partie des souvenirs irremplaçables que je garde de ces quatre années dans l'équipe de Microbiologie Moléculaire des Actinomycètes (MMA) à l'Institut de Biologie Intégrative de la Cellule.

Je souhaite exprimer toute ma reconnaissance à Sylvie Lautru, ma directrice de thèse. Tout au long du projet, tu m'as encadrée avec beaucoup de patience et de disponibilité. Tu m'as témoigné une grande confiance concernant la réalisation des expériences et m'as encouragée à gagner en autonomie. Tu as toujours écouté mes suggestions avant de me donner ta vision des choses, puis de discuter ensemble de la suite. Tu as également été présente pour écouter mes doutes et me rassurer dans les moments plus difficiles. Merci de m'avoir guidée tout en me laissant libre de choisir mon chemin et ma manière de faire les choses.

Je remercie les membres de mon jury de thèse, Yanyan Li, Stéphane Cociancich, Annick Méjean, Hasna Boubakri et Matthieu Jules, d'avoir accepté de consacrer du temps à évaluer mes travaux de recherche. Annick Méjean a également fait partie de mes comités de suivi de thèse, avec Muriel Gondry, Christiane Elie et Jean-Luc Pernodet, et je les remercie pour leurs conseils et leur bienveillance concernant mon projet.

Concernant les procédures administratives, j'ai bénéficié de l'aide toujours efficace de Muriel Decraëne et de Catherine Drouet, ainsi que de Marie-Hélène Sarda et de Martine Denis, et de Blandine Champion-Grosjean. Votre expertise des multiples procédures m'a permis de gagner un temps précieux.

Je tiens également à remercier Paolo Clérico, qui a réalisé la synthèse chimique d'un précurseur de l'anthelvencine spécialement pour une de mes expériences, mais aussi Laurent Micouin et son équipe, qui se sont penchés et se penchent encore aujourd'hui sur la caractérisation structurale de l'anthelvencine avec beaucoup de ténacité. Un grand merci également à Zhilai Hong, Yanyan Li et Soizic Prado, qui ont permis l'analyse en spectrométrie de masse à haute résolution des anthelvencines, et m'ont donné les clefs nécessaires à la compréhension des résultats.

D'autres petites mains ont directement participé à mon projet, celles de mes stagiaires Jennifer Perrin et Yacine Sellah. Merci beaucoup, d'une part vous m'avez aidée à faire avancer mon projet, et d'autre part, j'ai beaucoup appris en vous encadrant. Jennifer, tu as été ma première stagiaire, et j'ai été impressionnée par ta vivacité d'esprit et ta motivation. Tu t'es aussi très bien intégrée à l'équipe et je me rappelle la période de ton stage comme un moment plein de rires et de bonne humeur. Yacine, ton projet portait sur des parties plus difficiles de ma thèse, mais tu n'as jamais abandonné et ta persévérance a porté ses fruits, merci d'avoir souhaité autant que moi la réussite de ce projet.

Mon implication dans l'encadrement de l'équipe iGEM GO-Paris-Saclay 2018 a également consisté une expérience inestimable. Merci aux encadrants, Stéphanie, Philippe et Mahnaz, j'ai apprécié le résultat de notre coopération ! Merci aussi à tous les étudiants, pour leur enthousiasme et leur créativité, je remercie en particulier les étudiants que j'ai accompagnés à Boston pour présenter les résultats, une aventure pleine de rebondissements et de bons souvenirs.

J'aimerais également remercier tous les membres passés et présents des équipes MES et MMA, nos interactions ont rendu mon séjour ici inestimable. Merci en particulier à Jean-Luc, pour ton oreille attentive et ta connaissance incroyable des événements qui méritent un apéritif, à Christiane pour les anecdotes invraisemblables et les conseils pragmatiques, à Laetitia pour tous les jeudis aux expériences ratées et ta fidélité presque sans faille aux repas du CESFO, à Luisa pour ton efficacité et pour oser aborder les problèmes sans détours, à Alba pour les soirées au dehors et la complicité en 105, à Jerzy pour les petites astuces de labo et les pauses « café » chez Sylvain, à Laura, pour le partage paisible de l'espace et du projet pyrrolamides, à Sylvain, pour l'accueil dans ta cuisine et les discussions aux sujets divers, à Armel, pour les connaissances sur tout et les bonbons, à Manue, pour ta prévenance à mon égard et la gestion infatigable des soucis de séquençage, à Soumaya, pour les conseils spécial doctorants qui facilitent la vie, à Mathieu, pour les bonjours impromptus et le cactus, à Brittany, pour ta gaieté contagieuse et le voyage à Boston, à Corinne, pour la bonne humeur et la musique, à Stéphanie, pour ton dynamisme et cette incroyable capacité à remotiver les gens, à Hervé, pour ton dévouement à toujours ressusciter « mamie » et ton recul sur la science, à Marc, pour ne pas avoir jeté l'HPLC par la fenêtre malgré la tentation, à Aaron, pour des barbecues mémorables, et à Marie-Joëlle et Michelle, pour avoir toujours eu froid pour moi. Merci aussi à Nelly, pour avoir repris le flambeau en l'absence de Sylvain, et m'avoir permis de poursuivre mes expériences en toute sérénité.

Il a fait bon vivre au bâtiment 400, où les gens sont toujours prêts à apporter leur aide sur un appareil, ou à partager un moment convivial autour d'un repas ou d'un verre. Merci à tous pour cette ambiance des plus agréables ! Merci aussi aux « anciens » de l'équipe MMA qui sont restés dans les environs, Audrey, Florence et Drago, pour leurs conseils avisés.

Un énorme merci à Clara, ma voisine de palier. A nous deux, nous avons égayé le couloir, parfois un peu bruyamment, et acheté les vivres nécessaires aux apéritifs du vendredi soir. Mais bien plus que ça, tu es devenue une véritable amie, et les activités en dehors qui ont commencé par des sessions piscines se sont bien vite diversifiées. Je ne compte plus nos multiples discussions, ni les sorties à l'Opéra ou au cinéma. Ma vie pendant le doctorat n'aurait pas été la même sans toi !

Je souhaite adresser un remerciement spécial à Lucile. Sans toi, cette aventure n'aurait jamais commencé. Merci de m'avoir conseillée ce projet que tu avais initié pendant ton stage !

Je souhaite aussi exprimer une reconnaissance toute particulière à Valentin T., un chercheur en génétique et linguistique des populations humaines exceptionnel. Nos discussions sur le milieu de la recherche scientifique m'ont permis de prendre du recul, et d'élargir ma vision du monde. Tes réflexions sur l'éthique et le sens de la vie ont également généré en moi beaucoup de questions, le genre de questions qu'on ne peut pas se permettre d'ignorer. Je continue à chercher mes réponses.

Merci aux amis de tous horizons qui ont manifesté de l'intérêt pour mon projet, et de l'empathie pour mes mésaventures. Merci surtout à ma famille, pour toute sa patience, alors que je ruminais mes problèmes techniques ou parlais dans des divagations impossibles à suivre. Merci d'avoir été un soutien sans faille, à tout moment, pendant ces années d'études qui aboutissent maintenant. Ces quelques mots ne sauraient exprimer tout l'amour que j'éprouve pour vous.

# Index

Acknowledgements .....	1
Index .....	4
List of introduction figures.....	5
List of introduction tables .....	6
List of abbreviations .....	7
Introduction .....	8
1. Natural products and synthetic biology .....	8
1.1. Microbial natural products in human health.....	8
1.2. Strategies to find new natural products.....	13
1.3. Synthetic biology as a tool to produce natural products and expand their scope.....	18
2. Non-ribosomal peptide synthetases (NRPSs), a class of complex modular enzymes.....	22
2.1. NRPS assembly lines and facilitators.....	22
2.2. NRPS domains: structure and substrate specificity.....	26
2.3. Conformational changes and interactions inside NRPS modules .....	31
2.4. NRPS subunit structure .....	36
3. Combinatorial biosynthesis experiments of NRPSs, knowledge from trial and error on the modifications of NRPSs.....	38
3.1. Modifications of A domains .....	38
3.2. Swapping modules or domains to modify NRPS structure.....	41
3.3. Modification of the length of NRPS.....	50
3.5. Directed evolution to restore functionality of the chimeric NRPS .....	53
3.6. Conclusions about points to keep in mind when modifying the NRPSs.....	54
4. The pyrrolamides, a family of metabolites synthesized by NRPSs.....	56
4.1. The pyrrolamides, a family of minor groove binders .....	56
4.2. Congocidine biosynthesis .....	59
4.3. Biosynthesis of distamycin, congocidine and disgocidine in <i>Streptomyces netropsis</i> DSM40846.....	63
Objectives of the thesis project: .....	66
Chapter I - Revised structure of anthelvencin A and characterization of the anthelvencin biosynthetic gene cluster from <i>Streptomyces venezuelae</i> ATCC 14583.....	67
Chapter I - Supplemental Material.....	82
Chapter II - Modular and Integrative Vectors for Synthetic Biology Applications in <i>Streptomyces</i> spp. ....	99
Chapter II - Supplemental Material.....	127
Chapter III - Refactoring of the congocidine biosynthetic gene cluster: from gene cassettes to gene cluster.....	139
Chapter III - Supplemental Material.....	158
General Conclusion .....	181
References .....	184
French summary of the thesis / Résumé de la thèse en Français.....	197

# List of introduction figures

Figure 1: Examples of the different classes of specialized metabolites .....	9
Figure 2: All small-molecule approved drugs from 1981 to 2014; n = 1202 (adapted from Newman and Cragg, 2016).....	10
Figure 3: Decomposition of biosynthetic gene cluster diversity among all sequenced prokaryotic genomes (Cimermanic et al., 2014).....	11
Figure 4: Structure of specialized metabolites with promising biological activities obtained from recently explored environments.....	14
Figure 5: Exemples of DNA assembly methods.....	19
Figure 6: Biosynthetic gene cluster refactoring principle.....	19
Figure 7: Structures of balhimycin (a) and derivatives (b) (adapted from Winn et al., 2016) .....	21
Figure 8: Exchange of tailoring genes to produce novobiocin/clorobiocin analogs (adapted from Pickens et al., 2011).....	22
Figure 9: NRPS biosynthesis model.....	23
Figure 10: The different NRPS categories .....	24
Figure 11: Model of the position of an MbtH-like protein within an NRPS (Herbst et al., 2013).....	26
Figure 12: Adenylation domain structure (Hur et al., 2012).....	27
Figure 13: Conserved motifs and crystallization of the Phe-adenylation domain PheA (Stachelhaus et al., 1999).....	27
Figure 14: PCP domain structure (Tufar et al., 2014).....	28
Figure 15: X-ray crystal structure of the stand-alone C domain, VibH, from the <i>Vibrio cholerae</i> vibriocin synthetase (Hur et al., 2012) .....	29
Figure 16: Crystal structures of the surfactin thioesterase domain, SrfTE (Hur et al., 2012).....	30
Figure 17: Termination module of SrfA-C (Tanovic et al., 2008) .....	31
Figure 18: Four structures of the linear gramicidin synthetase (LgrA) initiation module representing every major conformation of the module in the catalytic cycle (Reimer et al., 2016).....	33
Figure 19: Dynamics of the revised NRPS cycle.....	33
Figure 20: Linkers of the domains of the termination module of SrfA-C (Tanovic et al., 2008).....	34
Figure 21: Schematic of a proposed regular helical structure for multi-module NRPS enzymes (Lott and Lee, 2017).....	36
Figure 22: Sequence alignment of putative COM domains (Hahn and Stachelhaus, 2004).....	37
Figure 23: Identification of a flavodoxin-like subdomain in GrsA responsible for substrate binding (Kries et al., 2015).....	40
Figure 24: Possibilities of domain substitution in the NRPSs .....	41
Figure 25: Structures of daptomycin, A54145 and CDA (Calcium-Dependent Antibiotic), and corresponding NRPSs .....	46
Figure 26: Identification of the fusion point used for swapping A-PCP-C tridomains (Bozhüyük et al., 2018) .....	47
Figure 27: A-PCP-C (XU) exchange experiments.....	48
Figure 28: Module or domain deletions of plipastatin .....	50
Figure 29: Module insertion in balhimycin NRPS .....	51
Figure 30: Evolution of a PCP domain and modification of its role.....	54
Figure 31: Chemical structures of the members of the pyrrolamide family and name of their <i>Streptomyces</i> producer.....	56
Figure 32: Representation of congocidine binding to DNA (Kopka et al., 1985; Goodsell et al., 1995). ....	58
Figure 33: Structure of some pyrrolamide derivatives.....	59

Figure 34: Modifications of the pyrrole group to target the four DNA base pairs .....	59
Figure 35: <i>S. ambifaciens</i> ATCC 23877 <i>cgc</i> biosynthetic gene cluster and congoicidine structure.....	60
Figure 36: Biosynthetic pathway of the precursor, 4-acetamidopyrrole-2-carboxylate (Lautru et al., 2012)	60
Figure 37: Biosynthetic pathways of the precursor, 3-amidinopropionamide and guanidinoacetate.....	61
Figure 38: Proposed mechanism for the assembly of congoicidine in <i>S. ambifaciens</i> .....	62
Figure 39: Biosynthetic gene clusters responsible for the production of distamycin, congoicidine and disgoicidine in <i>S. netropsis</i> .....	63
Figure 40: Biosynthetic pathways proposed for the assembly of distamycin, disgoicidine and congoicidine	65

## List of introduction tables

Table 1: Examples of bioactive molecules produced by <i>Streptomyces</i> .....	10
Table 2: Examples of approaches activating silent biosynthetic gene clusters.....	15
Table 3: Outcomes of the swapping experiments of PvdD .....	42
Table 4: Examples of daptomycin combinatorial biosynthesis outcome.....	45
Table 5: Members of the pyrrolamide family, producer and biological activity reported.....	57
Table 6: Effects of the deletion of <i>dst</i> genes on the production of congoicidine, distamycin and disgoicidine .....	64

## List of abbreviations

A domain = Adenylation domain  
 AA = amino acid  
 AMP = Adenosine MonoPhosphate  
 ANL family = Acyl-CoA synthetase, NRPS adenylation domain, and Luciferase family  
*ant* genes = anthelvicin biosynthetic genes  
 AntiSMASH = antibiotics and secondary metabolite analysis shell  
 ATP = Adenosine TriPhosphate  
 ArylCP domain = Aryl Carrier Protein domain  
 BAC = Bacterial Artificial Chromosome  
 BGC = Biosynthetic Gene Cluster  
 bp = base pairs  
 C domain = Condensation domain

CATCH = Cas9-Assisted Targeting of Chromosome segments  
 CDA = Calcium Dependent Antibiotic  
 CDPS = CycloDiPeptide Synthase  
*cgc* genes = congocidine biosynthetic genes  
 CGCL = strain of *S. lividans* containing part of (or complete) *cgc* cluster  
 COM domain = Communication Mediating domain  
 Cy domain = heterocyclisation domain  
 DNA = Desoxyribonucleic Acid  
*dst* genes = distamycin/disgocidine/congocidine biosynthetic genes  
 DSTL = strain of *S. lividans* containing part of (or complete) *dst* cluster  
 E domain = Epimerisation domain  
 EPR analysis = Electron Paramagnetic Resonance  
 ESKAPE bacteria = *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*,  
*Pseudomonas aeruginosa* and *Enterobacter* species  
 F domain = Formylation domain  
 FDA = US Food and Drug Administration  
 HDAC = Histone deacetylase  
 HPLC = High Performance Liquid Chromatography  
 HR-MSMS = High Resolution Mass Spectrometry with Fragmentations  
 kb = kilobases  
 LAL family regulator = Large ATP-binding regulators of the LuxR family regulator  
 LCR = Ligase Cycling Reaction  
 LLHR = Linear to Linear Homologous Recombination  
 MLP = MbtH-like proteins  
 MS = Mass Spectrometry  
 NMR = Nuclear Magnetic Resonance  
 NP = Natural Product  
 NRP = Non Ribosomal Peptide  
 NRPS = Non Ribosomal Peptide Synthetase  
 OSMAC approach = “One Strain-MANy Compounds” approach  
 PCP domain = Peptidyl Carrier Protein domain (or Thiolation (T) domain)  
 PCR = polymerase chain reaction  
 PKS = polyketide synthase  
 ppant arm = phosphopantetheinyl arm  
 SAM = (S)-Adenosyl-Methionine  
 SARP = *Streptomyces* antibiotic regulatory protein  
 SP = synthetic promoter  
 R domain = reductase domain  
 RBS = Ribosome Binding Site  
 RiPP = Ribosomally synthesized and Post translationally modified Peptide  
 RXP = rhabdopeptides and xenortide peptide  
 SLIC = Sequence- and Ligation-Independent Cloning  
 T domain = Thiolation domain (or Peptidyl Carrier Protein (PCP) domain)  
 TAR cloning = Transformation-Associated Recombination cloning  
 TE domain = Thioesterase domain  
 tRNA = transfer ribonucleic acid  
 WHO = World Health Organisation  
 WT = wild-type  
 XU = Exchange Unit  
 XUC = Exchange Unit Condensation domain

# Introduction

## 1. Natural products and synthetic biology

### 1.1. Microbial natural products in human health

#### 1.1.1. Historical role of natural products

The simplest definition of “natural product” (NP) as stated in the editorial of *Nature Chemical Biology* in July 2007 (2007) is “a small molecule that is produced by a biological source”. Natural products consist in chemicals not involved in basal metabolism, and not necessary for growth in a nutrient-rich environment. They may have pharmacological properties or commercial use. The main different groups of natural products are presented briefly in [Box 1](#). In this manuscript, the term “anti-infective” will include antibacterial, antiparasitic, antifungal and antiviral agents, while the term “antibiotic” itself will be used in a stricter sense, only to describe antibacterial compounds.

Natural products have been used in traditional medicine even before the bioactive molecules were identified. A record from 2600 BC listed approximately 1000 plant-derived substances used in Mesopotamia (Cragg and Newman, 2013). Chinese, Egyptian, Greek and Roman civilizations all have documents referring to medicinal plants (Demain, 2009). Even today, a substantial part of the world population relies on plant-derived medicine. One of the most famous recent examples is the antimalarial drug artemisinin (Figure 1). Artemisinin was extracted from *Artemisia annua* used in traditional Chinese medicine, and artemisinin analogs are now used to treat malaria patients.

#### Box 1: Classes of natural products

Natural products, also called specialized metabolites, are usually classified by their structure or the enzymes directing the biosynthesis (Figure 1). Polyketides are assembly of decarboxylated (alkyl)-malonyl thioesters (Rutledge and Challis, 2015). They are synthesized by polyketide synthases (PKSs), and are usually highly modified and decorated during the biosynthesis or afterwards. For instance, macrolides such as erythromycin are assembled by PKS. Terpenes such as the antimalarial compound artemisinin are constituted of isoprene units assembled by terpene synthases (Gao et al., 2012). Alkaloids, such as caffeine, are specialized metabolites containing nitrogen, very often on a heterocyclic ring, derived from amino acids (Rutledge and Challis, 2015). Peptides, derived from different biosynthetic pathways, can be specialized metabolites. Some of them are ribosomally synthesized and post-translationally modified peptides (RiPPs), such as the thiopeptide thiostrepton (Arnison et al., 2013). Non-ribosomal peptides are made of amino acids, possibly non proteogenic, linked by amide bonds by non-ribosomal peptide synthetases (NRPSs). An example is the molecule of penicillin. Finally, some of the cyclodipeptides are derived from two amino acids joined by cyclodipeptide synthases (CDPS), as is albonoursin (Lautru et al., 2002).

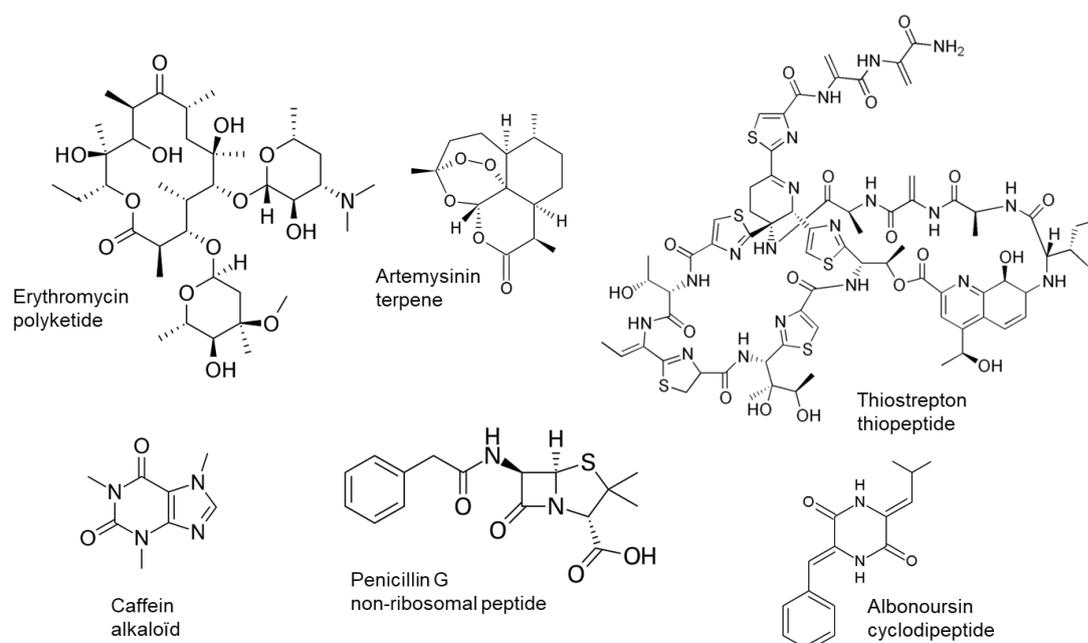


Figure 1: Examples of the different classes of specialized metabolites

### End of Box 1.

While microbial natural products, also named microbial specialized metabolites, were hardly accessible before the 20<sup>th</sup> century, they now constitute an important source of pharmaceuticals. The discovery of the antibiotic penicillin (Figure 1) produced by the fungi *Penicillium* is the first example which led to industrial production: by the 1940s, penicillin was in regular clinical use (Lyddiard et al., 2016). Actinomycin discovery, produced by an *Actinomyces* species, was soon followed by the discovery of streptomycin in 1943. It marked the beginning of a “Golden era” for anti-infective discovery. For more than 20 years, dozens of classes of compounds were discovered. One half of today’s antibiotics were discovered during that period (Davies, 2006).

#### 1.1.2. Current place of the natural products in the recently approved drugs

Since the 1970s, the number of natural products reaching the clinical market has slowed down. Newman and Cragg have analyzed the origin of the drugs approved by the US Food and Drug Administration (FDA) from 1981 to 2014, and they showed that still 2/5 of the small molecules approved are natural products or natural product-derived molecules coming from plants and microorganisms (Figure 2) (Newman and Cragg, 2016). To this number can be added the natural product-inspired molecules, which amount to another 25% of all small molecules. Altogether, NP and their derivatives correspond to 45% of the anti-infectives, including 58% of the approved antibacterial drugs. They also correspond to 65% of the anticancer agents approved in the past 30 years (Newman and Cragg, 2016). Natural products and their derivatives are thus still an important source of anti-infective and anticancer agents.

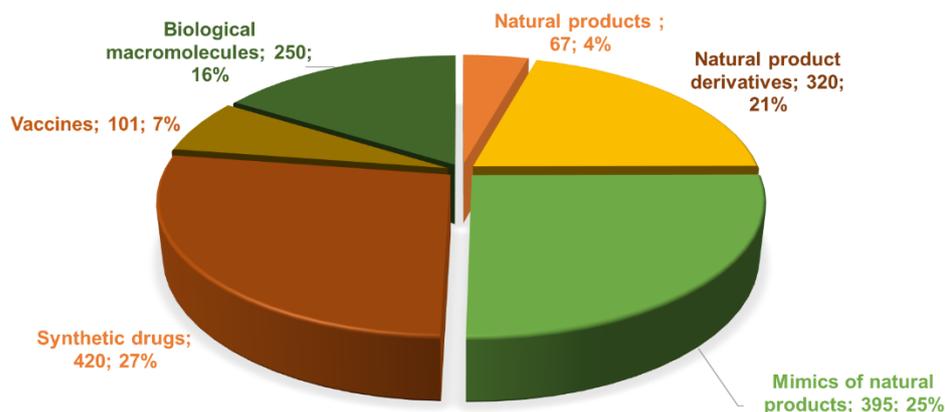


Figure 2: All small-molecule approved drugs from 1981 to 2014; n = 1202 (adapted from Newman and Cragg, 2016)

### 1.1.3. Microbial natural product producers

A minority of microorganisms are responsible for the production of more than 80% of known microbial specialized metabolites. In fact, historically, almost all antibacterial compounds were isolated from actinobacteria and, among this phylum, from bacteria of the *Streptomyces* genus. Altogether, over 9,000 bioactive compounds were isolated from actinobacteria, and 60 are used in medicine, agriculture or research. 80% of these 60 compounds are from *Streptomyces* species (Demain, 2009). Nowadays, actinobacterial specialized metabolites represent about 25%, of anti-infective specialized metabolites. Examples of bioactive compounds produced by *Streptomyces* species are listed in Table 1.

Table 1: Examples of bioactive molecules produced by *Streptomyces*

Type of compound	Producing species	Bioactive agent(s)	Source or reference
Antibacterial agent producers	<i>Streptomyces venezuelae</i>	Chloramphenicol	(Ehrlich et al., 1947)
	<i>Streptomyces roseosporus</i>	Daptomycin	(Mchenney et al., 1998)
	<i>Streptomyces fradiae</i>	Neomycins	(Dulmage, 1953)
	<i>Streptomyces griseus</i>	Streptomycin	(Schatz and Waksman, 1944)
	<i>Streptomyces aureofaciens</i>	Tetracycline	(Darken et al., 1960)
	<i>Streptomyces clavuligerus</i>	Cephalosporin	(Brannon et al., 1972)
Antifungal agent producers	<i>Streptomyces noursei</i>	Nystatin	(Zotchev et al., 2000)
	<i>Streptomyces kasugaensis</i>	Kasugamycin	(Umezawa et al., 1965)
Bioherbicide/ biopesticide producers	<i>Streptomyces hygrosopicus</i>	Herbimycin	(Omura et al., 1979)
Antiparasitic agent producers	<i>Streptomyces avermitilis</i>	Avermectins	(Burg et al., 1979)

Antiviral agent producers	<i>Streptomyces hygroscopicus</i>	Hygromycin	(González et al., 1978)
Immunosuppressant agent producers	<i>Streptomyces hygroscopicus</i>	Rapamycin	(Chen et al., 1999)
Antitumor agent producers	<i>Streptomyces peucetius</i>	Doxorubicin (adriamycin)	(Arcamone et al., 2000)
	<i>Streptomyces verticillus</i>	Bleomycin	(Shen et al., 2001)
	<i>Streptomyces caespitosus</i>	Mitomycine C	(Wakaki et al., 1958)

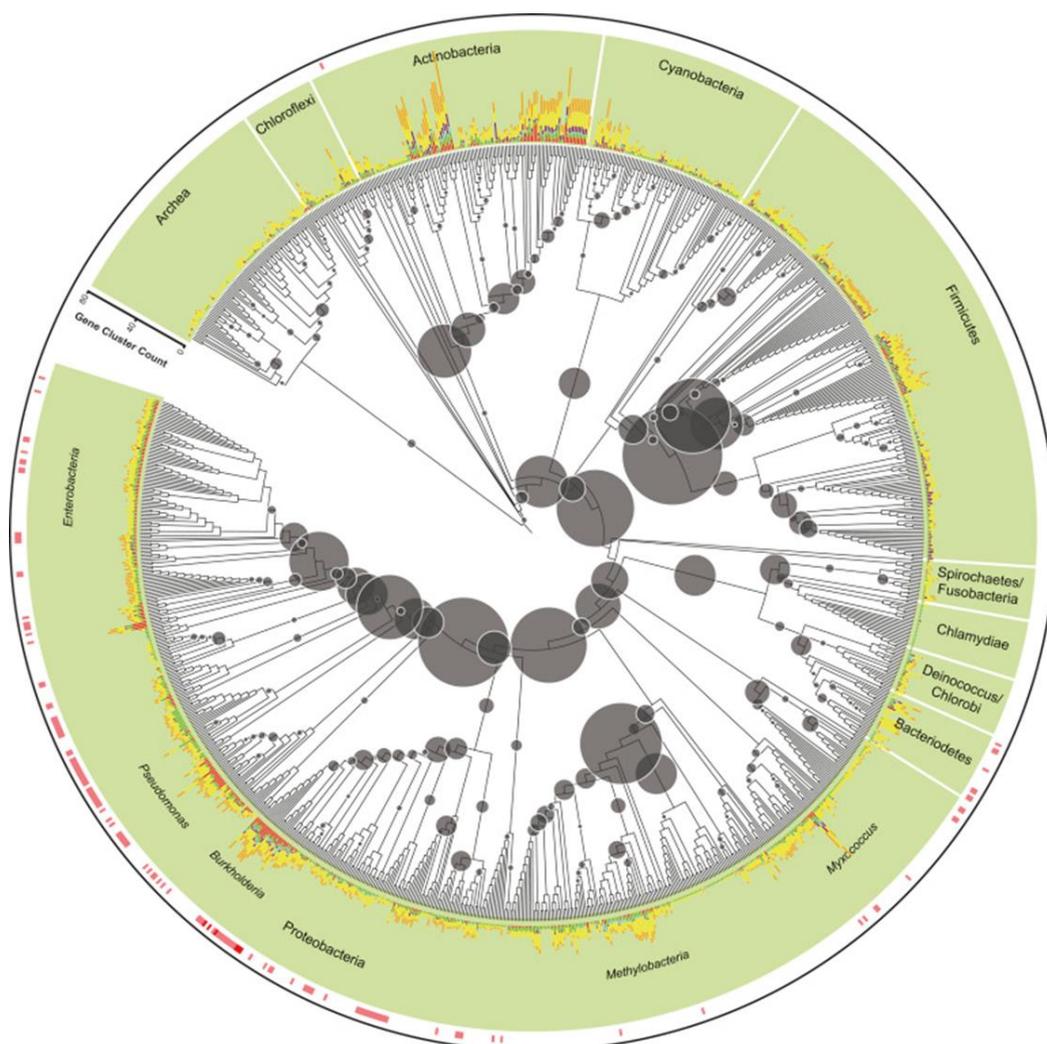


Figure 3: Decomposition of biosynthetic gene cluster diversity among all sequenced prokaryotic genomes (Cimermanic et al., 2014)

The diversity of each node in the phylogenetic tree is represented by the size of the circle (larger circle defines higher degree of diversity).

The biosynthesis of microbial specialized metabolites is most of the time directed by genes physically grouped together in the genome, called Biosynthetic Gene Clusters (BGCs). Cimermanic and co-workers (2014) have analyzed the distribution of BGCs of 1,154 sequenced genomes among the bacterial phylogenetic tree. Figure 3 shows that apart from actinobacteria, confirmed to be remarkably prolific specialized metabolite producers, other important producers

can be found in the cyanobacteria, proteobacteria (myxobacteria, *Pseudomonas*, *Burkholderia*), and firmicutes (*Bacillus*) phyla. Among fungi, specialized metabolite producers are in particular found in the ascomycota (*Penicillium*, *Aspergillus*) phylum.

#### 1.1.4. Current situation: a crucial need for new pharmaceutical compounds

In the 1950s, geneticists believed that the development of microbial pathogenic strains resistant to antibiotic treatments was highly unlikely (Davies, 2006). And yet, for almost all antibiotic treatments, pathogen bacteria resistant to the antibiotic can be detected only a few years after the introduction of the antibiotic on the clinical market (Davies and Davies, 2010). Resistance to antibiotics arose fast partly because they were used in large quantities irresponsibly, for instance for agricultural applications, and partly because we underestimated microorganisms' capacity to adapt (Procópio et al., 2012). Antimicrobial resistance is now considered by many organizations (World Health Organisation, European Centre for Disease Prevention and Control ...) as a major public health threat (Ferri et al., 2017). In 2014, the Review on Antimicrobial Resistance UK Commission estimated that antimicrobial resistance caused 700,000 deaths worldwide and that this figure was likely to reach 10 million by 2050 (Review on Antimicrobial Resistance). This worrying situation led the World Health Organisation (WHO) to establish a list of bacteria for which new antibiotics are urgently needed in February 2017 (WHO publishes list of bacteria for which new antibiotics are urgently needed, 2017). Bacteria of this list are classified according to three levels of priority, critical, high and medium. In the critical and high levels can be found all the so-called "ESKAPE" bacteria (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* species) (Fair and Tor, 2014; Lewis, 2013). A study commissioned by the Wellcome Trust in 2016 aimed at evaluating alternatives to antimicrobial compounds (Czaplewski et al., 2016). The most advanced approaches were shown to be antibodies, probiotics and vaccines now in Phase II or Phase III trials. However, in the medium term, the commission confirmed that conventional antibiotics would still be needed, as these approaches would mainly serve as adjunctive or preventive therapies.

Meanwhile, the discovery of new microbial natural products with promising antibiotic activity has slowed down. There are three main reasons for this current decline in antibiotic compounds discovery: new compounds are harder to find, industrials have turned away from antibiotic research, and regulation became stricter (Bérdy, 2012). The discovery of new compounds, which seemed to be never ending in the 1960s, slowed down drastically while rediscovery of already known molecules became more and more frequent (Lewis, 2013). Research expenses increased for companies, while the number of leads decreased, and newly discovered antibacterial agents were restricted to last-resort use in hospitals only, which resulted in low profits. This led the big pharmaceutical companies to first turn to synthetic combinatorial chemistry in the 1990s. However, these approaches had very limited success, probably because the "chemical space" of NP and synthetic drugs are different (Harvey et al., 2015). In addition, several drugs approved by the FDA in the past, such as streptomycin, tetracycline, and most aminoglycosides, would not pass the regulation tests today (Bérdy, 2012). For all these reasons, big pharmaceutical companies have now abandoned antibiotic research to join the more profitable chronic disease drug market. Most of the antibiotic drug lead research nowadays is done by start-up companies or academic laboratories.

Apart from microbial infections, efficient treatments are still needed for numerous diseases. Cancer was responsible for 9.6 million deaths in 2018 and is representing the second leading cause of deaths worldwide (Cancer, 2108). Even for well-treated cancers, new compounds, as potent but less toxic for the patients, are highly desirable. Parasitic and helminthic infections also remain a worldwide problem, especially in developing regions. Malaria, dengue and leishmaniasis are of particular concern. Soil-transmitted infections affect about 1.5 billion people in the world, and infected children suffer from nutrition and physical impairment (Soil-transmitted helminth infections, 2019). Fungal diseases pose a real threat for people with weakened immune system such as patients with HIV (Human Immunodeficiency Virus) or cancer (Global fungal diseases, 2018).

In conclusion, bioactive compounds, whether it is for antibacterial, antifungal, antiparasitic or anticancer therapies, are dearly needed. The next section of this introduction covers the strategies presently employed to discover new natural products with pharmaceutical potential.

## 1.2. Strategies to find new natural products

### 1.2.1. Studying new specialized metabolite-producing strains from underexplored environments

Traditionally, scientists isolated microorganisms from the soil, because it was of easy access and relatively easy to reproduce growth conditions. Today, more and more environments are explored, environments that are bound to procure new species of microorganisms, hence maybe new kinds of natural products (Hug et al., 2018). In particular, aquatic environments have attracted increased attention since the 1970s. Oceans contain approximately 87% of life on earth (Bérdy, 2012), they constitute the largest pool of microorganisms. Marine actinomycetes were proven to be remarkable for their specialized metabolite production (Subramani and Aalbersberg, 2012). For instance, the cancer cell cytotoxic salinosporamide A (Figure 4), a proteasome inhibitor, was isolated from *Salinispora tropica* (Feling et al., 2003).

Extreme environments, such as deserts or polar areas, inhabited by extremophiles including acidophiles, alkalophiles, halophiles, and hyperthermophiles, are also explored. They have led to interesting discoveries (Masand et al., 2018; Tian et al., 2017). Thus, more than 20 new specialized metabolites were identified from *Penicillium* species isolated from an abandoned copper mine water basin, Berkeley Pit Lake, contaminated with high concentrations of dissolved metal sulfites (Pettit, 2011). Among them are two new polyketide terpenoids berkeleydione and berkeleytrione (Figure 4), with promising activities against cancer and Huntington disease.

Endophytes and symbionts are also a source of specialized metabolites. Bérdy (2012) reported that 80 % of endophytic fungi produce a bioactive compound of some kind, one of the best-known examples being the production of the anti-cancer drug taxol (Paclitaxel) from *Taxomyces* (Figure 4) (Cragg and Newman, 2013).

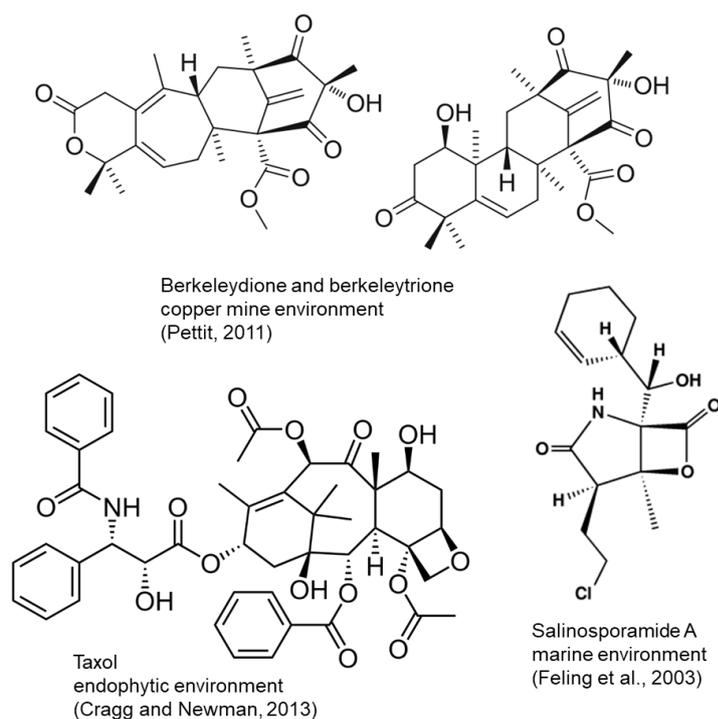


Figure 4: Structure of specialized metabolites with promising biological activities obtained from recently explored environments

Although there has been an increasing interest in “exotic” environments in the scientific community, a recent study has shown there may be no need to wander so far: the parks of New York contain plethora of yet unknown microorganisms and compounds (Nothias et al., 2016). Altogether, there are still plenty of microorganisms to study, and we will without doubt discover many new natural products by tapping into these resources (Cragg and Newman, 2013; Demain, 2009).

*Streptomyces* are probably among the best studied bacteria for their specialized metabolism. They are prolific natural product producers and numerous studies have been carried out to explore their specialized metabolism repertoire. For this reason, the next two sections will be centered on this genus, although the methods that have been used to isolate and characterize *Streptomyces* metabolites could probably be applied to other genera.

#### 1.2.2. Expressing *Streptomyces*' specialized metabolism full potential in the native host

*Streptomyces* genomes usually contain several dozen biosynthetic gene clusters (BGCs) that can be predicted by bioinformatics tools such as antiSMASH (antibiotics and secondary metabolite analysis shell) (Blin et al., 2017). For instance, *Streptomyces avermitilis* genome contains 25 potential BGCs, which correspond to 6% of its genome (Ömura et al., 2001). *Streptomyces ambofaciens* genome contains 23 clusters potentially involved in specialized metabolism, and yet, it was known for more than 40 years to produce only spiramycin and congocidine (Aigle et al., 2014). Most of *Streptomyces* specialized metabolites are not expressed, or not detected, in standard laboratory conditions. The corresponding BGCs are called “cryptic”, or “silent”. Various methods have been employed to

activate the expression of these clusters (Rutledge and Challis, 2015), some examples are listed in Table 2.

Table 2: Examples of approaches activating silent biosynthetic gene clusters

Approach	Principle	Compound discovered	Reference
Variation in growth conditions	Cultivation of <i>Streptomyces armeniacus</i> on a malt-containing medium	armeniaspirols	(Hug et al., 2018)
Co-culturing	Cocultivation of <i>S. endus</i> S-522 with <i>Tsukamurella pulmonis</i> TP-B0596	alchivemycin A	(Rutledge and Challis, 2015)
Addition of chemical elicitors	Addition of subinhibitory concentrations of trimethoprim to <i>Burkholderia thailandensis</i> culture	malleilactone	(Hug et al., 2018)
General regulation	Induction by an allele of <i>absA1</i> (from <i>S. coelicolor</i> ) in <i>Streptomyces flavopersicus</i>	pulvomycin	(Rutledge and Challis, 2015)
Knock out of one biosynthetic gene cluster	Knocking out the <i>rifA</i> PKS gene responsible for rifampicin biosynthesis from <i>Amycolatopsis mediterranei</i> S699	amexanthomycins A–J	(Hug et al., 2018)
Pathway specific transcriptional regulation	Inactivation of the repressor <i>gbrR</i> in <i>S. venezuelae</i>	gaburedin A	(Hug et al., 2018)
Heterologous expression	Expression in <i>E. coli</i> of the terpene synthase encoded by the <i>saw76</i> gene of <i>S. avermitilis</i>	avermilol	(Rutledge and Challis, 2015)

The empirical approach called “the OSMAC approach” (One Strain-MAny Compounds), is based on the fact that a strain will not express all its spectrum of specialized metabolites in a given condition (Bode et al., 2002). By modifying the culture conditions (nutrient sources, medium components in general, pH, aeration, temperature), different compounds may be produced. The addition of metal ions may also have an effect (Hug et al., 2018; Liu et al., 2013). When *in silico* data is available to predict the structure or the role of the compound of interest, these modifications may be made rationally. For instance, an iron-depleted medium was used to induce the production of a likely siderophore predicted in *Streptomyces coelicolor* genome, and this resulted in isolation of coelichelin (Lautru et al., 2005). Another method not requiring any genetic knowledge is the co-cultivation with other species, as interspecies cross talks may induce metabolite production (Liu et al., 2013; Zarins-Tutt et al., 2016). Histone deacetylase (HDAC) inhibitors modulate gene expression by deacetylating histone proteins and they have been especially useful in fungi natural product research. They have also successfully been used for bacteria (Hug et al., 2018; Zarins-Tutt et al., 2016). Finally, chemical elicitors such as sub-inhibitory concentrations of antibiotics may also induce antibiotic production (Rutledge and Challis, 2015; Zarins-Tutt et al., 2016).

The methods described above are empirical and do not rely on any knowledge of the mechanisms governing the production of specialized metabolites by the strains. As an alternative to this approach, genetic methods have been developed based on knowledge of specialized metabolism regulation. Specialized metabolites production is under tight regulation in *Streptomyces* species. Global regulation involves master regulators. It is extremely complex and coordinated with morphological developments (Bibb, 2005; Bibb and Hesketh, 2009). A metabolic switch is observed in fermentors from exponential growth to stationary growth, when most specialized metabolites are produced. During the switch, there are signaling cascades, regulation by small ligands and phosphorylation state (Liu et al., 2013). There are pleiotropic regulators involved in both antibiotic production and aerial hyphae development. It is for example the case of the gene *bldA*, which codes for the unique tRNA for the rare leucine codon UUA (van Wezel et al., 2009). Regulatory genes, specialized metabolite genes and morphology changing-genes containing the rare codon can only be translated when *bldA* is expressed. There are also pleiotropic regulators of several antibiotic pathways, such as the *absA* operon, a two-component system used to repress antibiotic production in *S. coelicolor* and *Streptomyces griseus* (van Wezel et al., 2009).

In addition to this global level of regulation, the expression of genes directing the biosynthesis of a given metabolite is often controlled locally by transcription regulators located in biosynthetic gene clusters (Hug et al., 2018). The over-expression of pathway-specific activators or deletion of repressors can trigger the production of the expected metabolite. For instance, deleting the *tetR* repressor encoded in the gene cluster led to the production of kinamycin in *S. ambofaciens* (Bunet et al., 2011), while stambomycins were only observed after constitutively expressing a Large ATP-binding regulator of the LuxR (LAL) family regulator (Lauret et al., 2011). Pathway-specific *Streptomyces* antibiotic regulatory protein (SARP) control the production of many specialized metabolites. The overexpression of the SARP *caaR* allowed for instance to detect clavulanic acid in *Streptomyces clavulagerus* (Zarins-Tutt et al., 2016).

Finally, it should be mentioned that knocking down pathways of known metabolites can also be helpful: some compounds may be present in smaller amount, and they will be detected more easily in the absence of the major compounds (Rutledge and Challis, 2015). Knocking down gene clusters may also alleviate competition for common precursors.

The genetic approaches described above rely on the ability to genetically manipulate the strain of interest. When this is not the case, or when no genetic tools have been developed for the strain, another possibility is the heterologous expression of the gene cluster, that is the insertion of the biosynthetic gene cluster in a host strain (Zarins-Tutt et al., 2016).

### 1.2.3. Producing specialized metabolites by heterologous expression

There are many examples in the literature of *Escherichia coli* and *Saccharomyces cerevisiae* used as heterologous hosts because their genetic toolbox is well developed, but they may not be ideal for all actinomycetes natural products (Pickens et al., 2011). Firstly, the high GC-content of actinomycetes genomic DNA often impedes correct translation. Adjusting codon usage requires the synthesis of DNA, which is often problematic in the case of large NRPS or PKS genes. Secondly, there is often a need for chaperone or helper proteins, such as phosphopantetheinyl

transferase or MbtH-like proteins, which are encoded in actinomycetes genome, but often not included in the biosynthetic gene cluster of interest (Ongley et al., 2013). Thirdly, precursors from primary metabolism, such as branched-chain acyls, may not be produced in *E. coli* or *Sa. cerevisiae*. Historically, *Streptomyces albus*, and *S. coelicolor* have been extensively used as heterologous hosts (Baltz, 2010), and they remain among the laboratory favorite pets. Industrial producers have also been used as hosts, such as *S. avermitilis* or *Streptomyces roseosporus* (Baltz, 2016). In recent years, various *Streptomyces* strains have been engineered to constitute good chassis for the production of specialized metabolites. Thus, endogenous gene clusters have been deleted (*S. coelicolor*, (Gomez-Escribano and Bibb, 2011), *S. avermitilis* (Komatsu et al., 2010), *S. albus* (Kallifidas et al., 2018)). These strains present a low background noise as they do not produce specialized metabolites anymore. These strains have often been further optimized for the expression of biosynthetic gene clusters, for example by introducing mutations known to be favorable for this expression (in *rpoB* or *rpsL* in *S. coelicolor*), or by increasing the resistance to oxidative stress (deletion of *pjfk* in *S. albus*).

Most BGCs span from 10 to 120 kilobases (kb). To introduce them in a tractable host imply to be able to manipulate and retrieve DNA fragments of these sizes from the native producer (Ongley et al., 2013; Rutledge and Challis, 2015). The cluster can then be maintained on a stable plasmid or integrated within the host genome. The traditional method to capture a biosynthetic gene cluster is to construct genomic libraries, but the complete process is quite tedious and for very large clusters, it is often difficult to capture the whole cluster on one vector (cosmid, BAC...). It is then necessary to reassemble the cluster from two or three vectors (Perlova et al., 2006). New techniques have been developed recently: Linear to Linear Homologous Recombination (LLHR) allows to bring together two linear pieces of DNA with sequence identity at the extremities in *E. coli* (Fu et al., 2012). Another technique of interest is the transformation-associated recombination (TAR cloning), which is based on yeast natural capacities of recombination. Yamanaka et al. (2014) reported first the use of this method in 2014. They cloned a 67-kb gene cluster directing the biosynthesis of the lipopeptide taromycin A in one step, which would have been difficult using a genomic library. Another very recent technique is CATCH (Cas9-Assisted Targeting of Chromosome segments), which combines the use of RNA-guided Cas9 nuclease to cut the cluster from its genome, and the use of Gibson assembly to ligate the cluster to a linear plasmid (Jiang et al., 2015). Using this technique, the authors were able to clone up to 100-kb DNA.

The heterologous expression of a biosynthetic gene cluster is sometimes sufficient to afford the production of a specialized metabolite. This was for example the case of collinone, a polyketide antibiotic that was not detected in the native producer, *Streptomyces collinus*, but was produced when the biosynthetic gene cluster was transferred in *S. coelicolor* CH999 (Martin et al., 2001). Yet, the heterologous expression of a gene cluster is often insufficient on its own and further manipulations of the gene cluster, such as the deletion of transcriptional repression (Yamanaka et al., 2014) or the replacement of native promoters by strong and constitutive ones (pathway refactoring, developed in the next section) are often required.

### 1.3. Synthetic biology as a tool to produce natural products and expand their scope

#### 1.3.1. Synthetic biology, a new toolbox for natural product engineering

Synthetic biology has been described as “an engineering approach to improve or completely create systems and organisms with specific or desirable functions” (Guzmán-Trampe et al., 2017). One of the principles of synthetic biology is to rely on fundamental biology, chemistry, and bioinformatics to improve or construct new biological parts, devices, and systems. Engineering can have a role at different scales: protein engineering to modify protein properties, metabolic engineering to implement a biosynthetic pathway, strain engineering to identify and optimize high titer producers (Pickens et al., 2011; Smanski et al., 2016). Synthetic biology permits for instance to control space (from protein scaffold to compartmentalization and bacterium consortia) and time (from allosteric control to regulatory cascades and molecular clock) at different scales in a designed system (Medema et al., 2011). It now plays a prominent role in antibiotic discovery and biosynthetic pathway engineering.

Biological DNA basic parts are small DNA fragments whose sequence confers a specific function. For example, these DNA basic parts include promoters, ribosome binding sites (RBS), coding sequences, and regulators among others. In order to modify a cluster and replace some of its parts, one must have at his disposal libraries containing parts available for replacement. Many libraries of characterized parts are available for *Sa. cerevisiae* and *E. coli* (Pickens et al., 2011), and recently, some libraries have been reported for *Streptomyces* species as well (Smanski et al., 2016). Shao and collaborators (2013) tested several heterologous promoters in *Streptomyces lividans* when they engineered the spectinabilin pathway. These promoters, however, were not well characterized, limiting their usefulness in other studies. Other libraries were derived from well-characterized promoters, such as ermEp1 (Siegl et al., 2013) or kasOp (Bai et al., 2015). In this latter case, the library constructed is based on the already optimized promoter kasOp\*. The synthetic promoters derived from kasOp\* have a strength varying between 1 to 190% of kasOp\*. The authors also characterized 15 native and 174 synthetic RBSs that cover a 200-fold strength range. In contrast, there are not many characterized terminators actually available, though some recent studies aim at filling this gap (Horbal et al., 2018). Some studies have, however, underlined that these DNA parts are characterized in a specific context, including surrounding DNA sequences and the host strain itself, and that their characterization was not systematically transferable outside of this context (Vilanova et al., 2015; Yeung et al., 2017).

In addition to libraries of standard DNA parts, synthetic biology requires performant DNA assembly methods. New DNA assembly technologies have been developed in the past years, and they constitute an extremely useful toolbox for biosynthetic gene cluster capture, (re)assembly and modification (Figure 5)(Kim et al., 2015; Luo et al., 2016; Sands and Brent, 2016). Traditionally, DNA assembly was made by digestion by restriction enzymes and ligation (Sands and Brent, 2016). Since then, more sophisticated methods still based on the use of restriction enzymes have been developed, such as the Biobrick assembly (Knight, 2003), or the Golden Gate assembly (Engler and Marillonnet, 2014). Ligase cycling reaction (LCR) is a technique based on the use of a thermostable ligase and multiple cycle of denaturation-annealing-ligation temperatures (de Kok et al., 2014). Other assembly techniques are based on homologous recombination *in vivo*, such as DNA assembler (Shao et al., 2009) and Red/ET recombineering (Gust et al., 2004) or *in vitro* such

as Gibson assembly (Gibson et al., 2009), sequence- and ligation-independent cloning (SLIC) (Li and Elledge, 2012), or Gateway system (Sands and Brent, 2016). Many other techniques not described here are available, and allow the assembly of several DNA parts, forming a modified biosynthetic gene cluster (Sands and Brent, 2016).

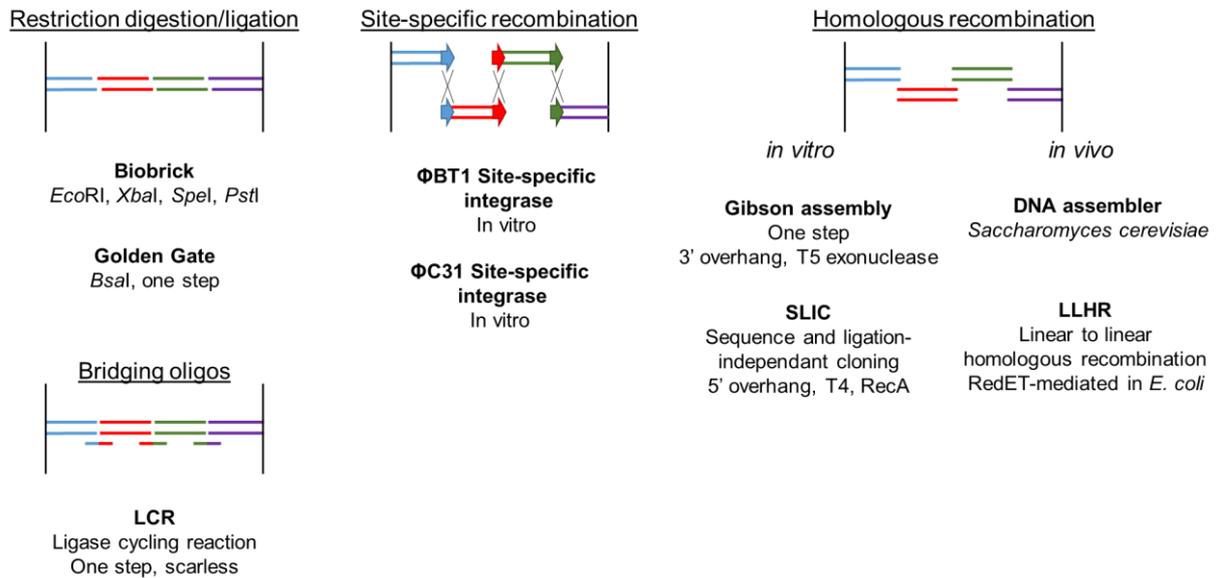


Figure 5: Examples of DNA assembly methods

### 1.3.2. Refactoring of specialized metabolite biosynthetic gene clusters

Refactoring consists in rewriting the DNA sequence without changing its functionality. It may be done to erase all native regulation, to optimize the sequence for heterologous expression or as a first step towards the generation of synthetic pathways within a cell (Figure 6). A pioneering refactoring work is the refactoring of the nitrogen fixating gene cluster (20 genes) from *Klebsiella oxytoca* (Temme et al., 2012). In this study, the authors aimed at (i) removing all native regulation and non-essential genes, (ii) re-organizing the genes into synthetic operons using well-characterized synthetic biological parts (promoters, ribosome binding sites (RBS), terminators) and (iii) randomizing/optimizing codon usage for *E. coli* expression. Their refactored gene cluster, constituted of 89 genetic parts, was functional, although with a reduced activity.

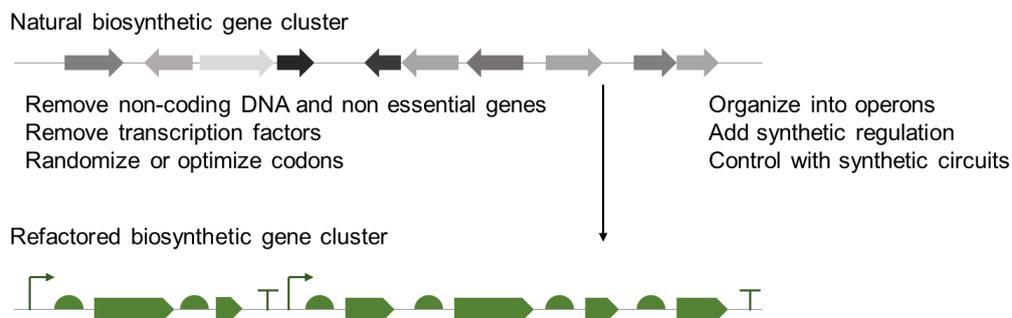


Figure 6: Biosynthetic gene cluster refactoring principle

This figure represents the different steps to follow to refactor a biosynthetic gene cluster

In addition to modifying transcriptional/translational elements to better control the expression of a set of genes, the refactoring of a gene cluster can also be used to introduce or remove genetic elements that will facilitate the re-assembly of the cluster. Thus, when Osswald and co-workers (2014) refactored the epothilone BGC (56 kb, 7 genes) of *Sorangium cellulosum* for expression in *Myxococcus xanthus*, they added unique restriction sites, while subtracting about 700 unwanted restriction sites.

The refactoring of the nitrogen fixing gene cluster and of the epothilone gene cluster involved extensive modifications of the original DNA sequence. This could only be obtained through the synthesis of DNA fragments that were next assembled. Indeed, DNA synthesis is becoming an increasingly attractive option, though still expensive (Kim et al., 2015). However, such an extensive refactoring may not always be required, and there are many examples of simpler refactoring, consisting mainly in replacing native promoters by constitutive or synthetic ones, especially in the case of rather small clusters (Rutledge and Challis, 2015). Such examples include the refactoring of spectinabilin (Shao et al., 2013). The spectinabilin cluster from *Streptomyces orinoci* remained silent when expressed in *S. lividans*, even when a gene encoding a transcriptional repressor was deleted. The authors chose nine strong promoters and one inducible promoter to refactor the cluster, and after assembly using DNA assembler method, they observed production of spectinabilin, though with a yield of 10% compared to the production in the WT strain. Using the same assembly method, three novel polycyclic tetramate macrolactams were identified when the BGC refactored with strong promoters was expressed in *S. lividans* (Luo et al., 2013). Very recently, combining TAR cloning and red/ET recombineering, Moore and colleagues refactored the *spz* cluster and detected the production of more than a hundred of compounds related to streptophenazine (Bauman et al., 2019).

Once a pathway is refactored, it is usually much easier to replace one part by another one, to refine the knowledge of the biosynthetic pathway (Luo et al., 2013) or to obtain a higher yield when the functions are equivalent (Smanski et al., 2014). It is also possible to obtain a new compound by adding a part with a different function (Smanski et al., 2016). Refactoring thus leads the way to the modification of specialized metabolites to produce new analogs.

### 1.3.3. Production of non-natural analogs and expansion of the range of specialized metabolites

Once a metabolite of interest has been isolated, it may be interesting to try to improve its properties by generating analogs. Derivatives of natural products can be produced by a number of chemical or biological methods, or by a combination of these methods. Traditionally, microbial natural products were obtained by fermentation and then chemically modified (hemi-synthesis). In the last decades, new methods, based on the metabolic capacities of microorganisms, have been developed. Thus, chemically synthesized precursors analogs can be fed to the producing strain. This method relies on enzymatic substrate promiscuity, but may sometimes be successful, as it was the case for a derivative of balhimycin, bromobalhimycin (Sun et al., 2015). However, the natural metabolite is still produced, as there is a competition between the native substrate and the added one. To avoid such competition, it is possible to resort to genetic engineering to knock out the production of the natural precursor in the strain, prior to the feeding of the precursor analog

(mutasynthesis). For instance, new derivatives of balhimycin were obtained when the gene responsible for the synthesis of  $\beta$ -hydroxytyrosine was deleted and the strain fed with fluorinated  $\beta$ -hydroxytyrosine analogs (Figure 7) (Winn et al., 2016).

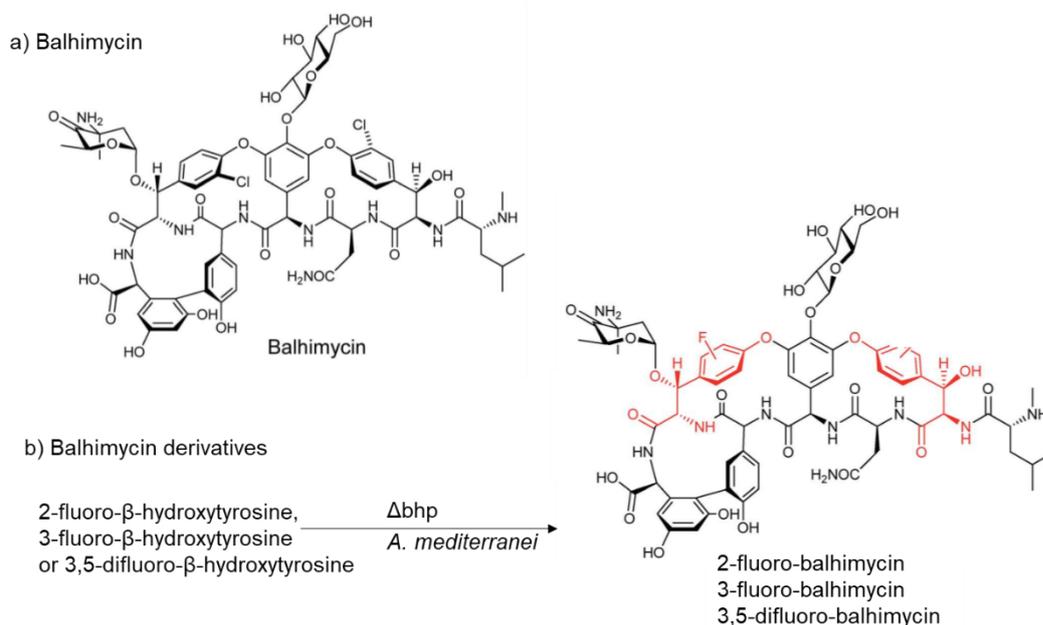


Figure 7: Structures of balhimycin (a) and derivatives (b) (adapted from Winn et al., 2016)

Another synthetic approach, called combinatorial biosynthesis, consists in combining (subtracting, adding or replacing) biosynthetic genes from various gene clusters. The engineered organism then produces analogs of the original natural product (Goss et al., 2012). For instance, some enzymatic domain exchanges allowed the biosynthesis of ivermectin (22,23-dihydroavermectins), a derivative of the natural product avermectin (Pickens et al., 2011).

Combinatorial biosynthesis can be coupled to mutasynthesis and chemoenzymatic synthesis to increase further the chemical diversity generated. Thus using this combination of methods, Heide (2009) reports the generation of more than a hundred derivatives of the aminocoumarins novobiocin, clorobiocin and coumermycin A1 (Figure 8). Structurally, novobiocin and clorobiocin are similar, except for the group at the C-8 position of the aminocoumarin moiety (methyl or chlorine group) and the 3-OH group of the desoxysugar (a carbamoyl or a methylpyrrol-2-carboxyl moiety). All the nine possible hybrids of novobiocin and clorobiocin were tested and it was shown that the better antibiotic activity of clorobiocin was mainly due to the methylpyrrol-2-carboxyl moiety attached to the desoxysugar.

Although the refactoring and the genetic engineering of biosynthetic gene clusters have encountered some success, it has often been at the expense of the yield of the obtained metabolite(s) (Osswald et al., 2014; Shao et al., 2013). This highlights the necessity of a greater understanding of the fundamental biological processes governing the biosynthesis of natural products (Goss et al., 2012; Kim et al., 2015).

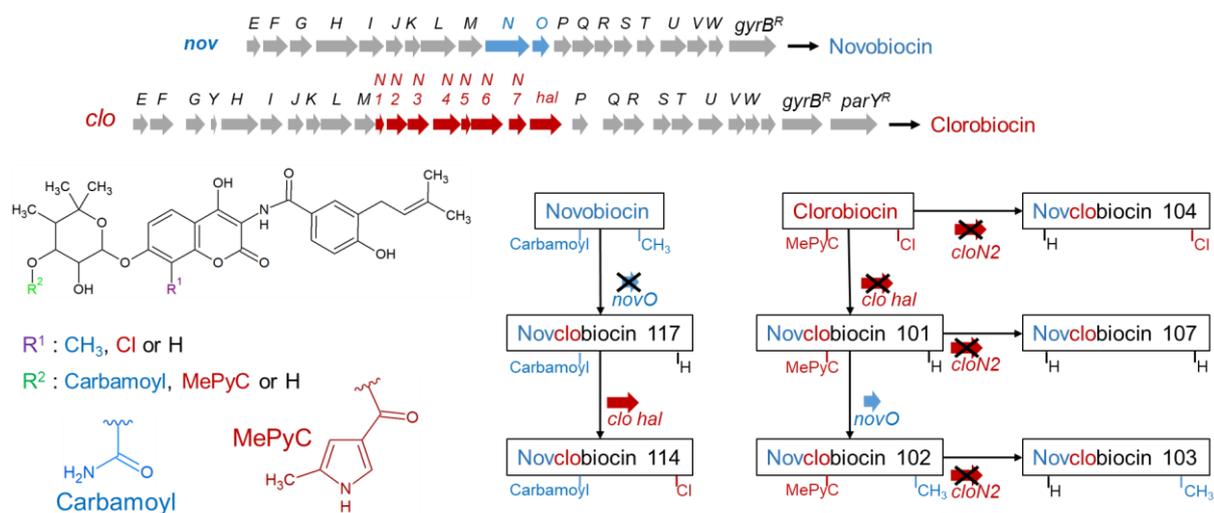


Figure 8: Exchange of tailoring genes to produce novobiocin/clorobiocin analogs (adapted from Pickens et al., 2011)

The two clusters are shown in parallel, with the genes responsible for the structure differences colored. MePyC = methyl-pyrrol-2-carboxyl.

Combinatorial biosynthesis has been mainly applied to two families of metabolites, non-ribosomal peptides (NRPs) and polyketides. The work carried out on the polyketide biosynthetic systems is out of the scope of this manuscript and will not be addressed here. In the next sections, I will detail our knowledge concerning the non-ribosomal peptide synthetases (NRPSs), and present the combinatorial biosynthetic approaches that were conducted on this family of enzymes.

## 2. Non-ribosomal peptide synthetases (NRPSs), a class of complex modular enzymes

The number of non-ribosomal peptides (NRPs) exhibiting anti-infective properties is important. One reason for this lies in the diversity of incorporated monomers: approximately five hundreds, including non-proteogenic amino acids, fatty acids, and sugars (McErlean et al., 2019; Strieker et al., 2010). But this comes with a price: the enzymes synthesizing the NRPs are huge; for instance cyclosporine, an 11-residue peptide, requires an enzyme of about 1.5 mega daltons. An extensive review on NRPS notably describing the incorporated monomers has recently been published (Süssmuth and Mainz, 2017).

### 2.1. NRPS assembly lines and facilitators

#### 2.1.1. Principle of NRP biosynthesis

NRPSs are large multi-modular enzymes responsible for the biosynthesis of a non-ribosomal peptide (NRP). Several subunits may be needed, each of them being constituted of modules. The model of assembly is presented on Figure 9. Each module incorporates one monomer to the final peptide. Each module is divided in domains. There are three core domains. The adenylation (A) domain recognizes the amino acid, activates its carboxylate moiety under the

form of an amino acid adenylate at the expense of one molecule of ATP, and covalently binds it as a thioester to the 4'-phosphopantetheinyl (ppant) arm of the peptidyl carrier protein (PCP) domain, also called thiolation (T) domain (Keller and Schauwecker, 2003). The PCP domain presents the substrate tethered to its cofactor to the other domains. The condensation (C) domain catalyzes the formation of an amide bond between two amino acids and, thus, the elongation of the peptidyl chain. The initiation module usually only contains A and PCP domains, while the extension modules contain C, A and PCP domains. At the end of the assembly chain, the termination module also usually contains a thioesterase (TE) domain, which releases the product by hydrolyzing the thioester bond, sometimes through intramolecular cyclization. Release of the product can also be catalyzed by a C domain, a reductase (R) domain or even be non-enzymatic (McErlean et al., 2019).

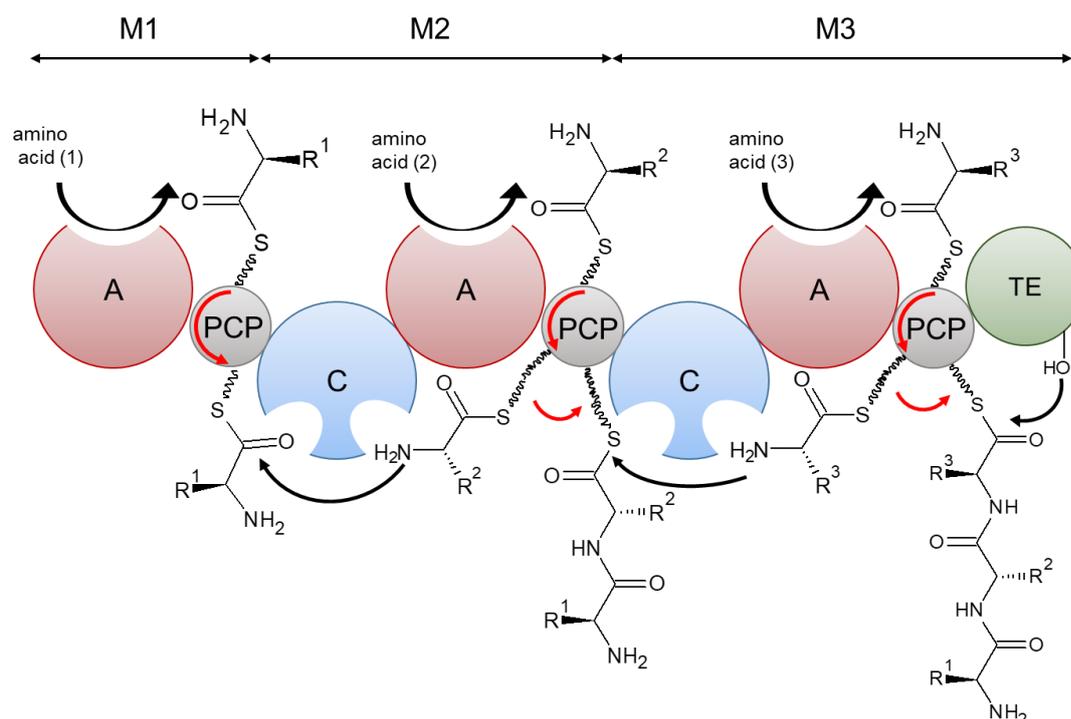


Figure 9: NRPS biosynthesis model

Amino acid substrates are recognized by adenylation domains (A). The aminoacyl-AMP intermediate formed is then loaded on the thiol group of a 4'-phosphopantetheine arm tethered to the peptidyl carrier protein domain (PCP). Condensation domains (C) catalyze successive peptide bond formation. The first module is known as the initiation module (M1) and subsequent modules (M2) are known as elongation modules. The final module (M3) contains an additional thioesterase domain (TE) which catalyses hydrolysis or cyclisation to release the peptide from the NRPS.

In addition to the core domains, optional domains can be included in the modules, such as epimerization domains, methylation domains or cyclization domains (Hur et al., 2012; McErlean et al., 2019; Winn et al., 2016). Epimerization domains catalyze the epimerization of L-amino-acids into their D-form. They are only active on substrates tethered to the PCP domain. The presence of heterocyclic rings in the NRP is explained by the action of the heterocyclization (Cy) domain. Cy domains exhibit a strong specificity, and they produce thiazoline rings from the thiol of cysteine residue, or oxazoline ring from the hydroxyl group of serine or threonine residue. The cycles can be further oxidized or reduced by the corresponding oxidation or reduction domains, which are often stand alone proteins. Methyltransferase domains transfer a methyl group from its cosubstrate

(S)-adenosyl methionine (SAM). While N-methyltransferases act *in cis* during the biosynthesis or *in trans* on the complete product, C-methyltransferases tend to methylate precursors before the assembly of the final molecule. Formylation (F) domains, which add a formyl group, have been little studied until now, except for the F domain of gramicidin NRPS, which exhibits high specificity. Finally, halogenase domains are frequent in NRPSs, and halogen groups play an important role in the antibiotic properties (such as for the antibiotic balhimycin and antifungal syringomycin E). The peptide can also be modified by other tailoring enzymes after being released from the NRPS.

NRPSs are monomeric (Weissman, 2015). An NRPS can be organized as one protein, and then it is called type I NRPS, or as several interacting subunits, which is type II NRPS. Type II NRPS is preponderant in bacteria (Hur et al., 2012). There are three categories of NRPSs (Figure 10). Type A corresponds to linear NRPS: the assembly chain is followed strictly, there are as many monomers as modules, and the order is maintained. This type is often used as a canonical example, and knowing the sequence, one can predict the final NRP. Tyrocidine is synthesized by a type A NRPS. Type B NRPS is called iterative, some of the modules can be reused several times, and the peptide is made of repetitive sequences. Enniatin is an example of type B NRP. Type C is non-linear NRPS, the arrangement of the modules does not correspond to the sequence of amino acids obtained, and one domain, not one module, may be reused. Myxochelin is an example of type C NRP.

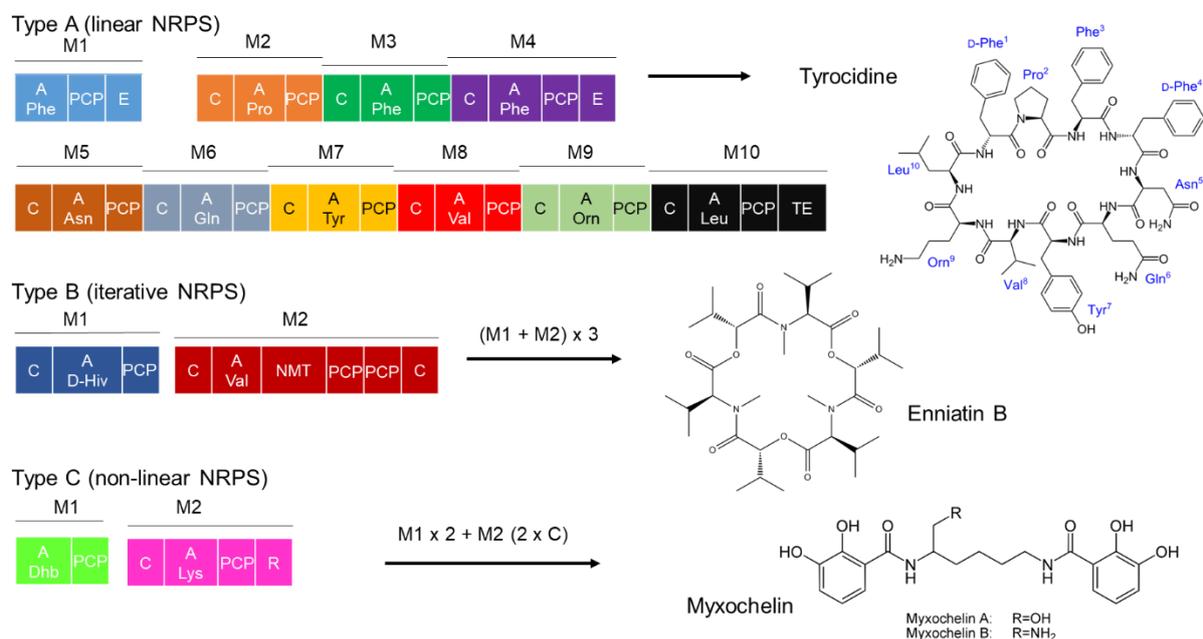


Figure 10: The different NRPS categories

NMT= N-methyltransferase domain ; R = reductase cleavage ; D-Hiv = D-2-hydroxyisovaleric acid ;

Dhb = dihydroxybenzoyl

### 2.1.2.PCP domain priming by the PPtases

The attachment of the 4'phosphopantetheinyl (ppant) arm to the PCP domain is done by the Sfp-type phosphopantetheine transferases (PPtases) from a Coenzyme A in a Mg<sup>2+</sup>-dependant

reaction (Hur et al., 2012; Strieker et al., 2010). PCP domain is converted from the inactive *apo* state to the active *holo* state. Since there is a large amount of acylated Coenzyme A in the bacteria, the PCP domain is often misprimed with an inactive acylated-ppant. Type II-thioesterases then function as repair enzymes and hydrolyze the acyl group, yielding a functional *holo*-PCP domain. PPTases and type II-thioesterases are usually not included in a specific biosynthetic gene cluster, they are present on the genome, and play a pleiotropic role, priming PCP domains from different BGCs. Sfp was one of the first described PPTases, and it exhibits an important promiscuity. Bunet *et al.* (2014) have found a Sfp-type PPTase in *S. ambofaciens*, associated to no specialized metabolite cluster, with a pleiotropic role. The deletion of the encoding gene abolished the production of congocidine and coelichelin, synthesized by NRPSs, and of spiramycin, stambomycin and grey-spore pigment, all polyketides synthesized by polyketide synthases. This shows that this PPTase is involved in the priming of the peptidyl carrier and acyl carrier proteins of several of the biosynthetic pathways, and is likely involved with all the NRPS and PKS clusters of the strain.

### 2.1.3. Role of MbtH-like proteins (MLP) as helpers

MbtH-like proteins (MLP) are small proteins of about 70 amino acids found in some NRP gene clusters (Hur et al., 2012). They were named after the MbtH protein encoded in the BGC of the siderophore mycobactin in *Mycobacterium tuberculosis*. The function of these proteins is not fully understood yet, but they associate with A domains during NRP biosynthesis and they are considered as chaperones or facilitators. MLP may be needed for the correct solubility and activity of the A domain, or only for its solubility. It may enhance both solubility and activity of an A domain that is functional on its own as well (Schomer and Thomas, 2017). For instance, the purification of Cgc18 involved in congocidine biosynthesis required the MLP partner SAMR23877 to obtain a soluble fraction, and the authors reported many other cases for which solubility and/or activity was impeded in the absence of MLP (Al-Mestarihi et al., 2015). Associated MLP and A domain are bound tightly and copurified, and stoichiometric amounts of 1:1 of MLP:A didomains have been reported (Baltz, 2011).

MLP structure consists of three  $\beta$ -strands, which interact with one adjacent  $\alpha$  helix (Miller et al., 2016). There is no obvious catalytic group in MLP structure (Schomer et al., 2018). The structure of SlgN1, the NRPS of streptolydigin, made of MbtH-like domain at the N terminus and adenylation domain, was recently crystallized (Herbst et al., 2013). The MLP interacts with the big N terminal part of the A domain (Figure 11). It is worth noting that MLP has no direct contact to the substrate of the A domain. The full module of EntF containing C-A-PCP-TE domains has also been crystallized bound to its native MLP from *E. coli*, or to a non-cognate MLP from *Pseudomonas aeruginosa* (Miller et al., 2016). The interaction surface is similar to the one reported by Herbst and collaborators (2013). The presence or absence of the MLP had no visible impact on the structure of the A domain, which suggests that the activation of A domain is not achieved by a conformational change (Miller et al., 2016). However, in the structure of Dhbf domain A crystallized with its MLP (required for adenylation activity but not for folding), the A domain adopted a more compact form than its structure in absence of MLP (Tarry et al., 2017). Even the smaller C terminal part of the A domain ( $A_{sub}$ ), which is not in direct contact with MLP, seemed impacted.

Not all A domains are dependent on MLP to function correctly. For instance, the A domain CmnO involved in capreomycin biosynthesis is not active without the CmnN MLP, while the A domain CmnF is unaffected by the absence of MLP (Miller et al., 2016). So far, attempts to predict the dependency of A domains to MLPs based on sequence analysis have failed (Miller et al., 2016).

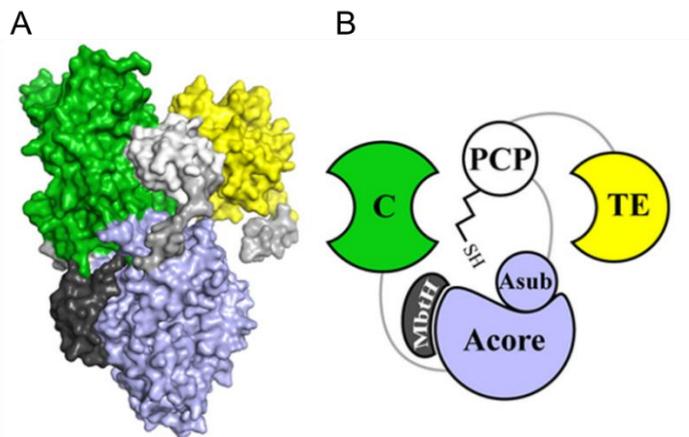


Figure 11: Model of the position of an MbtH-like protein within an NRPS (Herbst et al., 2013).

A) protein structure B) scheme of the domain organization

A domain is separated in two parts, the N terminal core part and the C terminal smaller subdomain. The MbtH-like domain of SlgN1 (dark gray) was crystallized with the core part of the A domain. The remaining domains were positioned by superposing SlfN1 A and SrfA-C structures.

MLPs are usually encoded within the BGC containing the gene encoding their NRPS A domain partner, but a recent study showed that orphan MLPs can be encoded in bacterial genomes (Esquilín-Lebrón et al., 2018). In the case of the orphan and only MLP encoded in *M. xanthus* DK1622 genome, the authors showed that this MLP interacts with NRPSs from at least seven distinct BGCs. This suggests that MLP are not specific of given A domain or a given cluster. This is indeed confirmed by the observation that MLP can activate non-cognate A domains. It was observed in *S. coelicolor*, where CdaX can complement the deletion of CchK and restore coelicelin production, and *vice versa* (Lautru et al., 2007). Schomer and Thomas (2017) have also studied the impact of 7 non-cognate MLPs on EntF activity, involved in enterobactin biosynthesis. EntF native MLP is YbdZ. It copurifies with EntF and improves both its solubility and its affinity for its substrate L-Serine. The authors observed that 5 of the 7 non-cognate MLPs could restore enterobactin production (Schomer and Thomas, 2017). Another study also suggested that the interaction of a MLP with a non-cognate A domain could broaden the A domain substrate promiscuity (Mori et al., 2018).

## 2.2. NRPS domains: structure and substrate specificity

### 2.2.1.A domain structure and specificity

A domain is a well-defined globular structure of 550 to 600 amino acids, which consists in two subdomains connected by 5-10 residues: a big N terminal domain of about 450 amino acids ( $A_{core}$ ), and a smaller C terminal domain of about 100 amino acids ( $A_{sub}$ ) (Figure 12). The active site is located at the junction between the two subdomains. A domains belong to the ANL superfamily

of adenyating enzymes (Acyl-CoA synthetase, NRPS adenylation domain, and Luciferase) (Gulick, 2009). All the enzymes of this family catalyze two catalytic reactions (Hur et al., 2012; Strieker et al., 2010). For A domains, the first reaction is the formation of the adenylate by the  $Mg^{2+}$ -dependent reaction of an amino acid with ATP to yield an acyl-AMP, thus releasing pyrophosphate. The second is the formation of a thioester by reaction of the adenylate with the sulfhydryl group of the ppant arm of a PCP domain. A change of conformation (from adenylate conformation to thioester conformation) consisting in a  $140^\circ$  rotation of the C terminal subdomain ( $A_{sub}$ ) is observed between the two reactions, and as a result the opposing face of  $A_{sub}$  is presented to the active site (Sundlov et al., 2012).

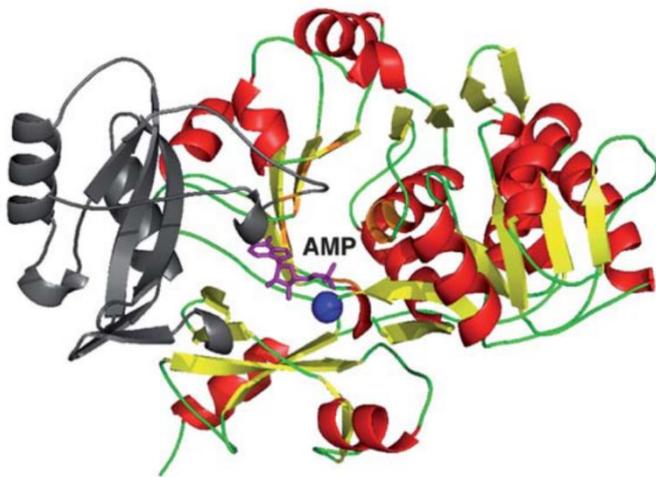


Figure 12: Adenylation domain structure (Hur et al., 2012)

The large N-terminal domain  $A_{core}$  is represented in red, and the small C-terminal domain  $A_{sub}$  in gray. AMP and  $Mg^{2+}$  (blue sphere) are represented on the structure

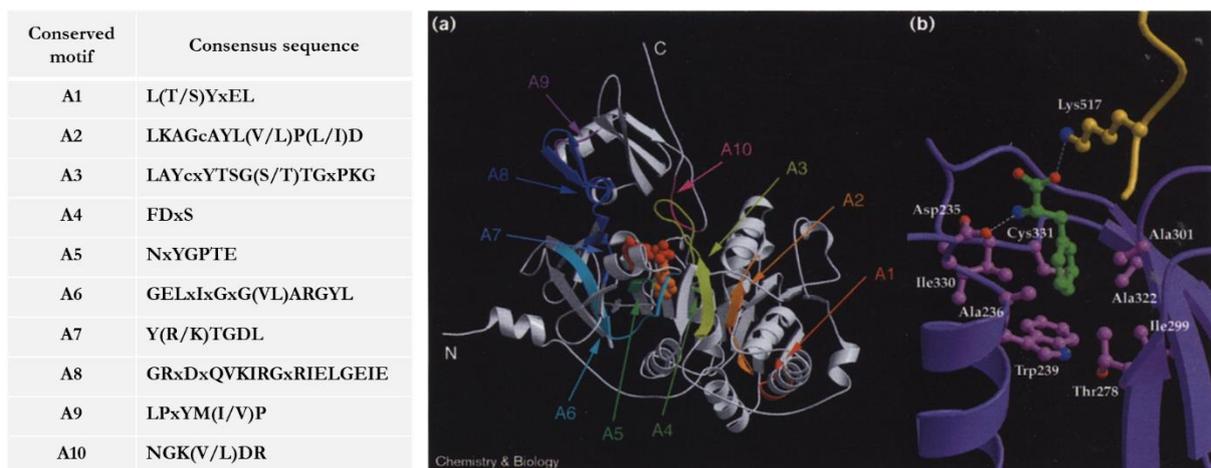


Figure 13: Conserved motifs and crystallization of the Phe-adenylation domain PheA (Stachelhaus et al., 1999)

- The structure is represented with all the conserved motifs annotated (see table), in orange is represented Phe.
- The amino acid involved in Phe recognition and binding are represented, with Phe in green

A domains are the gate keepers of the assembly line, and they present a high specificity for their substrate (Strieker et al., 2010). There are highly conserved sequences, named A1 to A10,

which have a role in the recognition of ATP, its binding and the adenylate formation (Figure 13a) (Keller and Schauwecker, 2003). While these motifs are conserved in all A domains, the residues involved in binding the A domain substrates are variable between various A domains but mostly conserved for a given substrate. PheA, an adenylating domain activating phenylalanine (Phe), was the first to be crystallized and its structure was solved with Phe and AMP (Stachelhaus et al., 1999). Stachelhaus and coworkers (1999) analyzed 10 contact making residues (Figure 13b), and classified them depending on degree of variation by comparing PheA sequence to more than 100 A domains. The highly variant residues were then used to predict the specificity, and derive a signature sequence for 20 substrates of NRPSs. The authors report a predicting success rate of 86%, with only 26 of the 160 sequences unmatched (Stachelhaus et al., 1999). Based on the same structure PheA, another group proposed a very similar approach based on 8 residues (Challis et al., 2000).

There is now a code determining the specificity of each A domain, made of about 10 amino acids, referred as the NRP synthesis “codons”. Specificity is determined by these codons, as well as the cavity of the substrate. Hydrophobicity and side chain size are criteria which may play an important role (Hur et al., 2012). It is interesting to note that several signatures may lead to the same selectivity, the NRP synthesis “codons” present some degeneracy (Lautru and Challis, 2004). Two outputs of this work are the possible prediction of a domain substrate, and the possibility to engineer a domain to change its specificity. For instance, starting from a Phenylalanine A domain, Stachelhaus and coworkers were able to accommodate Leucine with only two mutations. Since the first predictions, the models have been refined, automated methods were developed (Rausch et al., 2005; Röttig et al., 2011). One of the most recent is SANDPUMA, a prediction model available online and integrated to the latest version of AntiSMASH (Chevrette et al., 2017). It is worth noting that this mainly concerns proteogenic amino acids (Kudo et al., 2019). Indeed, specificity of A domains accepting nonproteogenic amino acids is not quite as understood, and more protein structural analyses will be necessary to better understand the substrate recognition mechanisms.

### 2.2.2.PCP domain structure

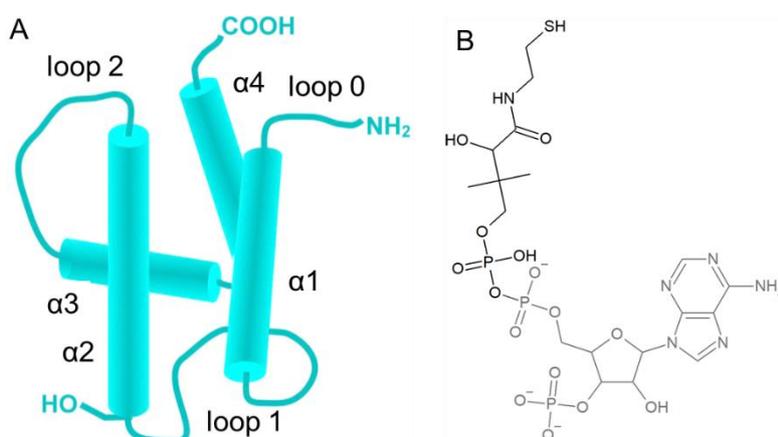


Figure 14: PCP domain structure (Tufar et al., 2014)

A, PCP structure, B, coenzyme A

The 4'phosphopantetheinyl arm is loaded on the hydroxyl group of a conserved serine, at the N terminus of the second  $\alpha$  helix. The 4'phosphopantetheinyl arm comes from a coenzyme A, the part in gray is left as a side product and the part in black is loaded on the PCP.

PCP domains are very small structures of about 80 amino acids (Keller and Schauwecker, 2003), made of 4  $\alpha$  helices (Figure 14). Though the structure is well conserved, the sequence is variable, and shape and charge distribution vary as well, which must affect the PCP domain interactions with other domains (Kittilä et al., 2016). PCP domains have a 4'phosphopantetheinyl (ppant) cofactor bound to the hydroxyl group of a conserved serine residue, in a conserved GGxS motif at the N terminus of the second helix (Figure 14) (Strieker et al., 2010). The reactive sulfhydryl group at the extremity of the cofactor reacts with the adenylated amino acid bound to the A domain to yield the thioester-bound amino acid. During elongation, substrates are shuttled along the modules from one PCP domain to another.

### 2.2.3. C domain structure and specificity

C domains are located at the N terminus of each module, they catalyze bond formation between two consecutive amino acids. C domains are also able to condense an amino acid with another molecule, such as a polyketide, or an acid. C domains are about 500 amino acid long, and are constituted of two subdomains that form a V-shape (Figure 15)(Hur et al., 2012). They have conserved core sequences (C1 to C6), and C3 (sequence HHxxxDG) plays a prominent role in the condensation reaction (Keller and Schauwecker, 2003; Samel et al., 2007). The catalytic center is at the junction of the two subdomains, and contains the second Histidine of the conserved motif C3. There is a channel, leading from one side of the enzyme to the other, through the active site. This channel allows the entrance of the ppant arms to which are tethered the two substrates. The peptide bond formation is believed to depend on electrostatic interactions with the conserved histidine rather than acid-base catalysis (Samel et al., 2007).

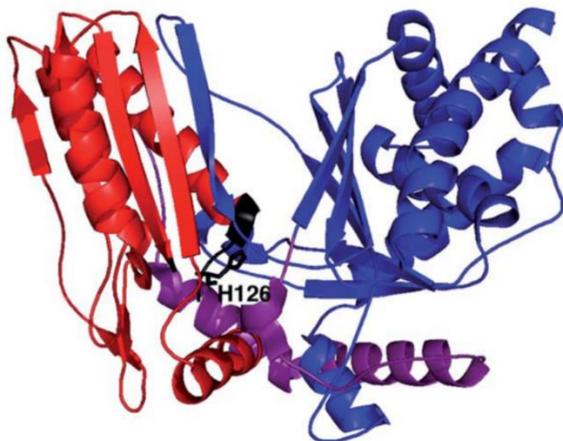


Figure 15: X-ray crystal structure of the stand-alone C domain, VibH, from the *Vibrio cholerae* vibrioactin synthetase (Hur et al., 2012)

The two lobes are represented with different colors, the histidine indicated (His126) is part of the catalytic center.

C domains present some substrate specificity to some extent. This selectivity appears to be higher at the acceptor side (binding site of the downstream residue) than at the donor side (binding site of the upstream residue) (Lautru and Challis, 2004). The stereo selectivity (L or D amino acid) for the acceptor amino acid is really high. There is also some selectivity for the side chain (Lautru and Challis, 2004). Yet, some NRPSs synthesize several variants of a given peptide, so some

promiscuity must exist, at least for some C domains. This strong specificity for the acceptor substrate suggests however that the C domain and its downstream A domain should be kept as an item in the case of engineering NRPS. For the donor side, though some stereospecificity may be observed, the substrate specificity is much more relaxed and there are examples of non-cognate substrates incorporated (Lautru and Challis, 2004). There was still some selectivity observed in the case of larger intermediates (Brown et al., 2018; Calcott and Ackerley, 2014). Interestingly, a C domain highly selective for its donor group was reported in a hybrid fungal PKS-NRPS (Kakule et al., 2014). This C domain accepts a polyketide as donor substrate, and fusion experiments showed that the length of the chain had a significant impact: while octaketides were accepted, nonaketides were included only in a small minority by the condensation domain. In this case where the NRPS does not contain any A domain, the authors concluded that the C domain was responsible for the primary selectivity. Similarly, for the family of microcystins, *in vitro* studies revealed that while the A-PCP didomain was multi-specific, the C-A-PCP module was mono-specific, arguing in favor of a major role of the C domain in substrate control, and consisting in the first instance of C domain directly modulating the substrate specificity of A domain (Meyer et al., 2016).

#### 2.2.4. TE domain structure and specificity

TE domains are found only at the end of the assembly chain, they allow the release of the final compound. This domain is about 250 residue-long, and via a serine residue used as a nucleophilic catalyst (conserved motif GxSxG), it catalyzes the hydrolysis or macro cyclization of the compounds (Hur et al., 2012). Macro cyclization is the most common outcome, and given the different ring sizes observed among NRP, TE domain must present some substrate specificity. The catalytic residues are located inside a hydrophobic cavity with the shape of a bowl, and a “lid” region is on top of the cavity (Figure 16). In the first crystallized structure of a TE domain, the TE domain of the surfactin synthetase (SrfTE), the “lid” consists in three  $\alpha$  helices. The “lid” can cover the active site with the  $\alpha$  helices parallel, when it is in “closed” position, excluding water. In “open” position, the “lid” is located aside, because the first helix is angled upward, making an opening on the face opposed to PCP domain (Miller et al., 2016). The “lid” is thought to be responsible for the substrate recognition.

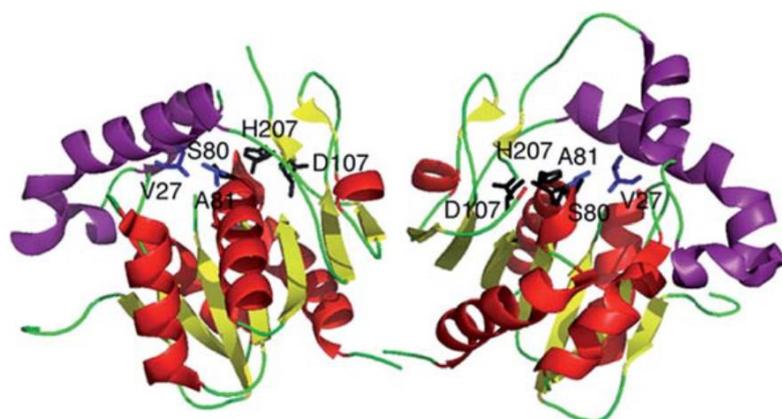


Figure 16: Crystal structures of the surfactin thioesterase domain, SrfTE (Hur et al., 2012) On the left, the violet part is the « lid » observed in closed conformation, on the right it is presented in open conformation. Residues S80, D107, and H207 in black form the catalytic triad.

Study of the TE domains from surfactin and from tyrocidin NRPSs showed that the TE domain exhibits side chain specificity for the residue next to the catalytic serine, and side chain selectivity and enantioselectivity for the residues next to the intramolecular nucleophile (Lautru and Challis, 2004). However, study of the influence of the length of the chain or the nature of the cyclization nucleophile yields no clear results, since SrfTE is very specific, while TycTE has a broader substrate specificity. In some cases, high substrate specificity might limit the range of analogs accepted by the TE domains.

Altogether, A, C and TE domains hence all present some substrate specificity which must be taken into account before modifying the assembly chain. Yet, attention should also be focused on the interactions among domains, modules and subunits, which must be respected for the partners to interact correctly. To precisely understand NRPS assembly, and be able to engineer it (Jenke-Kodama and Dittmann, 2009), we need to know: (i) the structural arrangements of domains within modules, (ii) the role of linker regions between domains and modules, (iii) how the order of interactions is controlled (which partner to interact with), (iv) how proteins associate with each other correctly. These points will be developed in the next subsection.

### 2.3. Conformational changes and interactions inside NRPS modules

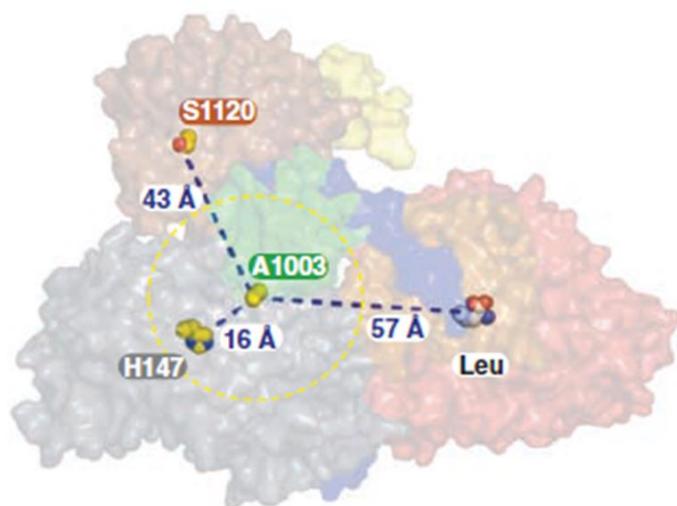


Figure 17: Termination module of SrfA-C (Tanovic et al., 2008)

A1003 corresponds to the mutated serine residue where the ppant arm is bound. The yellow circle has a radius of 20 Å and corresponds to the position the ppant arm can reach, and the catalytic residues of each domain are represented (Leu A domain, His C domain, S TE domain). In this conformation, the substrate can only attain the C domain catalytic site.

During NRP biosynthesis, the PCP domain has to interact with several other domains. Its ppant arm has to access at least three different domain catalytic sites: the catalytic site of the A domain, to be loaded with the monomer, the catalytic site of the upstream C domain to serve as an acceptor and the catalytic site of the downstream C domain to serve as a donor (Gulick, 2016). Although the ppant arm has little contact with the core PCP domain and can move freely, it only spans 20 Å (Samel et al., 2007). The first structure of an entire module, the terminal module of

SrfA-C (C-A-apoPCP-TEI, (Tanovic et al., 2008)), shows that the catalytic sites of A and TE domains are out of reach (Figure 17). Therefore, domain rearrangements are necessary to allow access of the substrate to the different catalytic sites (Izoré and Cryle, 2018).

### 2.3.1. Conformational changes of A and PCP domains during NRP biosynthesis

Recently, the first module of gramicidin, LgrA (F-A-PCP), was crystallized at different catalytic steps (Schmeing, 2016), and four distinct conformations have been observed (Figure 18). In all conformations, the formylation domain, which adds a formyl group to the substrate bound to the PCP domain, forms a rigid elongated shape with the core adenylation domain ( $A_{core}$ ). In contrast,  $A_{sub}$  and PCP domains undergo huge movements during the cycle. In open conformation (a, no substrate bound),  $A_{sub}$  is located away from  $A_{core}$ . When ATP and the amino acid substrate reach the active site,  $A_{sub}$  rotates of  $30^\circ$ , yielding the closed conformation (b, also called adenylate-forming conformation). Once the adenylate is formed,  $A_{sub}$  rotates of  $140^\circ$  and presents its opposite face to the active site. This movement induces a displacement of the PCP domain, which brings the ppant arm into the A domain, in the thiolation conformation (c, also called thioester conformation). Another rotation of  $180^\circ$  of  $A_{sub}$  brings PCP domain in reach of the F domain, to the formylation state (d) (Reimer et al., 2016). It is possible that the PCP domain interacts with the downstream C domain during the first two states (a,b), where PCP domain structure is not well resolved. During the whole cycle, the movement of the module is coordinated by  $A_{sub}$ ; PCP domain and therefore the ppant arm move because of the movement in the A domain (Gulick, 2016).

A terminal module from an uncharacterized NRPS, AB3404 (C-A-PCP-TE), was crystallized at the same time and shows the ppant arm of the PCP domain located in the C domain active site (Gulick, 2016). The A domain is in “closed” or adenylate-forming conformation in this structure, which shows that both C and A domains can be in catalytic state simultaneously. Based on this structure, the structure of SrfA-C (Figure 17, (Tanovic et al., 2008)) and a third terminal module structure (EntF, with the ppant arm oriented in the A domain in thioester-forming conformation (Miller et al., 2016)), Drake and collaborators (2016) proposed a 3-step catalytic cycle (Figure 19). In state I, the A domain is in the thioester-forming conformation, with the ppant arm of the PCP domain located in the active site of the A domain (crystal structure of EntF). In state II, the A domain is in the adenylation-forming conformation while the ppant arm of the PCP domain is located in the acceptor site of the upstream C domain (simultaneous condensation reaction and adenylation of the next amino acid substrate, increasing the efficiency of NRPS catalysis). In the final state, state III, the PCP domain now loaded with the peptidyl chain is oriented towards the downstream C domain in elongation modules, or TE or R domains in termination modules. There is no crystal structure available for the state III yet.

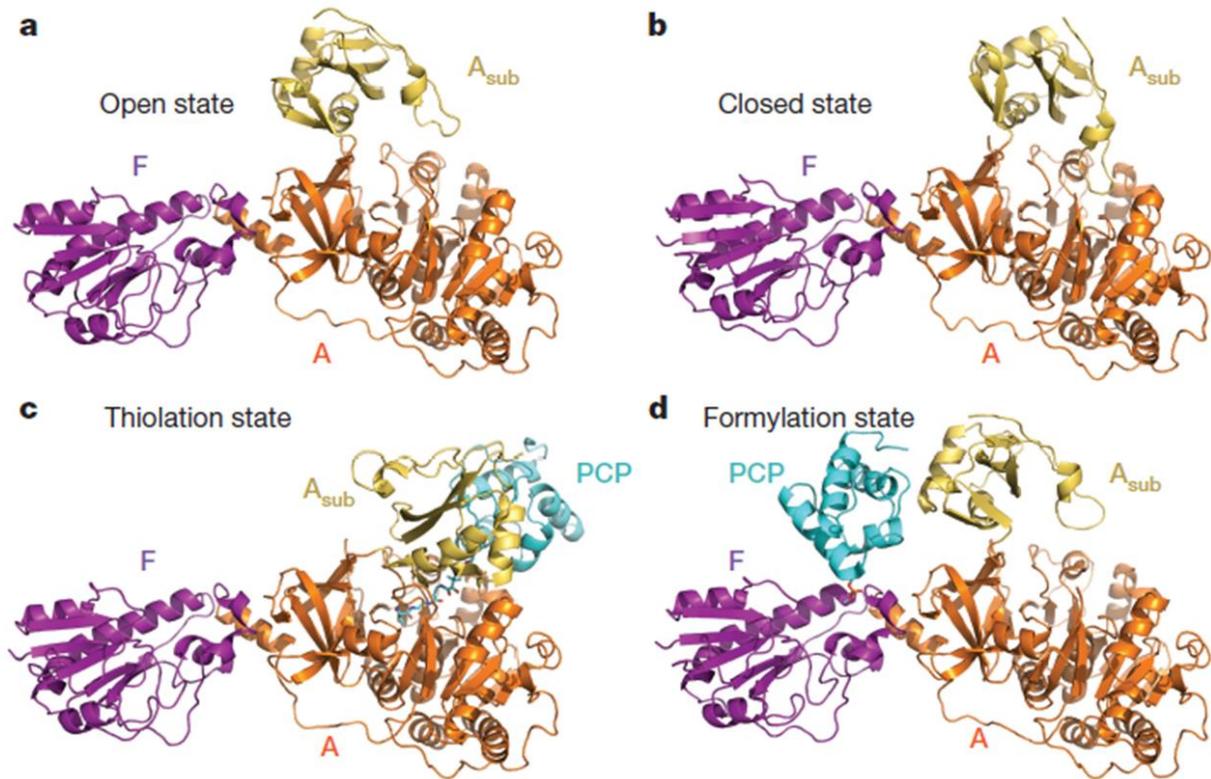


Figure 18: Four structures of the linear gramicidin synthetase (LgrA) initiation module representing every major conformation of the module in the catalytic cycle (Reimer et al., 2016)  
The PCP domain is not necessary for the open and closed states and is disordered in b and c.

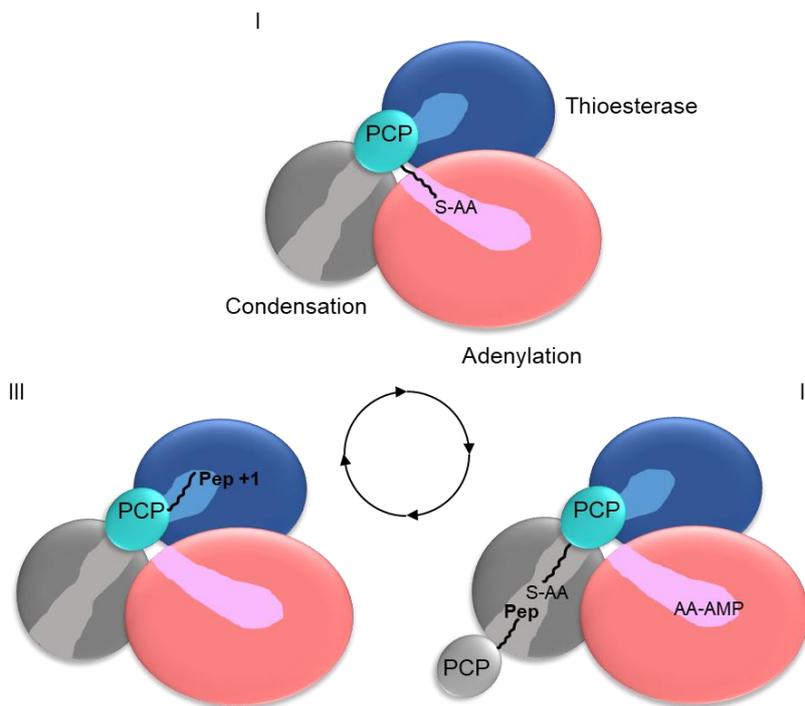


Figure 19: Dynamics of the revised NRPS module cycle

The pantetheine cofactor is represented by the wavy line with a terminal thiol, AA-AMP = amino-acyl-adenylate, Pep = Peptide, S-AA = amino acid bound to the 4'-phosphopantetheinyl arm of the PCP domain

The commonly accepted hypothesis is that C domain opens and closes to accommodate PCP domain, but until now, only the closed conformation has been observed (Miller et al., 2016). Interestingly, the  $\alpha$  helix 1 from the C domain of EntF is unresolved in one of the structures observed by Miller and coworkers, which means that the tunnel for the ppant arm is bigger than in other observed structures. The larger tunnel comes with destabilized interactions, and unwinding of  $\alpha$  helix 1 may be a mechanism to bind and release downstream PCP domain, which still remains to be confirmed (Kittilä et al., 2016). As for the TE domain, its structure is often disordered, it seems that TE domain is able to adopt variable positions (Gulick, 2016; Miller et al., 2016).

As demonstrated by the description of the conformational changes during the catalytic cycles, NRPSs are highly dynamic structures.  $A_{core}$ - $A_{sub}$  movements are the most important observed during the NRP biosynthesis (Gulick, 2016). These movements imply modifications of the protein/protein interactions between the different NRPS domains and modules.

### 2.3.2. Interdomain linkers constrain domain movements

Interdomain linkers are essential, notably because they maintain protein interactions and affect protein stability, orientation and folding. However, they must also allow the domain movements necessary for the different catalytic cycle conformations. The termination module of SrfA-C is the first example where interdomain linkers were highlighted (Figure 20) (Tanovic et al., 2008). Among the different interdomain linkers, C-A linkers have been described as the most rigid: they are made of 32 residues, are L-shaped and are associated with both domains (Tanovic et al., 2008).

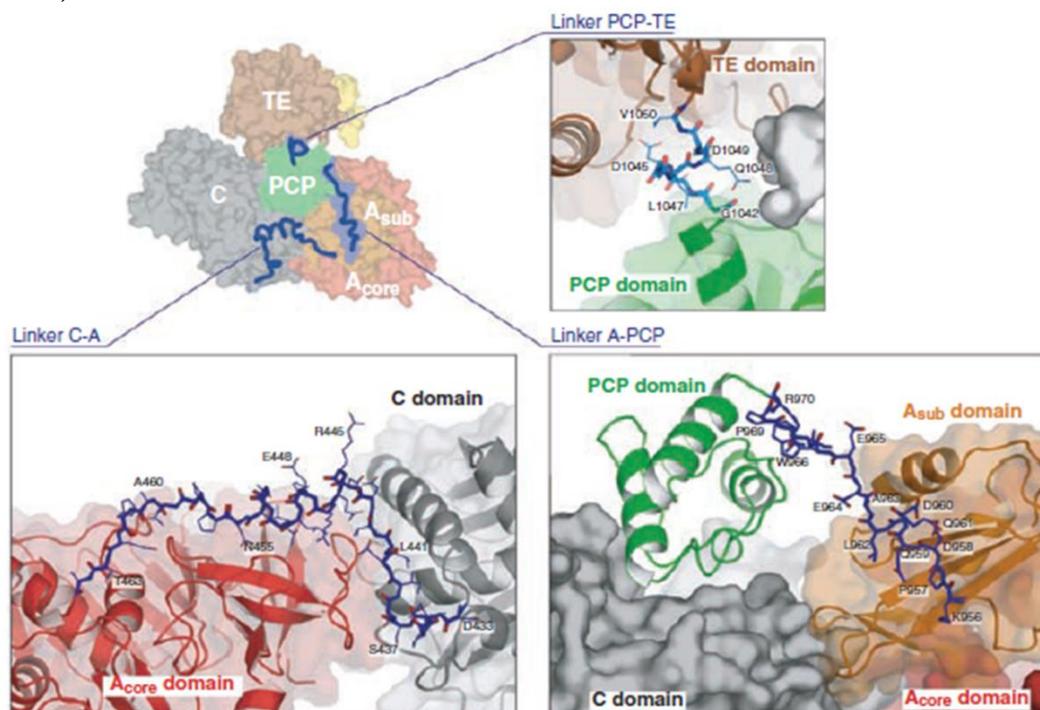


Figure 20: Linkers of the domains of the termination module of SrfA-C (Tanovic et al., 2008) Linkers are shown in blue. C-A linker is 32-residue long, and 11 of them form an  $\alpha$ -helix. A-PCP linker is 15-residue long and PCP-TE linker is 9-residue long, both are disordered, with little interactions with their respective domains.

In contrast, the other linkers described are usually quite mobile. SrfA-C A-PCP linker between  $A_{\text{sub}}$  and PCP domains is only 15 residues (Figure 20). It contains an ordered LPxP motif both maintaining the proper position of the catalytic Lysine of A10 (Figure 13A) and anchoring PCP domain to  $A_{\text{sub}}$  domain (Miller et al., 2014). The rest of the linker has no contact with either A or PCP domains, hence a free rotation of  $A_{\text{sub}}$  and PCP domains is possible. The C-PCP linker of Tyc C5-6 is only also partially involved in the interactions between the two domains, 7 residues out of 15 are mobile and allow an important conformational flexibility (Samel et al., 2007). The same is observed for the PCP-TE linker, both in SrfA-C and EntF, which is disordered, showing there are several conformations adopted during this state (Miller et al., 2016).

The linker flexibility allows movements of the domains, hence the modification of protein/protein interactions during the catalytic cycle.

### 2.3.3. Protein/protein interaction surfaces vary during the catalytic cycle

Because of the movements of NRPS domains during the catalytic cycle, protein/protein interaction surfaces must vary during this cycle. The C- $A_{\text{core}}$  interface described in SrfA crystal structure, however, is a really stable interface with a rigid linker between the two domains. Thus, it was thought to remain unchanged during the catalytic cycle (Tanovic et al., 2008). Yet, the C-A interfaces are slightly different for the three termination module structures, EntF, SrfA-C and AB3403 (Drake et al., 2016). This shows that C domain can move relative to A domain, and that the C-A platform is, thus, more dynamic than we previously thought, though it remains by far the most constant interface (Miller et al., 2016; Reimer et al., 2018).

About all the surface residues of the PCP domain are used for the interaction with other domains at some point of the cycle in the initiation module LgrA (Schmeing, 2016). PCP domain residues around the ppant arm (Figure 14), especially on  $\alpha$  helix II,  $\alpha$  helix III and the loop 2 in between, are notably important for the recognition of the catalytic E, C or TE domains (Gulick, 2016; Kittilä et al., 2016; Sundlov et al., 2012; Drake et al., 2016; Chen et al., 2016). Loop 1, located between the  $\alpha$  helices I and II, has a key role in A domain recognition (Jaremko et al., 2017). Loop 0 was also shown to stabilize the core fold of PCP domain, and to have an impact on the conformation of PCP domain, hence probably on communications with the other domains (Harden and Frueh, 2017).

$A_{\text{sub}}$  and PCP domains are very flexible in the NRPS structure (Strieker et al., 2010). Kittilä and co-workers (2016) suggest, however, that  $A_{\text{sub}}$  domain movements are not sufficient to explain PCP domain movements during the catalytic cycle, and that conformational changes are also due to covalent modifications (attachment of the ppant arm and of the amino acid/peptidyl chain) along the cycle. Usually adding the ppant arm does not alter significantly PCP domain structure, but on an atypical instance, PCP conformation was modified upon the ppant arm binding, and A domain affinity for the carrier protein increased (Goodrich et al., 2017). This remains an atypical example, but changes in electrostatic interactions and solvent accessibility may impact the course of the catalytic reaction and the change of conformation (Gulick, 2016; Sundlov et al., 2012).

The articulation of domains in a module is remarkably dynamic, and leads us to wonder how inter modular structure interacts. The next section will report knowledge concerning the NRPS inter modular structure and interactions.

## 2.4. NRPS subunit structure

### 2.4.1. Intermodular linkers

Compared to interdomain linkers, intermodular linkers remain little studied. After establishing a database containing nearly 40,000 intermodular linkers, Farag and collaborators (2019) observed that intermodular linkers were specific of a pair of amino acids, which means that they connect two modules that activate a specific pair of substrates. Therefore, intermodular linkers could also be gatekeepers of the specificity of the NRPSs.

### 2.4.2. NRPS multimodular structure

The first multi-modular structure obtained consists in part of the two-module NRPS Dhbf: the A and PCP domains from the module 1, and C domain from the module 2 (Tarry et al., 2017). Contrary to expectations (Reimer et al., 2018), A1-PCP1-C2 crystals showed that there was no contact between A1 and C2. Hence, the PCP domain must play an important role as a mediator of intermodular contacts.

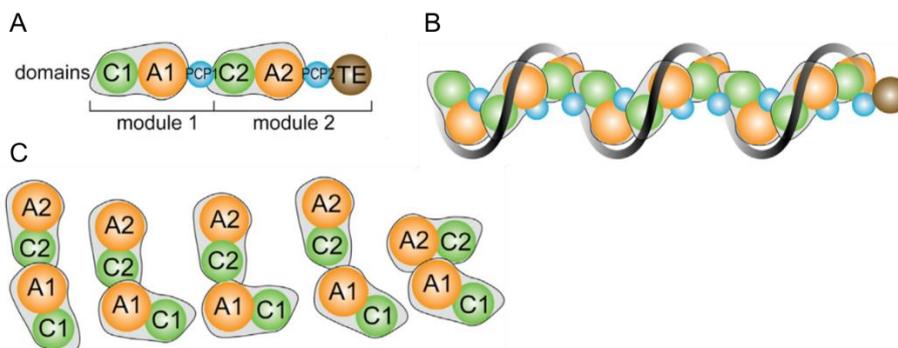


Figure 21: Schematic of a proposed regular helical structure for multi-module NRPS enzymes (Lott and Lee, 2017)

A. Representation of an NRPS made of 2 modules. B. Hypothetical multimodular structure of a nine-module NRPS enzyme, forming a helix. C. Electron microscopy of the two-module structure of A, representing different forms. The observed structure does not correspond to the model proposed in B.

Using the structure of the termination module SrfA-C and the structure of di-module TycC5-6 PCP-C, Marahiel proposed a model based on a helical organization (Figure 21B), where each module is rotated of  $120^\circ$  relative to its neighbor (Marahiel, 2016). However, electron microscopy of the two full modules Dhbf (C1-A1-PCP1-C2-A2-PCP2) revealed that while C - A didomains always form a stable platform, the overall form of the two modules is L-shaped, with a variable angle between the two modules (Figure 21A and C) (Tarry et al., 2017). The results suggest that there is not a single module-module conformation and no consistent module-module interface,

but only transient interactions. Though the orientation is somewhat constrained, there is probably no regular repeating supramodular architecture in NRPSs.

### 2.4.3. Communication mediating (COM) and docking domains between NRPS subunits

In type II NRPSs, the various subunits constituting the NRPS have to establish functional and specific interactions with their cognate partners to produce the expected NRP. Short communication-mediating domains (COM), mediating these interactions, have been detected in NRPSs catalyzing the formation of cyclic (lipo)peptides (Hahn and Stachelhaus, 2004; Liu et al., 2016). The COM domains are defined as the most C terminal 20 to 30 amino acids of TycA, and the 15 to 25 N-terminal amino acids of TycB (Figure 22). Matching pairs of COM domains are decisive to allow the formation of the product, though the core part of the subunits also slightly contributes to the interaction (Dehling et al., 2016). Indeed, COM domain swapping experiments led to successful interaction between non-cognate subunits, *in vitro* (Hahn and Stachelhaus, 2006), or *in vivo* (Chiocchini et al., 2006; Liu et al., 2016).

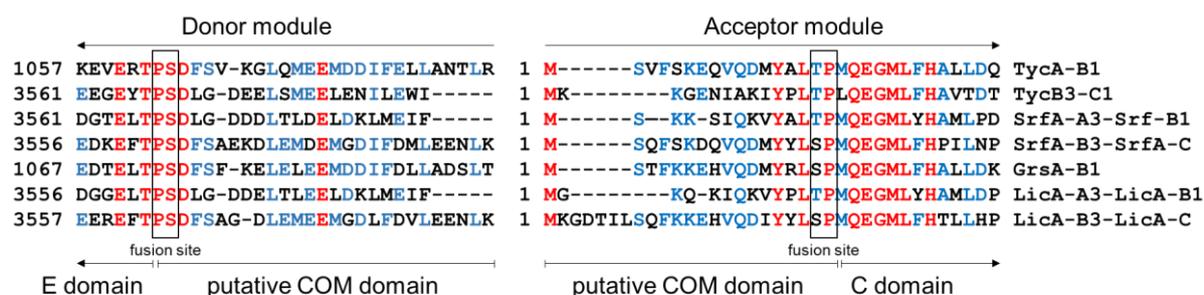


Figure 22: Sequence alignment of putative COM domains (Hahn and Stachelhaus, 2004)

Conserved residues (in blue: quite conserved; in red: always conserved) and fusion sites used for swapping experiments are indicated.

Using the NRPSs GrsA and TycB1, which functionally interact, Dehling and colleagues (2016) attempted to determine the structure adopted by the COM domains. They concluded that it was most likely that the TycB1 acceptor COM domain adopted a hand-shaped structure with a hydrophobic core while the GrsA donor COM domain exhibited a helix pattern. Further experiments are yet still necessary to confirm this helix-hand model.

Although COM domains are often quoted in NRPS reviews, they remain an atypical feature of NRPSs, mainly shared by lipopeptide NRPSs. They might just be one recognition system among several orthogonal systems (Süssmuth and Mainz, 2017). For instance, the rhabdopeptides and xenortide peptides (RXP), made of 2 to 3 monomodular iterative NRPSs, have N and C terminal docking domains with no homology to the COM domains reported above (Hacker et al., 2018). The N terminal docking domains are about 65 amino acid long and are quite structured, while the C terminal docking domains are about 20 amino acid long and rather unstructured. In other cases, protein-protein linkers may exist, but may be less conserved. More studies are required on this topic.

A tremendous amount of knowledge has been gained on the structures of the NRPS megaenzymes during the last decade. We now acknowledge that NRPSs are highly dynamic, with multiple conformations and transient interactions. Yet much remains to be deciphered, as for instance the mechanisms of iterative (type B) or non-linear (type C) NRPSs are not understood (Kim et al., 2015). The multiple elements that come into play for the proper functioning of NRPSs most likely explain the difficulty met to engineer the assembly lines. However, even before our current understanding of NRPS dynamics and mechanisms, engineering experiments have been carried out and have contributed to our knowledge of NRPSs. In the next section, several examples of combinatorial biosynthesis will be discussed, highlighting the fundamental knowledge gained via these experiments.

### **3. Combinatorial biosynthesis experiments of NRPSs, knowledge from trial and error on the modifications of NRPSs**

NRPS biosynthetic systems are responsible for the production of a huge diversity of compounds. Yet, modification of these biosynthetic systems could lead to the development of natural products analogs with improved pharmaceutical properties, or to the generation of entirely new compounds. The manipulation of NRPS biosynthetic pathways can be conceived at various levels: the precursors, the tailoring enzymes and the NRPS biosynthetic systems themselves can be modified. The first two approaches have been introduced earlier in this manuscript (see section 1.3.3) and will not be developed here. This section will focus on the NRPSs themselves. This can have a tremendous impact on the NRP diversity, but also has its importance in fundamental research: knowledge can be acquired from trial and error on the modifications of NRPSs. Knowledge acquired on NRPS enzymes through examples of combinatorial biosynthesis experiments are reported in this section.

#### **3.1. Modifications of A domains**

Modifying the primary structure of a peptide synthesized by an NRPS necessarily implies to modify one or several A domains. This can be achieved by different methods: modifying residues (site-directed mutagenesis) or regions (sub-domains) of an A domain, or entirely replacing one A domain by another one. This last method will be treated in the next section. This section is focused on the first two methods, which have the advantage of leaving the global structure of the A domain intact, thus potentially preserving regions important for interactions with other domains.

##### **3.1.1. Modifications of A domain specificity by mutagenesis**

The discovery of the “nonribosomal code” opened the way to site-directed mutagenesis to change an A domain substrate specificity, by targeting the 10 residues identified as conferring the specificity. The first experiment was reported by the team of Mohamed Marahiel (Eppelmann et

al., 2002). They changed the substrate specificity of the A domain of the first module of the surfactin synthetase from Glu to Gln (one residue mutated) and of the fifth module from Asp to Asn (one to three mutations). In all cases, a complete switch of A domain substrate specificity was observed. However, when the substrate specificity of the A5 domain was changed from Asp to Asn, this was at the expense of the catalytic efficiency, which decreased 10 fold compared that of the wild type A5.

Site-directed mutagenesis has also been performed on Calcium Dependent Antibiotic (CDA) NRPS in *S. coelicolor*, on the A domain of the module 10, to change its substrate specificity from (2S,3R)methyl glutamate (mGlu) and glutamate to (2S,3R)methyl glutamine (mGln) and glutamine (Thirlway et al., 2012). In this case again, only one mutation was required to observe *in vivo* production of a CDA variant incorporating Gln instead of Glu in position 10, or, when the mutant was fed with Gly-mGln, of a variant incorporating mGln instead of mGlu. Regrettably in both cases, the yield of the variants compared to the natural products were not reported.

Another approach, still aiming at modifying the substrate binding pocket of A domains, was designed by Evans and collaborators (2011). It consisted in targeting by saturation mutagenesis the three most highly variant residues of the residues conferring specificity to replace valine by a non-polar residue in AdmK, a subunit of the hybrid PKS-NRPS involved in andrimid biosynthesis. Four clones isolated from a library of 14,000 clones produced three new derivatives of andrimid (Ile/Leu, Ala or Phe instead of Val), and one already known derivative. One of these mutants contained four mutated residues, and the 4<sup>th</sup> residue corresponded to a surface residue far from the catalytic site, showing that mutations outside of the specificity-conferring amino acids should also be considered. Yet, in all cases but one, the titers of the andrimid variants produced were far lower (between 4 and 1900 fold) than the production of andrimid by the wild-type strain, even though the culture media were supplemented with 50 mM excess of the amino acid replacing valine.

Using a similar approach, the group of Hilvert undertook to modify the substrate specificity of the A domain of TycA from L-Phe to (S)- $\beta$ -Phe (Niquille et al., 2018). They proceeded by random modifications of four residues in the active site, combined with the reduction of a loop between two  $\beta$ -sheets that has been suggested to be important for  $\alpha/\beta$  specificity. They obtained a variant with a 220:1 preference for (S)- $\beta$ -Phe over L-Phe, while maintaining high catalytic efficiency. Moreover, the author reconstituted in *E. coli* a functional NRPS composed of the engineered TycA module with GrsB. When the 5 amino acid substrates were fed to the strain, the expected peptide was obtained with a remarkable titer of  $120 \pm 20 \text{ mg l}^{-1}$ .

With the exception of the engineering of TycA A domain by Niquille and colleagues (2018), the A domain mutagenesis experiments conducted so far have been moderately successful. On the one hand, complete switch of substrate specificity has often been obtained. However, this is almost always at the expense of the global efficiency of the NRPSs. Different reasons may explain this limited success. One of them is that residues not located within the binding pocket defined by the 10 residues first identified may contribute to A domain substrate specificity and catalytic efficiency. This was already suggested by Marahiel team in 2002 (Eppelmann et al., 2002) and seems to be confirmed by the andrimid experiment (Evans et al., 2011). The next subsection presents experiments that were carried out taking this aspect into consideration.

### 3.1.2. Subdomain swapping

A domain subdomain swapping consists in exchanging a substantial region of the A domain encompassing (part of) the substrate binding pocket. Indeed, it was observed in the homoaomycin NRPS gene that the gene sequences of the A domains present extremely high identity (90%) except for 400 base pairs (bp) around the catalytic site (Crüsemann et al., 2013). Exchange of the identified subdomain of A led to *in vitro* active A domains with modified specificity when hormaomycin NRPS subdomains were used, confirming the evolutionary origin of the diversification of hormaomycin NRPS A domains. However, these exchanges led to inactive A domains when CDA A subdomains were used.

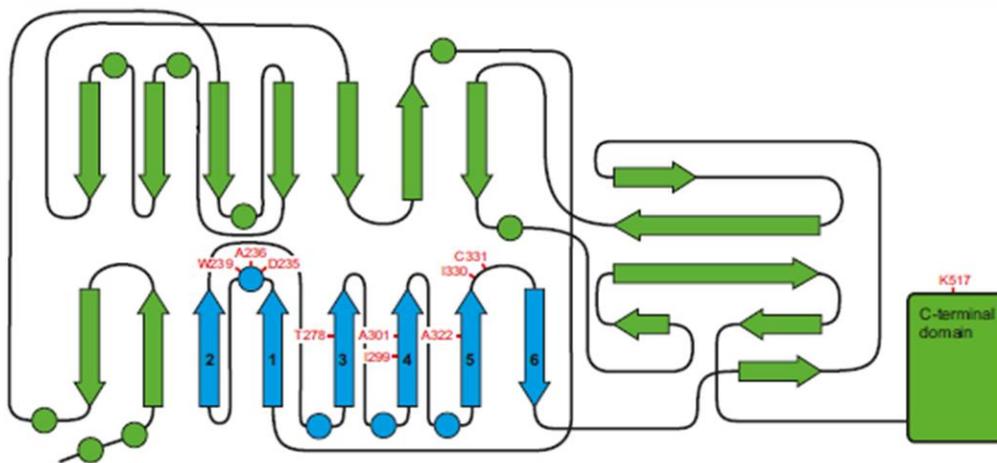


Figure 23: Identification of a flavodoxin-like subdomain in GrsA responsible for substrate binding (Kries et al., 2015)

Circles and arrows symbolize  $\alpha$  helices and  $\beta$ -strands respectively, specificity conferring residues are indicated in red, and the flavodoxin-like subdomain is in blue

In a similar way, Kries and collaborators (2015) attempted to reprogram the A domain specificity of the A(Phe) domain of GrsA. They identified a compact fold, a flavodoxin-like subdomain (132 amino acids) that contains the active site and 9 of the 10 specificity-conferring residues (Figure 23). This subdomain was replaced by 9 other subdomains coming from GrsB or NRPSs from other organisms. The resulting hybrid A domains all adopted the *holo*-form *in vitro*, but only four of them exhibiting adenylating activity. When the flavodoxin-like subdomain of GrsB2 was used, the chimeric A domain activated valine as expected, but with a 15-fold decrease in catalytic efficiency compared to GrsB2, its original module. When GrsA(Val) and GrsB1 were tested for the production of the expected cyclic D-Val-L-Pro was observed, although the reaction was 300-times slower than with the native GrsA-GrsB1 system.

Mutagenesis experiments aim at modifying the substrate specificity of A domains without touching to the general structure of these domains, to avoid disrupting the necessary interactions with other NRPS domains. However, these approaches do not take into account the substrate specificity exhibited by other NRPS domains, and especially the substrate specificity exhibited by

C domains at their acceptor sites. Thus, in parallel to these targeted modification of A domains, approaches were developed to swap entire domains or modules.

### 3.2. Swapping modules or domains to modify NRPS structure

Swapping experiments, which consist in replacing one or several domains or modules to modify the sequence of the synthesized peptide, are particularly tempting in modular enzymes such as NRPSs. To replace one amino acid (or more generally, an NRPS substrate) by another one, such swapping experiments must include an A domain. This A domain, however, can be replaced on its own, or with its associated PCP (A-PCP) or C (C-A) domains (Figure 24), or both (entire module, C-A-PCP).

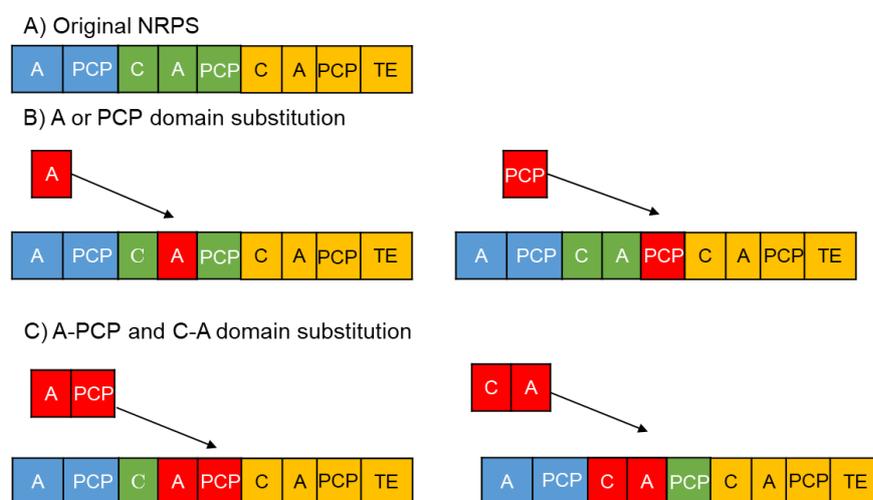


Figure 24: Possibilities of domain substitution in the NRPSs

#### 3.2.1. Domain exchanges

##### - A domains

Not many A domain exchanges have been reported, and they have encountered various degrees of success. In a review published in 2014, Richard Baltz (2014) mentioned that early works on cyclic lipopeptides combinatorial biosynthesis at Cubist involved A domain swapping. These experiments failed and were never published. In a more recent work, the team of David Ackerley replaced the A domain of the last module of the NRPS PvdD involved in the biosynthesis of the pyoverdine siderophore in *Pseudomonas aeruginosa* (Calcott et al., 2014). This A domain is Thr-specific and was replaced by three Thr-specific A domains, as well as six A domains of various substrate specificity (Ser, Lys, Asp and Gly), originating from various modules of pyoverdine synthesizing NRPSs (Table 3). When Thr-specific A domains were used, pyoverdine production was observed, although at a reduced titer for one of the mutants. No new products were detected (fluorescence assay) when non-Thr specific A domain were used. Once again, these results suggested that ignoring the substrate specificity at the C domain acceptor site was likely to result in failures in NRPS combinatorial biosynthesis.

Table 3: Outcomes of the swapping experiments of PvdD

Swapping experiment	Domains introduced	Outcome
A domain swapping of the second module	A(Thr) domain	Pyoverdine produced (3 cases out of 3)
	A(other) domain	No expected product observed, traces of pyoverdine
C-A didomain swapping of the second module	C-A(Thr)	Pyoverdine produced in 1 case (out of 3)
	C-A(other)	Truncated product except for one C-A(Lys) and one C-A(Ser)
PCP-C-A swapping		Results identical to C-A swapping (3 cases)
PCP domain swapping of the first module	PCP domain associated to C domains in <i>cis</i>	Pyoverdine produced (6 cases out of 6)
	PCP domain associated to other domains	Variable outcome: Correct pyoverdine production (5), impaired production (3) or no production (3)

- PCP exchanges

PCP domains are central in NRP biosynthesis, as they interact with many NRPSs domains (A, upstream C, downstream C, TE, optional domains) and free-standing enzymes (PPTases, substrate-modifying enzymes acting on PCP-loaded substrates). A few teams have attempted to examine the portability of PCP domain across NRPS systems. Thus, the Marahiel group examined *in vitro* the interactions of PCP domains with A and epimerization (E) domains, using A/PCP-E or A-PCP/E fusions of gramicidin, tyrocidine and bacitratin NRPS domains (the slash indicates the fusion site) (Linne et al., 2001). They observed aminoacylation by A domains in all constructions although the efficiency of this aminoacylation was impaired at various degrees in A/PCP-E constructs. The effects of separating PCP-E pairs were more dramatic, as epimerization was observed only once out of A-PCP/E constructs. This suggested that the disruption of the interactions between PCP and E domains was more detrimental than the disruption of the interactions between PCP and A domains.

More recently, Calcott and Ackerley (2015) studied the effect of NRPS context on PCP substitutions. They replaced the PCP domain from the first module of the last subunit of the pyoverdine synthetase PvdD of *P. aeruginosa* by 18 other PCP domains from various pyoverdine synthetases, but originally associated with downstream C domains, in *cis* (within the same subunit) or in *trans* (in different subunits), E domains or TE domains (Table 3). The six PCP domains originally associated with C domain in *cis* all allowed the production of pyoverdine at wild-type levels (NRPS context conserved). On the contrary, when PCP domains with different NRPS contexts (*Ctrans*, E, TE domains) were used, the titers of pyoverdine achieved were highly variable, from no production (three PCP domains, associated with either *Ctrans*, E or TE domains) to impaired production rates (two PCP domains associated to *Ctrans* domains, two associated to TE domains) to close to wild-type production levels (three PCP domains associated with E domains, one with a *Ctrans* domain and one with TE domain). The same type of observation was made by Owen and collaborators (2016): a *cis*-associated PCP domain could not replace TE-associated PCP domains. This suggested that it was probably important, when exchanging PCP domains, to

respect the PCP type, i.e. the nature of the domain (*Ctrans*, E or TE domains for example) found downstream of the PCP domain in the native NRPS.

### 3.2.2. Didomain exchanges

#### - C-A replacements

Because of the substrate specificity exhibited by C domains at their acceptor site, and also based on crystallographic structures that suggested that the C-A domains constituted a rigid platform, several teams have tested the swapping of cognate C-A pairs in combinatorial biosynthesis experiments. In their series of experiments on cyclic lipopeptides, the team of Richard Baltz at Cubist successfully replaced the C-A(activating kynurenine) didomain of the last module of the daptomycin synthetase by the C-A(activating asparagine) didomain of module 11 of the A54145 synthetase (Doekel et al., 2008). The expected cyclic lipopeptide was obtained with 43% yield compared to daptomycin production.

In a similar but more extensive experiment, Calcott and colleagues (2014) replaced the C-A(activating threonine) didomain of the last module of the *P. aeruginosa* pyoverdine synthetase by nine C-A(activating serine, threonine, lysine, aspartate and glycine) didomains of various modules of different pyoverdine synthetases (Table 3). Only three strains produced the expected product (pyoverdine or pyoverdine analog) with a good yield (close to wild-type levels for one C-A(Thr) and a C-A(Lys) exchanges, and 50% of the wild-type level for one C-A(Ser) exchange). All other constructs, including two C-A(Thr) and two C-A(Ser) exchanges, resulted in the production of truncated products. For one of the C-A(Thr) replacement that failed to yield pyoverdine, this result could have been anticipated as the C domain is of the <sup>D</sup>C<sub>L</sub> type, i.e. with a growing peptide chain ending with a D-amino acid at the donor site. In some of the other replacements that failed, the C-A didomain used was located at the N-terminal extremity of an NRPS subunit. Thus, the N-terminal extremity of the C domains may have included some kind of docking domains that may have impaired interactions with the upstream PCP domain.

From these experiments, it appears that swapping of C-A didomains may be possible in combinatorial biosynthesis experiments, if attention is paid to certain important points, including the nature and the NRPS context of the C domains. It should be underlined nonetheless that the experiments reported in these two studies were carried out with closely related domains and in terminal modules, which does not allow to evaluate the potential difficulties linked to possible donor site substrate specificities of the C domains.

#### - A-PCP

Very few A-PCP exchanges have been carried out, and these were achieved mainly before the C domain substrate specificities were known. As early as 1995, the team of Marahiel reported the production in *Bacillus subtilis* of four variants of surfactin obtained by replacing the A(Leu)-PCP didomain of the last module of the surfactin NRPS by A-PCP didomains of bacterial or fungal origin, with Phe, Orn, Cys and Val A domain substrate specificities (Stachelhaus et al., 1995). The titers of the surfactin analogs, especially with regards to the natural metabolite surfactin, were not

reported, but in a recent review, Brown and colleagues (2018) mentioned that these titers were lower than 1% of the initial surfactin titers. A few years later, the same team replaced the A(Leu)-PCP didomain of the second module of the first surfactin synthetase subunit (SfrA-A) with seven A-PCP domains from gramicidin (A domains with Phe, Leu, Orn, Val substrate specificities) and from the ACV (A domains with Cys and Val substrate specificities) (Schneider et al., 1998). *In vitro* analysis of the substrate specificities of the SfrA-A mutants were as expected, demonstrating the functionality of the imported A domains. The supernatant of only one mutant strain was analyzed (replacement with an A(Orn)-PCP didomain). Only truncated products were observed, yet with an ornithine incorporated at the second position of the peptide. At the light of our current knowledge of NRPS mechanisms, this suggests once more the existence of other domains of the NRPSs, most likely C domains, exhibiting a quite strict specificity for the growing peptide chain.

### 3.2.3. Modules or module-like exchanges

#### - Modules (C-A-PCP)

Because modules constitute the NRPS units responsible for the incorporation of one amino acid, exchanges of NRPS modules are very tempting and indeed, they have been attempted by several teams. One of the first experiments was carried out by the team of Mohamed Marahiel (Mootz et al., 2000). In this experiment, the TycA (A(Phe)-PCP-E) subunit as well as the first module (C-A(Pro)-PCP) of the TycB subunit of the tyrocidine synthetase were used. The C-A(Pro)-PCP module was fused with the 10<sup>th</sup> and last module (C-A(Leu)-PCP-TE) or with the 9<sup>th</sup> module (C-A(Orn)-PCP) fused with the TE domain of the synthetase. The proteins were expressed and purified and the system was tested for the production of a tripeptide. In both cases, the expected tripeptide was observed.

Following this first *in vitro* experiment, *in vivo* replacements of modules have been achieved. The team of Richard Baltz, for example, carried out nine module exchanges in the daptomycin synthetase (Doekel et al., 2008; Nguyen et al., 2006). Notably, they replaced the last module (module 13, C-A(Kyn)-PCP-TE) of the synthetase by the last module of the A54145 and of the CDA synthetase (Table 4 and Figure 25). These replacements respected the two “rules” established so far: the respect of the nature of the C and PCP domains. The mutant strains produced the expected daptomycin analogs with good yields (76% and 119% of the daptomycin titer). These experiments suggested that the three TE domain of the daptomycin, A54145 and CDA synthetases have a relaxed substrate specificity.

The team also exchanged only the three C-A(Kyn)-PCP domains of module 13. They replaced it with the C-A(Asn)-PCP domains of the module 11 of the A45145 synthetase. No production of daptomycin analog was observed, which may be explained by the exchange of the PCP type: the PCP of the module 11, usually interacting with a C domain, possibly could not interact correctly with the TE domain of module 13. Other experiments respecting the PCP type yielded daptomycin analogs with yields varying between 3 and 50 % of daptomycin titers. No obvious explanation can be offered for the decreased yield of these module exchange experiments. It may suggest, nonetheless, that C domains exhibit more substrate specificity at the donor site than usually thought. Another hypothesis, suggested from Farag and collaborators (2019), is that

the yield is further reduced due to intermodular linker incompatibility, when the number of incompatible intermodular linkers increases, or when the species providing the linkers are different.

Table 4: Examples of daptomycin combinatorial biosynthesis outcome

Replaced element from Dpt BGC	Replacing element	Type of modification	Resulting amino acid change	Yield (%)	Reference
M13 C-A	C-A from M11 of LptD	C-A exchange	Asn11 for Kyn13	43	(Doekel et al., 2008)
M13 C-A-PCP	C-A-PCP from M11 of LptD	C-A-PCP exchange	Asn11 for Kyn13	0	(Doekel et al., 2008)
M13 C-A-PCP-TE	M13 of LptD	C-A-PCP-TE exchange	Ile13 for Kyn13	76	(Doekel et al., 2008)
M13 C-A-PCP-TE	Last module of cdaPS3	C-A-PCP-TE exchange	Trp13 for Kyn13	119	(Doekel et al., 2008)
M8 C-A-PCP	M11 C-A-PCP of DptBC	C-A-PCP exchange	D-Ser11 for D-Ala8	18	(Nguyen et al., 2006)
M11 C-A-PCP	M8 C-A-PCP of DptBC	C-A-PCP exchange	D-Ala8 for D-Ser11	50	(Nguyen et al., 2006)
M8 C-A-PCP	M11 C-A-PCP of LptD	C-A-PCP exchange	D-Asn11 for D-Ala8	10	(Nguyen et al., 2006)
M11 C-A-PCP	M11 C-A-PCP of LptD	C-A-PCP exchange	D-Asn11 for D-Ser11	17	(Nguyen et al., 2006)
M8 C-A-PCP-E	M11 of LptD C-A-PCP-E	C-A-PCP-E exchange	D-Asn11 for D-Ala8	3	(Nguyen et al., 2006)
M11 C-A-PCP-E	M11 of LptD C-A-PCP-E	C-A-PCP-E exchange	D-Asn11 for D-Ser11	10	(Nguyen et al., 2006)
Modules 8-11	LptC	Multi module exchange	D-Lys8-OmAsp9-Gly10-D-Asn11 for D-Ala8-Asp9-Gly10-D-Ser11	<0.5	(Nguyen et al., 2006)
DptD	LptD	Subunit exchange	Ile13 for Kyn13	25	(Miao et al., 2006)
DptD	cdaPS3	Subunit exchange	Trp13 for Kyn13	50	(Miao et al., 2006)

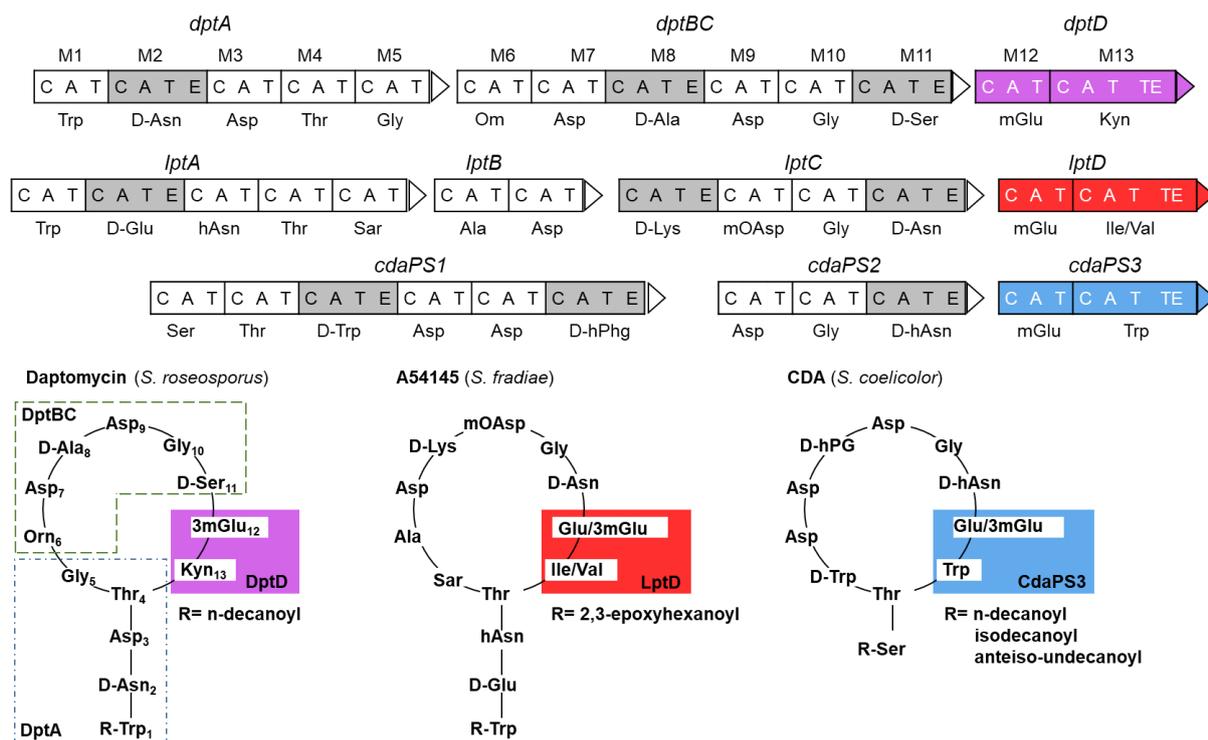


Figure 25: Structures of daptomycin, A54145 and CDA (Calcium-Dependent Antibiotic), and corresponding NRPSs

For reasons of space constraints, PCP domains are written as T (thiolation) domains in this figure.

Daptomycin, A54145 and CDA are closely related structures: they all contain a 10-membered ring and a lipid tail at the N-terminal end. The NRPSs are also similar, the monomers incorporated by the modules 4, 7, 10 and 12 (numbers based on daptomycin nomenclature) are identical among the three lipopeptides, and the modules 8 and 11 all contain an E domain.

- PCP-C-A exchange

Classically, NRPS modules are defined as C-A-PCP units. Yet, experiments described earlier in this manuscript (PCP exchanges, section 3.2.1) suggest that A-PCP interfaces are more permissive than PCP-C interfaces. For this reason, the team of Ackerley undertook to exchange a PCP-C-A(Thr) unit overlapping the two modules of the last subunit of the *P. aeruginosa* pyoverdine synthetase (PvdD) by PCP-C-A units originating from various pyoverdine synthetases (Calcott and Ackerley, 2015). The two exchanges that respected the C/PCP rules previously mentioned led to the production of pyoverdine analogs, with yields roughly of 30% and 55% of the natural pyoverdine (Table 3). No significant differences in analog titers were observed when PCP-C-A versus C-A exchanges were compared.

- A-PCP-C exchange (XU)

Going against the generally admitted rule that C-A domains form a rigid catalytic platform and should not be separated, the team of Helge Bode decided to use the C-A linker as a fusion point (Bozhüyük et al., 2018). Analyzing C-A linker sequences and available structures, they observed that C-A linker sequences are more conserved than the sequences of other shorter linkers, and that the N-terminal part of this linker is structured and mainly associated with the C domain

whereas the C-terminal part form no secondary structure and mostly interact with the A domain (Figure 26). Thus, they targeted the four residues located at the beginning of the C-terminal part of the linker and in a conformationally flexible loop as fusion points to construct ambactin hybrid NRPS.

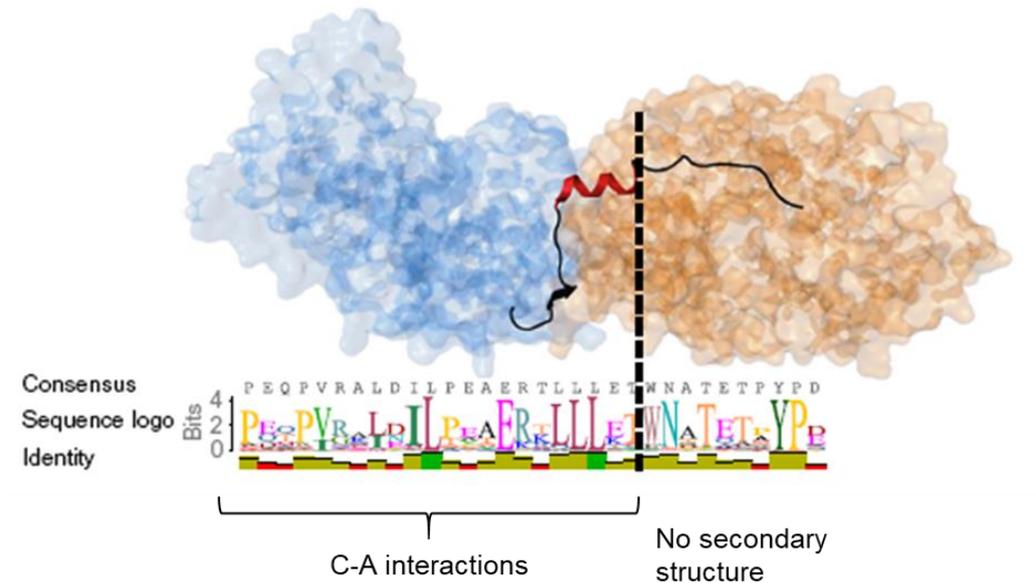


Figure 26: Identification of the fusion point used for swapping A-PCP-C tridomains (Bozhüyük et al., 2018)

C-A didomain excised from the SrfA-C crystal structure (Protein Database ID: 2VSQ) with the C-A linker depicted in a ribbon representation (top). C domain, blue; A domain, orange. C-A linker sequence logo of linkers excised from *Photobabidus* and *Xenorhabdus* NRPSs (bottom). Dashed line shows the used fusion point of the C-A hybrid linker.

They defined Exchange Units (XUs) as A-PCP-C or A-PCP-C/E domains. Using this approach, they were first able to successfully replace one or two XU units from the ambactin synthetase by one or two “homologous” (same NRPS context, and substrate specificity for the A domain) XU units from the GameXPeptide synthetase (Figure 27A). Replacements failed, however, when the acceptor site substrate specificity of the C domain of the XU was not respected.

Using XUs from three various *Photobabidus* and *Xenorabidus* NRPSs, they next constructed a chimeric NRPS producing the same xenotetrapeptide as the natural NRPS with reasonable yield (about 50% of the xenotetrapeptide production by the natural NRPS) (Figure 27B). They applied the same principle for the construction of a chimeric GameXPeptide synthetase (XU from up to four different NRPSs) (Figure 27C). However, production titers sharply decreased with increasing numbers of heterologous XU.

Although interesting as clearly showing that C-A linkers can constitute points of fusion for domain exchanges, these types of exchange constrain the choice of the following unit (to respect the substrate specificity of the acceptor site of the C domain), and thus necessitate a type of domino approach, as mentioned by Brown and colleagues (2018) in their recent review.

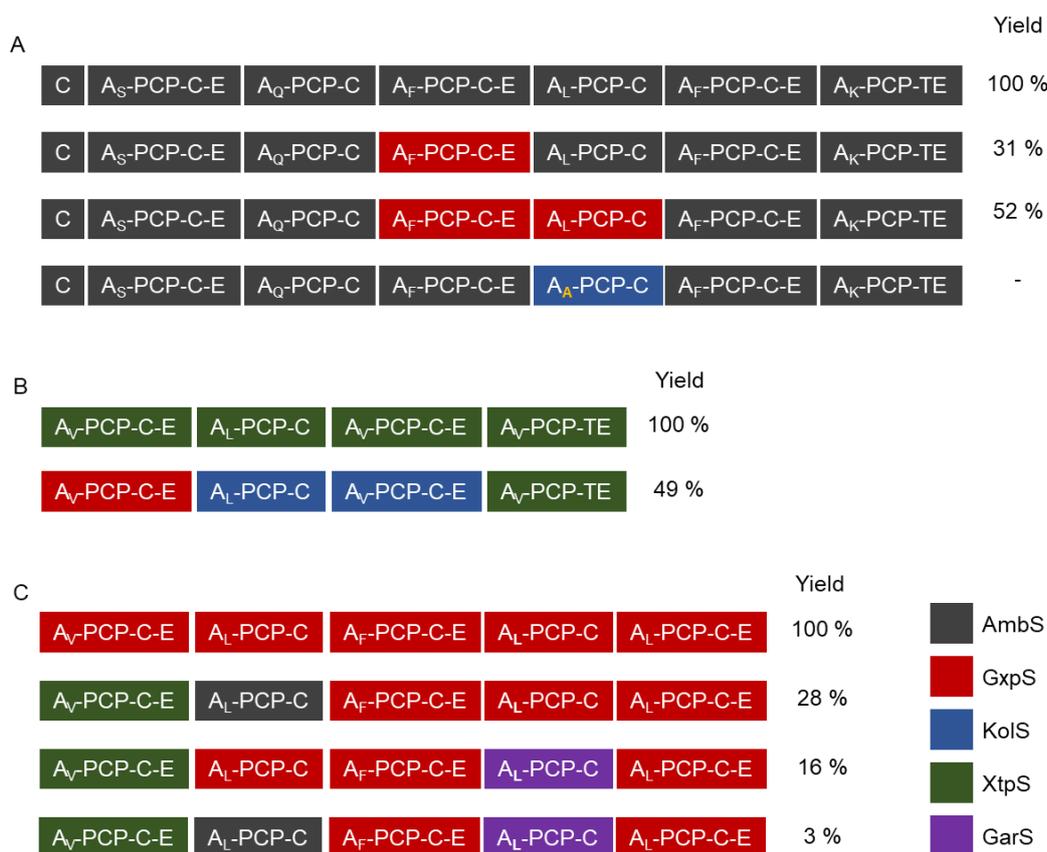


Figure 27: A-PCP-C (XU) exchange experiments

A. Exchanges of one or two XU from the ambactin NRPS

B. Construction of a xenotetrapeptide hybrid NRPS

C. Construction of a GameXPeptide hybrid NRPS

The spaces separate the different XU, and the color informs on the origin of the XU (Ambactin NRPS (AmbS), GameXPeptide NRPS (GxpS), Kolossin NRPS (KoLS), Xenotetrapeptide NRPS (XtpS) or gargantuanin NRPS (GarS).

### 3.2.4. Intradomain fusions

Although the vast majority of NRPS engineering achieved so far involved cutting and pasting complete domain(s) or module(s), a few groups reported the utilization of fusion points located within various domains. The first of that type of experiments was carried out by the group of Frank Bernhard on the surfactin synthetase (Symmank et al., 1999). They fused various domains or modules of the surfactin synthetase together using intradomain fusions. The chosen fusion points were in the A domain (between A<sub>core</sub> and A<sub>sub</sub>), the PCP domain (within the conserved sequence containing the serine residue to which the ppant arm is attached) and the C domain (several site tested, including the conserved sequence containing the catalytic histidine). Only the adenylation capacity of the resulting hybrid enzymes were tested *in vitro*. Hybrids with fusions carried out within the A domain retained adenylation activity with the substrate specificity of the N-terminal (A<sub>core</sub>) part of the enzyme. For fusions done within the PCP domains, the authors showed that the amino acid substrate was correctly loaded on the hybrid PCP domain. Intra C-domain fusions resulted in inactive enzymes, except when the fusion was carried out within the conserved sequence containing the catalytic histidine. In that case, the authors showed that the substrate was correctly loaded on the PCP domain.

Following this first *in vitro* experiment, Yakimov and colleagues (2000) carried out the same type of intra C-domain fusion, this time in *in vivo* experiments. In particular, they replaced the first module (incorporating Glu) of the surfactin synthetase by the equivalent module (incorporating Gln) of the lichenysin synthetase using the conserved sequence containing the catalytic histidine of the C domains as fusion points. The resulting mutant strain produced the surfactin analog with the same titer as the wild type strain.

Very recently, the team of Helge Bode carried out some very similar experiments, with the idea of controlling the acceptor site specificity of C domains (Bozhüyük et al., 2019). The fusion point was chosen this time within the four amino acids of the loop separating the two subdomains of C domains (Figure 15). The concept was named Exchange Unit Condensation Domain (XUC), the units to exchange are composed of C (subpart acceptor)-A-PCP-C(subpart donor). Using 5 XUC units coming from 4 NRPSs, the authors managed to produce GameXPepptide compounds to up to 66% of the yield of the native GxpS NRPS. The combination of XU and XUC units also yielded a functional NRPS, showing that both strategies are compatible. Exchanging XUCs from closely related genera seems to be a requirement as well, stricter than for XU exchanges. Using XUC concept and the TAR cloning method, Bode and colleagues generated a peptide library by randomizing different residues of GxpS (Bozhüyük et al., 2019). This new concept of XUC units, possibly associated to the XU units, could prove very valuable for future exchange experiments, and lead to the production of numerous novel compounds.

### 3.2.5. Subunit exchanges

Subunit exchanges have rarely been reported, except for lipopeptide NRPSs. One of the reported cases consists in the exchange of the last subunit of daptomycin NRPS, DptD, with LptD or cdaPS3 (Miao et al., 2006). The three subunits contain two modules, with the first incorporated amino acid being mGlu in all cases, and the second amino acid being variable (Kyn for DptD, Ile/Val for LptD, Trp for cdaPS3) (Figure 25). The daptomycin derivatives generated by the subunit swapping are therefore identical to the derivatives obtained by swapping of the module 13 (described in the C-A-PCP swapping section). However, the disrupted interface differs: while it was between the module M12 and M13 previously, the disrupted interface corresponds now to the docking domains between DptBC and DptD. The mutant strains produced the expected analogs, but with a decreased yield (25% and 50% of the daptomycin titer) compared to the experiment of module M13 exchange (76% and 119%) (Table 4). This reduced production may be explained by impaired communication between the subunits. Baltz and collaborators indeed identified COM domains at the extremities of the subunits, but they did not attempt to engineer these docking domains (Miao et al., 2006).

Several other studies on lipopeptides, mentioned in the section 2.4.3., actually report that COM domain swapping experiments led to successful interactions between non-cognate subunits. For instance, using the fusion sites indicated on Figure 22, a tripeptide (L-Phe-D-Orn-L-Leu) was produced *in vitro* resulting from successful interactions between three NRPSs derived from different pathways (tyrocidine, bacitracin and surfactin pathways) (Hahn and Stachelhaus, 2006). *In vivo*, Chiocchini and coworkers (2006) reprogrammed the COM domains to establish a productive interaction between the subunits of surfactin NRPS, SrfA-A and SrfA-C, generating a

shortened lipotetrapeptide while keeping titers similar to the WT production (70% of the surfactin titer). Liu and coworkers (2016) similarly re-ordered in *B. subtilis* the five NRPS subunits of plipastatin through COM domain modifications, resulting in five new products of different lengths.

### 3.3. Modification of the length of NRPS

#### 3.3.1. Modules and domains insertion / deletions

Other than NRPS exchanges, deletion or insertion of domains / modules may yield new derivatives. In those cases, to maintain functional enzyme interactions and respect the specificity of the downstream domains is again a challenge, and the TE domain has an important role. Several experiments on SrfA NRPS indicated for instance that the thioesterase was specific of a certain ring size.

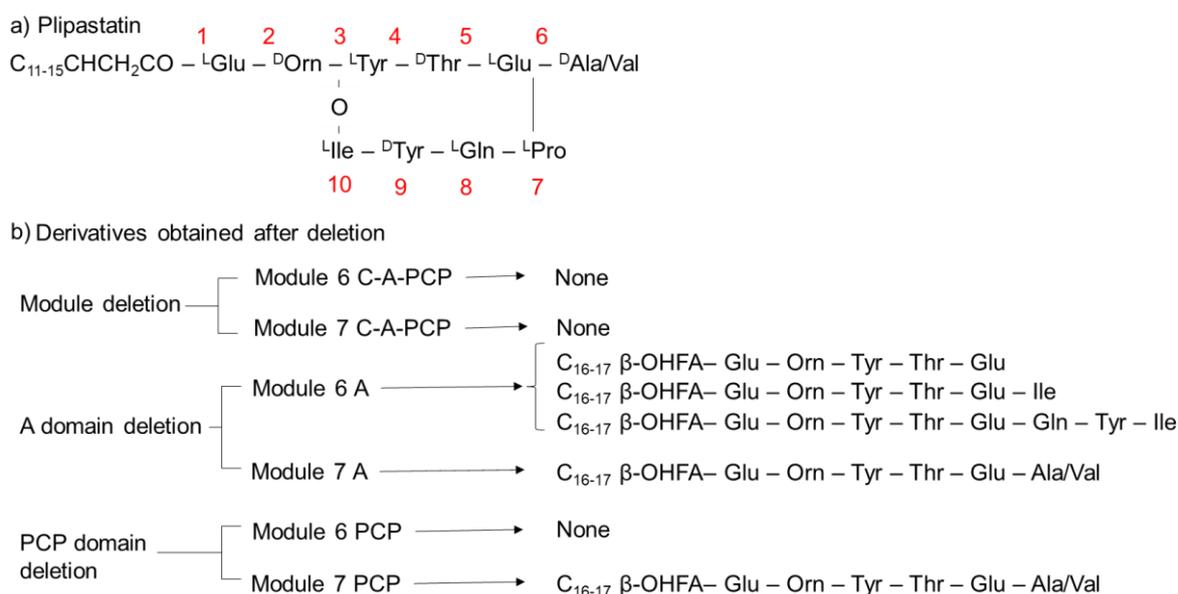


Figure 28: Module or domain deletions of plipastatin

Module and domain deletions were attempted to obtain new plipastatin derivatives (Gao et al., 2018). Plipastatin is an 8-membered ring molecule (Figure 28), synthesized by 5 NRPSs. As module 6 or 7 deletions were unsuccessful, even with retained linkers, experiments were pursued with domain deletions. The results obtained were puzzling. While deletion of C6 (C domain from the module 6) or PCP6 was followed by an absence of production, deletion of A6 gave three novel derivatives of plipastatin. One of them is a pentapeptide, a truncated product made by the first 5 modules. The two others are a hexapeptide and an octapeptide, and they derive respectively from the skipping of the module 6 and 7, or the module 6, 7, 8 and 9. Though skipping of the module 6 only was expected, skipping of two or four modules was observed. On the contrary, deletion of PCP7 or A7 had as a consequence the production of a truncated product, a linear hexapeptide. These results obtained recently confirm, if ever a confirmation was needed, that we still do not clearly understand the way the NRPSs interact.

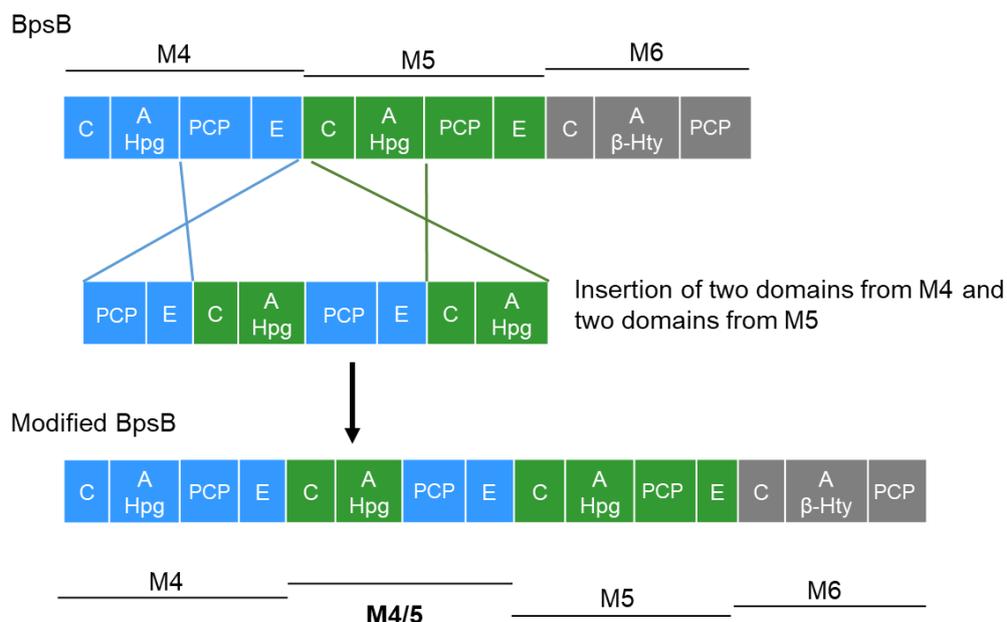


Figure 29: Module insertion in balhimycin NRPS

Hpg: hydroxyphenylglycine; β-Hty: β-hydroxytyrosine

The first and only experiment reporting a module insertion was done on balhimycin from *Amycolatopsis balhimycina* (Butz et al., 2008). Balhimycin is constructed from 3 NRPS subunits BspA, B and C, made of 7 modules. The modules 4 and 5 both allow the incorporation of a D-hydroxyphenylglycine (D-Hpg), and it was decided to introduce a hybrid module between modules 4 and 5, incorporating an extra D-Hpg (Figure 29). This hybrid module is constituted from the domains C5 and A5, and the domains PCP4 and E4, hence the only non-natural transition is between A5 and PCP4. The authors detected the expected cyclic octapeptide, but it was a minor compound, corresponding the 1/5 in yield compared to a linear heptapeptide (which contained the three D-Hpg, but not the first monomer). Other truncated products were observed as well, implying some specificity issues downstream the assembly line. Though the experiment was carefully planned to avoid new enzyme interfaces, and to be as little disruptive as possible concerning the specificities of substrate by inserting a monomer that was already present twice, unexpected compounds were observed. In general, outcomes of insertion or deletion of elements remain quite difficult to predict.

### 3.3.2. Variation of the length of NRP generated by iterative NRPS

The rhabdopeptides and xenortide peptides (RXPs) are produced by *Xenorhabdus* and *Photorhabdus* species, symbionts of an entomopathogenic nematode, and they constitute the largest class of NRP to date (Cai et al., 2017). Indeed, RXPs biosynthetic gene clusters, constituted of 2 to 3 mono-modular NRPSs, can generate diverse RXPs of two to eight amino acids. This diversity can be explained by the iterative and flexible use of the stand-alone modules, combined with a relaxed selectivity of the domains.

The terminal module of RXP NRPSs often consists in a stand-alone C domain, involved in the release of the peptide *via* attack of a free amine. Cai et al. (2017) showed that the

stoichiometry between the elongation module and the C terminal domain controls the length of the RXPs: longer chains are favored in excess of the elongation module, and only shorter chains are generated when the elongation module and the C terminal domain are in equivalent ratio.

Hacker et al. (2018) considered that, if the ratio of the modules impacted the length of the RXPs produced, then another way to influence the length of the RXPs was to modify the affinity between modules (subunits here). They identified docking domains that mediate the selective interactions between RXP NRPSs, and differ from the classical COM domains observed in lipopeptide NRPSs. Modifications of these docking domains resulted in altered interaction affinities and allowed to increase the length of the compounds obtained (Hacker et al., 2018). Conversely, replacement of the RXPs NRPS docking domains by “classical” or collinear NRPS docking domains generated specified peptides with defined length, but at the price of a decreased yield, suggesting that more complex internal domain-domain interactions exist (Cai et al., 2019).

Altogether, this work emphasizes the importance of the docking domains in iterative NRPSs. The authors report that several other orthogonal docking domain systems most likely exist (Hacker et al., 2018). Their structural and chemical study would be of high interest, as it would enable their future use in NRPS engineering or understanding the basic principles of these megasynthase pathways.

### **3.4. Choice of fusion sites for combinatorial biosynthesis experiments**

Except for the C-A linker, most inter domain linkers are flexible, and as such, they were often selected as fusion sites for NRPS exchanges, deletions or insertions. However, very few studies report analyses of the linker modification themselves. Baltz and collaborators are among the rare groups to have spent significant effort on the modification of a linker (Nguyen et al., 2006). During their study of the daptomycin NRPS DptD, they showed that the PCP-C linker could tolerate amino acid substitutions at three different positions, as well as addition or subtraction of up to four amino acids. Their successful exchanges of C-A didomains suggest that the A-PCP linker is also flexible enough to be used as a fusion point.

However, despite their flexibility, linkers can be involved in transient protein interactions and as such have important roles during the NRP biosynthesis. For instance, in the case of the yersiniabactin NRPS, the linkers upstream and downstream of the PCP domain were shown to stabilize the correct folding of the domain (Harden and Frueh, 2017). Gullick and collaborators also reported that the LPxP motif in the A-PCP domain maintains the correct folding of the A domain catalytic site and couples the movement of the PCP to the A domain (Miller et al., 2014). Indeed, when Di Ventura and collaborators exchanged the PCP of IndC with that of BpsA, maintaining the BpsA A-PCP linker together with the incoming PCP domain was necessary to obtain a functional indigoidine synthetase, confirming the importance of the A-PCP linker (Beer et al., 2014).

A consensus concerning the fusion points to use has yet to emerge. An alternative to splicing in poorly conserved regions is to cut in contrary at highly conserved sites. For now, two

studies reported indeed the successful use of a conserved region in the C domain as a fusion site (Bozhüyük et al., 2019; Yakimov et al., 2000).

### 3.5. Directed evolution to restore functionality of the chimeric NRPS

An ever-present issue observed for the chimeric enzyme obtained after NRPS engineering is the decrease of the biological activity or of the NRP production yield. Rounds of directed evolution may restore the NRPS functionality, based on a selective pressure or a screening method such as growth, inhibition screening, fluorimetric screening or mass spectrometry (MS) screening. For instance, Fischbach and collaborators replaced the A(Ser) domain from EntF of enterobactin by a A(Ser) domain from syringomycin, SyrE, and observed a 30-fold loss of activity, due to poor solubility (Fischbach et al., 2007). From a library of  $2 \cdot 10^4$  clones, they obtained a clone with a production yield similar to the one of the WT using growth as a screening assay.

The same team also constructed a hybrid of the NRPS AdmK from the hybrid polyketide-NRP andrimid (Fischbach et al., 2007). They replaced the AdmK-A(Val) by an A domain selecting 2-aminobutyrate, and observed a 32-fold reduced production compared to native andrimid production. They equally replaced AdmK-A(Val) by BacA-A1(Ile) and observed this time a 7-fold reduction. In both cases, a small library of  $10^4$  clones and 3 rounds of selection based on inhibition screening allowed to obtain clones with productivity similar to the one of the WT. Remarkably, in all cases, the mutations were distributed along the A domain, and hardly predictable. It is worth noting that there are no C domains in andrimid biosynthesis, the condensation is effected by transglutaminase-like enzymes, hence there was no substrate specificity question including the C domains (Calcott and Ackerley, 2014).

Directed evolution was also used to replace EntB, an Aryl Carrier Protein (ArylCP) domain from enterobactin biosynthesis, by the ArylCP VibB from vibriobactin or HMWP2 from yersiniabactin (Zhou et al., 2007). As enterobactin is a siderophore, selection could be easily done by growth measurements in an iron-depleted medium. Four convergent mutations were observed, with at least three of them involved in interactions with different domains (one with the PPase, one with A domain and downstream C domain, and one with A domain).

Directed evolution can also be done on colored compounds, which allow an easy screening for production. For instance, Owen and collaborators (2016) attempted to replace the PCP domain of the NRPS BpsA, single module responsible for indigoidine production, a violet compound (Figure 30). The PCP domain from the first module of PvdD (PCP1), usually associated to a *Cis* domain, could not replace either PCP domain from the second module of PvdD (PCP2), nor BpsA PCP, usually associated to a TE domain. However, after mutagenesis of PCP1 in the inactive BpsA hybrid system, the evolved PCP1, now functional in BpsA, could also replace successfully PCP2 in PvdD. One to three mutations were sufficient to allow PCP1 to interact correctly not only with BpsA TE, but with TE domains in general. The authors conclude that while PCP and TE domains should be kept together whenever possible, one positive selection round might be enough to change the outcome of the experiment (Owen et al., 2016). They suspect that more often than not, functional interactions may be impeded just by a few residue positions.

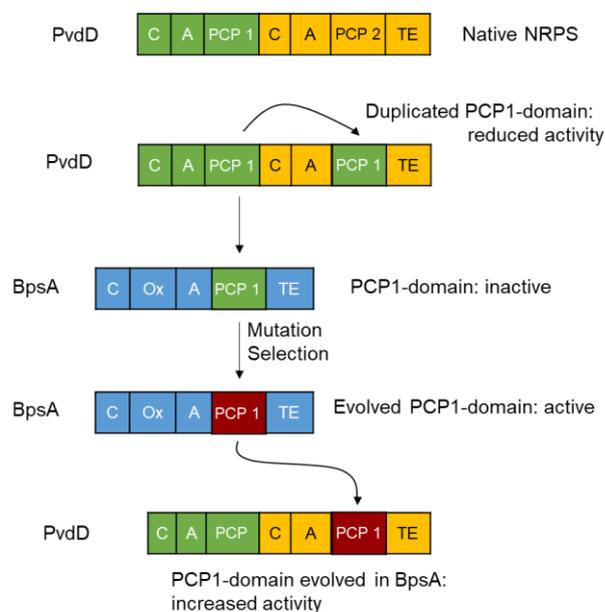


Figure 30: Evolution of a PCP domain and modification of its role

Altogether, in cases where the productivity of the mutant is very low, directed evolution may allow to restore the functionality of the chimeric NRPS. It has not been done much in practice, even if numerous altered NRPSs were constructed to obtain new derivatives, partly because of the need of a selection pressure.

### 3.6. Conclusions about points to keep in mind when modifying the NRPSs

Modifying the number or the nature of the monomers incorporated by the NRPSs could lead to the development of molecules with therapeutic applications, but is impeded by our limited understanding of the NRPS biosynthetic processes.

In all the experiments performed until now, one common point for combinatorial experiments is the use of parts of NRPSs not only from phylogenetically close organisms (avoiding genera crossing), but also from NRPSs synthesizing structurally related metabolites. This is increasing the chances of a successful outcome (Brown et al., 2018). In other respects, the consensus is far from being reached, and many different approaches were followed.

All in all, two main strategies were employed to modify the NRPS core structure. The first one is to target the A domains, which are responsible for the main substrate specificity. In some rare cases, A domains have been reported to be rather promiscuous, which may allow generation of unnatural products *in vitro* (Zhu et al., 2019). Otherwise, A domains can be modified, notably by site-directed mutagenesis or subdomain swapping, which keep a majority of the assembly line intact and minimize the interface disruptions, or by A domain swapping. However, this approach is often limited by the substrate specificity of the C domains, particularly at the acceptor site of the upstream C domain. These modifications should therefore be favored in cases of C domains with relaxed acceptor site substrate specificity (Thirlway et al., 2012). Apart from these specific cases, they have a limited potential.

The second strategy involves engineering of multiple catalytic domains. Among the different multi-domain swapping approaches, C-A domain and C-A-PCP module swapping have been the most frequently used (Calcott et al., 2014; Doekel et al., 2008; Mootz et al., 2000; Nguyen et al., 2006). They were first selected because they maintain the C-A interface, which was thought to be rigid, but their success more likely resides in the respect of the substrate specificity of the upstream C domain acceptor site. C-A and C-A-PCP swapping were also preferred to subunit swapping, possibly because they avoid the disruption of docking domains, which are not always well identified. One constrain for such exchanges, identified by the team of Richard Baltz (2018), is to maintain the C domain type, which means that the substitute C domain should catalyze the same kind of reaction, whether linking fatty acid, D-amino acid or L-amino acid to L-amino acids. The variation of the observed outcomes in terms of production might be explained by some substrate specificity at the downstream C domain donor site, due to steric or other constraints, but has not been quite pinpointed yet. Similarly, constraints coming from the TE domains are yet to be finely deciphered, as shown by the experiments involving deletions or insertions (Butz et al., 2008; Gao et al., 2018).

While using the C-A linker as fusion point has generally been avoided, Bode and collaborators showed that the precise point of fusion was essential (Bozhüyük et al., 2018). Indeed, targeting a flexible region in the C-A linkers that accepts recombination, they managed to perform successful A-PCP-C exchanges, though limited by the strict substrate specificity of the C domain acceptor site. In order to avoid this issue, they then proceeded to exchanges by splicing C domains within a conserved region located between the two lobes constituting these domains (Bozhüyük et al., 2019). This example is particularly remarkable, as it potentially allows to respect both the substrate specificities of the upstream C domain acceptor site and the downstream C domain donor site. Moreover, it shows conserved intra domain regions may be alternative fusion sites to the linkers.

To fill the gaps in our understanding of the NRPSs, we have to perform more experiments analyzing the substrate specificities and the protein/protein interactions of these systems. However, one of the main drawbacks in NRPS engineering is technique: it is quite challenging to engineer the mega enzymes. Another problem results from NRPS complexity: it is nearly impossible to vary only one parameter, and the frequent failures can usually have several origins.

In order to gain theoretical knowledge on these enzymes, it might thus be interesting to work with a model NRPS system, such as the extensively studied pyoverdine dimodule PvdD (Table 3), which is easier to manipulate. Some atypical NRPSs made of stand-alone enzymes have been described (Binz et al., 2010; Süßmuth and Mainz, 2017), such as the NRPS of streptothricin, containing two stand-alone A domains and one PCP-C didomain. Another family of NRP is synthesized by atypical NRPSs: the pyrrolamides. Due to the features of its NRPS (stand-alone modules and domains) and the existence of several members of the family synthesized by homologous enzymes, it is quite adapted to combinatorial experiments to interrogate our modular enzymes and decipher the factors impeding production upon genetic engineering. The characteristics of the pyrrolamide family and their NRPSs will be detailed in the next section.

## 4. The pyrrolamides, a family of metabolites synthesized by NRPSs

### 4.1. The pyrrolamides, a family of minor groove binders

#### 4.1.1. Structure, biological activities and mode of action

Pyrrolamides are specialized metabolites characterized by the presence of one or several monomers of 4-aminopyrrole-2-carboxylic acid, their structure is presented on Figure 31. Interestingly, they are constituted of a few chemical moieties, which seem to have been combined in different manners. The production of members of the family has been reported in different *Streptomyces* species and other related actinobacteria, all Gram-positive soil bacteria with high GC DNA content.

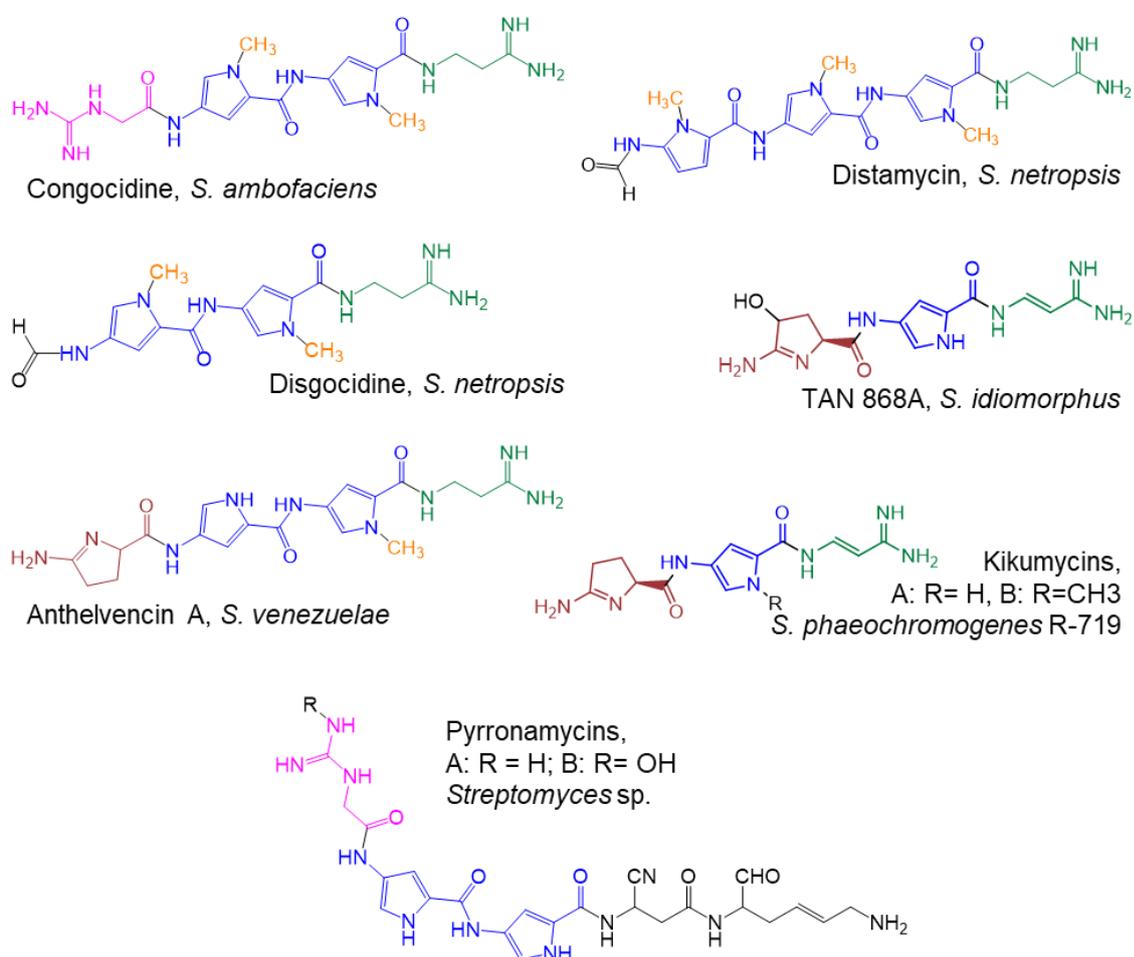


Figure 31: Chemical structures of the members of the pyrrolamide family and name of their *Streptomyces* producer  
4-amino-pyrrole-2-carboxylic acid groups are displayed in blue. Groups which are common to several molecules are colored specifically.

Table 5: Members of the pyrrolamide family, producer and biological activity reported

Pyrrolamides	<i>Streptomyces</i> producers	Biological activities	References
Congocidine (=Netropsin)	<i>S. netropsis</i> <i>S. ambofaciens</i>	Antibacterial, antiviral, antitumor, cytotoxic	(Cosar et al., 1952; Finlay et al., 1951; Julia and Preau-Joseph, 1967)
Distamycin	<i>S. netropsis</i>	Antibacterial, antiviral, antitumor, cytotoxic	(Arcamone et al., 1964; Casazza et al., 1965)
Disgocidine	<i>S. netropsis</i>	uncharacterized	(Vingadassalon et al., 2015)
Anthelvencins A and B	<i>S. venezuelae</i> ATCC14583	Antibacterial, anthelmintic, cytotoxic	(Probst et al., 1965)
Kikumycins A and B	<i>S. phaeochromogenes</i>	Antibacterial, antiviral	(Kikuchi et al., 1965; Takaishi et al., 1972)
Pyrronamycins	<i>S.</i> KY11678	Antibacteriophage, antitumor, cytotoxic	(Asai et al., 2000)
TAN 868 A	<i>S. idiomorphus</i>	Antibacterial, antiviral, cytotoxic	(Takizawa et al., 1987)

Biological activity has been reported for most pyrrolamides isolated so far (Table 5). For instance, distamycin has been reported as a potential antiviral agent against herpes simplex virus and some adenovirus (Casazza et al., 1965; Matteoli et al., 2008). It also exhibits mild antibacterial activity. Anthelvencin was also reported to control nematode infections in mice and swine and inhibit a broad spectrum of microorganisms *in vitro* (Probst et al., 1965). Congocidine, also called netropsin, was described as an antibacterial compound, and reported for its action on trypanosomal infection (notably by *Trypanosoma congolense*) in mice (Cosar et al., 1952). Despite these numerous activities, none of the pyrrolamides is used today in human or animal medicine. Indeed, mild to important toxicity was always reported in parallel to the biological activities of interest (Asai et al., 2000; Finlay et al., 1951; Matteoli et al., 2008; Probst et al., 1965; Takizawa et al., 1987).

The cytotoxicity of the pyrrolamides most likely results from their mode of action. Pyrrolamides bind to the minor groove of DNA (Figure 32A), and interfere with replication and transcription processes (Kopka et al., 1985). Congocidine and distamycin are the most studied members of this family. The two molecules stabilize the DNA helix, and they show an affinity for A-T-rich domains (Zimmer et al., 1971). The X-ray analysis of the complex congocidine-DNA 5'CGCGAATTCGCG shows that congocidine is centered on the AATT region of the minor groove (Goodsell et al., 1995). It binds to the 4 A-T base pairs by displacing water molecules. It makes hydrogen bonds between the NH of the amide and adenine N3 and thymine O2 atoms in adjacent position and opposite strands (Figure 32B). Distamycin has an extended binding site compared to congocidine, it covers 5 of the 6 A-T base pairs from the sequence 5'CGCAAATTTGCGC (Neidle, 2001). The affinity of congocidine and distamycin to A-T-base

pairs can be explained by space constraints. Indeed, pyrrole groups are packed against the C2 position of adenine, leaving no space for the amine group of guanine (Goodsell et al., 1995).

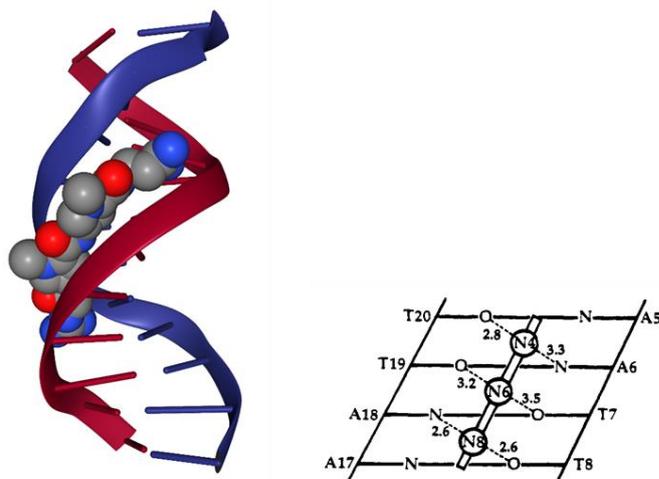


Figure 32: Representation of congoicidine binding to DNA (Kopka et al., 1985; Goodsell et al., 1995).

A. number 6BNA, 3D view. B. Schematic view of the structure, with hydrogen bonds represented by dashed lines.

#### 4.1.2. Synthetic derivatives of pyrrolamides

The unwanted cytotoxicity of pyrrolamides has hindered their use in human medicine, but many derivatives have been chemically synthesized to overcome this issue. Design of analogs led to a range of very effective antimicrobial compounds (Bolhuis and Aldrich-Wright, 2014), as well as anti-viral, antifungal and antiparasitic compounds (Rahman et al., 2019). One pyrrolamide analog with a stilbene-like fragment as a head group, MGB-BP-3 (Figure 33), was selected for treatment of *Clostridium difficile* infections, and is currently in phase II of clinical trials (Bhaduri et al., 2018). Derivatives of pyrrolamides with potent anti-cancer activity were also obtained (Barrett et al., 2013). Tallimustine (Figure 33) is a derivative of distamycin with an alkylating functional group, it is also A-T-rich sequence-specific and exhibits a broad anti-tumor activity. However, the clinical studies were stopped because of severe myelotoxicity (Bhaduri et al., 2018). Brostallicin (Figure 33) is another derivative with anti-cancer properties and an improved cytotoxicity/myelotoxicity ratio. It acts as an alkylator agent in presence of high levels of thiols (such as glutathione) and is currently in phase II of clinical studies for soft sarcoma (Rahman et al., 2019).

The specificity of binding sequence displayed by congoicidine and distamycin convinced some researchers that it was possible to use them to target specific DNA regions, with a potential application in gene expression extinction. To reach this objective, a requirement was the ability to target C/G base pairs as well. It was shown that replacing pyrrole groups by imidazoles allows the recognition of G-C base pairs (Figure 34) (Kopka et al., 1985; Bolhuis and Aldrich-Wright, 2014). Indeed, the extra nitrogen in imidazole groups can form a hydrogen bond with the amine group of guanine. Four ring pairings (Imidazole/Pyrrole, Pyrrole/Imidazole, Hydroxypyrrole/Pyrrole and Pyrrole/Hydroxypyrrole) then make it possible to distinguish all four base pairs in the minor groove of DNA (Bhaduri et al., 2018). Analogs targeting transcription factor binding sequences

were developed (Bhaduri et al., 2018; Rahman et al., 2019). For instance, a compound targeting 5' GGGACT was shown to inhibit binding of the transcription factor NF- $\kappa$ B (which regulates genes involved in immune and inflammatory responses) (Bolhuis and Aldrich-Wright, 2014).

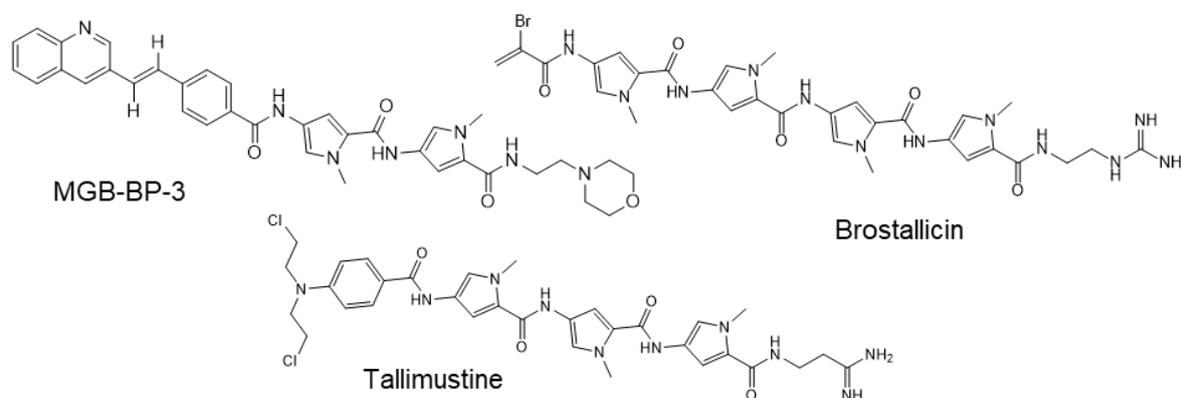


Figure 33: Structure of some pyrrolamide derivatives

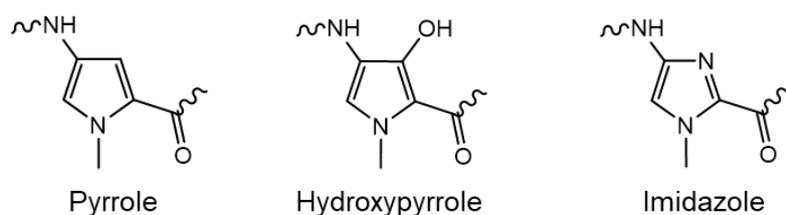


Figure 34: Modifications of the pyrrole group to target the four DNA base pairs

## 4.2. Congocidine biosynthesis

### 4.2.1. Congocidine biosynthetic gene cluster

While congocidine/DNA binding has been extensively studied since the molecule discovery in 1951, the biosynthetic gene cluster of congocidine remained unknown until 2009 when Juguet et al. reported its isolation and characterization from *Streptomyces ambofaciens* ATCC 23877 (Juguet et al., 2009). This article also consists in the first report of any pyrrolamide biosynthetic pathway.

The cluster of genes directing the biosynthesis of congocidine consists of 22 genes and spans about 30 kb (Figure 35). Functional analysis of the cluster indicated that one gene is related to the transcriptional regulation of the *cgc* genes, two gene are involved in congocidine resistance, 13 are responsible for precursor biosyntheses, and the remaining 6 genes encode enzymes that assemble the precursors or tailor the pyrrolamide backbone.

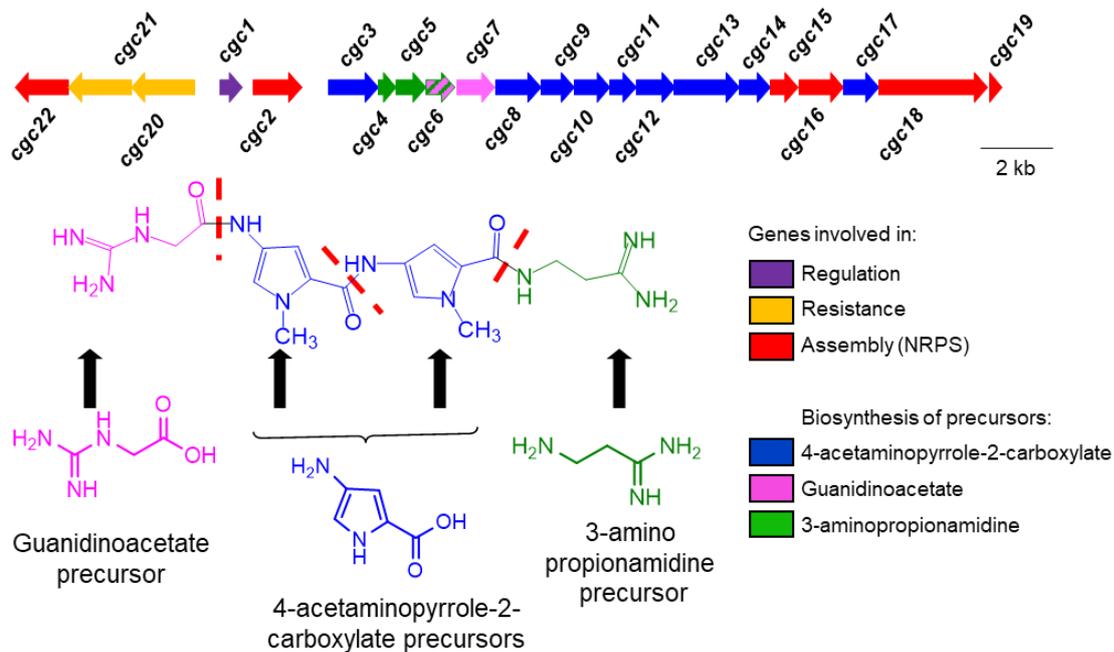


Figure 35: *S. ambifaciens* ATCC 23877 *cgc* biosynthetic gene cluster and congoicidine structure  
Red dashed lines separate the different monomers of congoicidine

#### 4.2.2. Biosynthesis of the precursors of congoicidine

Congoicidine is assembled from three precursors: guanidinoacetate, 4-acetamidopyrrole-2-carboxylate and 3-aminopropionamide (Figure 35). Their biosynthetic origins were established using genetics, biochemistry and analytic chemistry (Lautru et al., 2012; , Elie et al., unpublished).

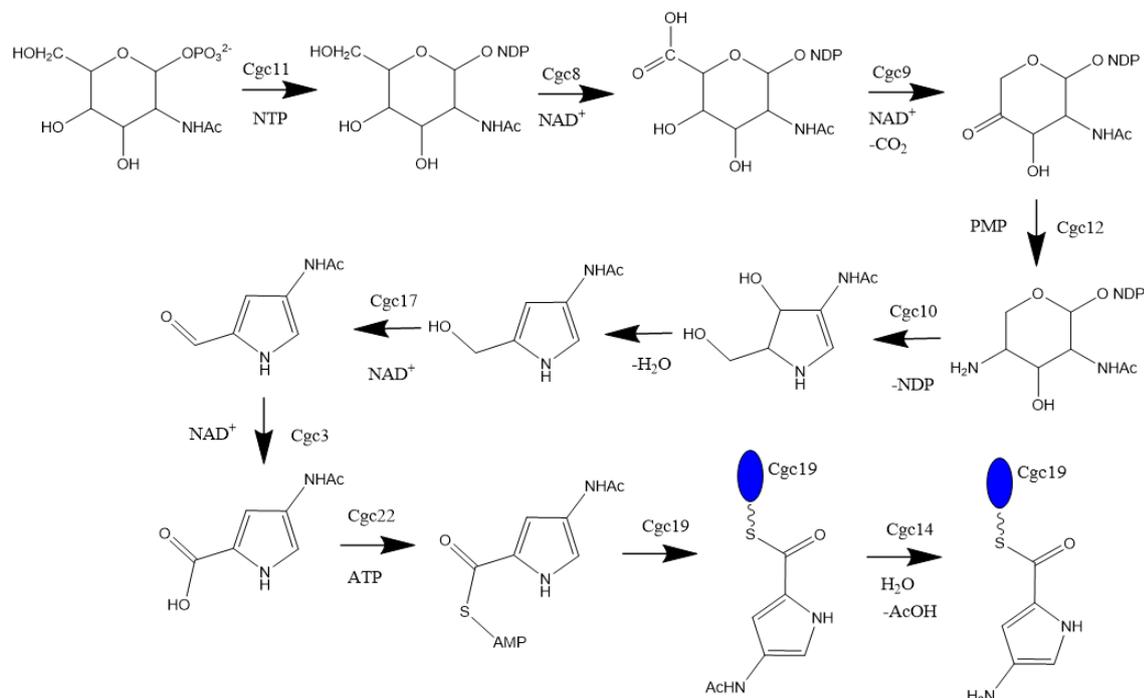


Figure 36: Biosynthetic pathway of the precursor, 4-acetamidopyrrole-2-carboxylate (Lautru et al., 2012)  
PMP, pyridoxamine phosphate

The 4-acetamidopyrrole-2-carboxylate precursor of congocidine is assembled from N-acetylglucosamine-1-phosphate (Lautru et al., 2012), and the biosynthetic pathway involves carbohydrate-metabolizing enzymes (Figure 36). This pathway differs from all pathways leading to the formation of pyrrole rings described so far (Walsh et al., 2006). Although no clear role could be attributed to Cgc13, deleting *cgc13* led to a decreased production of congocidine, while feeding 4-acetamidopyrrole-2-carboxylate to the mutant strain restored the production to its wild-type level (Lautru et al., 2012). It is thus hypothesized that Cgc13 is also involved somehow in 4-acetamidopyrrole-2-carboxylate synthesis, possibly providing N-acetylglucosamine-1-phosphate.

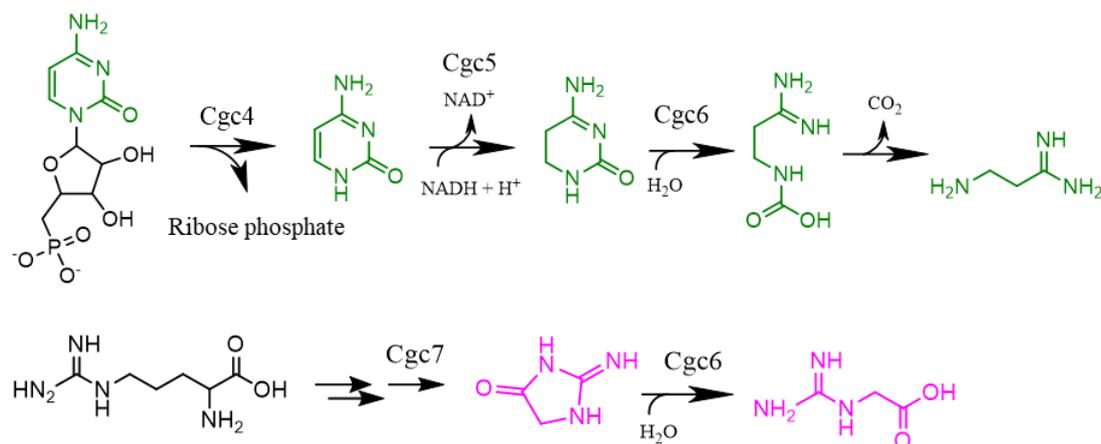


Figure 37: Biosynthetic pathways of the precursors 3-amidinopropionamidine and guanidinoacetate

3-amidinopropionamidine and its intermediary species are represented in green, guanidinoacetate and its precursor are represented in pink.

The guanidinoacetate precursor originates from L-arginine (Wildfeuer, 1964). Its biosynthesis is not fully understood but involves Cgc7 and Cgc6 (Figure 37) (Elie et al., unpublished). As for 3-aminopropionamidine, it originates from cytidine monophosphate and is synthesized by the Cgc4, Cgc5 and Cgc6 enzymes (Figure 37). Unexpectedly, Cgc6 is involved both in the biosynthesis of 3-aminopropionamidine and guanidinoacetate (Elie et al., unpublished).

#### 4.2.3. Assembly of congocidine by an atypical NRPS

Congocidine is assembled by an atypical NRPS made of one isolated and noncanonical module (Cgc18) and three stand-alone domains (two condensation domains - Cgc2 and Cgc16 – and one PCP domain – Cgc19) (Juguet et al., 2009). The PPTase responsible for the phosphopantetheinyl transfer of the PCP domain is a pleiotropic PPTase, involved in the activation of several acyl- and peptidyl-carrier protein domains, which is not located in the *cgc* cluster (Bunet et al., 2014).

A mechanism of congocidine assembly is proposed in Figure 38 (Al-Mestarihi et al., 2015; Juguet et al., 2009; Vingadassalon et al., 2015). Activation and adenylation of each of the two 4-acetamidopyrrole-2-carboxylate precursors is made not by an A domain, but by an Acyl-CoA synthetase Cgc22 (Figure 36). Acyl-CoA synthetases belong to the ANL superfamily (Acyl-CoA

synthetase, NRPS adenylation domain, and Luciferase), as the adenylation domains of NRPSs. It was suggested that Cgc22 activates 4-aminopyrrole-2-carboxylate by catalyzing ATP-dependent adenylation (Al-Mestarihi et al., 2015). Then the AMP-activated 4-aminopyrrole-2-carboxylate is loaded onto the Cgc19 PCP domain. It is thought that the pyrrole precursor is deacylated by Cgc14 once loaded on Cgc19, yielding tethered-4-aminopyrrole-2-carboxylate. Indeed, 4-aminopyrrole-2-carboxylate is never observed in culture supernatant of *cgc* deletion mutants (Lautru et al., 2012). As aromatic amines are usually toxic and as acetylation of the amine is often used as a protection mechanism, keeping the pyrrole precursor under the N-acetylated form while in solution could constitute a mechanism of protection for the cells.

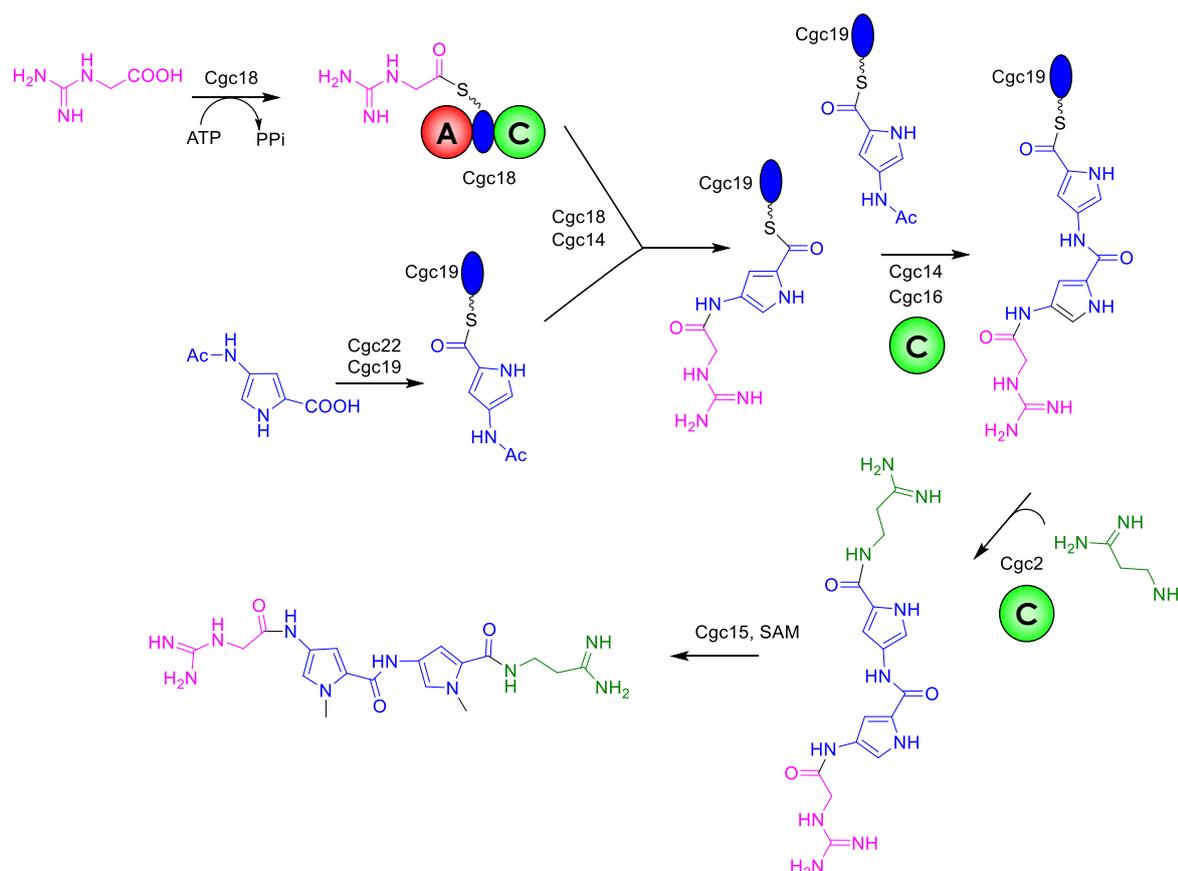


Figure 38: Proposed mechanism for the assembly of congocidine in *S. amboufaciens*

Guanidinoacetate is activated by the A domain of Cgc18, and loaded onto the PCP domain of Cgc18. Cgc18 A domain requires the presence of an MbtH-like protein encoded outside of the *cgc* gene cluster as a partner to be functional (Al-Mestarihi et al., 2015). The C domain of Cgc18 then catalyzes the condensation of the guanidinoacetate with the Cgc19-bound 4-aminopyrrole-2-carboxylate. The second 4-aminopyrrole-2-carboxylate is next condensed by the Cgc16 C domain. 3-aminopropionamide is finally added to the molecule by the Cgc2 C domain. This has for consequence the release of di-demethyl-congocidine (congocidine without any methyl group on the nitrogen of the pyrrole groups). The last step of the biosynthesis involves Cgc15, a SAM-dependant N-methyltransferase that catalyzes the methylation of the nitrogen of the pyrroles (Juguet et al., 2009).

#### 4.2.4. Resistance mechanism and regulation of congocidine production

A transcriptional regulator, Cgc1, is encoded within the *cgc* gene cluster. This regulator has been shown to activate the transcription of all *cgc* genes (Vingadassalon et al., unpublished). Two genes, *cgc20* and *cgc21*, encode two proteins belonging to the ABC-type multidrug resistance proteins (Stumpp et al., 2005). These genes confer resistance to congocidine and export of congocidine is likely the only mechanism of resistance in *S. ambofaciens* ATCC 23877 (Juguet et al., 2009).

#### 4.3. Biosynthesis of distamycin, congocidine and disgocidine in *Streptomyces netropsis* DSM40846

*S. netropsis* was known to produce distamycin since 1964 (Arcamone et al., 1964). In 2015, two studies showed that it produces two other pyrrolamides, congocidine, and a distamycin/congocidine hybrid, named disgocidine (Figure 39) (Hao et al., 2014; Vingadassalon et al., 2015).

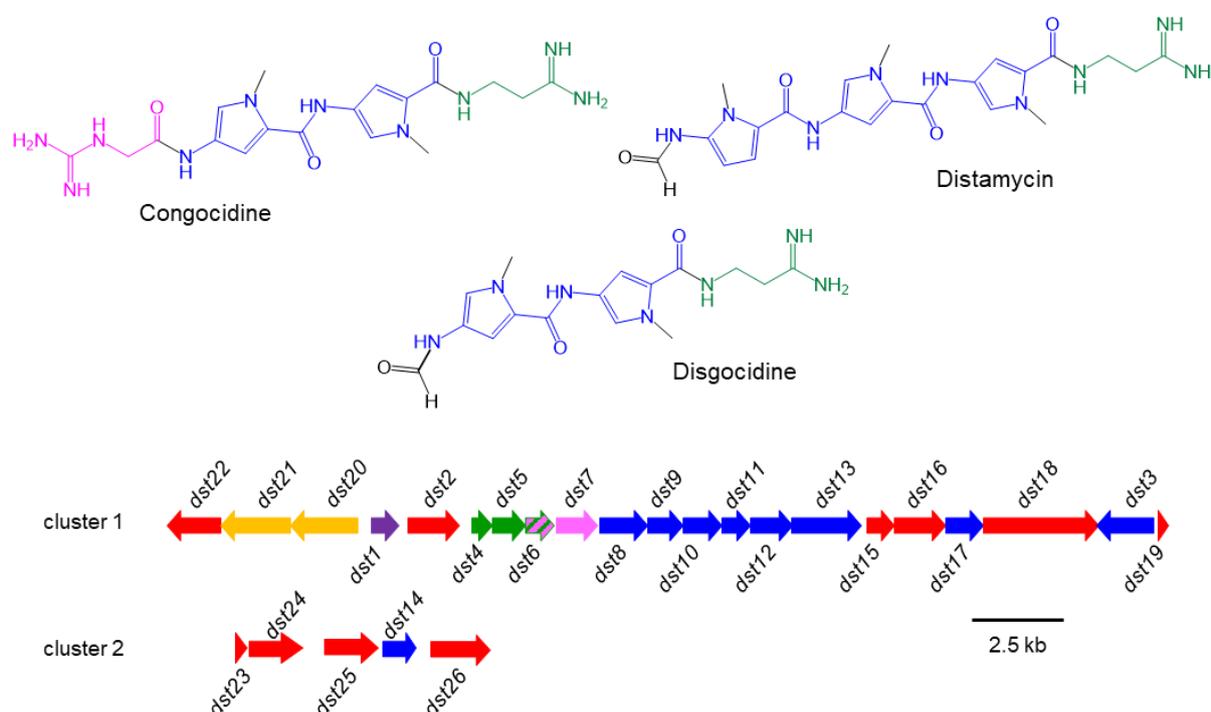


Figure 39: Biosynthetic gene clusters responsible for the production of distamycin, congocidine and disgocidine in *S. netropsis*. *dst* genes were numbered following *S. ambofaciens* *cgc* cluster nomenclature when applicable.

Two clusters, physically distant on *S. netropsis* chromosome, are responsible for the production of the three pyrrolamides (Figure 39). Genes from both clusters are necessary for the production of each of the molecules. Indeed, cluster 1 contains all the homologs of the *cgc* genes from *S. ambofaciens*, except for the homolog of *cgc14*. It thus contains all the genes necessary for the biosynthesis of the precursors of the three pyrrolamides, for the resistance to these pyrrolamides

and for the transcriptional regulation of the cluster. All genes necessary for the assembly of congocidine (but *cgc14*) are also encoded within this cluster.

Table 6: Effects of the deletion of *dst* genes on the production of congocidine, distamycin and disgocidine

Genotype	Effect on Congocidine production	Effect on Disgocidine production	Effect on Distamycin production
Clusters 1 and 2	++	++	++
$\Delta dst25$	++	++	+
$\Delta dst24$	++	+	-
$\Delta dst23$	++	+	-
$\Delta dst26$	++	-	-
$\Delta dst22$	-	-	-
$\Delta dst2$	++	++	++
$\Delta dst16$	-	-	+
$\Delta dst19$	-	-	-
$\Delta dst18$	-	++	++
$\Delta dst2/\Delta dst25$	-	-	-
$\Delta dst2/\Delta dst24$	++	+	-
$\Delta dst24/\Delta dst25$	++	+	-

As for cluster 2, it contains the homolog of *cgc14* and 4 extra genes, encoding: two condensation domains, *dst24* and *dst25*, one PCP domain *dst23*, and a formylation enzyme *dst26*. The effects of the deletion of the assembly genes on the production of distamycin, congocidine and disgocidine are summarized on Table 6. It was observed that *dst22* and *dst19* are necessary for the production of each molecule. In contrast, *dst23* which is a PCP domain homolog to *dst19* is only necessary to produce distamycin, and improves the production of disgocidine. *dst2* can be fully replaced by *dst25*, and can replace *dst25* almost as equally (production of distamycin is decreased in absence of *dst25*), both genes are almost exchangeable. It is not the case for the couple *dst16/dst24*. Indeed, *dst16* is necessary for congocidine and disgocidine production, and improves distamycin production, whereas *dst24* is necessary for distamycin production, and improves disgocidine production. The difference in production in those cases of homolog enzymes is likely due to high substrate specificities or impaired protein interactions. It is worth noting that no COM-domain could be detected in the sequence of the *dst* NRPS. Based on these data summarized in Table 6, a mechanism of biosynthesis was proposed for the three molecules (Figure 40). Interestingly, disgocidine production seems to result from the interaction of the two clusters (Vingadassalon et al., 2015). Several biosynthetic pathways can explain the production of disgocidine, in what seems to be a case of “natural combinatorial biosynthesis”. Moreover, the presence of “gene scars” in

cluster 2 suggests that originally both clusters were functional on their own, and that genes were lost during evolution due to functional redundancy.

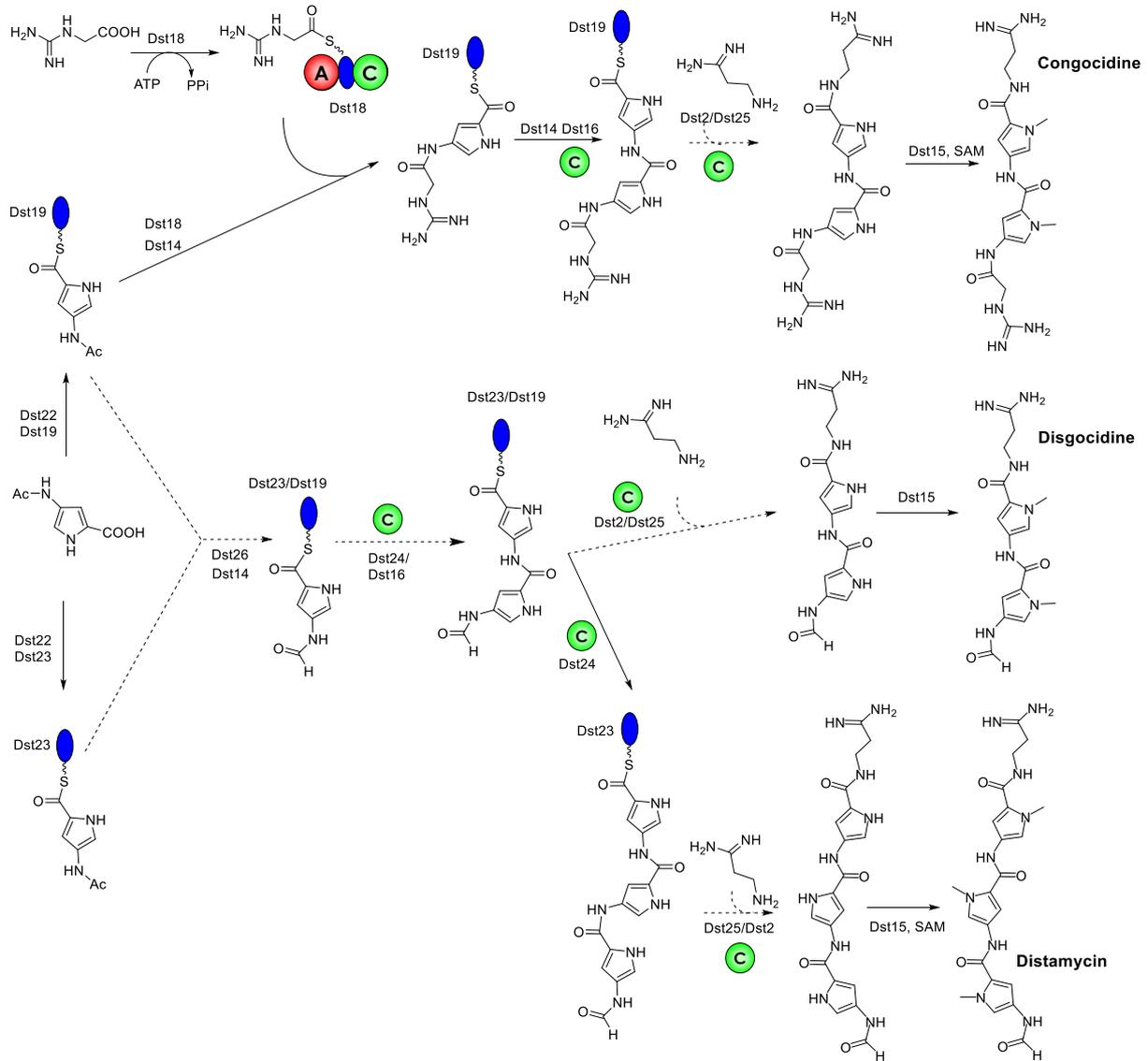


Figure 40: Biosynthetic pathways proposed for the assembly of distamycin, disgocidine and congocidine  
Dashed arrows represent reactions for which the enzymes are not uniquely defined

## Objectives of the thesis project:

The review of the literature on NRPS mechanisms and synthetic biology presented in sections 2 and 3 of this introduction clearly shows that, if the general principles of non-ribosomal peptide biosynthesis are well understood, much work is still needed to decipher the fine mechanisms allowing the coordinated functioning of the numerous (enzymatic) domains constituting these mega-complexes. Structural and biochemical studies will undoubtedly be necessary, but using combinatorial biosynthesis to tackle these questions could also bring important information. In this respect, the NRPSs directing the biosynthesis of pyrrolamide could constitute a good model. Indeed, these atypical NRPS systems are constituted of stand-alone modules and domains only, much smaller objects than classical NRPS subunits and thus easier to manipulate genetically or biochemically. Thus, with the aim of contributing to a better understanding of NRPS systems, we decided to build on the expertise our team has acquired on pyrrolamide biosynthetic systems to elaborate a combinatorial biosynthesis approach based on these systems. My PhD project consisted in constructing the tools required for future combinatorial biosynthesis of pyrrolamides. The project was divided in three axes, each developed in a distinct thesis chapter:

- (i) A prerequisite to do combinatorial biosynthesis is to have at your disposal genes from different biosynthetic gene clusters. Indeed, these genes are the basic bricks which provide the precursors and the enzymes that are to be exchanged. At the beginning of my project, the laboratory had characterized the biosynthetic pathways of congocidine (in *S. ambofaciens* (2009) and *S. netropsis* (unpublished)), and of distamycin / disgocidine / congocidine (in *S. netropsis* (2015)). However, biosynthetic genes of the other pyrrolamides were not identified. I thus undertook the **characterization of the biosynthetic gene cluster of anthelvencin, a pyrrolamide produced by *S. venezuelae* ATCC 14583**, which is presented in Chapter I.
- (ii) Combinatorial biosynthesis implies to have backbones that allow genetic manipulations of numerous gene constructions. Previously constructed integrative plasmids are still much used today, but they are not standardized and do not easily fit this purpose. I hence **developed a series of 12 integrative vectors**. These modular plasmids were designed to facilitate the construction of gene cassettes. They were also constructed to allow multiple or iterative integrations in *Streptomyces* chromosome and an excision system was set up to recycle the resistance markers and delete superfluous elements after integration. The construction of these vectors is presented in Chapter II.
- (iii) Exchange of genes supposes the existence of a bank of standardized gene cassettes. Therefore, I designed gene cassettes constituted of a synthetic promoter associated to a RBS, the pyrrolamides gene(s) and a terminator as standard bricks to be assembled. A logical first step before proceeding to combinatorial biosynthesis consisted in reconstructing a known biosynthetic pathway and confirming the production of the expected pyrrolamide. I undertook the **refactoring of the congocidine gene cluster by constructing and assembling all the *cgc* gene cassettes** necessary for production and assessed congocidine production in the host strain *S. lividans* TK23. This refactoring process is presented in the third and last chapter of this thesis.

Chapter I - Revised structure of anthelvencin A  
and characterization of the anthelvencin  
biosynthetic gene cluster from *Streptomyces  
venezuelae* ATCC 14583

## Chapter I introduction:

In this first chapter, I present my work on the characterization of the gene cluster directing the biosynthesis of anthelvencins in *Streptomyces venezuelae* ATCC 14583. These studies allowed to revise the structure of anthelvencin A, to identify a new anthelvencin metabolite, and to show the involvement of an enzyme from the ATP-grasp ligase family in the assembly of these pyrrolamides. Furthermore, the non-ribosomal peptide synthetase assembling anthelvencins is composed of stand-alone domains only, as it is the case for congocidine and distamycin NRPS. The new uncovered pyrrolamide genes therefore constitute an addition to our NRPS gene library, and will likely be valuable later on to proceed to NRPS exchanges for combinatorial biosynthesis experiments.

This work, presented using the format of an article, will be published soon and a short perspective at the end of the chapter discusses the remaining points that have to be considered before submission.

# Revised structure of anthelvencin A and characterization of the anthelvencin biosynthetic gene cluster from *Streptomyces venezuelae* ATCC 14583

Céline Aubry<sup>a</sup>, Paolo Clerici<sup>b</sup>, Claude Gerbaud<sup>a</sup>, Laurent Micouin<sup>b</sup>, Jean-Luc Pernodet<sup>a</sup>, and Sylvie Lautru<sup>a#</sup>

<sup>a</sup> Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

<sup>b</sup> Nouvelles méthodes de synthèse pour l'interface chimie-biologie, CNRS, Université Paris Descartes, 75006, Paris, France

# Corresponding author: Sylvie LAUTRU, sylvie.lautru@i2bc.paris-saclay.fr

## ABSTRACT

Anthelvencins A and B are pyrrolamide metabolites produced by *Streptomyces venezuelae* ATCC 14583. In this study, we revise the structure of anthelvencin A and identify a third anthelvencin metabolite, bearing two N-methylated pyrrole groups, which we named anthelvencin C. Using the genome sequence of *S. venezuelae*, we isolated the gene cluster directing the biosynthesis of anthelvencins and functionally characterized it. As observed for the biosynthesis of the other pyrrolamides congocidine and distamycin, the non-ribosomal peptide synthetase assembling anthelvencins is composed of stand-alone domains only. The assembly of anthelvencins also involves an enzyme from the ATP-grasp ligase family, Ant23. We propose that Ant23 uses a PCP-loaded 4-aminopyrrole-2-carboxylate as substrate.

**KEYWORDS** *Streptomyces*, pyrrolamide

## INTRODUCTION

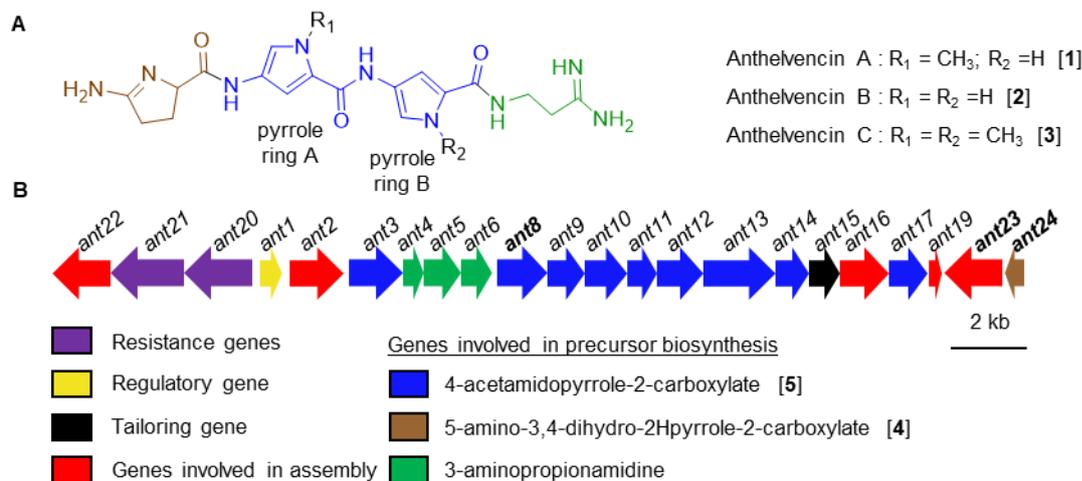
Anthelvencins A [1] and B [2] (Figure 1A) are specialized metabolites that were isolated in 1965 from cultures of *Streptomyces venezuelae* ATCC 14583-14585 and exhibit moderate antibacterial and anthelmintic activities (Probst et al., 1965). They belong to the family of pyrrolamide metabolites, the best characterized members of which are congocidine and distamycin. These metabolites are DNA minor groove binders that exhibit some sequence specificity, binding in regions of four (or more) A or T bases (Neidle, 2001). During the last decade, the biosynthetic gene clusters of congocidine and distamycin have been identified and the biosynthesis of these metabolites has been elucidated (Al-Mestarihi et al., 2015; Hao et al., 2014; Juguet et al., 2009; Lautru et al., 2012; Vingadassalon et al., 2015). One remarkable aspect of this biosynthesis is that it involves non-canonical non-ribosomal peptide synthetases (NRPSs), solely constituted of stand-alone modules or domains.

A structural analysis of anthelvencins shows that these metabolites most likely share two precursors with congocidine and distamycin: 4-acetamidopyrrole-2-carboxylate [5] and 3-aminopropionamide. The remaining precursor is probably 5-amino-3,4-dihydro-2H-pyrrole-2-carboxylate [4], a precursor shared with other pyrrolamides such as kikumycins (Takaishi et al., 1972), TAN 868A (Takizawa et al., 1987) or noformycin (Diana, 1973) (See Figure S1 in the supplemental). In fact, members of the pyrrolamide family seem to be assembled from a limited number of precursors that are combined in some kind of natural combinatorial manner. Understanding how these precursors are assembled and combined may improve our comprehension of NRPS enzymatic mechanisms and help to design functional synthetic NRPSs using synthetic biology. For these reasons, we undertook to isolate and characterize the biosynthetic gene cluster of anthelvencins of *S. venezuelae* ATCC 14583. In this study, we show that *S. venezuelae* ATCC 14583 produces, in addition to the already isolated anthelvencin A and B, a third anthelvencin (methylated on the two pyrrole groups) that we named anthelvencin C. Based on HR-MS<sup>2</sup> data, we revise the structure of anthelvencin A. We also identify the gene cluster directing the biosynthesis of anthelvencins in *S. venezuelae* ATCC 14583 genome and functionally characterize it.

## RESULTS AND DISCUSSION

### *In silico identification of a gene cluster putatively involved in anthelvencin biosynthesis in S. venezuelae ATCC 14583*

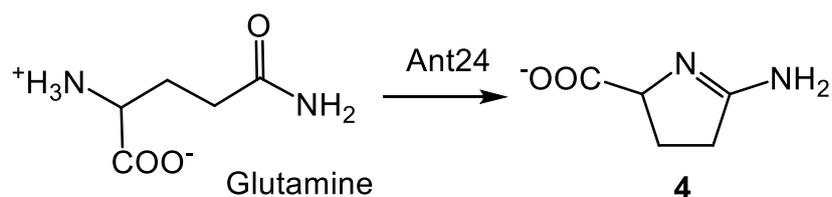
To isolate the gene cluster directing anthelvencin biosynthesis, we sequenced the genome of the *S. venezuelae* ATCC 14583 strain by the Illumina technology, using a paired-end genomic library. The 5.45 million reads of 301 bps were assembled using Velvet v1.2.10, resulting in 63 contigs with a total length of 9.08 Mbps (180-fold coverage).



**Figure 1.** Structure of anthelvencins A, B and C (A) and genetic organization of the anthelvencin biosynthetic gene cluster in *S. venezuelae* ATCC 14583 (B).

Genes in boldface are genes that have been replaced by a resistance cassette in this study.

The gene cluster directing the biosynthesis of anthelvencins was identified by mining the genome of *S. venezuelae* ATCC 14583 for homologs of genes involved in the biosynthesis of congocidine (Juguet et al., 2009). We identified a gene cluster (*ant*) that spans 26 kb and contains 22 genes (Figure 1B). Twenty of the Ant proteins exhibit a high amino acid sequence identity with Cgc proteins (from 64 to 84 % sequence identity, Table 1) and they most likely have a similar function to their Cgc homologs. Thus, the gene numbers attributed to the *ant* genes were chosen to follow the *cgc* nomenclature whenever possible. The genetic organization of the *ant* cluster is remarkably similar to the one of the *cgc* cluster (Figure S2, (Juguet et al., 2009)). Two *cgc* genes (*cgc7* and *cgc18*) involved in the biosynthesis of the guanidinoacetate precursor of congocidine (absent in anthelvencins) and its assembly have no homologs in the *ant* gene cluster. Instead, the cluster contains two genes, *ant24* and *ant23*, likely involved in the biosynthesis of 5-amino-3,4-dihydro-2H-pyrrole-2-carboxylate [4] and its assembly with the first pyrrole precursor respectively. Indeed, a protein blast and a conserved domain searches (Altschul et al., 1990; Marchler-Bauer et al., 2017) on the Ant24 sequence suggested that Ant24 belongs to the L-ectoine synthase (EC 4.2.1.108) family of enzymes. L-ectoine synthases catalyze the ring closure of N $\gamma$ -acetyl-L-2,4-diaminobutyric acid, yielding the osmolyte ectoine, a metabolite structurally related to [4]. In 2011, Witt and collaborators reported that the ectoine synthase from *Halomonas elongata* can catalyze the intramolecular condensation of glutamine to form [4] as a side reaction (Witt et al., 2011). Thus, it appears likely that Ant24 catalyzes the same reaction (Scheme 1).



**Scheme 1:** Proposed biosynthesis of 5-amino-3,4-dihydro-2H-pyrrole-2-carboxylate [4] by Ant24

**Table 1.** Sequence identities between Ant and Cgc proteins

Protein	Putative protein function	Cgc orthologue	% sequence identity
Ant1	Transcriptional regulator	Cgc1	71
Ant2	NRPS, C domain	Cgc2	66
Ant3	4-acetamidopyrrole-2-carboxaldehyde dehydrogenase	Cgc3	74
Ant4	Cytosine monophosphate hydrolase	Cgc4	83
Ant5	cytosine reductase	Cgc5	77
Ant6	dihydrocytosine hydrolase	Cgc6	78
Ant8	nucleotidyl N-acetylglucosamine dehydrogenase	Cgc8	84
Ant9	nucleotidyl-2-acetamido-2-deoxyglucopyranuronate decarboxylase	Cgc9	84
Ant10	glycosyltransferase	Cgc10	81
Ant11	N-acetylglucosamine-1-phosphate nucleotidyltransferase	Cgc11	76
Ant12	Nucleotidyl threo-2-acetamido-2-deoxy-pentopyran-4-ulose aminotransferase	Cgc12	79
Ant13	glycoside hydrolase	Cgc13	78
Ant14	4-acetamidopyrrole-2-carboxylate deacetylase	Cgc14	80
Ant15	methyltransferase	Cgc15	84
Ant16	NRPS, C domain	Cgc16	68
Ant17	4-acetamidopyrrole-2-carboxaldehyde dehydrogenase	Cgc17	83
Ant19	NRPS, PCP domain	Cgc19	64
Ant20	ABC transporter	Cgc20	81
Ant21	ABC transporter	Cgc21	81
Ant22	acyl co-A synthetase	Cgc22	72
Ant23	ATP-grasp domain-containing protein	/	
Ant24	Ectoine synthase-like protein	/	

Ant23 contains an ATP-grasp domain. ATP-grasp enzymes usually catalyze the ATP-dependent ligation of a carboxylate-containing molecule to an amino or thiol group-containing molecule (Galperin and Koonin, 1997). Some of these enzymes are encoded in specialized

metabolism gene clusters (Goswami and Van Lanen, 2015). They can be used as an alternative to or in combination with non-ribosomal peptide synthetases (NRPS), to elongate a peptide chain (Goswami and Van Lanen, 2015; Hollenhorst et al., 2009). Thus, it appears plausible that Ant23 catalyzes the amide bond formation between [4] and a PCP (Ant19)-bound 4-aminopyrrole-2-carboxylate.

### ***Abolition of the production of four metabolites in a *S. venezuelae* ATCC 14583 ant8 replacement mutant***

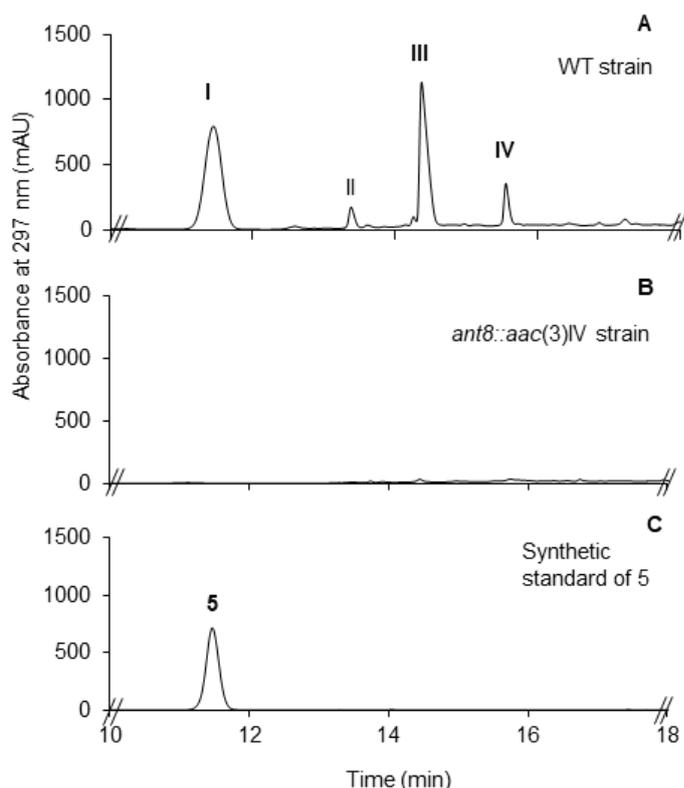
To verify that the *ant* gene cluster is involved in the biosynthesis of anthelvencins, we inactivated *ant8*. This gene is the ortholog of *ggc8* that is involved in the biosynthesis of the 4-acetamidopyrrole-2-carboxylate [5], precursor of congocidine (Lautru et al., 2012) and likely precursor of anthelvencins. The *ant8* gene was replaced by an *aac(3)IV* resistance cassette by homologous recombination using the pANT007 suicide plasmid, yielding the *S. venezuelae* ANT007 strain. This strain and the wild type *S. venezuelae* strain were cultivated for three days in MP5 liquid medium. The culture supernatants were then filtered and analysed by HPLC. The chromatograms (Figure 2) show that four metabolites present in the wild type strain supernatant (peaks I to IV) are absent in the supernatant of the ANT007 mutant strain. The first metabolite (peak I, retention time of 11.5 min) corresponds to 4-aminopyrrole-2-carboxylate [5], identified by its UV spectrum and by comparison with an authentic standard (Figure 2 and (Lautru et al., 2012)). The three peaks II (retention time of 13.3 min), III (retention time of 14.3 min) and IV (retention time of 15.5 min) have UV absorption spectra typical of pyrrolamides (Figure S3, (Vingadassalon et al., 2015)).

### ***Identification of metabolites II, III and IV***

To determine the chemical nature of the metabolites II, III and IV, we partially purified them. For that purpose, we used ANT012, a strain that expressed a second copy of the genes *ant23* and *ant24* under the promoter *rpsL(TP)* (Shao et al., 2013), as this strain produces compounds III and IV in slightly higher titers (data not shown). The ANT012 culture supernatant was recovered after three days of culture in MP5 medium and the compounds of interest were partially purified on a XAD16 resin. The elution fraction was concentrated to dryness solution, resuspended in water and analyzed by LC-HR-MS<sup>2</sup>.

The exact mass and fragmentation pattern of compound II (Figure S4) are consistent with II being anthelvencin B [2] ( $[M+H]^+$   $m/z = 414.1998$ ; calculated 414.1997). The exact mass of compound III (Figure S5) is consistent with III being anthelvencin A [1] ( $[M+H]^+$   $m/z = 428.2151$ ; calculated 428.2153). The fragmentation pattern however (Figure S5), indicates that the position of the methyl group is not on the B pyrrole ring, as previously proposed (but never experimentally established, (Probst et al., 1965)) but rather on the A pyrrole ring (Figure 1). To confirm the structure of anthelvencin A, we purified compound III and carried out NMR experiments. Unfortunately, the quality of the data obtained so far have not permitted to determine the exact position of the methyl group (Figure S7).

The exact mass and fragmentation pattern of compound IV (Figure S6) are consistent with IV being an anthelvencin metabolite methylated on both pyrrole groups ( $[M+H]^+$   $m/z = 442.2311$ ; calculated 442.2310), metabolite that we named anthelvencin C ([3], Figure 1A). We tried to purify anthelvencin C to confirm its chemical structure with NMR analyses but this metabolite turned out to be highly unstable, as already observed by M. Lee and coworkers (Lee et al., 1988).

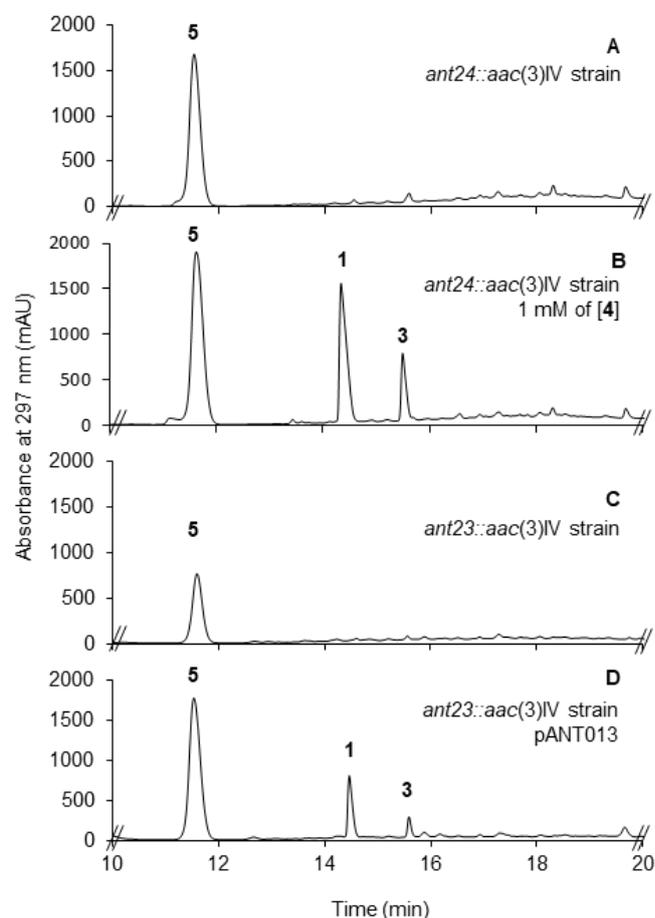


**Figure 2:** HPLC analysis of culture supernatants of A) *S. venezuelae* ATCC 14583 WT and B) ANT007 (*S. venezuelae* ATCC 14583 *ant8::aac(3)IV*). C) Standard of 4-acetamidopyrrole-2-carboxylate [5].

Samples were analyzed on a reverse phase  $C_{18}$  column, eluted in isocratic conditions with 0.1% HCOOH in  $H_2O$  (solvent A)/ 0.1% HCOOH in  $CH_3CN$  (solvent B) (95:5) at  $1 \text{ ml}\cdot\text{min}^{-1}$  for 7 min, followed by a gradient to 40:60 A/B over 23 min.

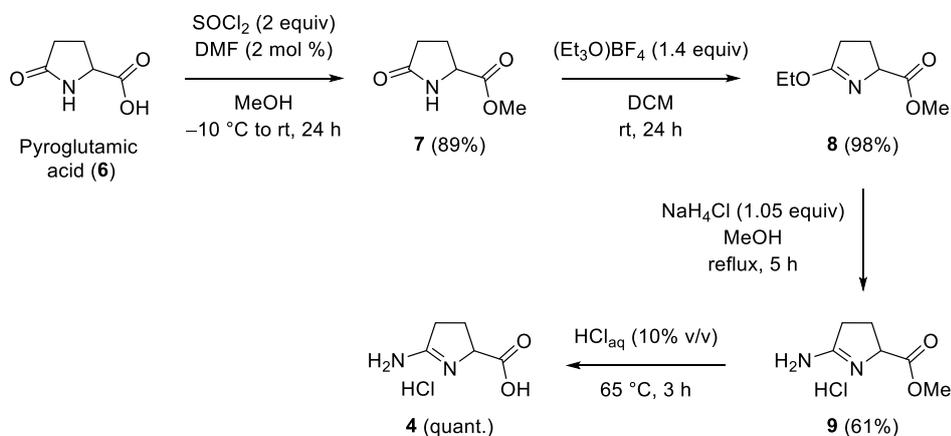
### ***Involvement of ant24 in 5-amino-3,4-dihydro-2H-pyrrole-2-carboxylate [4] biosynthesis***

To verify that *ant24* is involved in the biosynthesis of [4], we replaced it by an *aac(3)IV* resistance cassette by homologous recombination, following the same procedure as described above. The supernatant of the resulting mutant strain, called ANT009, was analysed by HPLC (Figure 3A). No production of anthelvencins was observed, confirming that *ant24* is necessary for production of these metabolites. To examine Ant24 putative function in the biosynthesis of [4], we chemically synthesized [4] according to a previously described synthetic procedure (Lee et al., 1988) (Scheme 2). We next fed the ANT009 strain with [4]. As shown in Figure 3B, this resulted in the restoration of the production of anthelvencins A and C, hence confirming the involvement of *ant24* in the biosynthesis of the anthelvencin precursor [4].



**Figure 3:** HPLC analysis of culture supernatants of (A) ANT009 (*S. venezuelae* ATCC 14583 *ant24::aac(3)IV*), (B) ANT009 (*S. venezuelae* ATCC 14583 *ant24::aac(3)IV*) cultivated in presence of 1mM of [4], (C) ANT008 (*S. venezuelae* ATCC 14583 *ant23::aac(3)IV*) and (D) ANT013 (*S. venezuelae* ATCC 14583 *ant23::aac(3)IV* pANT013) (genetic complementation of ANT008).

Numbers above peaks correspond to the metabolite numbers in the text. Samples were analyzed on a reverse phase C<sub>18</sub> column, eluted in isocratic conditions with 0.1% HCOOH in H<sub>2</sub>O (solvent A)/ 0.1% HCOOH in CH<sub>3</sub>CN (solvent B) (95:5) at 1 ml.min<sup>-1</sup> for 7 min, followed by a gradient to 40:60 A/B over 23 min.

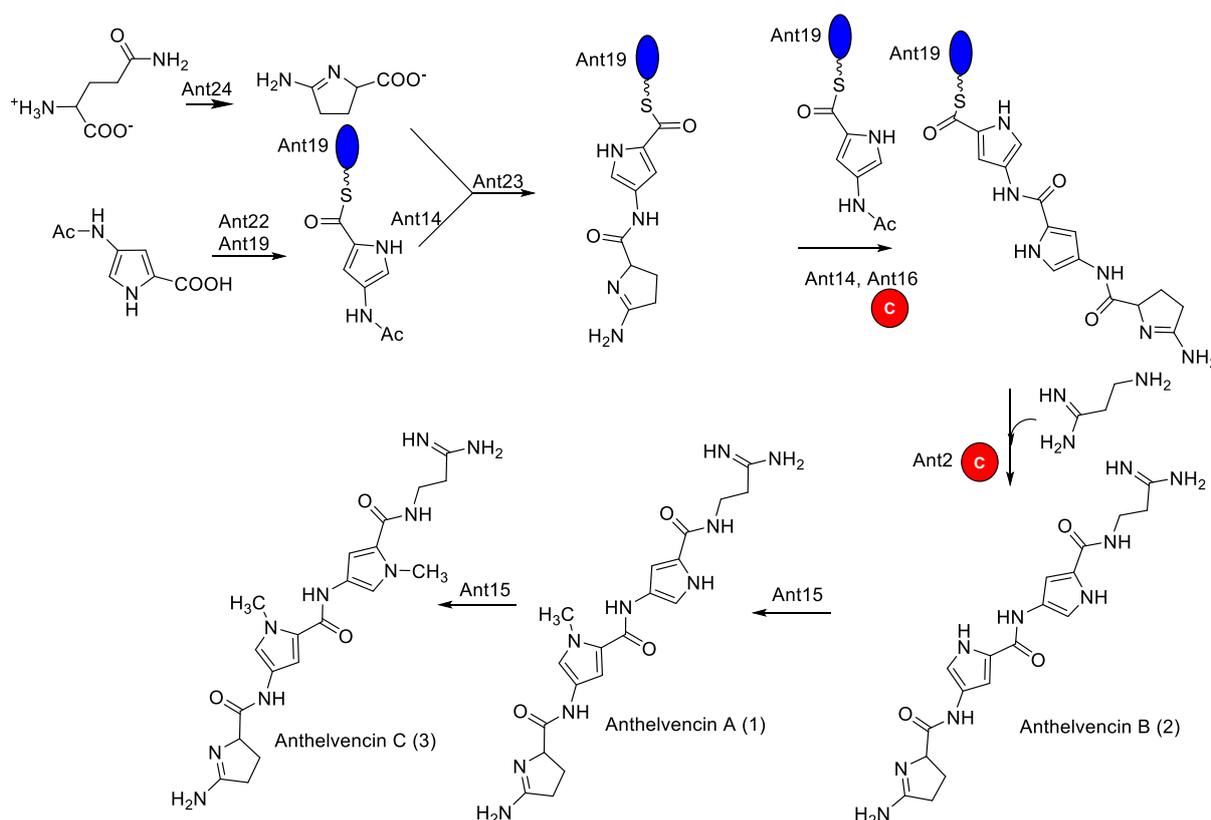


**Scheme 2:** Synthesis of 5-amino-3,4-dihydro-2Hpyrrole-2-carboxylate [4]

### Involvement of *ant23* in the biosynthesis of anthelvencins

To confirm that *Ant23* is involved in anthelvencin biosynthesis, we replaced *ant23* by the *aac(3)IV* resistance cassette following the previously described protocol. The resulting mutant strain was called ANT008. It was cultivated for three days in MP5 medium at 28°C and the culture supernatant was analysed by HPLC. Figure 3C shows that no anthelvencin is produced by the ANT008 mutant. To ensure that the observed phenotype was due to the replacement of *ant23* by the *aac(3)IV* cassette, we genetically complemented the ANT008 strain using a plasmid expressing *ant23* and *ant24* under a constitutive promoter. The production of anthelvencins was restored in the complemented strain, named ANT013 (Figure 3D), thus confirming that *ant23* is involved in anthelvencin biosynthesis.

### Proposed biosynthetic pathway for anthelvencin biosynthesis



**Figure 4:** Proposed biosynthetic pathway for anthelvencins A, B and C

Based on the results presented above and on previous characterizations of pyrrolamide biosyntheses (Al-Mestarihi et al., 2015; Juguet et al., 2009; Lautru et al., 2012; Vingadassalon et al., 2015), we proposed that anthelvencins are assembled from 3-amidinopropionamide, 4-aminopyrrole-2-carboxylate and 5-amino-3,4-dihydro-2H-pyrrole-2-carboxylate following the biosynthetic pathway presented in Figure 4. As already observed for the biosynthesis of other pyrrolamides (congozidine, distamycine), the non-ribosomal peptide synthetase involved in

anthelvencins is constituted solely of stand-alone domains (C and PCP domains). No adenylation domain is involved in the activation of the carboxylate groups of the precursors. Instead, activation of the carboxylate group of the pyrrole precursor [5] and the covalent attachment of the activated precursor to the PCP domain Ant19 is catalyzed by Ant22, which belongs to the family of acyl-CoA synthetases. The formation of the first amide bond between [4] and Ant19-bound [5] is likely catalyzed by Ant23, an enzyme from the ATP-grasp ligase family of enzymes, which form acylphosphate intermediates. Two stand-alone condensation domains, Ant16 and Ant2, catalyze the formation of the other amide bonds, adding a second pyrrole precursor and the 3-aminopropionamide respectively.

## CONCLUSIONS

We have identified and characterized the gene cluster directing the biosynthesis of anthelvencins in *Streptomyces venezuelae* ATCC 14583. We showed that this cluster directs the biosynthesis of two known metabolites, anthelvencin A, for which we propose a revised structure, anthelvencin B, and new anthelvencin that we named anthelvencin C.

## MATERIALS AND METHODS

### ***Bacterial strains, plasmids and growth conditions.***

Strains and plasmids used in this study are listed in Table S1 and S2. *Escherichia coli* strains were grown at 37 °C in LB or SOB complemented with MgSO<sub>4</sub> (20 mM final), supplemented with appropriate antibiotics as needed. The Soya Flour Mannitol (SFM) medium (Kieser et al., 2000) was used for genetic manipulations of *Streptomyces* strains and spore stocks preparations at 28°C. *Streptomyces* strains were grown at 28°C in MP5 (Pernodet et al., 1993) for anthelvencins [1-3] production.

### ***DNA Preparation and manipulations.***

All oligonucleotides used in this study were purchased from Eurofins and are listed in Table S3. The High fidelity DNA polymerase Phusion (Thermo Fisher Scientific) was used to amplify the DNA fragments for the construction of the suicide plasmids. Dreamtaq polymerase (Thermo Fisher Scientific) was used for PCR verification of plasmids and of the replacement of the targeted genes by the resistance cassette. DNA fragments were purified from agarose gels using the Nucleospin Gel and PCR clean-up kit from Macherey-Nagel. *E. coli* transformations and *E. coli*/*Streptomyces* conjugations were performed according to standard procedures (Sambrook and Russell, 2001; Kieser et al., 2000).

### ***S. venezuelae ATCC 14583 sequencing and assembly.***

Total DNA of *S. venezuelae* ATCC 14583 was extracted following standard protocols (Kieser et al., 2000). A paired-end library of the whole genome was constructed and sequenced at the high throughput sequencing core facility of I2BC with a MiSeq M01342 instrument (Illumina), generating 5.45 million 301 bp reads that were assembled using Velvet v1.2.10. The GenBank accession number of the anthelvencin A gene cluster is MK483114.

**Construction of the replacement mutants.**

The suicide plasmid pANT007 was constructed to replace the *ant8* gene by an *aac(3)IV* resistance cassette in *S. venezuelae*. This vector was constructed by assembling in pOSV400 the three following inserts: a 1.8 kb fragment upstream of *ant8*, the resistance cassette *aac(3)IV* and a 2.0 kb DNA fragment downstream of *ant8*. The 1.8 kb and 2.0 kb DNA fragments from *S. venezuelae* ATCC 14583 were amplified by PCR with the primers CEA001/CEA002 and CEA003/CEA004 respectively. The PCR products were purified and ligated into pCR® Blunt, yielding pANT001 and pANT002. Both plasmids were verified by sequencing. The *aac(3)IV* resistance cassette was obtained by digestion of pW60 (Corre et al., 2008) by *HindIII*. The 1.8 kb *HindIII*-*XbaI* fragment from pANT001, the 1.0 kb *HindIII* *aac(3)IV* fragment, and the 2.0 kb *HindIII*-*SpeI* fragment from pANT002 were ligated next into the *XbaI*-*SpeI*-digested pOSV400, yielding pANT007. The pANT007 plasmid was verified by restriction digestion using *StuI/XbaI* and *HindIII/XbaI/SpeI* and introduced into *S. venezuelae* ATCC 14583 by intergeneric conjugation from the *E. coli* ET12567/pUZ8002/pANT007 strain. Double-recombinant mutants were selected on SFM plates with 50 µg/mL apramycin and screened for hygromycin sensitivity. The resulting strain was named ANT007 and verified using the primers A5, A6, and CEA013-CEA016. The same protocol was used for the construction of the ANT008 (replacement of *ant23*) and ANT009 (replacement of *ant24*) mutants (see Tables S2 and S3 for plasmid names and for primer sequences).

**Construction of the ANT012 strain overexpressing *ant23* and *ant24*.**

The DNA region containing *ant23-ant24* was amplified by PCR from *S. venezuelae* ATCC 14583 genomic DNA using the primers CEA034/CEA035. The PCR product was purified and cloned into pCR® Blunt, yielding pANT011, and the sequence of the insert was confirmed by sequencing. The 2.0 kb *NheI/AflII* fragment from pANT011 was ligated in the *SpeI/AflII*-digested pCEA005 (21). The obtained plasmid was named pANT012 and confirmed by restriction digestion using *HindIII/KpnI* and *XbaI/XbaI*. This plasmid was introduced into *S. venezuelae* ATCC 14583 by intergeneric conjugation. The correct integration of pANT012 was verified using the primers CEA\_vec\_seq14 and CEA\_vec\_seq15 and strain was named ANT012.

**Genetic complementation of ANT008**

The ANT008 strain bearing the *aac(3)IV* resistance marker, the pANT012 plasmid previously constructed could not be used for the genetic complementation of the strain. Thus, the 2.4 kb *NsiI/AflII* DNA fragment of pANT012 containing *ant23* and *ant24* under the control of the *rpsL*(IP) promoter was ligated into the *NsiI/AflII*-digested pOSV806 (Aubry et al., 2019). The resulting plasmid was named pANT013 and introduced into ANT008 by intergeneric conjugation. The strain obtained was named ANT013.

**Chemical synthesis of 5-amino-3,4-dihydro-2H-pyrrole-2-carboxylate [4].**

Compound [4] was prepared according to a previously described synthetic procedure (Lee and Lown, 1987) (Scheme 2). Commercially available DL-pyroglutamic acid [6] was first converted into the corresponding methyl ester [7] by treatment with thionyl chloride (2 equiv.), and DMF (2 mol %) in methanol. Derivative [7] was then submitted to a reaction with triethyloxonium tetrafluoroborate (Meerwein's salt, 1.4 equiv.) in DCM to form carboximidate [8] in quantitative yield. This compound subsequently reacted with ammonium chloride (1.05

equiv.) in refluxing methanol to provide product [9] in 61% yield. Hydrolysis of the ester moiety of compound [9] finally afforded the desired acid [4] in a quantitative yield. Detailed synthesis available in the supplemental material.

#### ***Chemical complementation of ANT009.***

*S. venezuelae* ANT009 strain was cultivated in 50 mL of MP5. After 24 h, the cultures were separated in two 25 mL cultures, and 1 mM of [4] (final concentration) was added to one of the cultures. After a total of 48 h of culture, culture supernatants were analysed by HPLC as described below.

#### ***HPLC analysis of culture supernatants.***

*S. venezuelae* ATCC 14583 and its derivatives were cultivated in MP5 medium for three days at 28°C. The supernatants were filtered using Mini-UniPrep syringeless filter devices (0.2 µm, Whatman). The samples were analysed on an Atlantis C<sub>18</sub> T3 column (250 mm x 4.6 mm, 5 µm, column temperature 28°C) using an Agilent 1200 HPLC instrument with a quaternary pump. Samples were eluted in isocratic conditions with 0.1% HCOOH in H<sub>2</sub>O (solvent A)/ 0.1% HCOOH in CH<sub>3</sub>CN (solvent B) (95:5) at 1 mL·min<sup>-1</sup> for 7 min, followed by a gradient to 40:60 A/B over 23 min. Anthelvencins were detected by monitoring absorbance at 297 nm.

#### ***LC-HR-MS-MS analyses.***

The resuspended elution fraction obtained above was analysed by LC-HR-MS<sup>2</sup>. The analysis was performed using a Dionex Ultimate 3000 HPLC system coupled with a Maxis IIT<sup>TM</sup> QTOF mass spectrometer (Bruker, MA, USA) fitted with an electrospray ionization (ESI) source.

Chromatographic analysis was performed using a C<sub>18</sub> Acclaim<sup>TM</sup> RSLC PolarAdvantage II (2.1 x 100 mm, 2.2 µm pore size) column (Thermo Scientific, MA, USA). Column temperature was set at 40 °C and 2 µL of each sample was injected via an autosampler cooled to 4 °C. A flow rate of 0.3 mL/min was used, and the eluent was introduced directly into the MS for ion detection. Elution was conducted with a mobile phase consisting of 0.1% HCOOH in H<sub>2</sub>O (solvent A) and 0.1% HCOOH in CH<sub>3</sub>CN (solvent B) following the gradient elution profile: 0 min, 5% solvent B; 2 min, 5% solvent B; 9 min, 50% solvent B; 15 min 90% solvent B; 17 min 90% solvent B; 19 min 5% solvent B; 21 min 5% solvent B. In the first half minute of each run, a sodium formate solution was injected directly as an internal reference for calibration. The acquisition parameters of the ESI source were set up as follows: electrospray voltage for the ESI source: 3500V, nebulising gas (N<sub>2</sub>) pressure: 35 psi, drying gas (N<sub>2</sub>) flow: 8 L/min, and drying temperature: 200°C. Mass spectra were recorded over the m/z range 100-1300 at a frequency of 2 Hz, in positive ion mode. For MS/MS analysis, the cycle time was of 3 sec. Mass spectra were recorded over the m/z range 100-1300 at a frequency of 2 Hz, in positive ion mode. Selected parent ion at m/z 442.23 was fragmented at a fixed collision energy value of 40 eV and an isolation window of 0.5 amu.

#### ***Acknowledgements***

We acknowledge the High-throughput sequencing facility of I2BC for its sequencing and bioinformatics expertise (Centre de Recherche de Gif – <http://www.i2bc.paris-saclay.fr/>). We thank Zhilai Hong, Yanyan Li and Soizic Prado for their help with the LC-HRMS<sup>2</sup> analysis. The

research received funding from ANR-14-CE16-0003-01. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

## References

- Al-Mestarihi, A.H., Garzan, A., Kim, J.M., and Garneau-Tsodikova, S. (2015). Enzymatic evidence for a revised congocidine biosynthetic pathway. *Chembiochem Eur. J. Chem. Biol.* *16*, 1307–1313.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* *215*, 403–410.
- Aubry, C., Pernodet, J.-L., and Lautru, S. (2019). A set of modular and integrative vectors for synthetic biology in *Streptomyces*. *Appl. Environ. Microbiol.* Aug 1;85(16).
- Corre, C., Song, L., O'Rourke, S., Chater, K.F., and Challis, G.L. (2008). 2-Alkyl-4-hydroxymethylfuran-3-carboxylic acids, antibiotic production inducers discovered by *Streptomyces coelicolor* genome mining. *Proc. Natl. Acad. Sci. U. S. A.* *105*, 17510–17515.
- Diana, G.D. (1973). Synthesis of noformycin. *J. Med. Chem.* *16*, 857–859.
- Galperin, M.Y., and Koonin, E.V. (1997). A diverse superfamily of enzymes with ATP-dependent carboxylate-amine/thiol ligase activity. *Protein Sci. Publ. Protein Soc.* *6*, 2639–2643.
- Goswami, A., and Van Lanen, S.G. (2015). Enzymatic strategies and biocatalysts for amide bond formation: tricks of the trade outside of the ribosome. *Mol. Biosyst.* *11*, 338–353.
- Hao, C., Huang, S., Deng, Z., Zhao, C., and Yu, Y. (2014). Mining of the pyrrolamide antibiotics analogs in *Streptomyces netropsis* reveals the amidohydrolase-dependent “iterative strategy” underlying the pyrrole polymerization. *PloS One* *9*, e99077.
- Hollenhorst, M.A., Clardy, J., and Walsh, C.T. (2009). The ATP-dependent amide ligases DdaG and DdaF assemble the fumaramoyl-dipeptide scaffold of the dapdiamide antibiotics. *Biochemistry* *48*, 10467–10472.
- Juguet, M., Lautru, S., Francou, F.-X., Nezbedová, S., Leblond, P., Gondry, M., and Pernodet, J.-L. (2009). An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congocidine in *Streptomyces ambofaciens*. *Chem. Biol.* *16*, 421–431.
- Kieser, T., Bibb, M., Buttner, M., and Hopwood, D.A. (2000). *Practical Streptomyces genetics*, John Innes Foundation, Norwich NR47UH, UK.
- Lautru, S., Song, L., Demange, L., Lombès, T., Galons, H., Challis, G.L., and Pernodet, J.-L. (2012). A sweet origin for the key congocidine precursor 4-acetamidopyrrole-2-carboxylate. *Angew. Chem. Int. Ed Engl.* *51*, 7454–7458.

- Lee, M., and Lown, J.W. (1987). Synthesis of (4S)- and (4R)-methyl 2-amino-1-pyrroline-5-carboxylate and their application to the preparation of (4S)-(+)- and (4R)-(-)-dihydrokikumycin B. *J. Org. Chem.* *52*, 5717–5721.
- Lee, M., Coulter, D.M., and Lown, J.W. (1988). Total synthesis and absolute configuration of the antibiotic oligopeptide (4S)-(+)-anthelvencin A and its 4R-(-) enantiomer. *J. Org. Chem.* *53*, 1855–1859.
- Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., et al. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* *45*, D200–D203.
- Neidle, S. (2001). DNA minor-groove recognition by small molecules. *Nat. Prod. Rep.* *18*, 291–309.
- Pernodet, J.L., Alegre, M.T., Blondelet-Rouault, M.H., and Guérineau, M. (1993). Resistance to spiramycin in *Streptomyces ambofaciens*, the producer organism, involves at least two different mechanisms. *J. Gen. Microbiol.* *139*, 1003–1011.
- Probst, G.W., Hoehn, M.M., and Woods, B.L. (1965). Anthelvencins, new antibiotics with anthelmintic properties. *Antimicrob. Agents Chemother.* *5*, 789–795.
- Sambrook, J., and Russell, D.W. (2001). *Molecular cloning: a laboratory manual*, Third edition. CSHL Press, Cold Spring Harbor, NY.
- Shao, Z., Rao, G., Li, C., Abil, Z., Luo, Y., and Zhao, H. (2013). Refactoring the silent spectinabilin gene cluster using a plug-and-play scaffold. *ACS Synth. Biol.* *2*, 662–669.
- Takaishi, T., Sugawara, Y., and Suzuki, M. (1972). Structure of kikumycin A and B. *Tetrahedron Lett.* *13*, 1873–1876.
- Takizawa, M., Tsubotani, S., Tanida, S., Harada, S., and Hasegawa, T. (1987). A new pyrrole-amidine antibiotic TAN-868 A. *J. Antibiot. (Tokyo)* *40*, 1220–1230.
- Vingadassalon, A., Lorieux, F., Juguet, M., Le Goff, G., Gerbaud, C., Pernodet, J.-L., and Lautru, S. (2015). Natural combinatorial biosynthesis involving two clusters for the synthesis of three pyrrolamides in *Streptomyces netropsis*. *ACS Chem. Biol.* *10*, 601–610.
- Witt, E.M.H.J., Davies, N.W., and Galinski, E.A. (2011). Unexpected property of ectoine synthase and its application for synthesis of the engineered compatible solute ADPC. *Appl. Microbiol. Biotechnol.* *91*, 113–122.

# Revised structure of anthelvencin A and characterization of the anthelvencin biosynthetic gene cluster from *Streptomyces venezuelae* ATCC 14583

Céline Aubry<sup>a</sup>, Paolo Clerici<sup>b</sup>, Claude Gerbaud<sup>a</sup>, Laurent Micouin<sup>b</sup>, Jean-Luc Pernodet<sup>a</sup>, and Sylvie Lautru<sup>a#</sup>

<sup>a</sup> Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

<sup>b</sup> Nouvelles méthodes de synthèse pour l'interface chimie-biologie, CNRS, Université Paris Descartes, 75006, Paris, France

<sup>#</sup> Corresponding author: Sylvie LAUTRU, [sylvie.lautru@i2bc.paris-saclay.fr](mailto:sylvie.lautru@i2bc.paris-saclay.fr)

## Supplemental material

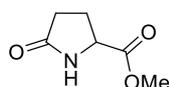
**Experimental part:** description of the synthetic strategy followed to synthesize of 5-amino-3,4-dihydro-2H-pyrrole-2-carboxylate [4]

Compound [4] was prepared according to a previously described synthetic procedure (Scheme 1 main manuscript)(Lee and Lown, 1987).

### **General remarks**

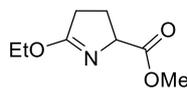
All reactions were carried out under inert atmosphere, in oven-dried glassware, using dry solvents unless otherwise specified. All commercially available compounds were purchased from Aldrich Chemical Co., Acros Organics, or Alfa Aesar and used as received. Analytical thin layer chromatography (TLC) was performed on silica gel plates (Merck 60F<sub>254</sub>) visualised either with a UV lamp (254 nm) or by using solutions of *p*-anisaldehyde/sulfuric acid/acetic acid (AcOH) in ethanol (EtOH) or KMnO<sub>4</sub>/K<sub>2</sub>CO<sub>3</sub>/AcOH in water followed by heating. Flash chromatography was performed on silica gel (60-230 mesh) unless otherwise specified. Organic extracts were dried over anhydrous MgSO<sub>4</sub>. <sup>1</sup>H (250 or 500 MHz), and <sup>13</sup>C (125 MHz) NMR spectra were recorded on a Bruker Nanobay Avance III 250 or a Bruker Avancell 500 in CDCl<sub>3</sub> or DMSO-*d*<sub>6</sub>, and calibrated using residual undeuterated solvent as an internal reference. Chemical shifts are reported in ppm, multiplicities are indicated by s (singlet), d (doublet), t (triplet), q (quartet), p (pentet), and m (multiplet or overlap of nonequivalent resonances), dd (doublet of doublets), td (triplet of doublets), and br (broad signal). Coupling constants, *J*, are reported in hertz (Hz). All NMR spectra were obtained at 300 K unless otherwise specified.

### **Synthesis of 5-oxoproline methyl ester [7]**



DL-Pyrroglutamic acid (20 g, 154 mmol, 1 eq.) was dissolved in dry methanol (70 mL). The solution was cooled to 10 °C using an ice-salt water bath, then thionyl chloride (22 mL, 308 mmol, 2 eq.) was added dropwise via a syringe. Dry DMF (0.3 mL, 3.5 mmol, 2 mol%) was finally added. The reaction was allowed to warm up to rt and stirring was continued for 24 h. The solvent was finally removed under reduced pressure and the crude product was purified by distillation (130-150 °C, 20 mbar). Pure compound [7] was isolated as a colorless oil (19.7 g, 138 mmol, 89% yield). Racemic compound. <sup>1</sup>H NMR (250 MHz, CDCl<sub>3</sub>): δ 7.38 (s br, 1H), 4.25 – 4.19 (m, 1H), 3.70 (s, 3H), 2.45 – 2.27 (m, 3H), 2.20 – 2.12 (m, 1H) ppm. Spectroscopic data were consistent with the literature data for this compound (Drauz et al., 1986).

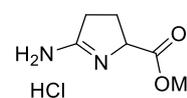
### **Synthesis of methyl 5-ethoxy-3,4-dihydro-2H-pyrrole-2-carboxylate [8]**



To a stirred solution of ester 7 (5.3 g, 37 mmol, 1 eq.) in dry DCM (50 mL) was added triethyloxonium tetrafluoroborate (1 M solution in dry DCM, 50 mL, 53 mmol, 1.4 eq.). The resulting mixture was stirred at room temperature for 48 h under an argon atmosphere. The reaction was then quenched with a saturated solution of NaHCO<sub>3</sub> (40 mL). Once the effervescence had subsided, the organic layer was separated and the aqueous phase was extracted with DCM (2 x 30 mL). The combined organic layers were dried over MgSO<sub>4</sub>, filtered and concentrated under reduced pressure to afford product [8] as a yellow oil (6.2 g, 36 mmol, 98% yield). This substrate was used in the following synthetic steps without any further purification. Racemic compound. <sup>1</sup>H NMR (250 MHz, CDCl<sub>3</sub>): δ 4.54 – 4.47 (m, 1H), 4.26 – 4.18 (m, 2H), 3.71 (s, 3H), 2.59 – 2.46 (m, 2H), 2.33 – 2.24 (m, 1H), 2.19 – 2.10 (m,

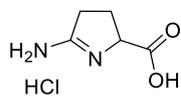
1H), 128 (t,  $J = 7.1$  Hz, 3H) ppm. Spectroscopic data were consistent with the literature data for this compound (Lee and Lown, 1987).

#### Synthesis of methyl 5-amino-3,4-dihydro-2H-pyrrole-2-carboxylate hydrochloride [9]



A stirred solution of compound [8] (5.7 g, 33 mmol, 1 eq.), and anhydrous  $\text{NH}_4\text{Cl}$  (1.9 g, 35 mmol, 1.05 eq.) in dry methanol (30 mL) was heated at reflux for 5 h under an argon atmosphere. The solvent was then removed under reduced pressure. The crude product was purified by recrystallization from DCM/cyclohexane. Pure compound [9] was isolated as a white solid (3.6 g, 20.1 mmol, 61% yield). Racemic product.  $^1\text{H NMR}$  (250 MHz,  $\text{DMSO-}d_6$ ):  $\delta$  4.59 (dd,  $J = 9.1, 4.8$  Hz, 1H), 3.68 (s, 3H), 2.82 (t,  $J = 8.6$  Hz, 2H), 2.45 – 2.38 (m, 1H), 2.11 – 2.04 (m, 1H) ppm. Spectroscopic data were consistent with the literature data for this compound (Lee and Lown, 1987).

#### Synthesis of 5-amino-3,4-dihydro-2H-pyrrole-2-carboxylate hydrochloride [4]



Derivative [9] (1.14 g, 6.38 mmol, 1 eq.) was dissolved in an aqueous solution of hydrochloric acid (10% v/v, 50 mL), and stirred at 50 °C for 3 h. Toluene (15 mL), was then added, and the mixture was concentrated under reduced pressure. The crude product was finally collected, and dried under high vacuum at 65 °C to afford acid [4] as a white solid (1.1g, 6.37 mmol, quantitative yield). Racemic product.  $^1\text{H NMR}$  (500 MHz,  $\text{DMSO-}d_6$ ):  $\delta$  13.27 (s br, 1H), 9.84 (s br, 1H), 9.53 (s br, 1H), 9.15 (s br, 1H), 4.49 (dd,  $J = 9.1, 4.9$  Hz, 1H), 2.83 – 2.79 (m, 2H), 2.46 – 2.28 (m, 1H), 2.10 – 2.07 (m, 1H) ppm.  $^{13}\text{C NMR}$  (125MHz,  $\text{DMSO-}d_6$ ): 172.1 (C), 171.5 (C), 60.1 (CH), 29.4 ( $\text{CH}_2$ ), 24.7 ( $\text{CH}_2$ ) ppm. Spectroscopic data were consistent with the literature data for this compound (Lee and Lown, 1987).

**Table S1:** Strains used in this study

Strain	Description	Reference
<i>Escherichia coli</i> DH5a	General cloning host	Promega
<i>E. coli</i> ET12567 pUZ8002	Host strain for conjugation from <i>E. coli</i> to <i>Streptomyces</i>	(Flett et al., 1997)
<i>Streptomyces venezuelae</i> ATCC 14583	Anthelvencin producer	(Probst et al., 1965)
ANT007	<i>S. venezuelae</i> replacement mutant of <i>ant8</i>	This study
ANT008	<i>S. venezuelae</i> replacement mutant of <i>ant23</i>	This study
ANT009	<i>S. venezuelae</i> replacement mutant of <i>ant24</i>	This study
ANT012	<i>S. venezuelae</i> with pANT012 overproducing anthelvencin and methylanthelvencin	This study
ANT013	ANT008 containing pANT013	This study

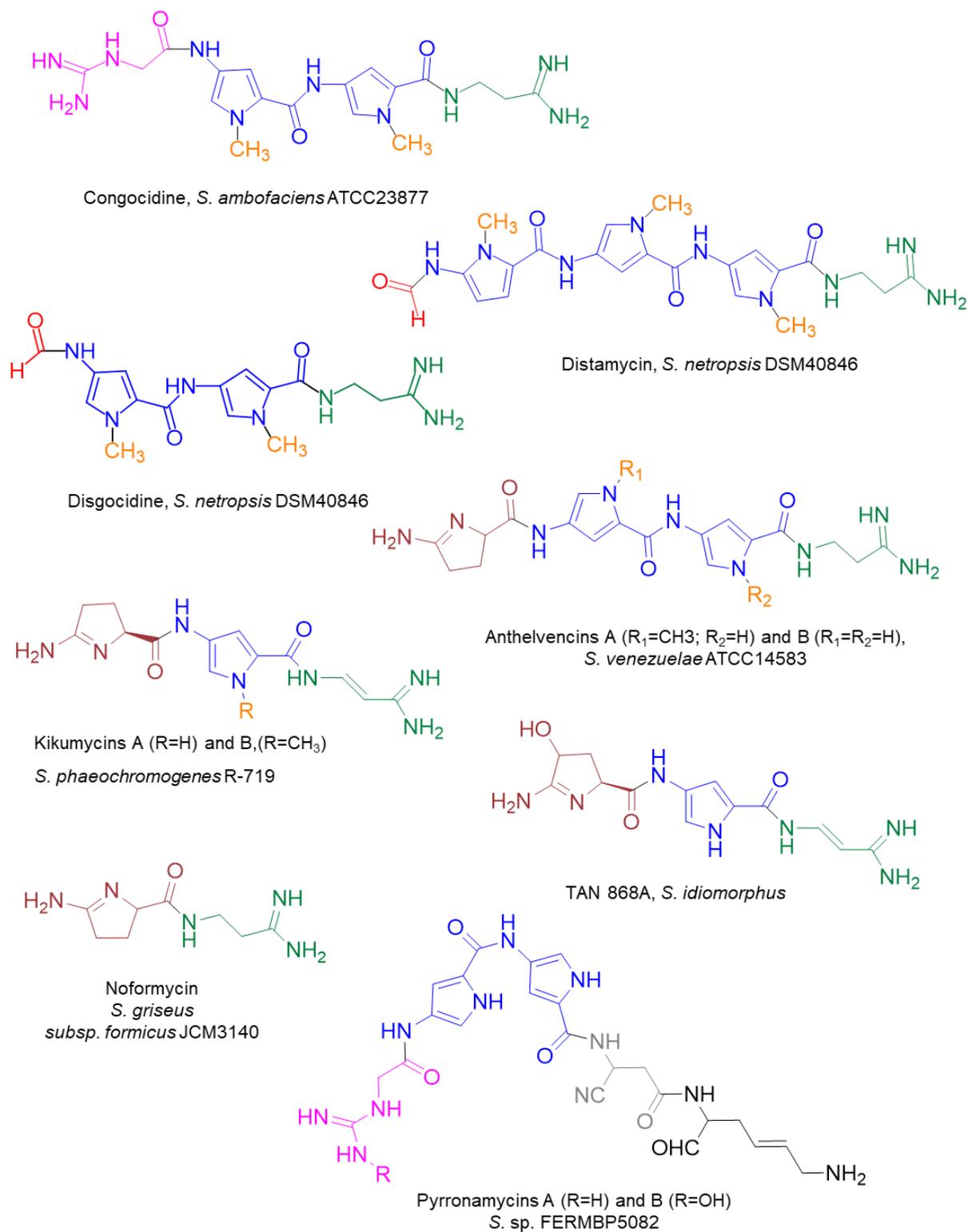
**Table S2:** Plasmids used in this study

Plasmid	Description	Reference
pCR®-Blunt	<i>E. coli</i> cloning vector	Invitrogen
pOSV400	Suicide vector for gene disruption in <i>Streptomyces</i>	(Boubakri et al., 2015)
pOSV802	Plasmid containing apramycin resistance and $\varphi$ C31 integrase	(Aubry et al., 2019)
pOSV806	Plasmid containing hygromycin resistance and $\varphi$ C31 integrase	(Aubry et al., 2019)
pW60	Source of the <i>aac(3)IV</i> cassette	(Corre et al., 2008)
pANT001	Plasmid pCR®-Blunt containing a 1.8 kb DNA fragment upstream of <i>ant8</i>	This study
pANT002	Plasmid pCR®-Blunt containing a 2.0 kb DNA fragment downstream of <i>ant8</i>	This study
pANT003	Plasmid pCR®-Blunt containing a 1.8 kb DNA fragment upstream of <i>ant23</i>	This study
pANT004	Plasmid pCR®-Blunt containing a 1.8 kb DNA fragment downstream of <i>ant23</i>	This study
pANT005	Plasmid pCR®-Blunt containing a 1.7 kb DNA fragment upstream of <i>ant24</i>	This study
pANT006	Plasmid pCR®-Blunt containing a 1.4 kb DNA fragment downstream of <i>ant24</i>	This study
pANT007	pOSV400 derivative used for the replacement of <i>ant8</i> by the <i>aac(3)IV</i> cassette	This study
pANT008	pOSV400 derivative used for the replacement of <i>ant23</i> by the <i>aac(3)IV</i> cassette	This study
pANT009	pOSV400 derivative used for the replacement of <i>ant24</i> by the <i>aac(3)IV</i> cassette	This study
pANT011	Plasmid pCR®-Blunt containing a 2.0 kb fragment containing <i>ant23</i> and <i>ant24</i>	This study
pCEA005	pOSV802 containing <i>rps(LTP)</i> p and the <i>tipA</i> RBS	(Aubry et al., 2019)
pANT012	pCEA005 derivative used for the overexpression of <i>ant23</i> and <i>ant24</i>	This study
pANT013	pOSV806 plasmid containing <i>ant23-ant24</i> under the <i>rps(LTP)</i> p with hygromycin resistance	This study

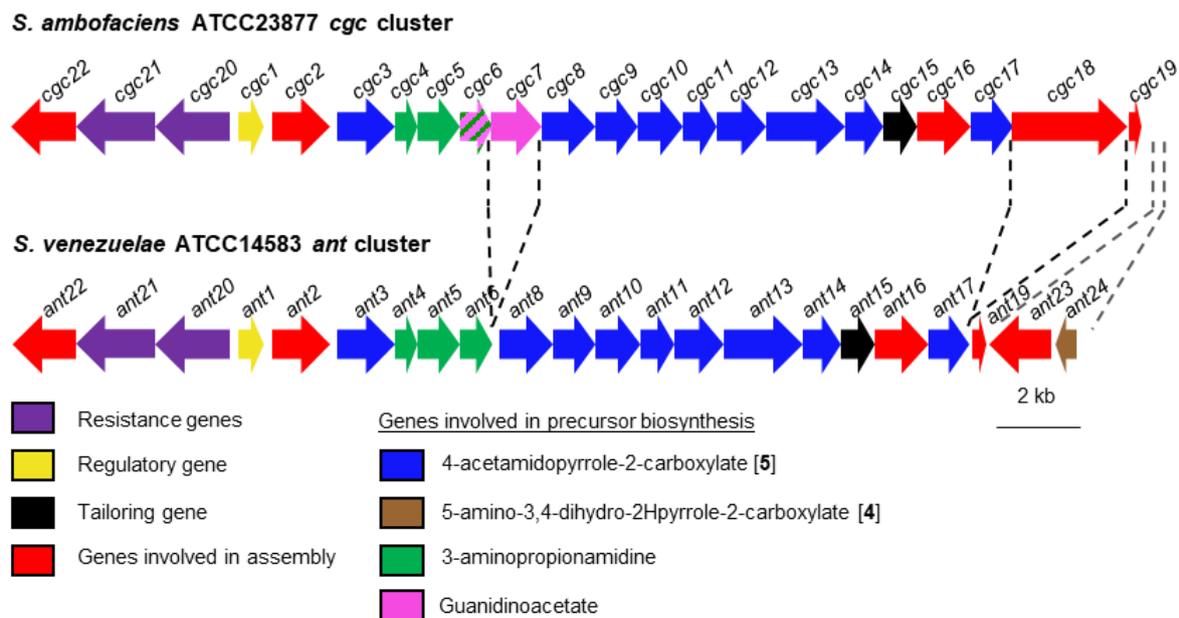
**Table S3:** Oligonucleotides used in this study

Name	Sequence	Description
CEA001	CAGTA <u>AAGCTT</u> CATGCGGTCGCGTACTGATG	Forward primer for region upstream of <i>ant8</i> , <i>Hind</i> III site underlined
CEA002	CAGT <u>CTCGAGT</u> GGGCCAGGAAGCAGTGATG	Reverse primer for region upstream of <i>ant8</i> , <i>Xho</i> I site underlined
CEA003	CAGT <u>ACTAGTCTT</u> GTCGTGGCCGTGTTCTC	Forward primer for region downstream of <i>ant8</i> , <i>Spe</i> I site underlined
CEA004	CAGTA <u>AAGCTT</u> GGCCGTGCGTAAGAAGATCC	Reverse primer for region downstream of <i>ant8</i> , <i>Hind</i> III site underlined
CEA005	CAGT <u>CTCGAGACCA</u> AGGGAGTCGAGGAATG	Forward primer for region upstream of <i>ant23</i> , <i>Xho</i> I site underlined
CEA006	CAGTA <u>AAGCTT</u> CCCTAGTAGCTCGAATGCAC	Reverse primer for region upstream of <i>ant23</i> , <i>Hind</i> III site underlined
CEA007	CAGTA <u>AAGCTT</u> TACATGCCGCTGCTCACAC	Forward primer for region downstream of <i>ant23</i> , <i>Hind</i> III site underlined
CEA008	CAGT <u>ACTAGTAA</u> CCTGATCGGCGCCTACAC	Reverse primer for region downstream of <i>ant23</i> , <i>Spe</i> I site underlined
CEA009	CAGT <u>CTCGAGCAC</u> CGAGATCGGTCTCTACC	Forward primer for region upstream of <i>ant24</i> , <i>Xho</i> I site underlined
CEA010	CAGTA <u>AAGCTT</u> CGCCCGGCTTCTATAAAACC	Reverse primer for region upstream of <i>ant24</i> , <i>Hind</i> III site underlined
CEA011	CAGTA <u>AAGCTT</u> TCTCACTCCCGGTGTGCATTCCG	Forward primer for region downstream of <i>ant24</i> , <i>Hind</i> III site underlined
CEA012	CAGT <u>ACTAGT</u> CGGCCGCCCTCTTCTGACC	Reverse primer for region downstream of <i>ant24</i> , <i>Xho</i> I site underlined
A5	CGACGTGGCAGGATCGAACG	Internal to <i>aac</i> (3)IV, used to confirm correct replacement in mutants
A6	GTCAACTGGGCCGAGATCCG	Internal to <i>aac</i> (3)IV, used to confirm correct replacement in mutants
CEA013	GTGAACTGATGCGCACCGAC	Control of the correct replacement of <i>ant8</i>
CEA014	GGGCTTTCTCCGTTTGCTTC	Control of the correct replacement of <i>ant8</i>
CEA015	AGAGCCTGTTCCGGCACCTG	Control of the correct replacement of <i>ant8</i> around the resistance cassette
CEA016	CCAGGTGCAGGCCGATGAAG	Control of the correct replacement of <i>ant8</i> around the resistance cassette
CEA017	TCGGCCTCTTCGTGAACCTG	Control of the correct replacement of <i>ant23</i>
CEA018	CACGGCATGACGCTGATGTG	Control of the correct replacement of <i>ant23</i> around the resistance cassette
CEA019	TTCCCTCGCGGAGAAGGGCTG	Control of the correct replacement of <i>ant23</i> around the resistance cassette

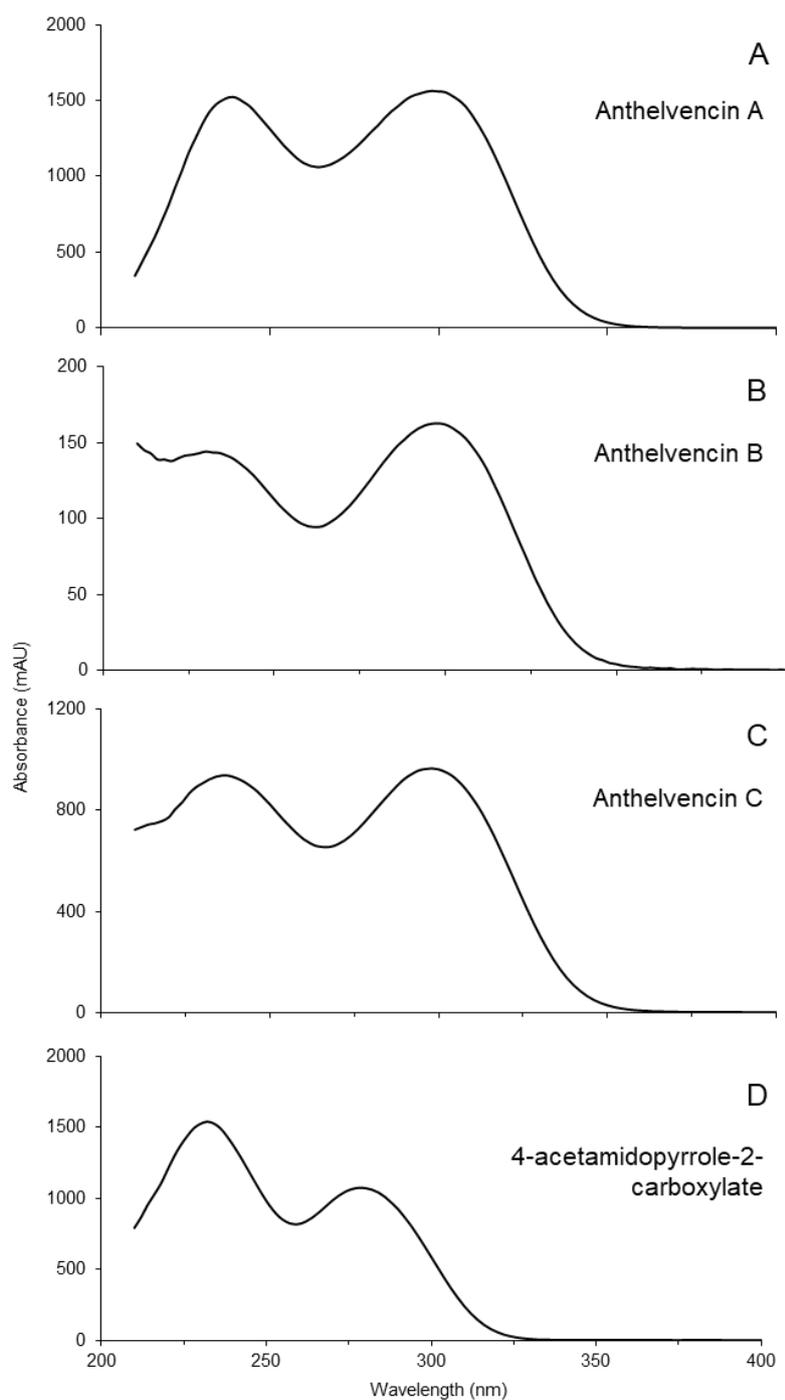
Name	Sequence	Description
CEA020	CGGGCACTTCAGTACCGGTC	Control of the correct replacement of <i>ant24</i>
CEA021	ATGCTGCGGAGACTCAGCAC	Control of the correct replacement of <i>ant24</i>
CEA022	GTGTCGGGCATGCTTTCCTG	Control of the correct replacement of <i>ant24</i>
CEA034	<u>ATGCATGCGGCCGCTGCTAGCGATGGCGAGG</u> TTTTATAGAAGCC	Amplification of the region <i>ant23-ant24</i> , <i>Nsi</i> I, <i>Not</i> I and <i>Nhe</i> I sites underlined
CEA035	<u>CCTAAGGCGGCCGCTACTAGTGTGTGAGCAG</u> CGGCATGTG	Amplification of the region <i>ant23-ant24</i> , <i>Afl</i> II, <i>Not</i> I and <i>Spe</i> I sites underlined
CEA_vec_ seq14	ATTTTCAGTGCAATTTATCTCTTC	Confirmation of the integration of pANT012 in <i>S. venezuelae</i>
CEA_vec_ seq15	TTCGATCACGTGGGCGAAGC	Confirmation of the integration of pANT012 in <i>S. venezuelae</i>



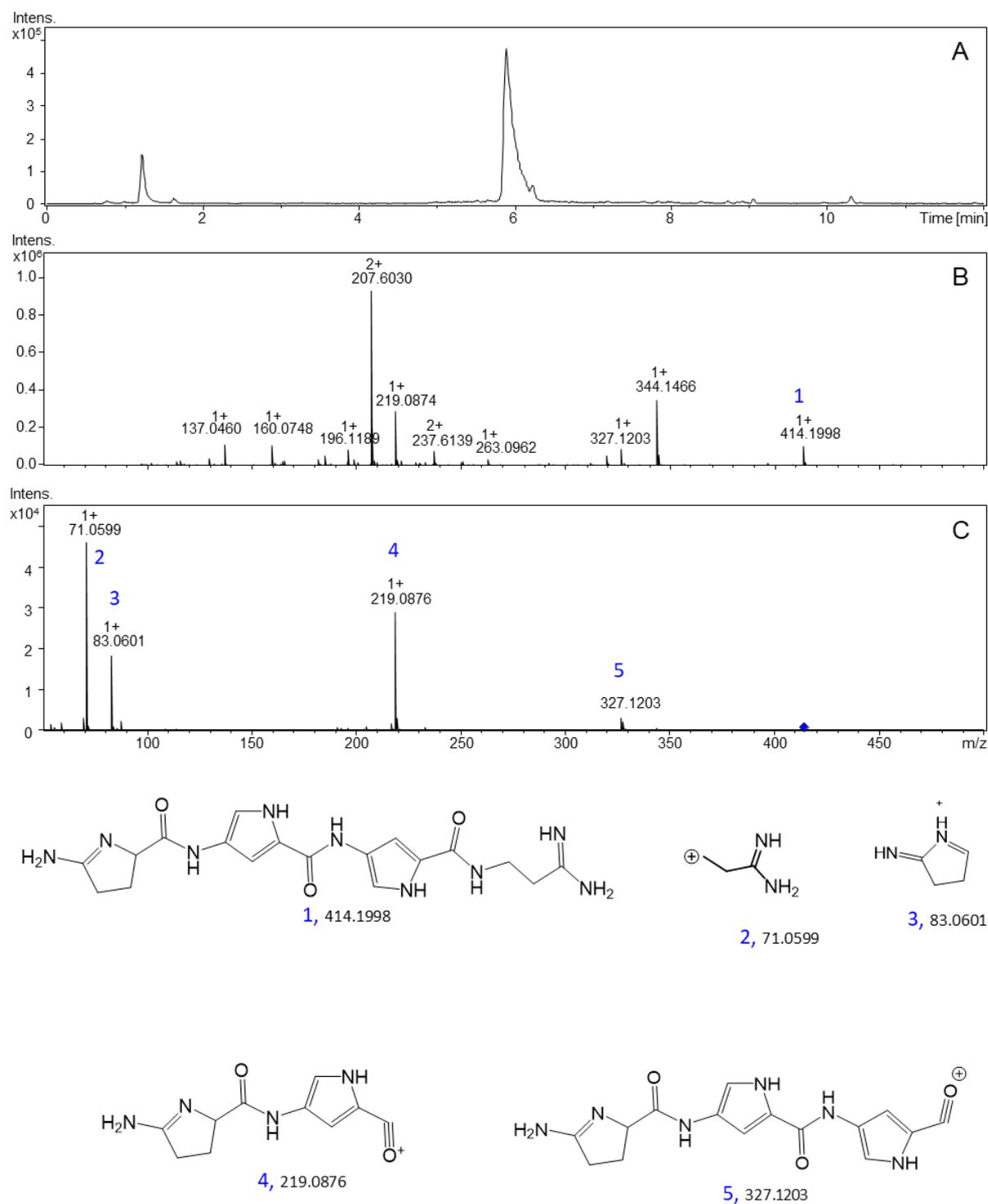
**Figure S1:** Structures of members of the pyrrolamide family and name of the *Streptomyces* producer



**Figure S2:** Genetic organization of the congocidine biosynthetic gene cluster in *S. ambofaciens* ATCC 23877, and genetic organization of the anthelvencin biosynthetic gene cluster in *S. venezuelae* ATCC 14583



**Figure S3: UV-visible spectra** of (A) anthelvencin A, Rt = 14,3 min, (B) anthelvencin B, Rt = 13,3 min, (C) anthelvencin C, Rt = 15,5 min, and (D) 4-acetamidopyrrole-2-carboxylate, Rt = 11,5 min.



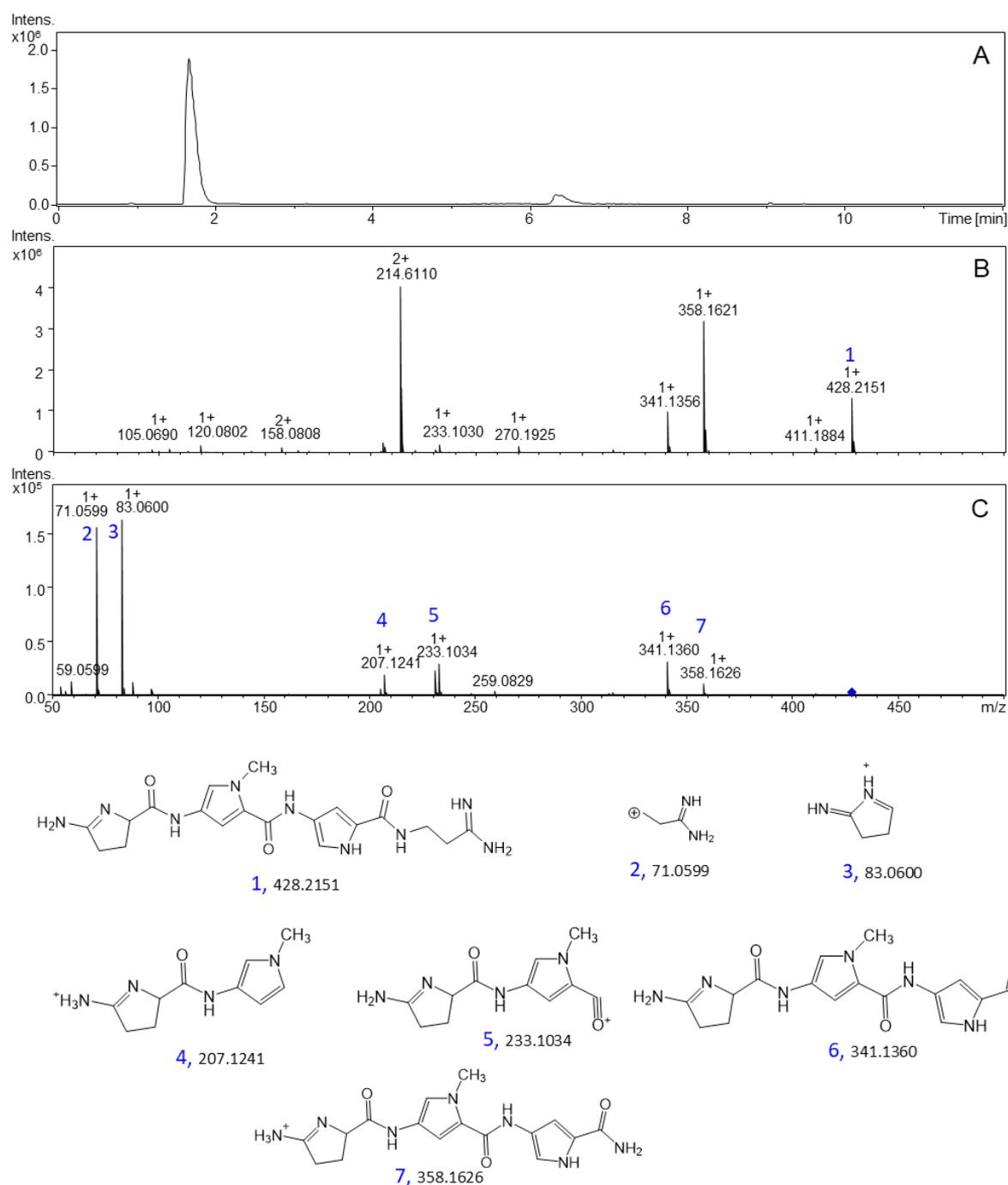
**Figure S4: Identification of anthelvencin B (Peak II) from HR-MS and HR-MS<sup>2</sup>**

(A) EIC 414.2 +All MS

(B) HR-MS spectrum of the peak at 1.3 min in the chromatogram (A)

(C) Fragmentation of peak (1) ( $m/z = 414.1998$ )

The putative structure of the obtained fragments are indicated below the spectra.



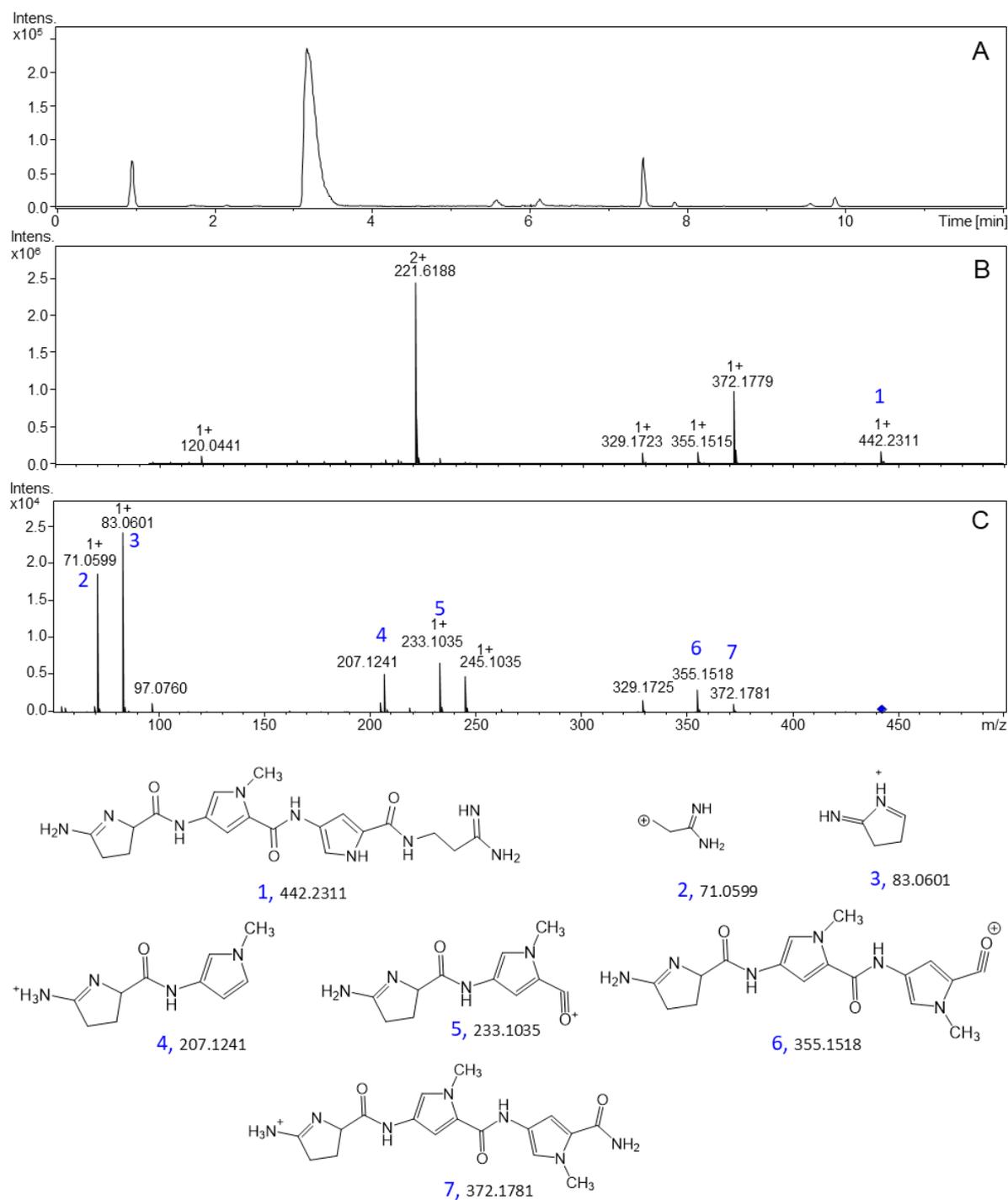
**Figure S5: Identification of anthelvencin A (peak III) from HR-MS and HR-MS<sup>2</sup>**

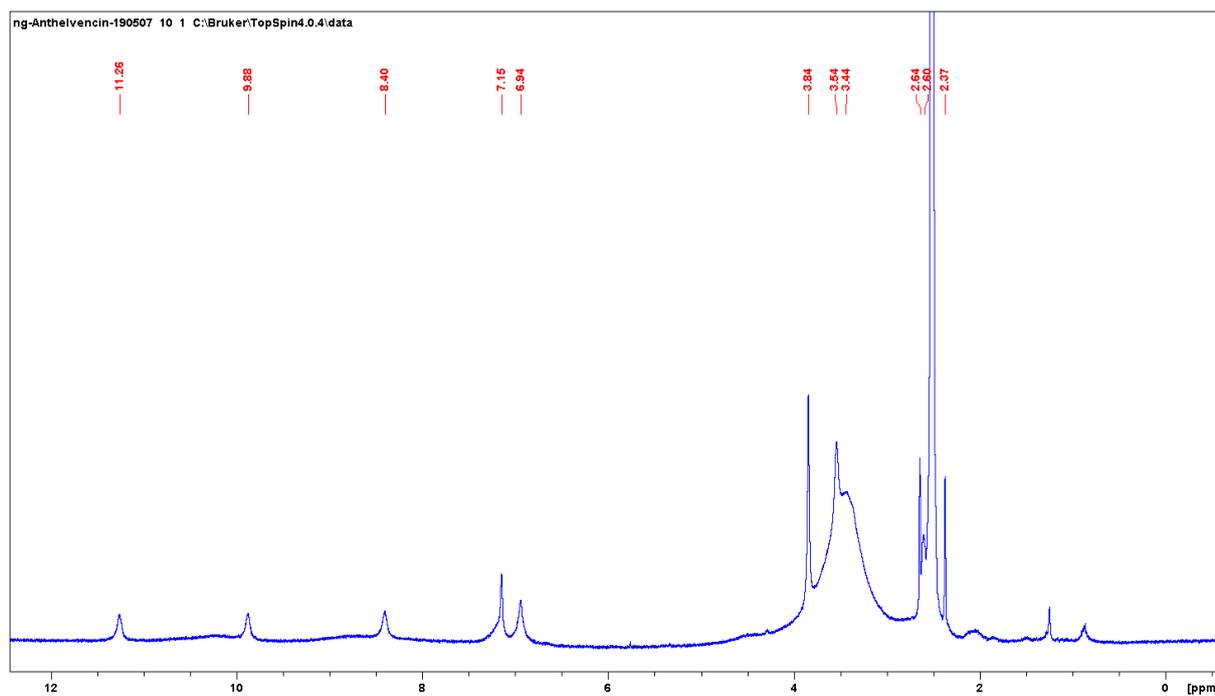
(A) EIC 428.2100 +All MS

(B) HR-MS spectrum of the peak at 1.8 min in the chromatogram (A)

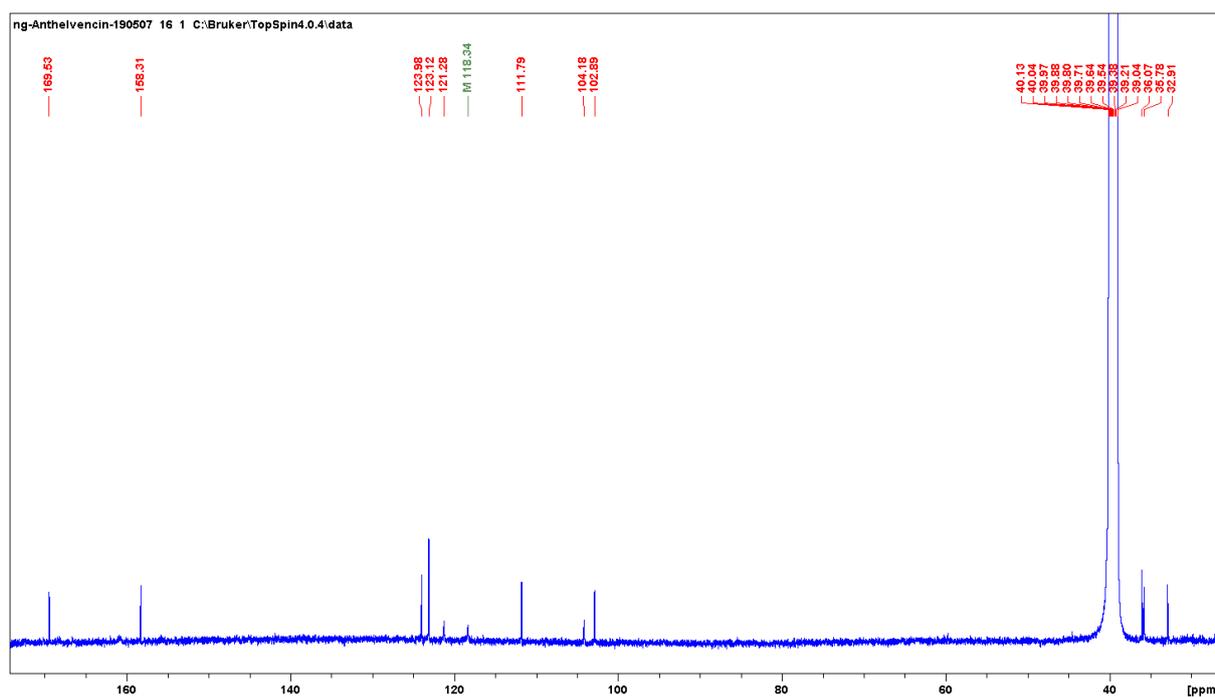
(C) Fragmentation of peak (1) (m/z = 428.2151)

The putative structure of the obtained fragments are indicated below the spectra.



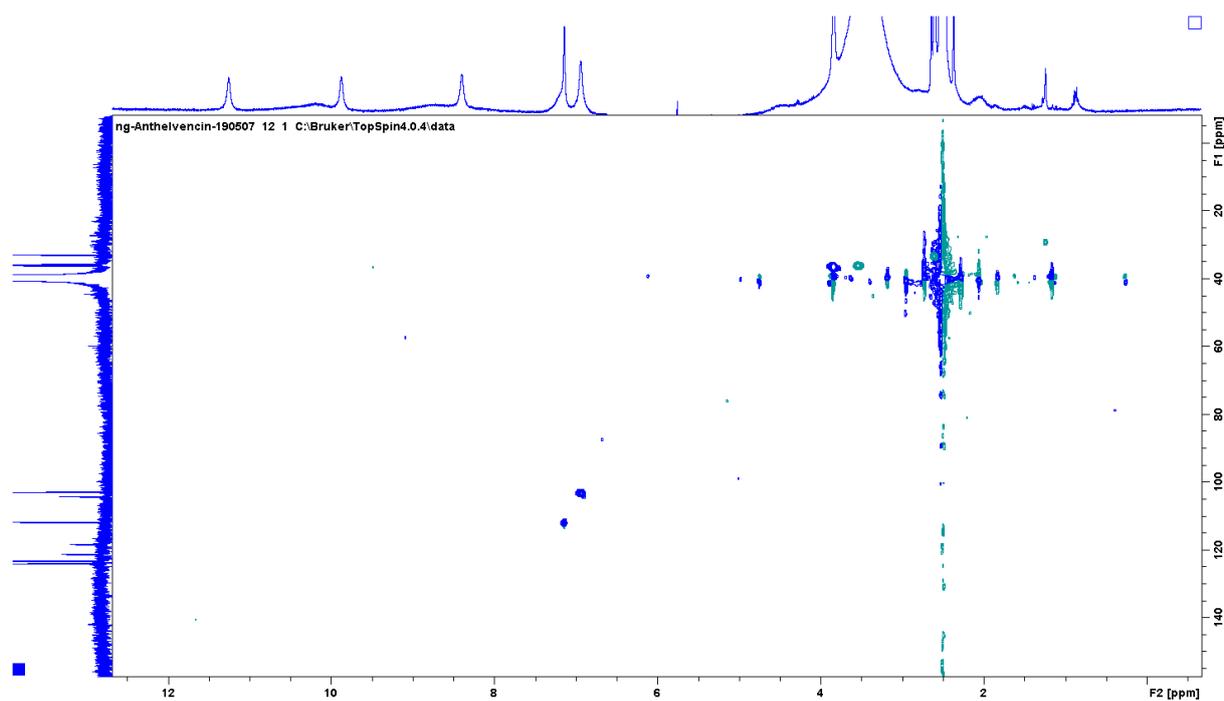


A)  $^1\text{H}$  NMR spectrum, anthelvencin in DMSO

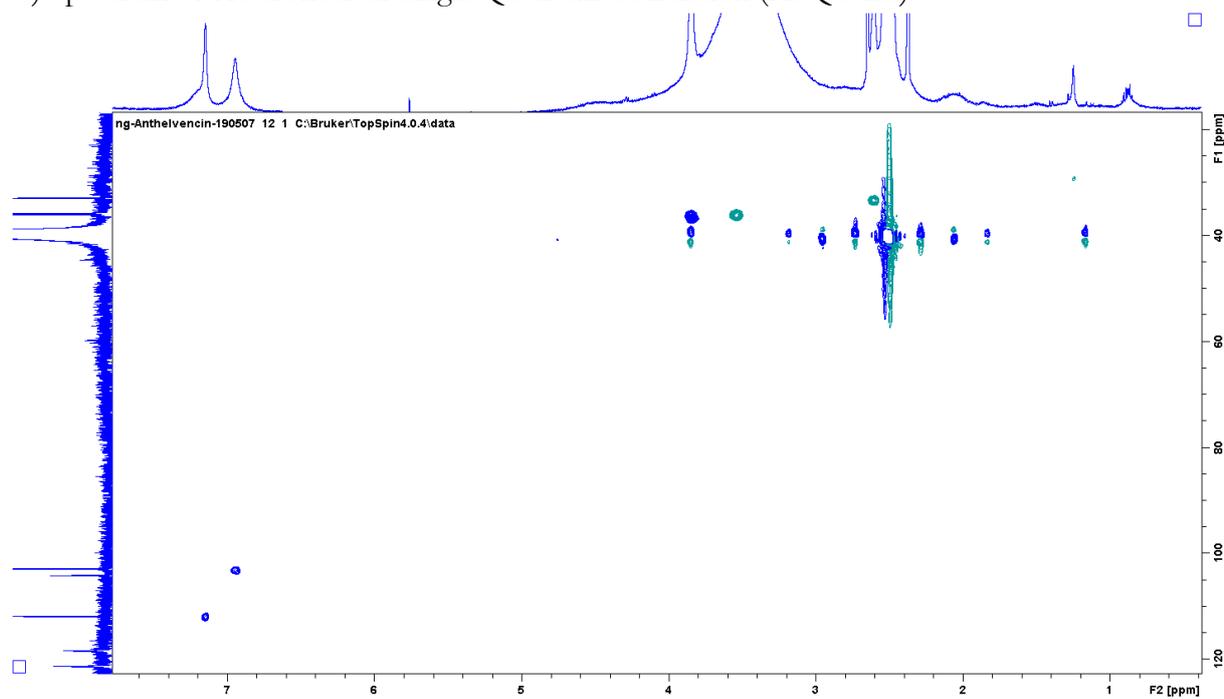


B)  $^{13}\text{C}$  NMR spectrum, anthelvencin in DMSO

Figure S7: Nuclear Magnetic Resonance (NMR) spectra of anthelvencin A

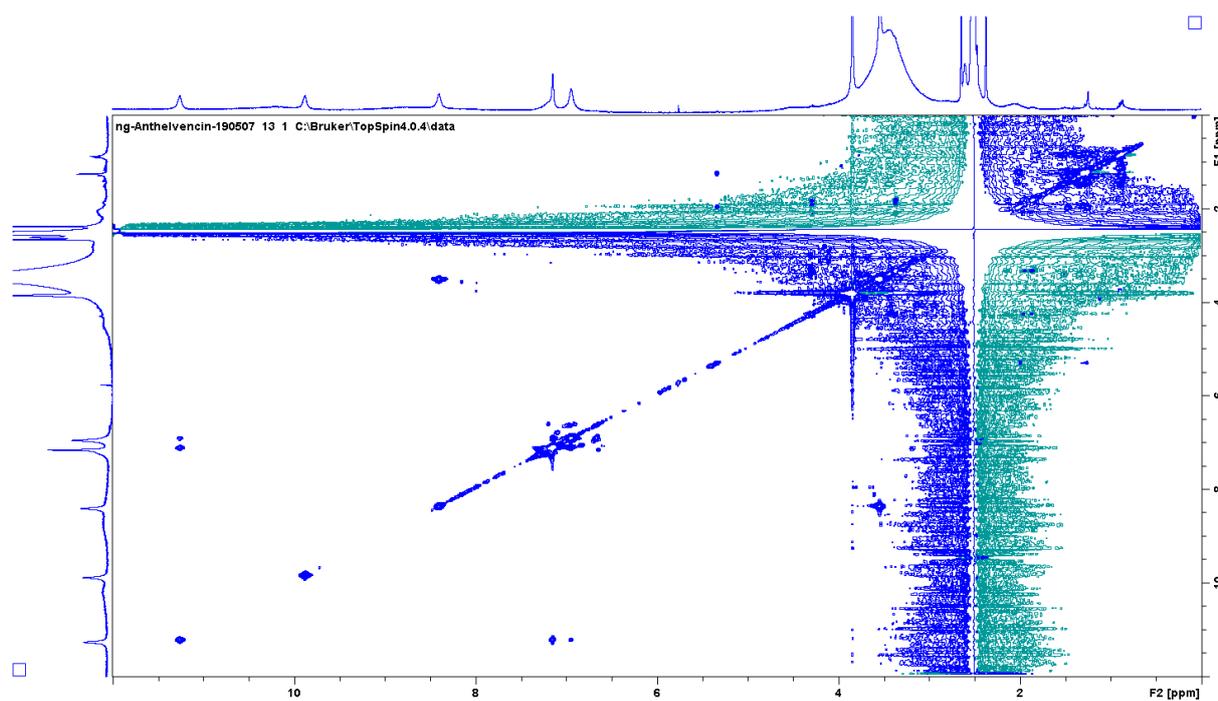


C) Spectrum of Heteronuclear Single Quantum Correlation (HSQC-Ed)

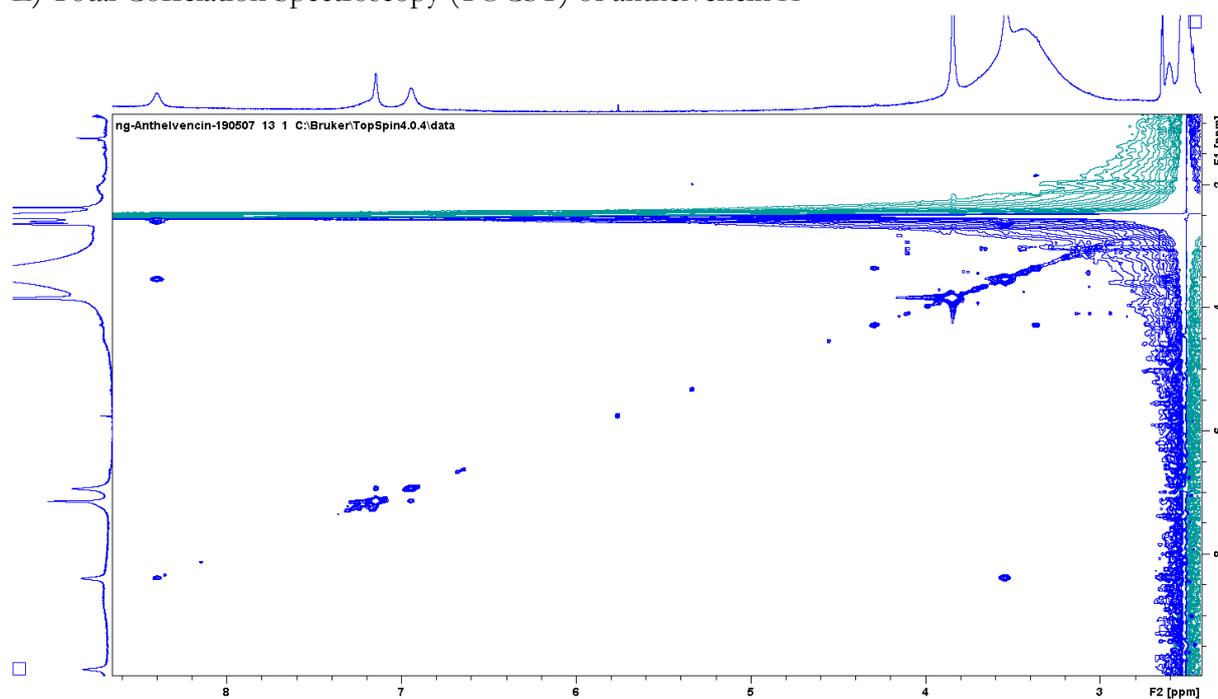


D) Detail of the spectrum of Heteronuclear Single Quantum Correlation (HSQC-Ed)

Figure S7: Nuclear Magnetic Resonance (NMR) spectra of anthelvencin A



E) Total Correlation Spectroscopy (TOCSY) of anthelvencin A



F) Detail of Total Correlation Spectroscopy (TOCSY) of anthelvencin A

**Figure S7: Nuclear Magnetic Resonance (NMR) spectra of anthelvencin A**

## **References:**

Aubry, C., Pernodet, J.-L., and Lautru, S. (2019). A set of modular and integrative vectors for synthetic biology in *Streptomyces*. *Appl. Environ. Microbiol.* Aug 1;85(16).

Boubakri, H., Seghezzi, N., Duchateau, M., Gominet, M., Kofroňová, O., Benada, O., Mazodier, P., and Pernodet, J.-L. (2015). The absence of pupylation (prokaryotic ubiquitin-like protein modification) affects morphological and physiological differentiation in *Streptomyces coelicolor*. *J. Bacteriol.* *197*, 3388–3399.

Corre, C., Song, L., O'Rourke, S., Chater, K.F., and Challis, G.L. (2008). 2-Alkyl-4-hydroxymethylfuran-3-carboxylic acids, antibiotic production inducers discovered by *Streptomyces coelicolor* genome mining. *Proc. Natl. Acad. Sci. U. S. A.* *105*, 17510–17515.

Drauz, K., Kleemann, A., Martens, J., Scherberich, P., and Effenberger, F. (1986). Amino acids. 7. A novel synthetic route to L-proline. *J. Org. Chem.* *51*, 3494–3498.

Flett, F., Mersinias, V., and Smith, C.P. (1997). High efficiency intergeneric conjugal transfer of plasmid DNA from *Escherichia coli* to methyl DNA-restricting streptomycetes. *FEMS Microbiol. Lett.* *155*, 223–229.

Lee, M., and Lown, J.W. (1987). Synthesis of (4S)- and (4R)-methyl 2-amino-1-pyrroline-5-carboxylate and their application to the preparation of (4S)-(+)- and (4R)-(-)-dihydrokikumycin B. *J. Org. Chem.* *52*, 5717–5721.

Probst, G.W., Hoehn, M.M., and Woods, B.L. (1965). Anthelvencins, new antibiotics with anthelmintic properties. *Antimicrob. Agents Chemother.* *5*, 789–795.

## Chapter I perspectives:

This chapter presented the characterization of the anthelvencin biosynthetic gene cluster and the isolation of new anthelvencin metabolite, anthelvencin C. In the course of our study, based on HR-MS<sup>2</sup> data, we realized that the published structure of anthelvencin A was most likely incorrect.

In an attempt to better characterize this structure, we purified this metabolite and analyzed it by NMR. However, the obtained NMR signals are of poor quality (broad peaks), suggesting the presence of a paramagnetic element. EPR analysis confirmed the presence of metal, possibly manganese. Due to time constraints, a new purification of anthelvencin A could not be carried out. In the next future, this would constitute the main priority, to repeat the NMR analysis and finish the work presented here. In a biological point of view, it would be interesting to determine whether the covalent binding of manganese participate to the biological function of anthelvencin.

# Chapter II - Modular and Integrative Vectors for Synthetic Biology Applications in *Streptomyces* spp.

## Chapter II introduction:

In this chapter, I report my work on the construction of modular integrative vectors. This set of vectors was built to facilitate the construction and assembly of gene cassettes necessary for combinatorial biosynthesis experiments. Since such standardized vectors are scarce in the field of actinobacterial specialized metabolism, we designed them to be flexible and easy to adapt to various synthetic biology applications in *Streptomyces* species.

This work was published in ‘Applied Environmental Microbiology’ journal, and I present here the published manuscript:

Aubry, C., Pernodet, J.-L., and Lautru, S. (2019). A set of modular and integrative vectors for synthetic biology in *Streptomyces*. *Appl. Environ. Microbiol.* Aug 1;85(16).

# Modular and Integrative Vectors for Synthetic Biology Applications in *Streptomyces* spp.

Céline AUBRY<sup>a</sup>, Jean-Luc PERNODET<sup>a</sup> and Sylvie LAUTRU<sup>a#</sup>

<sup>a</sup> Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

<sup>#</sup> Corresponding author: Sylvie LAUTRU, [sylvie.lautru@i2bc.paris-saclay.fr](mailto:sylvie.lautru@i2bc.paris-saclay.fr)

## ABSTRACT

With the development of synthetic biology in the field of (actinobacteria) specialized metabolism, new tools are needed for the design or refactoring of biosynthetic gene clusters. If libraries of synthetic parts (such as promoters or ribosome binding sites) and DNA cloning methods have been developed, to our knowledge, not many vectors designed for the flexible cloning of biosynthetic gene clusters have been constructed.

We report here the construction of a set of 12 standardized and modular vectors designed to afford the construction or the refactoring of biosynthetic gene clusters in *Streptomyces* species, using a large panel of cloning methods. Three different resistance cassettes and four orthogonal integration systems are proposed. In addition, FLP recombination target sites were incorporated to allow the recycling of antibiotic markers and to limit the risks of unwanted homologous recombination in *Streptomyces* strains when several vectors are used. The functionality and proper integration of the vectors in three commonly used *Streptomyces* strains, as well as the functionality of the Flp-catalyzed excision were all confirmed.

To illustrate some possible uses of our vectors, we refactored the albonoursin gene cluster from *Streptomyces noursei* using the Biobrick assembly method. We also used the seamless Ligase Chain Reaction cloning method to assemble a transcription unit in one of the vectors and genetically complement a mutant strain.

## IMPORTANCE

One of the strategies employed today to obtain new bioactive molecules with potential applications for human health (for example, antimicrobial or anticancer agents) is synthetic biology. Synthetic biology is used to biosynthesize new unnatural specialized metabolites or to force the expression of otherwise silent natural biosynthetic gene clusters. To assist the development of synthetic biology in the field of specialized metabolism, we constructed and are offering to the community a set of vectors that were intended to facilitate DNA assembly and integration in actinobacterial chromosomes. These vectors are compatible with various DNA cloning and assembling methods. They are standardized and modular, allowing the easy exchange of a module by another one of the same nature. Although designed for the assembly or the refactoring of specialized metabolite gene clusters, they have a broader potential utility, for example, for protein production or genetic complementation.

**KEYWORDS** *Streptomyces*, synthetic biology

## INTRODUCTION

Synthetic biology is a domain of biotechnology that emerged at the beginning of the 21<sup>st</sup> century. It aims, for one part, at the rational engineering of biological systems to confer on them new functions. In the field of specialized metabolism, synthetic biology aims first at cloning and refactoring of silent (cryptic) biosynthetic gene clusters, to afford the expression of genes and the production of metabolites that otherwise cannot be isolated and purified (1–3). Second, it is usually the method of choice for the synthesis of "unnatural natural products". In this case, it consists either in the design and assembly of new biosynthetic gene clusters (4) or in the engineering of biosynthetic enzymes such as the modular nonribosomal peptide synthetases (NRPS) (5–7) and polyketide synthases (PKS) (8, 9). Such approaches are often referred to as combinatorial biosynthesis.

The development of synthetic biology in the field of specialized metabolism requires the development of dedicated tools and methods. In particular, it requires hosts (chassis) optimized for the production of specialized metabolites, libraries of synthetic DNA parts such as promoters, ribosome binding sites (RBSs) or terminators, and vectors and DNA assembly methods for *de novo* assembly of gene clusters. Several *Streptomyces* strains, such as *Streptomyces calicolor* (10), *Streptomyces avermitilis* (11) or *Streptomyces albus* (12, 13) have been optimized as chassis for the heterologous production of specialized metabolites. High-producing industrial strains have also been reported for the successful heterologous production of specialized metabolites (14). In parallel, efforts have been made to construct libraries of synthetic promoters (15–18) and of RBSs (15).

Many DNA assembly methods have been proposed and used so far for the assembly of DNA fragments, and more specifically for the assembly of specialized metabolite biosynthetic gene clusters. These methods are mainly based on the existence of homology regions at the extremities of the fragments to be assembled, on the use of restriction enzymes or on the use of site-specific recombinases. Examples of homology-based methods include the one pot isothermal

assembly (19), the ligase cycling reaction (LCR, (20)) and the Direct Pathway Cloning (DiPaC, (3) for *in vitro* assembly, and the “DNA assembler” (21) based on transformation-associated recombination (TAR) in yeast or the “linear plus circular homologous recombination” (LCHR) method (used in the AGOS system, (22)) for *in vivo* assembly. The first restriction enzyme based DNA assembly method was the Biobrick assembly, based on the utilization of four restriction enzymes, two of which generate compatible cohesive ends (23). Other similar cloning methods based on the assembly of basic parts (promoter, coding sequence, terminator...) into transcriptional units that can then be assembled together have since been developed (Golden Gate (24); Modular Cloning “MoClo” (25); GoldenBraid 2.0 (26) ). Finally, Olorunniji and colleagues recently established a DNA assembly method based on the use of site-specific integrases and orthogonal pairs of *att* sites (27).

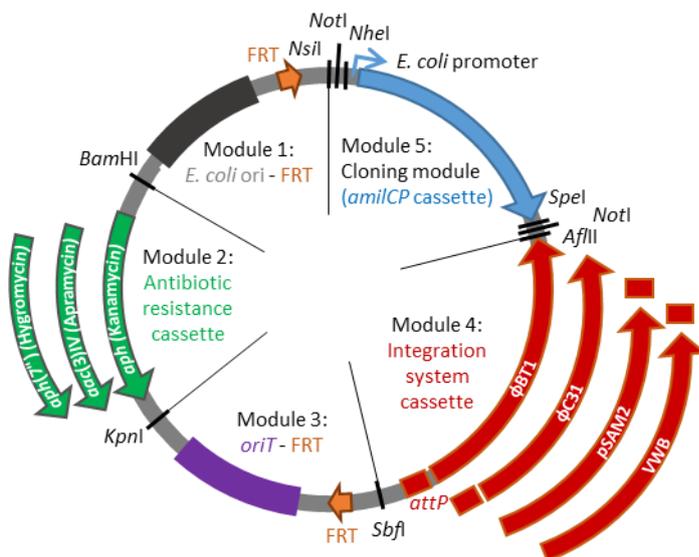
While many DNA assembly methods have been developed, none is universal and adapted to all experimental situations. Indeed, some methods are more suitable to the assembly of (large) transcriptional units together (restriction enzyme based methods, leaving a scar sequence but not requiring challenging PCRs of large and/or GC-rich fragments). Other are better suited to the assembly of the various elements of a transcriptional unit (homology-based methods allowing the precise positioning of the different elements without scar sequences). The size (from a few kilobases to more than 100 kb), the GC content and the presence and number of regions presenting relatively high degrees of sequence similarities (in NRPS or PKS genes for example) can vary a lot depending on the specialized metabolite gene cluster of interest. Thus, different experimental settings are likely to require different cloning approaches or even a combination of approaches. Therefore, the vectors used for cloning need to be flexible and adapted or easily adaptable to various assembly methods. It has been proposed that vectors built for synthetic biology should follow a standard and modular format (SEVA plasmids, (28)), allowing a rapid and easy exchange of a module by another one. Yet, in the field of specialized metabolite synthetic biology, not many of such vectors have been constructed. To our knowledge, one of the rare attempts was carried out by Phelan and colleagues (29) for the expression of genes in *Streptomyces* species. In their study, they describe the construction of 45 vectors based on three site-specific integration systems ( $\varphi$ BT1,  $\varphi$ C31 and VWB), four antibiotic resistance genes (apramycin, spectinomycin, thiostrepton/ampicillin) and 14 promoters. These vectors were mainly designed for monocistronic gene expression, although the presence of several restriction sites could allow the assembly of a few gene cassettes.

In this study, we describe the construction of a set of 12 standardized and modular vectors, designed to allow the assembly of biosynthetic gene clusters using various cloning methods in *Streptomyces* species, prolific producers of specialized metabolites. These vectors were designed on the model of the SEVA plasmids, although the exact architecture of these plasmids could not be used for our application. The 12 vectors were proven to be functional by the verified integration in the chromosome of three commonly used *Streptomyces* species. We also illustrate two possible uses of our vectors. We first refactored the albonoursin gene cluster using biobrick assembly. Second, we genetically complemented our *cgc22* mutant strain, CGCL030 (*cgc22* is involved in congocidine biosynthesis, (30)), by constructing a gene cassette constituted of a promoter, an RBS, *cgc22*, and a terminator using ligase chain reaction assembly.

## RESULTS AND DISCUSSION

### *Design of the vectors*

The vectors were designed to meet the following specifications. It should be possible to use several vectors in the same strain (orthogonality), so different antibiotic resistance cassettes and different systems of integration at specific sites in the chromosome of *Streptomyces* should be used for the construction of the vectors. The vectors should be *E. coli*/*Streptomyces* shuttle vectors so that genetic constructions can be prepared in *E. coli* before being introduced into *Streptomyces* strains; thus, an *E. coli* origin of replication has to be included. It should be possible to introduce the vectors into *Streptomyces* strains by *E. coli*/*Streptomyces* intergeneric conjugation, so the presence of an origin of transfer is necessary. The vectors should be compatible with several cloning methods, including homology and restriction enzyme based assembly methods. Finally, the vectors should be modular and flexible, so that each module can be easily replaced by another equivalent one if needed.



**Figure 1: Schematic representation of the set of modular and integrative vectors pOSV801-pOSV812.**

The various antibiotic resistance cassettes and integration systems used are indicated. Each restriction enzyme site indicated is unique, except *NotI* (two cutting sites). *E. coli* ori corresponds to the *E. coli* p15A origin of replication. *oriT* is the origin of transfer. *amilCP* is the gene coding for an *Acropora millepora* chromoprotein, a protein which exhibits blue color. FRT corresponds to the sites recognized by the Flp recombinase. The promoter of module 5 is only functional in *E. coli*. *attP* site are used by integrases to integrate the plasmid in *Streptomyces* genome at a specific site.

Each vector is made of five modules (Figure 1). The first module is constituted of the *E. coli* origin of replication and of an Flp recombination target (FRT) recognition site for the Flp recombinase. We chose the p15A *E. coli* origin of replication to limit the number of plasmid copies in the cell, and thus the metabolic burden induced by the vector, which could be important with large inserts. The second module consists in the antibiotic resistance marker. Three different resistance genes were chosen: *acc(3)IV* (conferring apramycin resistance), *aph(7<sup>+</sup>)* (conferring hygromycin resistance) and *aph* (conferring kanamycin resistance). The expression of

the resistance genes is under the control of a promoter that is functional both in *E. coli* and *Streptomyces*. The third module is constituted of the RP4 origin of transfer, *oriT*, and of a second FRT site. The two FRT sites have been positioned so that the *E. coli* origin of replication, the antibiotic resistance cassette and the origin of transfer can be excised once the vector is integrated in the chromosome of *Streptomyces*, allowing the recycling of the resistance marker and limiting the possibility of homologous recombination between two different vectors. The fourth module is the integration system cassette (integrase and their corresponding *attP* site) that allows site-specific integration into *Streptomyces* chromosomes after conjugation. Four different integration cassettes are used, derived from the integration systems of the actinophages  $\varphi$ BT1,  $\varphi$ C31 and VWB or of the integrative conjugative element pSAM2. Chromosomal integration sites for these systems are found in the genomes of *Streptomyces* species commonly used for heterologous expression (*Streptomyces coelicolor*, *Streptomyces lividans* or *Streptomyces albus* J1074 for exemple). The construction of plasmids with four different integrase systems moreover maximizes the likelihood of being able to use at least one of them in any given strain. The last module is the cloning module. Our objective for this module was to permit the cloning and assembly of genes or gene cassettes using a variety of cloning methods (based on homology regions or on the use of restriction enzymes), as different projects may require different cloning approaches. Thus, this module was designed to allow the iterative assembly of genes (or gene cassettes) using the Biobrick assembly method (23) (see Figure S1 in the supplemental material). We chose this assembly method rather than other methods based on the use of type IIS endonucleases (e.g. Golden Gate method (24)), as the latter enzymes cut *Streptomyces* genomic DNA with a high frequency (about 1 site every 1 to 1.4 kb for three of the most frequently used enzymes *BsaI*, *BsmBI* and *BpiI* in *S. coelicolor*, *S. avermitilis* and *S. albus* genomes). The Biobrick cloning system is based on the use of restriction enzymes generating compatible cohesive ends, here *NheI* and *SpeI* (Figure S1). Once ligated, the two DNA parts are separated by a 6-bp scar sequence devoid of the *NheI* and *SpeI* restriction sites. The *NheI* and *SpeI* sites were chosen to avoid the generation of a stop codon in the scar sequence, thereby allowing the fusion of protein domains if needed, and because they are relatively rare in *Streptomyces* genomes. The *NsiI*, *AflII* sites that are also used in the Biobrick cloning system are relatively scarce too in *Streptomyces* genomes (e.g. about one site every 70-80 kb for *NsiI* and one site every 200-300 kb for *AflII* in *S. coelicolor*, *S. avermitilis* and *S. albus* genomes). A *NoI* site is included between the *NsiI* and *NheI* sites and between the *SpeI* and *AflII* sites to facilitate the verification of the cloning. The cloning module includes between the prefix and suffix sequences an *amilCP* gene (31). This gene codes for a chromoprotein, giving a blue color to the cell. This cassette is meant to be replaced by the construction of interest and offers a convenient mean of screening the clones containing the new construction. The five modules are separated by unique restriction sites (*BamHI*, *KpnI*, *SbfI*, *AflII* and *NsiI*), so that each module (e.g. the antibiotic resistance cassette or the integration system) can easily be replaced by another one.

On one side of the insert, the sequence is the same in all plasmids and the primer on-ori (Table 4) has been designed in the origin of replication of p15A to facilitate the verification of the insert by sequencing. On the other side of the insert, the sequence is that of the various integrase cassettes and, thus, no universal primer could be designed.

**Construction of the vectors**

The first vector, pOSV800, was assembled by Gibson isothermal assembly (19) from five PCR-amplified DNA fragments, one for each module. The apramycin resistance gene and the  $\varphi$ BT1 integration system were used for this first assembly. The final twelve vectors all derive from pOSV800 (Table 1 and Figure S2). The *NheI* and the *SpeI* restriction sites present in the integration cassette of pOSV800 were removed by site-directed mutagenesis, yielding pOSV801. The vector pOSV802 was constructed by replacing the  $\varphi$ BT1 integration cassette of pOSV800 by the  $\varphi$ C31 integration cassette. The vectors pOSV806 (resistance to kanamycin) and pOSV810 (resistance to hygromycin) were next obtained by the replacement in pOSV802 of the *aac(3)IV* gene by the *aph* and *aph(7<sup>+</sup>)* genes respectively by  $\lambda$ -Red recombination (32).

Table 1: Description of the constructed vectors

Name of the vector	Accession numbers	Resistance to	Integration system
pOSV801	126044 <sup>(a)</sup> /LMBP 11369 <sup>(b)</sup>		$\varphi$ BT1
pOSV802	126595 <sup>(a)</sup> /LMBP 11370 <sup>(b)</sup>	Apramycin	$\varphi$ C31
pOSV803	126596 <sup>(a)</sup> /LMBP 11371 <sup>(b)</sup>		pSAM2
pOSV804	126597 <sup>(a)</sup> /LMBP 11372 <sup>(b)</sup>		VWB
pOSV805	126598 <sup>(a)</sup> /LMBP 11373 <sup>(b)</sup>		$\varphi$ BT1
pOSV806	126606 <sup>(a)</sup> /LMBP 11374 <sup>(b)</sup>	Hygromycin	$\varphi$ C31
pOSV807	126600 <sup>(a)</sup> /LMBP 11375 <sup>(b)</sup>		pSAM2
pOSV808	126601 <sup>(a)</sup> /LMBP 11376 <sup>(b)</sup>		VWB
pOSV809	126602 <sup>(a)</sup> /LMBP 11377 <sup>(b)</sup>		$\varphi$ BT1
pOSV810	126603 <sup>(a)</sup> /LMBP 11378 <sup>(b)</sup>	Kanamycin	$\varphi$ C31
pOSV811	126604 <sup>(a)</sup> /LMBP 11379 <sup>(b)</sup>		pSAM2
pOSV812	126605 <sup>(a)</sup> /LMBP 11380 <sup>(b)</sup>		VWB

(a): accession number in Addgene plasmid repository; (b) accession number in BCCM/GeneCorner Plasmid Collection.

The vector pOSV803 was constructed by replacing the  $\varphi$ BT1 integration cassette of pOSV800 by the pSAM2 integration cassette, after the removal of the *Bam*HI and *Kpn*I sites from this cassette by site-directed mutagenesis. The vectors pOSV807 (resistance to hygromycin) and pOSV811 (resistance to kanamycin) were next obtained by the replacement in pOSV803 of the apramycin resistance cassette by the hygromycin (from pOSV806) and kanamycin (from pOSV810) resistance cassettes, respectively.

Similarly, pOSV804 was constructed by replacing the  $\varphi$ BT1 integration cassette of pOSV800 by the VWB integration cassette after the removal of the *Bam*HI site from the VWB integration cassette by site-directed mutagenesis. The vectors pOSV808 (resistance to

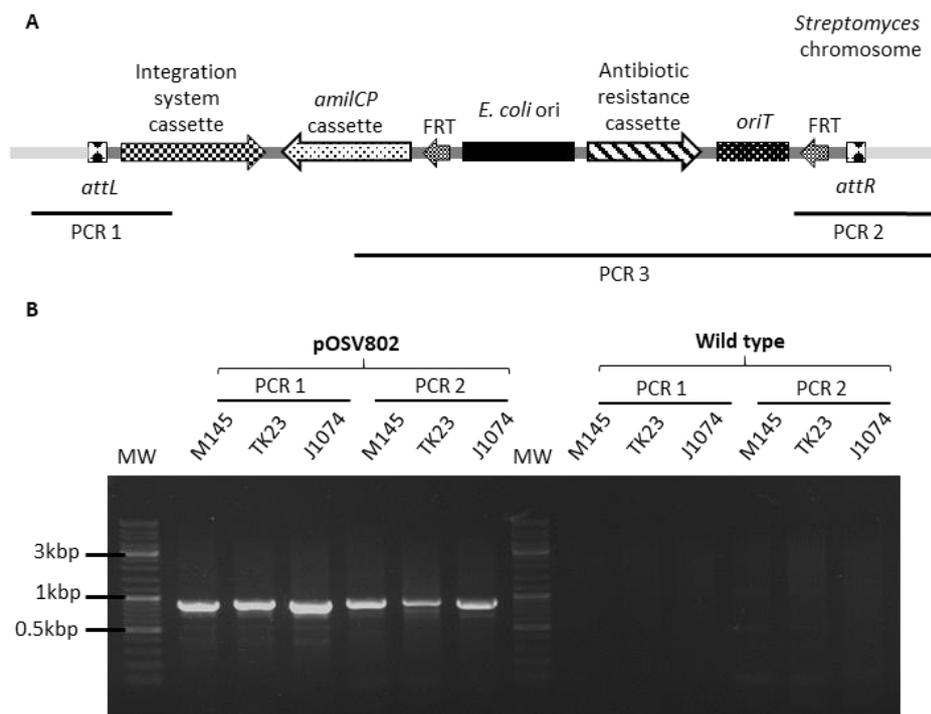
hygromycin) and pOSV812 (resistance to kanamycin) were next obtained by the replacement in pOSV804 of the apramycin resistance cassette by the hygromycin and kanamycin resistance cassettes, respectively.

Finally, pOSV805 (resistance to hygromycin) and pOSV809 (resistance to kanamycin) were next obtained by the replacement in pOSV801 of the apramycin resistance cassette by the hygromycin and kanamycin resistance cassettes, respectively.

### ***Verification of the functionality of the vectors: integration into Streptomyces chromosome***

To verify that the 12 vectors we constructed were all functional, we integrated them in the chromosome of three *Streptomyces* strains commonly used for heterologous expression: *Streptomyces calicolor* M145, *Streptomyces lividans* TK23 and *Streptomyces albus* J1074. The vectors were introduced in the *Streptomyces* strains by intergeneric conjugation from *E. coli*. The exconjugants were selected for using the appropriate antibiotics, and resistant clones were verified by PCR on extracted genomic DNA. The general principle for the PCR verification of the correct integration of the vectors at the expected chromosomal site is presented in Figure 2A. Briefly, two DNA fragments encompassing the *attL* and *attR* sites respectively were amplified by PCR (PCR 1 and PCR2). The results of these PCR verification for the integration of pOSV802 are presented in Figure 2B. DNA fragments with a size of roughly 900 bps were amplified as expected when using the genomic DNA of the *Streptomyces* strains bearing the pOSV802 plasmid as matrix. The sequences surrounding the *attL* and *attR* sites were verified. No PCR amplification was observed when the genomic DNAs of the wild type strains were used as matrix. Thus, these results confirmed the integration of the pOSV802 at the expected site in the chromosome of the three *Streptomyces* species.

Results of the PCR verification of the correct integration of the eleven other vectors are presented in the supplemental data (Figure S3 to Figure S9). All PCR products had the expected size, indicating that the vectors integrated at the expected location in the *Streptomyces* chromosomes. Altogether, these experiments demonstrate that the 12 plasmids (i) are replicative in *E. coli*, (ii) can be transferred by intergeneric conjugation into *Streptomyces*, (iii) confer the expected resistance and (iv) integrate at the expected location in the chromosome of *Streptomyces*.



**Figure 2: Verification of the integration of pOSV802 in *S. caelicolor* M145, *S. lividans* TK23, *S. albus* J1074 chromosomes.**

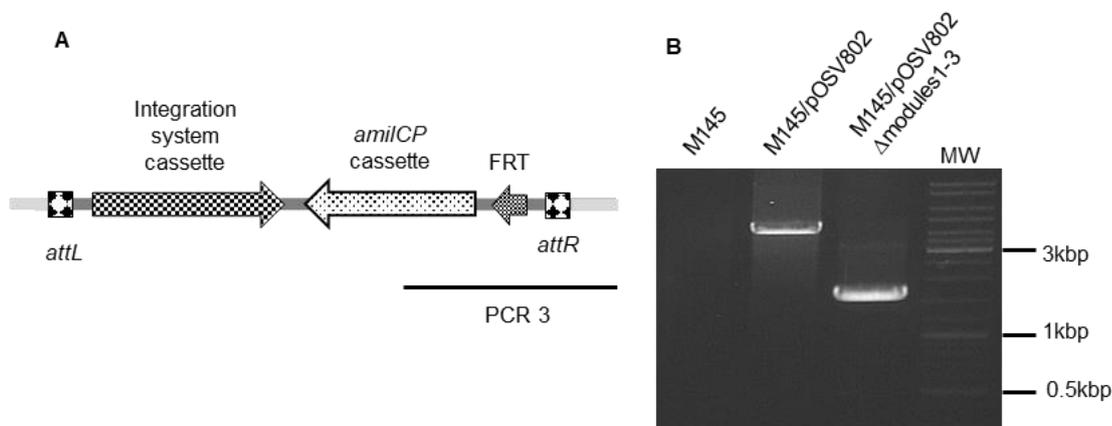
(A) Principle of the PCR verification of the integration of the pOSV801 to pOSV812 vectors in the *Streptomyces* chromosomes (PCR 1 & PCR 2) (PCR 3: PCR verification before excision of modules 1-3). (B) PCR fragments obtained by PCR 1 (*attL* region; expected sizes: 913 bps for M145 and TK23, 888 bps for J1074) and by PCR 2 (*attR* region; expected sizes: 911 bps for M145 and TK23, 907 bps for J1074) on the three *Streptomyces* strains bearing pOSV802. No PCR amplification is expected when the genomic DNA of the wild type *Streptomyces* strains is used as matrix. MW corresponds to the molecular weight ladder (Thermo Scientific™ GeneRuler™ DNA Ladder Mix)

### ***Excision of modules 1, 2 and 3 using the Flp recombinase***

One potential difficulty when multiple genetic constructions need to be integrated in *Streptomyces* chromosomes is the limited number of antibiotic resistance markers that are functional in a given strain. To allow the recycling of resistance markers, we included in our vectors FRT sites surrounding module 1 (*E. coli* origin of replication), module 2 (antibiotic resistance cassette) and module 3 (origin of transfer). Thus, once a vector has been integrated in a *Streptomyces* chromosome, these three modules, which are no longer necessary, can be excised using the Flp recombinase brought *in trans* by a replicative plasmid, leaving a scar of 34 base pairs (33).

To verify that modules 1, 2 and 3 could be excised using the Flp recombinase, we used the pUWLHFLP plasmid reported by Siegel and Luzhetskyy (34) and followed the protocol described in (33) to excise modules 1-3 in *S. caelicolor* M145/pOSV802 as an example. The pUWLHFLP plasmid is a replicative plasmid that allows the constitutive expression of a *flp* gene with a codon usage optimized for *Streptomyces* species. About one apramycin sensitive clone was obtained for each 100 clones screened, which is roughly ten times less than what was previously described (33). One sensitive clone was chosen for PCR verification of the excision of the modules 1 to 3 (Figure 3). As expected, a smaller (1.6 kb) fragment was amplified with the

genomic DNA of the sensitive clone M145/pOSV802 $\Delta$ modules1-3 compared to the 4.2 kb fragment obtained with *S. calicolor* M145/pOSV802 genomic DNA. The sequencing of the 1.6 kb fragment confirmed the correct excision of modules 1 to 3.



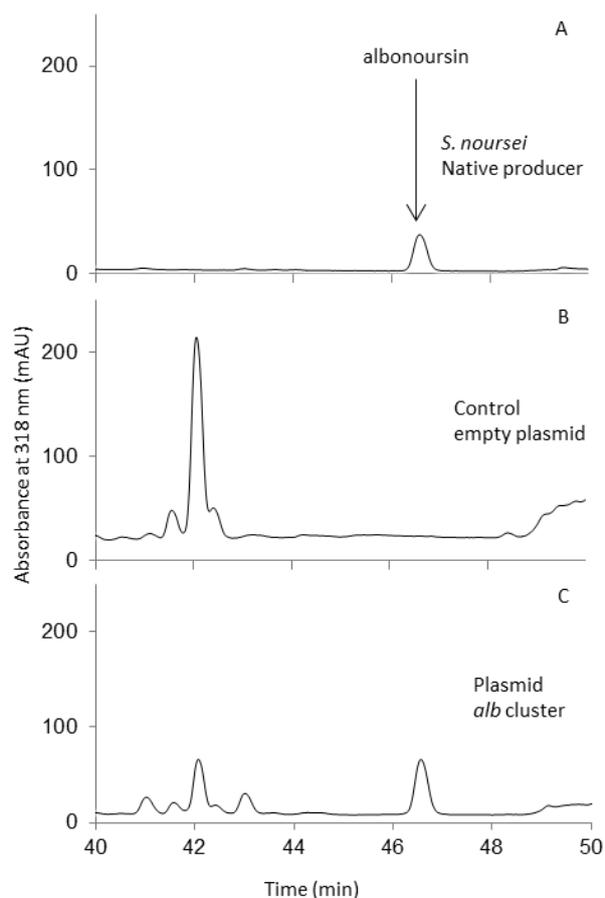
**Figure 3: Verification of the excision of modules 1, 2 and 3 by the FLP recombinase.**

(A) Principle of the PCR verification of the FLP-catalyzed excision of modules 1 to 3 (PCR 3; Figure 2A shows PCR3 on non-excised pOSV802). (B) PCR fragments obtained by PCR 3; expected sizes: 4192 pbs for M145/pOSV802 and 1,637 bps for M145 containing pOSV802 after excision of modules 1 to 3 by the FLP recombinase.

This experiment demonstrated the feasibility of the excision of modules 1-3 after the integration of one of our vectors in the chromosome of a *Streptomyces* species. As the pUWLHFLP plasmid is relatively unstable, it can be lost after two rounds of growth on solid medium soya flour mannitol (SFM) without selection pressure, allowing the integration of a second vector bearing the same resistance marker. It should be noted that it will not be possible to use the pUWLHFLP plasmid, which bears a hygromycin resistance gene when pOSV805-808 (bearing a hygromycin resistance gene) are used. However, other plasmids for the expression of FLP in *Streptomyces* have been constructed harboring different resistance markers, *e.g.* thiostrepton resistance (33).

### ***Refactoring the albonoursin gene cluster***

The pOSV801 to pOSV812 vectors were mainly designed for the assembly of gene cassettes to form new gene clusters or to refactor silent gene clusters, although their use may not be limited to these applications. To illustrate one of the possible uses of our vectors, we decided to refactor the albonoursin gene cluster. Albonoursin (cyclo( $\Delta$ Phe- $\Delta$ Leu)), produced by *Streptomyces noursei*, belongs to the family of diketopiperazine metabolites studied in our group. Its biosynthetic gene cluster consists of three genes, *albA*, *albB* and *albC* (35). We chose to express the *alb* gene under the control of the *rpL*(TP) constitutive promoter (2), and to assemble the required elements using the Biobrick assembly method.



**Figure 4: HPLC analysis of albonoursin production.**

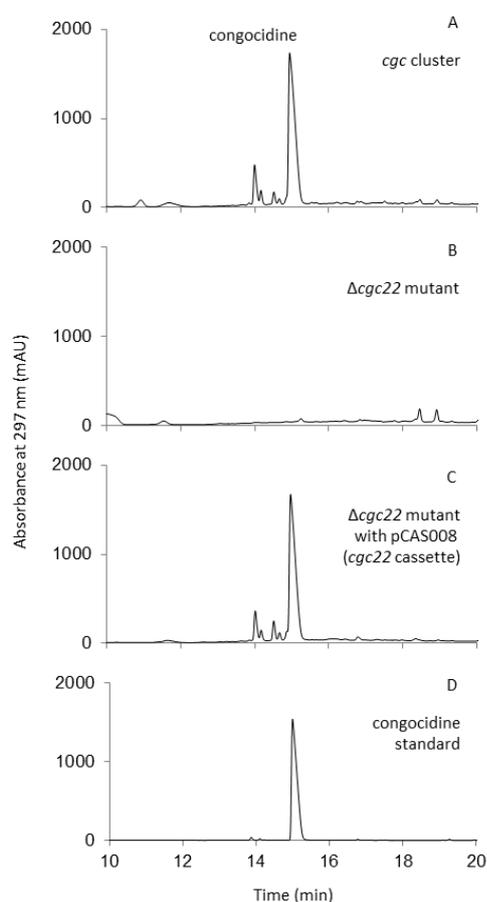
Chromatograms of the analysis of the culture supernatants of the native albonoursin producer *S. noursei* (A); the control *S. calicolor* M145/pOSV802 (B), and *S. calicolor* M145/pCEA007 (C).

The *rpsL*(TP) promoter followed by the ribosome binding site (RBS) sequence of *tipA* (36) was first cloned into pOSV802, yielding pCEA005. Similarly, the *alb* gene cluster was cloned in pOSV802, yielding pCEA006. The *NheI/AflII* fragment of pCEA006 containing the *alb* gene cluster was finally cloned into the *SpeI/AflII* digested pCEA005, and the resulting pCEA007 plasmid was introduced in *S. calicolor* M145 by intergeneric conjugation. To verify that *S. calicolor* M145/pCEA007 produced albonoursin, the culture supernatant of this strain, together with the culture supernatants of *S. noursei* (positive control) and of *S. calicolor* M145/pOSV802 (negative control) were analyzed by LC-MS. The chromatograms (Figure 4) and the MS spectra and fragmentation patterns (Figure S10 and (37)) confirmed that M145/pCEA007 produces albonoursin.

### ***Genetic complementation of mutant strain: assembly of a gene cassette using the Ligase Cycling Reaction (LCR) in pOSV812***

Cloning methods based on the use of restriction enzymes necessitate the presence or introduction of restriction sites in the sequence, which may sometimes be problematic (for example, for the fusion of protein domains, or for the cloning of an RBS sequence in front of a coding sequence). In these cases, the use of seamless cloning methods is preferable. To

demonstrate that gene cassettes could be assembled in our vectors using such seamless cloning methods, we undertook the genetic complementation of a mutant constructed previously, during the study of the congoicidine biosynthetic gene cluster ((30), mutant strain CGCL030). Congoicidine is a pyrrolamide antibiotic assembled by an atypical NRPS. The gene *cgc22*, deleted in the strain CGCL030, encodes an acyl-CoA synthetase that activates the pyrrole precursor during congoicidine assembly. To construct the plasmid for genetic complementation, we assembled three DNA fragments in pOSV802 by LCR (20): the SP22 constitutive promoter with the ribosome binding site (RBS) of the capsid  $\varphi$ C31 gene (15), the *cgc22* gene and the T4 terminator (38). The LCR method is based on the ligation of DNA fragments using bridging oligonucleotides whose sequences are complementary to the sequences of the extremities of the DNA fragments to be assembled (Figure S11). The assembly is achieved through multiple cycles of denaturation-annealing-ligation using a thermostable ligase. This method has the advantages of working for the assembly of very short fragments (< 100 bps) and does not necessitate the existence of homology regions at the extremities of the DNA fragments that will be assembled.



**Figure 5: HPLC analysis of the genetic complementation of the  $\Delta$ *cgc22* mutant.**

Chromatograms of the analysis of the culture supernatant of the CGCL006 strain expressing the complete *cgc* cluster (A); the culture supernatant of the CGCL030 mutant strain expressing the *cgc* cluster except for *cgc22* (B); the culture supernatant of the CGCL083 strain (CGCL030 genetically complemented with pCAS008) (C), and the congoicidine standard (D).

Each DNA fragment was amplified by PCR. The oligonucleotides used for the amplification of the promoter and RBS fragment and of the T4 terminator fragment were designed to reconstitute the prefix and the suffix sequences once all the fragments have been assembled in the vector. All PCR fragments were phosphorylated and assembled in one step with the *NotI*/Klenow-digested vector pOSV812. To verify that the constructed gene cassette was functional, the pCAS008 plasmid was introduced by intergeneric conjugation in the *S. lividans* CGCL030 strain expressing the whole *egc* gene cluster but *egc22* (30). The supernatants of 4-day cultures of the CGCL030/pCAS008, CGCL030 and of CGCL006 expressing the complete *egc* gene cluster were then analyzed by HPLC. Figure 5 shows that production of congocidine is restored in CGCL030/pCAS008, demonstrating the functionality of the constructed gene cassette.

In conclusion, we constructed a set of plasmids dedicated to DNA assembly and integration in *Streptomyces* chromosomes. We aimed at offering a modular and flexible platform that can be used in various experimental settings, from the assembly of small gene cassettes to the assembly of larger DNA fragments, and that will be compatible with a large variety of cloning methods. Varying the nature of the resistance cassette (resistance to three different antibiotics) and of the integration system (four different systems), we constructed a total of 12 plasmids. To increase our plasmid collection, we plan in the future to add new resistance cassettes (e.g. erythromycin) and integration systems (e.g. integration systems from TG1,  $\phi$ Joe or SV1 (39–41), but also to include new modules such as the CEN-ARS module (1) for DNA cloning and assembly in yeast. All our plasmids will be made available to the community through the deposit in plasmid collections such as Addgene or the BCCM/Genecorner plasmid collection.

## MATERIALS AND METHODS

### ***Bacterial strains, plasmids and growth conditions***

Strains and plasmids used in this study are listed in Table 2 and 3. *Escherichia coli* strains were grown at 37°C in LB or SOB medium complemented with MgSO<sub>4</sub> (20 mM final), supplemented with appropriate antibiotics as needed. The Soya Flour Mannitol (SFM) medium (42) was used for genetic manipulations of *Streptomyces* strains and spore stocks preparations. *Streptomyces* strains were grown at 28°C in MP5 (43) for congocidine or albonoursin production.

### ***DNA Preparation and manipulations***

All oligonucleotides used in this study were purchased from Eurofins and are listed in Table 4. The High fidelity DNA polymerase Phusion (Thermo Fisher Scientific) was used to amplify the fragments used for the construction of the vectors. DreamTaq polymerase (Thermo Fisher Scientific) was used for PCR verification of plasmid integration in *Streptomyces* strains. DNA fragments were purified from agarose gels using the Nucleospin Gel and PCR clean-up kit from Macherey-Nagel. DNA extractions and manipulations, *E. coli* transformations and *E. coli*/*Streptomyces* conjugations were performed according to standard procedures (44, 42).

Table 2: Strains used during the study

Strain	Description	Reference
<i>Escherichia coli</i> DH5a	General cloning host	Promega
<i>E. coli</i> ET12567/pUZ8002	Host strain for conjugation from <i>E. coli</i> to <i>Streptomyces</i>	(55)
<i>E. coli</i> ET12567/pUZ8003	Host strain for conjugation from <i>E. coli</i> to <i>Streptomyces</i> when using vectors containing the kanamycin resistance cassette (pUZ8003 is a modified pUZ8002 with <i>aph</i> replaced by <i>bla</i> )	Our unpublished data
<i>E. coli</i> S17-1	Host strain for conjugation from <i>E. coli</i> to <i>Streptomyces</i> when using vectors containing the kanamycin resistance cassette	(56)
<i>E. coli</i> BW25113/pIJ790	Host strain for PCR targeting	(32)
<i>S. calicolor</i> M145	<i>Streptomyces</i> host strain for heterologous expression	(42)
<i>S. lividans</i> TK23	<i>Streptomyces</i> host strain for heterologous expression	(42)
<i>S. albus</i> J1074	<i>Streptomyces</i> host strain for heterologous expression	(42)
<i>S. noursei</i> ATCC11455	Albonoursin native producer	ATCC
<i>S. calicolor</i> M145/pOSV801	M145 containing pOSV801	This work
<i>S. calicolor</i> M145/pOSV802	M145 containing pOSV802	This work
<i>S. calicolor</i> M145/pOSV803	M145 containing pOSV803	This work
<i>S. calicolor</i> M145/pOSV804	M145 containing pOSV804	This work
<i>S. calicolor</i> M145/pOSV805	M145 containing pOSV805	This work
<i>S. calicolor</i> M145/pOSV806	M145 containing pOSV806	This work
<i>S. calicolor</i> M145/pOSV807	M145 containing pOSV807	This work
<i>S. calicolor</i> M145/pOSV808	M145 containing pOSV808	This work
<i>S. calicolor</i> M145/pOSV809	M145 containing pOSV809	This work
<i>S. calicolor</i> M145/pOSV810	M145 containing pOSV810	This work
<i>S. calicolor</i> M145/pOSV811	M145 containing pOSV811	This work
<i>S. calicolor</i> M145/pOSV812	M145 containing pOSV812	This work
<i>S. lividans</i> TK23/pOSV801	TK23 containing pOSV801	This work
<i>S. lividans</i> TK23/pOSV802	TK23 containing pOSV802	This work
<i>S. lividans</i> TK23/pOSV803	TK23 containing pOSV803	This work
<i>S. lividans</i> TK23/pOSV804	TK23 containing pOSV804	This work
<i>S. lividans</i> TK23/pOSV805	TK23 containing pOSV805	This work
<i>S. lividans</i> TK23/pOSV806	TK23 containing pOSV806	This work
<i>S. lividans</i> TK23/pOSV807	TK23 containing pOSV807	This work
<i>S. lividans</i> TK23/pOSV808	TK23 containing pOSV808	This work
<i>S. lividans</i> TK23/pOSV809	TK23 containing pOSV809	This work
<i>S. lividans</i> TK23/pOSV810	TK23 containing pOSV810	This work
<i>S. lividans</i> TK23/pOSV811	TK23 containing pOSV811	This work
<i>S. lividans</i> TK23/pOSV812	TK23 containing pOSV812	This work
<i>S. albus</i> J1074/pOSV801	J1074 containing pOSV801	This work
<i>S. albus</i> J1074/pOSV802	J1074 containing pOSV802	This work
<i>S. albus</i> J1074/pOSV803	J1074 containing pOSV803	This work
<i>S. albus</i> J1074/pOSV804	J1074 containing pOSV804	This work
<i>S. albus</i> J1074/pOSV805	J1074 containing pOSV805	This work
<i>S. albus</i> J1074/pOSV806	J1074 containing pOSV806	This work
<i>S. albus</i> J1074/pOSV807	J1074 containing pOSV807	This work

<i>S. albus</i> J1074/pOSV808	J1074 containing pOSV808	This work
<i>S. albus</i> J1074/pOSV809	J1074 containing pOSV809	This work
<i>S. albus</i> J1074/pOSV810	J1074 containing pOSV810	This work
<i>S. albus</i> J1074/pOSV811	J1074 containing pOSV811	This work
<i>S. albus</i> J1074/pOSV812	J1074 containing pOSV812	This work
<i>S. calicolor</i> M145/pOSV802Δmodules1-3	M145 containing pOSV802 after excision with <i>flp</i>	This work
<i>S. calicolor</i> M145/pCEA007	M145 containing pCEA007	This work
CGCL006	TK23 containing pCGC002 ( <i>egg</i> cluster)	(30)
CGCL030	TK23 containing pCGC221 ( <i>egg</i> cluster with <i>egg22</i> deleted)	(30)
CGCL083	CGCL030 containing pCAS008	This work

### **Construction of pOSV800**

pOSV800 was constructed by assembling five fragments coming from five different vectors using the one-pot isothermal assembly developed by Gibson *et al.* (19). The first fragment ( $\varphi$ BT1 integrase gene and *attP* site) was amplified from pRT801 (45) using the CEA\_vec01 and CEA\_vec02 primers. The second fragment (*oriT* origin of transfer) was amplified from pOSV408 (46) using the CEA\_vec03 and CEA\_vec04 primers. The third fragment (apramycin resistance cassette *aac(3)IV*) was amplified from pSET152 (47) using CEA\_vec05 and CEA\_vec06 primers. The fourth fragment (p15A origin of replication) was amplified from pAC-BETA (48) using CEA\_vec07 and CEA\_vec08 primers. The fifth and last fragment (*amilCP* cassette surrounded by “biobrick”-like prefix (*NsiI*, *NotI* and *NheI* sites) and suffix (*SpeI*, *NotI* and *AflII*)) was amplified from pSB1C3-BBa-K1155003 (iGEM registry of standard biological parts) using CEA\_vec09 and CEA\_vec10 primers. Two FRT sites were introduced in the primer sequences of CEA\_vec03 and CEA\_vec08. The PCR products were purified and diluted to 100 ng/ $\mu$ L. 1  $\mu$ L of each of the PCR product was used for the assembly. A mix containing T5 exonuclease (New England Biolabs, NEB), Taq ligase (NEB) and Phusion High fidelity polymerase (Thermo Fisher Scientific) in the appropriate buffer was prepared following the protocol described by Gibson (49). The reaction was carried out by adding 5 $\mu$ L of DNA to 15  $\mu$ L of the mix and incubating at 50°C for one hour. 5  $\mu$ L were used for a standard transformation of *E. coli* DH5 $\alpha$ . The *amilCP* cassette, coding for a blue protein, allowed the easy screening of potential correct clones. Plasmid DNA was extracted from a blue clone and the sequence of the plasmid was confirmed by sequencing.

### **Construction of pOSV801**

The  $\varphi$ BT1 integrase gene in pOSV800 contains a *NheI* and a *SpeI* restriction sites that were chosen for the Biobrick type of cloning. To remove these sites, one base was modified by site directed mutagenesis following the protocol described by (50). CEA\_vec21 and CEA\_vec22 were used to remove the *NheI* site by replacing an A by a G at the position 123 in the integrase gene sequence (position 38926 of the  $\varphi$ BT1 bacteriophage genome sequence), conserving the amino acid leucine (CTA becoming CTG) in the protein. Similarly, CEA\_vec23 and CEA\_vec24 were used to remove the *SpeI* site in the terminator downstream of the  $\varphi$ BT1 integrase gene at position 40663 in the  $\varphi$ BT1 bacteriophage genome sequence, replacing a T by a G.

Briefly, the plasmid was amplified using the first pair of oligonucleotides with the Phusion polymerase. 1  $\mu$ L of *DpnI* was added to the reaction to digest the original vector for 2 hours at 37°C, and competent *E. coli* DH5 $\alpha$  cells were transformed with 5  $\mu$ L of the mixture. The second site directed mutagenesis was performed following the same protocol. The sequence of the resulting plasmid was verified by sequencing and the plasmid was named pOSV801.

### **Construction of the pOSV802-812**

The pOSV802 to pOSV812 vectors derived all from pOSV800, except for pOSV805 and pOSV809, which derive from pSV801 (See Figure S2 in the supplemental material). The eleven vectors were confirmed by restriction analyses, and by sequencing each fragment obtained by PCR. The  $\varphi$ BT1 integration cassette was replaced either by the  $\varphi$ C31, VWB or pSAM2 integration cassettes and the *aac(3)IV* gene was replaced by either the *aph* or the *aph(7'')* genes. The use of the pSAM2 (from pOSV554,(51)) and VWB integration (from pKT02, (52)) cassette necessitated the removal of a *KpnI* and a *BamHI* sites, and of a *BamHI* site respectively. Thus, these cassettes were first cloned into pCR®-Blunt following the procedure advised by Invitrogen, yielding pCEA003 and pCEA004 respectively. The *BamHI* site from the VWB integrase was removed by site-directed mutagenesis using the oligonucleotides CEA\_025 and CEA\_026, by changing the base 1008 of the integrase gene sequence from C to A, thus keeping the amino acid unchanged (ATC becoming ATA, Isoleucine). The mutation in the resulting plasmid pCEA004 was verified by sequencing. The *KpnI* and *BamHI* sites, located upstream of the integrase pSAM2 coding sequence and only three base pair apart, were removed in single round of site-directed mutagenesis, using the oligonucleotides CEA\_027 and CEA\_028. The mutations in the resulting plasmid pCEA003 were verified by sequencing.

To replace the  $\varphi$ BT1 integration cassette by the  $\varphi$ C31 integration cassette in pOSV800, the  $\varphi$ C31 integration cassette was amplified by PCR from pSET152 (47) using the oligonucleotides CEA\_vec11 and CEA\_vec12. The PCR product was digested by *SbfI* and *AflIII* and cloned into the *SbfI* and *AflIII*-digested pOSV800, yielding pOSV802. The replacement of the  $\varphi$ BT1 integration cassette by the pSAM2 integration cassette in pOSV800 was executed likewise, cloning the 1.6kb *SbfI*/*AflIII* fragment from pCEA003 into the *SbfI* and *AflIII*-digested pOSV800, yielding pOSV803. The same protocol was used to replace the  $\varphi$ BT1 integration cassette by the VWB integration cassette in pOSV800, yielding pOSV804.

The replacement of the *aac(3)IV* gene (apramycin resistance) by the *aph(7'')* gene (hygromycin resistance) or the *aph* gene (kanamycin resistance) in pOSV802 was carried out by  $\lambda$ -Red recombination as described by Gust and colleagues (32). The *aph(7'')* and *aph* genes were amplified by PCR using the oligonucleotides CEA\_vec\_017 and CEA\_vec\_018 for *aph(7'')* and CEA\_vec\_019 and CEA\_vec\_020 for *aph*, and the PCR products were used to replace the *aac(3)IV* gene in pOSV802, yielding pOSV806 and pOSV810 respectively. The joining sequences were confirmed by sequencing. Sequencing showed that the sequences of *aph* and *aph(7'')* were as predicted, except for the base 188 of *aph(7'')*, in which A was substituted by G, leading to the substitution of Asp (GAC) by Gly (GGC). Yet no functional difference has been observed, the plasmid confers full resistance to hygromycin.

Table 3: Plasmids used in this study

Plasmid	Description	Reference
pCR®-Blunt	<i>E. coli</i> cloning vector	Invitrogen
pRT801	Source of the $\varphi$ BT1 integrase fragment	(45)
pAC-BETA	Source of the origin of replication p15A	(48)
pOSV408	Source of the origin of transfer	(46)
pSET152	Source of the apramycin resistance cassette and of the $\varphi$ C31 integrase fragment	(47)
psB1C3 – Bba_K1155003	Source of the <i>amilCP</i> cassette	iGEM registry of standard biological parts
pKT02	Source of the VWB integrase fragment	(52)
pOSV215	Source of the T4 terminator	(54)
pOSV554	Source of the integrase pSAM2 fragment	Our unpublished data
pOSV400	Source of the ORF of hygromycin resistance gene	Our unpublished data
pOSV401	Source of the ORF of kanamycin resistance gene	Our unpublished data
pSL128	Source of the albonoursin cluster ( <i>albA</i> , <i>albB</i> and <i>albC</i> )	(35)
pCEA001	pUC57 containing <i>rpsL</i> (TP)p and <i>tipA</i> RBS	Genecust
pCEA002	pGEM-T easy containing <i>rpsL</i> (TP)p and <i>tipA</i> RBS with the last 6 nucleotides replaced by the <i>SpeI</i> site	This work
pCEA003	Plasmid pCR®-Blunt containing pSAM2 integrase, used for site-directed mutagenesis	This work
pCEA004	Plasmid pCR®-Blunt containing VWB integrase, used for site-directed mutagenesis	This work
pCEA005	pOSV802 containing <i>rpsL</i> (TP)p and <i>tipA</i> RBS with the last 6 nucleotides replaced by the <i>SpeI</i> site	This work
pCEA006	pOSV802 containing the genes <i>albA</i> , <i>albB</i> and <i>albC</i> instead of the <i>amilCP</i> cassette	This work
pCEA007	pOSV802 containing <i>rpsL</i> (TP)p and the albonoursin cluster instead of <i>amilCP</i>	This work
pOSV800	Plasmid constructed containing apramycin resistance and $\varphi$ BT1 integrase with two biobrick sites <i>NheI</i> and <i>SpeI</i> in $\varphi$ BT1 integrase	This work
pOSV801	Plasmid constructed containing apramycin resistance and $\varphi$ BT1 integrase	This work
pOSV802	Plasmid constructed containing apramycin resistance and $\varphi$ C31 integrase	This work
pOSV803	Plasmid constructed containing apramycin resistance and pSAM2 integrase	This work
pOSV804	Plasmid constructed containing apramycin resistance and VWB integrase	This work
pOSV805	Plasmid constructed containing hygromycin resistance and $\varphi$ BT1 integrase	This work
pOSV806	Plasmid constructed containing hygromycin resistance and $\varphi$ C31 integrase	This work
pOSV807	Plasmid constructed containing hygromycin resistance and pSAM2 integrase	This work
pOSV808	Plasmid constructed containing hygromycin resistance and VWB integrase	This work

pOSV809	Plasmid constructed containing kanamycin resistance and $\varphi$ BT1 integrase	This work
pOSV810	Plasmid constructed containing kanamycin resistance and $\varphi$ C31 integrase	This work
pOSV811	Plasmid constructed containing kanamycin resistance and pSAM2 integrase	This work
pOSV812	Plasmid constructed containing kanamycin resistance and VWB integrase	This work
pCAS008	pOSV812 with cassette SP22p- <i>gac</i> 22-T4 terminator instead of <i>ami</i> CP	This work

To replace the *aac*(3)IV gene cassette in pOSV801, pOSV803 and pOSV804 by the *aph*(7<sup>+</sup>) gene cassette, the 1.4 kb *Kpn*I/*Bam*HI fragment of pOSV806 was cloned into *Kpn*I/*Bam*HI-digested pOSV801, pOSV803 and pOSV804, yielding pOSV805, pOSV807 and pOSV808 respectively. Using the same protocol, the *aac*(3)IV gene was replaced in pOSV801, pOSV803 and pOSV804 by the *aph* gene cassette, yielding pOSV809, pOSV811 and pOSV812 respectively. The vectors obtained were verified by restriction analyses.

#### ***Verification of the integration of the vectors in Streptomyces species***

The 12 vectors constructed were introduced in three *Streptomyces* species (*Streptomyces calicolor* M145, *Streptomyces lividans* TK23 and *Streptomyces albus* J1074) by intergeneric conjugation following the standard procedure (42). *E. coli* ET12567/pUZ8002 was used as a donor strain for pOSV801 to pOSV808. For pOSV809 to pSOV812, which confer resistance to kanamycin, we used *E. coli* S17-1 as a donor strain to perform conjugation with *S. lividans* TK23 and *S. albus* J1074, and *E. coli* ET12567/pUZ8003 as a donor strain to perform conjugation with *S. calicolor* M145. Genomic DNA was extracted from the ex-conjugants obtained. To confirm that the vectors had been integrated into the host chromosomal DNA at the expected sites, PCR 1 and PCR 2 were performed as shown in Figure 2, using the primers CEA\_vec\_seq12, CEA\_vec\_seq\_16 – 20 and CEA\_42 – 58. These PCRs amplify a fragment of about 900 bp only if the plasmid is integrated at the expected chromosomal *attB* site.

#### ***Excision mediated by the Flp recombinase***

We used the M145/pOSV802 to verify that modules 1, 2 and 3 could be excised using the Flp recombinase once integrated into the chromosome of *Streptomyces*. For this purpose, we used the plasmid pUWLHFLP and followed the protocol described by (33). pUWLHFLP is similar to pUWLFLP, but the thiostrepton resistance cassette has been replaced by a hygromycin resistance cassette (34). Briefly, pUWLHFLP was introduced by intergeneric conjugation into the strain M145/pOSV802, and exconjugants were replicated on SFM plates containing 100 $\mu$ g/mL hygromycin. After one round of liquid cultures in TSB, stocks of spores were made. Spore dilutions were plated on SFM supplemented with nalidixic acid and the clones were screened for loss of apramycin resistance by replica-plating. The loss of the fragment of the vector was subsequently confirmed by amplifying the fragment around both FRT sites (PCR 3, primers CEA\_vec\_seq15 and CEA\_045 (Figure 3)), which was then sequenced. Stocks of spores of the confirmed clones were prepared on SFM supplemented with nalidixic acid and the loss of the helper vector pUWLHFLP was confirmed by PCR (primers thio-fwd and CEA\_seq24).

**Construction of pCEA007**

The albonoursin gene cluster, constituted of the three genes *albA*, *albB* and *albC*, was cloned into the pOSV802 and placed under the control of the *rpsL*(TP) promoter (2) by following Biobrick assembly procedure (Figure S1). The pCEA001 plasmid was used to amplify the *rpsL*(TP) promoter sequence followed by the *tipA* RBS sequence using the primers F\_pref\_rpslp\_TP and R\_suff\_rpslp\_TP. The PCR product was cloned into pGEM-T Easy and the resulting plasmid was named pCEA002. The 0.4 kb *NsiI*/*SpeI*-digested fragment of pCEA002 was ligated into *NsiI*/*SpeI*-digested pOSV802, yielding pCEA005. The insert sequence of pCEA005 was confirmed by sequencing. The albonoursin gene cluster was amplified from pSL128 (35) using the primers CEA036 and CEA038. The PCR product was digested by *NsiI* and *SpeI* and ligated into the *NsiI*/*AflIII*-digested pOSV802, yielding pCEA006. The sequence of the insert was confirmed by sequencing. pCEA006 was then digested by *AflIII* and *NbeI* and the 1.8 kb fragment was ligated into the *SpeI*/*AflIII*-digested pCEA005, yielding pCEA007. The resulting plasmid pCEA007 was confirmed by digestion by *NotI* and by *EcoRI*/*HindIII*. This plasmid was introduced in *S. coelicolor* M145 by intergeneric conjugation.

**Construction of the pCAS008 plasmid**

The pCAS008 plasmid, expressing the *cgc22* gene under the control the SP22 promoter (15) was assembled using the ligase cycling reaction as previously described (53). pOSV812 was digested by *NotI*, and Klenow was added to the mix in order to obtain blunt ends. The 5 kb fragment was purified on agarose gel. The gene *cgc22* was amplified from the cosmid pCGC002 (30) with the primers onCAS031 and onCAS032. The promoter SP22 was ordered from Eurofins Genomics as a synthetic gene fragment and amplified with the primers onCAS001bis and onCAS002. The T4 terminator was amplified from the plasmid pOSV215 (54) with the primers onCAS007 and onCAS008bis. The primers upstream of the promoter SP22 and downstream of the terminator were designed in order to recreate the prefix and suffix of the biobrick (*NsiI*, *NotI*, *NbeI* and *SpeI*, *NotI*, *AflIII*, respectively). All fragments were then phosphorylated and ligated *via* ligase cycling reaction. The sequence of the resulting plasmid pCAS008 was confirmed by sequencing. The pCAS008 plasmid was introduced in *S. lividans* CGCL030 by intergeneric conjugation.

**LC and LC-MS analyses**

For albonoursin production, *S. coelicolor* M145/pCEA006, M145/pOSV802 and *S. noursei* strains were cultivated for 5 days in MP5 medium at 30°C. Supernatants were filtered using the Mini-UniPrep syringeless filter devices (0.2 µm, Whatman). The samples were analyzed on an Atlantis C<sub>18</sub> T3 column (250 mm x 4.6 mm, 5 µm, column temperature 30°C) using an Agilent 1200 HPLC instrument equipped with a quaternary pump. The filtrates were eluted using a 0%-45% linear gradient of solvent B (solvent A: 0.1% HCOOH in H<sub>2</sub>O; solvent B: 0.1% HCOOH in CH<sub>3</sub>CN) for 45 min (flow rate 1 mL/min). Albonoursin was detected by monitoring absorbance at 318 nm (35). A Bruker Daltonics Esquire HCT ion trap mass spectrometer equipped with an orthogonal Atmospheric Pressure Interface-ElectroSpray Ionization (AP-ESI) source was used for LC-MS analyses. The LC flow was split 1/10 to the mass spectrometer and 9/10 to a diode array detector. The ESI source was operated in positive mode with the nebulizing gas set to a pressure of 241 kPa. The drying gas was set to 8 l.min<sup>-1</sup> and the drying temperature was set to

340°C. Nitrogen served as the drying and nebulizing gas and helium gas was introduced into the ion trap both for efficient trapping and cooling of the ions and for fragmentation processes. Ionization and mass analysis conditions (capillary high voltage, skimmer and capillary exit voltages and ion transfer parameters) were optimized for detection of compounds in the  $m/z$  range of 50-600. For structural characterization by fragmentation, an isolation width of 1 mass unit was used. A fragmentation energy ramp was used for automatically varying the fragmentation amplitude to optimize the MS/MS process. For LC-MS analyses, filtrates were eluted using a slightly modified gradient: after 5 min of isocratic run at 100 % of buffer A, the concentration of B was linearly increased over 50 min to reach 50%.

For congocidine production, *S. lividans* CGCL083, CGCL030 and CGCL006 strains were cultivated in MP5 medium for 4 days at 30°C. Supernatants were filtered using Mini-UniPrep syringeless filter devices (0.2  $\mu\text{m}$ , Whatman). The samples were analyzed on an Atlantis C<sub>18</sub> T3 column (250 mm x 4.6 mm, 5  $\mu\text{m}$ , column temperature 30°C) using an Agilent 1200 HPLC instrument with a quaternary pump. Samples were eluted with in isocratic conditions of 0.1% HCOOH in H<sub>2</sub>O (solvent A)/ 0.1% HCOOH in CH<sub>3</sub>CN (solvent B) (95:5) at 1 ml/min for 7 min, followed by a gradient to 40:60 A/B over 23 min. Congocidine was detected by monitoring absorbance at 297 nm (30).

Table 4: Primers used in this study

Name	Sequence	Description
CEA_vec01	<u>ACTAGTAGCGGCCGCTTAAGCGCTC</u> CCTGCCCGCTGTGG	Amplification integrase $\phi$ BT1, suffix biobrick (sites <i>SpeI</i> , <i>NoI</i> and <i>AflII</i> underlined)
CEA_vec02	AATAGGAACITCCCTGCAGGTGGCG CCGGACGGGGCTTC	Amplification integrase $\phi$ BT1, site <i>SbfI</i> underlined
CEA_vec03	<b>CCTGCAGGGAAGTTCCTATTCTCTA</b> <b>GAAAGTATAGGAACTTCGTCCACGA</b> CGCCCGTGATTTTG	Amplification <i>oriT</i> , FRT site added in boldface, site <i>SbfI</i> underlined
CEA_vec04	CTCACC GCGACGTGGTACCCTTTTCC GCTGCATAACCCTG	Amplification <i>oriT</i> , site <i>KpnI</i> underlined
CEA_vec05	GGGTACCACGTCGCGGTGAGTTCAG G	Amplification <i>aac(3)IV</i> , site <i>KpnI</i> underlined
CEA_vec06	<u>GGATCCGGTTCATGTGCAGCTCCATC</u> AG	Amplification <i>aac(3)IV</i> , site <i>BamHI</i> underlined
CEA_vec07	GCTGCACATGAACCGGATCCCCTAGC GGAGTGTATACTGG	Amplification of p15A origin of replication, site <i>BamHI</i> underlined
CEA_vec08	<u>GCTAGCAGCGGCCGC</u> <u>ATGCATGAAGTTCCTATACTTTCTA</u> <b>GAGAATAGGAACTTCACA</b> ACTTATAT CGTATGGGGCTGAC	Amplification of p15A origin of replication, FRT site added in boldface, prefix biobrick (sites <i>NbeI</i> , <i>NoI</i> and <i>NsiI</i> underlined)
CEA_vec09	TGCATGCGGCCGCTGCTAGCGTTTTT TGATCTCAATCAATAAAG	Amplification <i>amilCP</i> cassette, prefix biobrick (sites <i>NsiI</i> and <i>NoI</i> underlined)
CEA_vec10	<u>CCTAAGCGGCCGCTACTAGTATATAA</u>	Amplification <i>amilCP</i> cassette, suffix

	ACGCAGAAAGGC	biobrick (sites <i>Afl</i> III, <i>Not</i> I and <i>Spe</i> I underlined)
CEA_vec11	CAGT <u>CCTGCAGG</u> ATTCCAGACGTCCC GAAGG	Amplification integrase $\varphi$ C31, site <u><i>Sbf</i></u> I underlined
CEA_vec12	CAGT <u>CTTAAG</u> CAGGCT* <u>CCCGGG</u> TG TCTC	Amplification integrase $\varphi$ C31, site <u><i>Afl</i></u> III underlined
CEA_vec13	CAGT <u>CCTGCAGG</u> AACGGTTC <u>TGGCA</u> AATATTC	Amplification integrase pSAM2, site <u><i>Sbf</i></u> I underlined
CEA_vec14	CAGT <u>CTTAAG</u> GTCAGTCATGCGGGC AAC	Amplification integrase pSAM2, site <u><i>Afl</i></u> III underlined
CEA_vec31	CAGT <u>CCTGCAGG</u> TCTCGAGCTCGCG AAAG	Amplification integrase VWB, site <u><i>Sbf</i></u> I underlined
CEA_vec32	CAGT <u>CTTAAG</u> GTCGACCCGTCTGACG CGTGTG	Amplification integrase VWB, site <u><i>Afl</i></u> III underlined
CEA_vec17	CTATGATCGACTGATGTCATCAGCG GTGGAGTGCAATGTCGTGACACAAG AATCCCGTTACTTC	Amplification ORF hygromycin resistance for PCR targeting
CEA_vec18	CCTTGCCCCTCCAACGTCATCTCGTTC TCCGCTCATGAGCTCAGGCGCCGGG GGCGGTGT	Amplification ORF hygromycin resistance for PCR targeting
CEA_vec19	CTATGATCGACTGATGTCATCAGCG GTGGAGTGCAATGTCICGCATGATT GAACAAGATG	Amplification ORF kanamycin resistance for PCR targeting
CEA_vec20	CCTTGCCCCTCCAACGTCATCTCGTTC TCCGCTCATGAGCTCAGAAGAACTCG TCAAGAAG	Amplification ORF kanamycin resistance for PCR targeting
CEA_vec21	CCAACGCACGACCGGCC <u>GCCAG</u> CTG TGCTTCGGTTCGACACG	site directed mutagenesis of <i>Nhe</i> I site of $\varphi$ BT1 integrase, base changed underlined (T→C)
CEA_vec22	CGTGTTCGACCGAAGCACAGCT <u>GGCG</u> GCCGGTCGTGCGTTGG	site directed mutagenesis of <i>Nhe</i> I site of $\varphi$ BT1 integrase, base changed underlined (A→G)
CEA_vec23	GCTGTGGTGACGAAGGAACTACT <u>CG</u> TTAGCCTAACTAACG	site directed mutagenesis of <i>Spe</i> I site of $\varphi$ BT1 integrase, base changed underlined (A→C)
CEA_vec24	CGTTAGTTAGGCTAAC <u>GAG</u> TAGTTCC TTCGTCACCCACAGC	site directed mutagenesis of <i>Spe</i> I site of $\varphi$ BT1 integrase, base changed underlined (T→G)
CEA_vec25	CTCCGGCGCACATGGAT <u>ACCT</u> GCAA TCAAGGC	site directed mutagenesis of <i>Bam</i> HI site of VWB integrase, base changed underlined (C→A)
CEA_vec26	GCCTTGATTGCAGGTA <u>TCC</u> ATGTGC GCCGGAAG	site directed mutagenesis of <i>Bam</i> HI site of VWB integrase, base changed underlined (G→T)
CEA_vec27	CATGGAATTTCGAGCTCGGTA <u>ACCG</u> G <u>AAT</u> CCCCGGGTACGC	site directed mutagenesis of <i>Bam</i> HI and <i>Kpn</i> I sites of integrase pSAM2, bases changed underlined (C→A and G→A)

CEA_vec28	GCGTACCCGGGGAT <u>T</u> CCCGG <u>T</u> TACC GAGCTCGAATTCCATG	site directed mutagenesis of <i>Bam</i> HI and <i>Kpn</i> I sites of integrase pSAM2, bases changed underlined (C→T and G→T)
CEA_vec_seq_1 2	TCTGGCAGCACTTTGAGGAC	Verification primer, in pSAM2 integrase, towards <i>attP</i>
CEA_vec_seq_1 5	TTCGATCACGTGGGCGAAGC	Verification primer of <i>flp</i> excision
CEA_vec_seq16	TTGCCAAAGGGTTCGTGTAG	Verification primer in <i>oriT</i> , towards <i>attP</i> of $\varphi$ C31 or $\varphi$ BT1 integrases
CEA_vec_seq_1 7	TCAGGTCACTGTCTGTTC	Verification primer in $\varphi$ BT1 integrase, towards <i>attP</i>
CEA_vec_seq18	AATCTTCGCCGACTTCAGC	Verification primer in $\varphi$ C31 integrase, towards <i>attP</i>
CEA_vec_seq_1 9	GGTTTGAACITTCCTCCCAATG	Verification primer in <i>ami</i> CP cassette, towards <i>attP</i> of pSAM2 or VWV integrases
CEA_vec_seq_2 0	GGTGAAGAACCGGGACACC	Verification primer in VWB integrase, towards <i>attP</i>
CEA042	GTGGTGTGCGGGAACAGACG	Verification primer in M145 and TK23, upstream of $\varphi$ BT1 <i>attB</i> site
CEA043	TCCGCGACGATCCACGAC	Verification primer in M145 and TK23, downstream of $\varphi$ BT1 <i>attB</i> site
CEA044	GCGTGGCGTGGACCATC	Verification primer in M145 and TK23, upstream of $\varphi$ C31 <i>attB</i> site
CEA045	AATGACCTCCGGGCTTTCG	Verification primer in M145 and TK23, downstream of $\varphi$ C31 <i>attB</i> site
CEA046	ACCGGCACCGCATGGCAG	Verification primer in M145 and TK23, upstream of pSAM2 <i>attB</i> site
CEA047	ACGGCGCGTGCGGCATC	Verification primer in M145 and TK23, downstream of pSAM2 <i>attB</i> site
CEA048	GAAAGACGGCCGACCACC	Verification primer in M145 and TK23, upstream of VWB <i>attB</i> site
CEA049	TGCCCCGCCCTCTGCATC	Verification primer in M145, downstream of VWB <i>attB</i> site
CEA050	CTGTATGCCGCGTCCCG	Verification primer in TK23, downstream of VWB <i>attB</i> site
CEA051	GGTGGTGTCCCGGACCAG	Verification primer in J1074, upstream of $\varphi$ BT1 <i>attB</i> site
CEA052	CCGCGACGATCCAGGACC	Verification primer in J1074, downstream of $\varphi$ BT1 <i>attB</i> site
CEA053	GGCGTGGATCATGGTGATCG	Verification primer in J1074, upstream of $\varphi$ C31 <i>attB</i> site
CEA054	GGTTGCGGGTGGCAAGTAG	Verification primer in J1074, downstream of $\varphi$ C31 <i>attB</i> site
CEA055	CGGCCAGCTCTGCATCCC	Verification primer in J1074, upstream of pSAM2 <i>attB</i> site

CEA056	CGGATTGTTTGCCGCCTTC	Verification primer in J1074, downstream of pSAM2 <i>attB</i> site
CEA057	GCATGCACGGCGACCTG	Verification primer in J1074, upstream of VWB <i>attB</i> site
CEA058	GTGACCCTGCCGGGATGG	Verification primer in J1074, upstream of VWB <i>attB</i> site
CEA_seq24	ACCATCGCCCACGCATAAC	Verification of the loss of pUWLHFLP
Thio_fwd	T <sup>*</sup> TGGACACCATCGCAAATC	Verification of the loss of pUWLHFLP
CEA036	AAA <u>ATGCATGCGGCCGCTGCTAGCG</u> GTGAGGCGCCACCCATCG	Amplification albonoursin cluster (sites <i>Nsi</i> I, <i>Noi</i> I and <i>Nbe</i> I underlined)
CEA038	AA <u>ACTTAAGGCGGCCGCTACTAGTCC</u> GCACCATGAGCAAGTGTC	Amplification albonoursin cluster (sites <i>Afl</i> II, <i>Noi</i> I and <i>Spe</i> I underlined)
F_pref_rpslp_T P	<u>ATGCATGCGGCCGCTTCTAGAGACC</u> GGGTCCGCGATCGGCGG	Amplification <i>rps</i> l(TP)p (sites <i>Nsi</i> I, <i>Noi</i> I and <i>Xba</i> I underlined)
R_suff_rpslp_T P	<u>CTTAAGGCGGCCGCTACTAGTGCTCC</u> CTTCTCAGAAGCGCAGG	Amplification <i>rps</i> l(TP)p (sites <i>Afl</i> II, <i>Noi</i> I and <i>Spe</i> I underlined)
onCAS001bis	<u>GCTGCTAGCTGTTCACATTCGAACCG</u> TCTCTG	Amplification SP22 promoter forward (truncated <i>Noi</i> I and <i>Nbe</i> I underlined)
onCAS002	ATGGACACTCCTTACTTAGAC	Amplification SP22 promoter reverse
onCAS003	GTATAGGAACCTTCATGCATGCGGCC GCTGCTAGCTGTTCACATTCGAACCG	Bridging oligonucleotide between plasmid pOSV812 and SP22 promoter
Bridge4	ACGGTTTACAAGCATAACTAGTAGC GGCCGCTTAAGGTCGACCCGTCTG	Bridging oligonucleotide between T4 terminator and pOSV812
onCAS007	TGATCCGGTGGATGACCTTTTG	Amplification T4 terminator forward
onCAS008bis	<u>GCTACTAGTTATGCTTGTAACCGTT</u> TTG	Amplification T4 terminator reverse (truncated <i>Noi</i> I and <i>Spe</i> I underlined)
onCAS031	ATGGCCACCGAGTCCGCCACC	Amplification <i>gce</i> 22 forward
onCAS032	CTACCCGCCGTCGCCGTCGC	Amplification <i>gce</i> 22 reverse
onCAS033	GAATACGACAGTCTAAGTAAGGAGT GTCCATATGGCCACCGAGTCCGCC	Bridging oligonucleotide between SP22 promoter and <i>gce</i> 22
onCAS034	GACGGCGACGGCGGGTAGTGATCC GGTGGATGACCTTTTGAATGAC	Bridging oligonucleotide between <i>gce</i> 22 and T4 terminator
on-ori	AT <sup>*</sup> TCAGTGCAAT <sup>*</sup> TATCTCT <sup>*</sup> TC	Universal sequencing primer in p15A origin for verification of the insert

### Acknowledgements

We thank Hervé Leh for critical reading of the manuscript. The research received funding from ANR-14-CE16-0003-01. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

## References

1. Yamanaka K, Reynolds KA, Kersten RD, Ryan KS, Gonzalez DJ, Nizet V, Dorrestein PC, Moore BS. 2014. Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc Natl Acad Sci U S A* 111:1957–1962.
2. Shao Z, Rao G, Li C, Abil Z, Luo Y, Zhao H. 2013. Refactoring the silent spectinabilin gene cluster using a plug-and-play scaffold. *ACS Synth Biol* 2:662–669.
3. Greunke C, Duell ER, D'Agostino PM, Glöckle A, Lamm K, Gulder TAM. 2018. Direct Pathway Cloning (DiPaC) to unlock natural product biosynthetic potential. *Metab Eng* 47:334–345.
4. Bozhüyük KAJ, Fleischhacker F, Linck A, Wesche F, Tietze A, Niesert C-P, Bode HB. 2018. *De novo* design and engineering of non-ribosomal peptide synthetases. *Nat Chem* 10:275–281.
5. Awakawa T, Fujioka T, Zhang L, Hoshino S, Hu Z, Hashimoto J, Kozone I, Ikeda H, Shin-Ya K, Liu W, Abe I. 2018. Reprogramming of the antimycin NRPS-PKS assembly lines inspired by gene evolution. *Nat Commun* 9:3534.
6. Baltz RH. 2018. Synthetic biology, genome mining, and combinatorial biosynthesis of NRPS-derived antibiotics: a perspective. *J Ind Microbiol Biotechnol* 45:635–649.
7. Baltz RH. 2014. Combinatorial biosynthesis of cyclic lipopeptide antibiotics: a model for synthetic biology to accelerate the evolution of secondary metabolite biosynthetic pathways. *ACS Synth Biol* 3:748–758.
8. Yuzawa S, Mirsiaghi M, Jovic R, Fujii T, Masson F, Benites VT, Baidoo EEK, Sundstrom E, Tanjore D, Pray TR, George A, Davis RW, Gladden JM, Simmons BA, Katz L, Keasling JD. 2018. Short-chain ketone production by engineered polyketide synthases in *Streptomyces albus*. *Nat Commun* 9:4569.
9. Yuzawa S, Backman TWH, Keasling JD, Katz L. 2018. Synthetic biology of polyketide synthases. *J Ind Microbiol Biotechnol* 45:621–633.
10. Gomez-Escribano JP, Bibb MJ. 2011. Engineering *Streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. *Microb Biotechnol* 4:207–215.
11. Komatsu M, Komatsu K, Koiwai H, Yamada Y, Kozone I, Izumikawa M, Hashimoto J, Takagi M, Omura S, Shin-ya K, Cane DE, Ikeda H. 2013. Engineered *Streptomyces avermitilis* host for heterologous expression of biosynthetic gene cluster for secondary metabolites. *ACS Synth Biol* 2:384–396.
12. Kallifidas D, Jiang G, Ding Y, Luesch H. 2018. Rational engineering of *Streptomyces albus* J1074 for the overexpression of secondary metabolite gene clusters. *Microb Cell Factories* 17:25.
13. Myronovskiy M, Rosenkränzer B, Nadmid S, Pujic P, Normand P, Luzhetskyy A. 2018. Generation of a cluster-free *Streptomyces albus* chassis strains for improved heterologous expression of secondary metabolite clusters. *Metab Eng* 49:316–324.

14. Baltz RH. 2016. Genetic manipulation of secondary metabolite biosynthesis for improved production in *Streptomyces* and other actinomycetes. *J Ind Microbiol Biotechnol* 43:343–370.
15. Bai C, Zhang Y, Zhao X, Hu Y, Xiang S, Miao J, Lou C, Zhang L. 2015. Exploiting a precise design of universal synthetic modular regulatory elements to unlock the microbial natural products in *Streptomyces*. *Proc Natl Acad Sci U S A* 112:12181–12186.
16. Seghezzi N, Amar P, Koebmann B, Jensen PR, Viroille M-J. 2011. The construction of a library of synthetic promoters revealed some specific features of strong *Streptomyces* promoters. *Appl Microbiol Biotechnol* 90:615–623.
17. Siegl T, Tokovenko B, Myronovskyi M, Luzhetskyy A. 2013. Design, construction and characterisation of a synthetic promoter library for fine-tuned gene expression in actinomycetes. *Metab Eng* 19:98–106.
18. Sohoni SV, Fazio A, Workman CT, Mijakovic I, Lantz AE. 2014. Synthetic promoter library for modulation of actinorhodin production in *Streptomyces coelicolor* A3(2). *PloS One* 9:e99701.
19. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6:343–345.
20. de Kok S, Stanton LH, Slaby T, Durot M, Holmes VF, Patel KG, Platt D, Shapland EB, Serber Z, Dean J, Newman JD, Chandran SS. 2014. Rapid and reliable DNA assembly via ligase cycling reaction. *ACS Synth Biol* 3:97–106.
21. Shao Z, Zhao H, Zhao H. 2009. DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res* 37:e16.
22. Basitta P, Westrich L, Rösch M, Kulik A, Gust B, Apel AK. 2017. AGOS: a plug-and-play method for the assembly of artificial gene operons into functional biosynthetic gene clusters. *ACS Synth Biol* 6:817–825.
23. Shetty RP, Endy D, Knight TF Jr. 2008. Engineering BioBrick vectors from BioBrick parts. *J Biol Eng* 2:5.
24. Engler C, Kandzia R, Marillonnet S. 2008. A one pot, one step, precision cloning method with high throughput capability. *PloS One* 3:e3647.
25. Weber E, Engler C, Gruetzner R, Werner S, Marillonnet S. 2011. A modular cloning system for standardized assembly of multigene constructs. *PloS One* 6:e16765.
26. Sarrion-Perdigones A, Vazquez-Vilar M, Palací J, Castelijns B, Forment J, Ziarsolo P, Blanca J, Granell A, Orzaez D. 2013. GoldenBraid 2.0: a comprehensive DNA assembly framework for plant synthetic biology. *Plant Physiol* 162:1618–1631.
27. Olorunniji FJ, Merrick C, Rosser SJ, Smith MCM, Stark WM, Colloms SD. 2017. Multipart DNA assembly using site-specific recombinases from the large serine integrase family. *Methods Mol Biol Clifton NJ* 1642:303–323.
28. Silva-Rocha R, Martínez-García E, Calles B, Chavarría M, Arce-Rodríguez A, de Las Heras A, Páez-Espino AD, Durante-Rodríguez G, Kim J, Nikel PI, Platero R, de Lorenzo V.

2013. The Standard European Vector Architecture (SEVA): a coherent platform for the analysis and deployment of complex prokaryotic phenotypes. *Nucleic Acids Res* 41:D666-675.
29. Phelan RM, Sachs D, Petkiewicz SJ, Barajas JF, Blake-Hedges JM, Thompson MG, Reider Apel A, Rasor BJ, Katz L, Keasling JD. 2017. Development of next generation synthetic biology tools for use in *Streptomyces venezuelae*. *ACS Synth Biol* 6:159–166.
30. Juguet M, Lautru S, Francou F-X, Nezbedová S, Leblond P, Gondry M, Pernodet J-L. 2009. An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congocidine in *Streptomyces ambofaciens*. *Chem Biol* 16:421–431.
31. Alieva NO, Konzen KA, Field SF, Meleshkevitch EA, Hunt ME, Beltran-Ramirez V, Miller DJ, Wiedenmann J, Salih A, Matz MV. 2008. Diversity and evolution of coral fluorescent proteins. *PLoS ONE* 3.
32. Gust B, Chandra G, Jakimowicz D, Yuqing T, Bruton CJ, Chater KF. 2004. Lambda red-mediated genetic manipulation of antibiotic-producing *Streptomyces*. *Adv Appl Microbiol* 54:107–128.
33. Fedoryshyn M, Petzke L, Welle E, Bechthold A, Luzhetskyy A. 2008. Marker removal from actinomycetes genome using Flp recombinase. *Gene* 419:43–47.
34. Siegl T, Luzhetskyy A. 2012. Actinomycetes genome engineering approaches. *Antonie Van Leeuwenhoek* 102:503–516.
35. Lautru S, Gondry M, Genet R, Pernodet JL. 2002. The albonoursin gene cluster of *S. noursei*: biosynthesis of diketopiperazine metabolites independent of nonribosomal peptide synthetases. *Chem Biol* 9:1355–1364.
36. Holmes DJ, Caso JL, Thompson CJ. 1993. Autogenous transcriptional activation of a thiostrepton-induced gene in *Streptomyces lividans*. *EMBO J* 12:3183–3191.
37. Li Y, Lai Y-M, Lu Y, Yang Y-L, Chen S. 2014. Analysis of the biosynthesis of antibacterial cyclic dipeptides in *Nocardiopsis alba*. *Arch Microbiol* 196:765–774.
38. Prentki P, Krisch HM. 1984. In vitro insertional mutagenesis with a selectable DNA fragment. *Gene* 29:303–313.
39. Morita K, Yamamoto T, Fusada N, Komatsu M, Ikeda H, Hirano N, Takahashi H. 2009. The site-specific recombination system of actinophage TG1. *FEMS Microbiol Lett* 297:234–240.
40. Fayed B, Younger E, Taylor G, Smith MCM. 2014. A novel *Streptomyces* spp. integration vector derived from the *S. venezuelae* phage, SV1. *BMC Biotechnol* 14:51.
41. Fogg PCM, Haley JA, Stark WM, Smith MCM. 2017. Genome integration and excision by a new *Streptomyces* bacteriophage,  $\Phi$ Joe. *Appl Environ Microbiol* 83.
42. Kieser T, Bibb M, Buttner M, Hopwood DA. 2000. *Practical Streptomyces genetics*, John Innes Foundation, Norwich NR47UH, UK.
43. Pernodet JL, Alegre MT, Blondelet-Rouault MH, Guérineau M. 1993. Resistance to spiramycin in *Streptomyces ambofaciens*, the producer organism, involves at least two different mechanisms. *J Gen Microbiol* 139:1003–1011.

44. Sambrook J, Russell DW. 2001. Molecular cloning: A laboratory manual, Third edition. CSHL Press, Cold Spring Harbor, NY.
45. Gregory MA, Till R, Smith MCM. 2003. Integration site for *Streptomyces* phage  $\phi$ BT1 and development of site-specific integrating vectors. *J Bacteriol* 185:5320–5323.
46. Vingadassalon A, Lorieux F, Juguet M, Le Goff G, Gerbaud C, Pernodet J-L, Lautru S. 2015. Natural combinatorial biosynthesis involving two clusters for the synthesis of three pyrrolamides in *Streptomyces netropsis*. *ACS Chem Biol* 10:601–610.
47. Bierman M, Logan R, O'Brien K, Seno ET, Rao RN, Schonher BE. 1992. Plasmid cloning vectors for the conjugal transfer of DNA from *Escherichia coli* to *Streptomyces* spp. *Gene* 116:43–49.
48. Cunningham FX, Pogson B, Sun Z, McDonald KA, DellaPenna D, Gantt E. 1996. Functional analysis of the beta and epsilon lycopene cyclase enzymes of *Arabidopsis* reveals a mechanism for control of cyclic carotenoid formation. *Plant Cell* 8:1613–1626.
49. Gibson DG. 2011. Enzymatic assembly of overlapping DNA fragments. *Methods Enzymol* 498:349–361.
50. Laible M, Boonrod K. 2009. Homemade site directed mutagenesis of whole plasmids. *J Vis Exp JoVE*.
51. Nguyen HC, Karray F, Lautru S, Gagnat J, Lebrihi A, Ho Huynh TD, Pernodet J-L. 2010. Glycosylation steps during spiramycin biosynthesis in *Streptomyces ambofaciens*: Involvement of three glycosyltransferases and their interplay with two auxiliary proteins. *Antimicrob Agents Chemother* 54:2830–2839.
52. Van Mellaert L, Mei L, Lammertyn E, Schacht S, Anné J. 1998. Site-specific integration of bacteriophage VWB genome into *Streptomyces venezuelae* and construction of a VWB-based integrative vector. *Microbiol Read Engl* 144 ( Pt 12):3351–3358.
53. Chandran S. 2017. Rapid Assembly of DNA via Ligase Cycling Reaction (LCR). *Methods Mol Biol Clifton NJ* 1472:105–110.
54. Raynal A, Karray F, Tuphile K, Darbon-Rongère E, Pernodet J-L. 2006. Excisable cassettes: new tools for functional analysis of *Streptomyces* genomes. *Appl Environ Microbiol* 72:4839–4844.
55. Flett F, Mersinias V, Smith CP. 1997. High efficiency intergeneric conjugal transfer of plasmid DNA from *Escherichia coli* to methyl DNA-restricting streptomycetes. *FEMS Microbiol Lett* 155:223–229.
56. Simon R, Priefer U, Pühler A. 1983. A broad host range mobilization system for *in vivo* genetic engineering: transposon mutagenesis in gram negative bacteria. *Nat Biotechnol* 1:784–791.

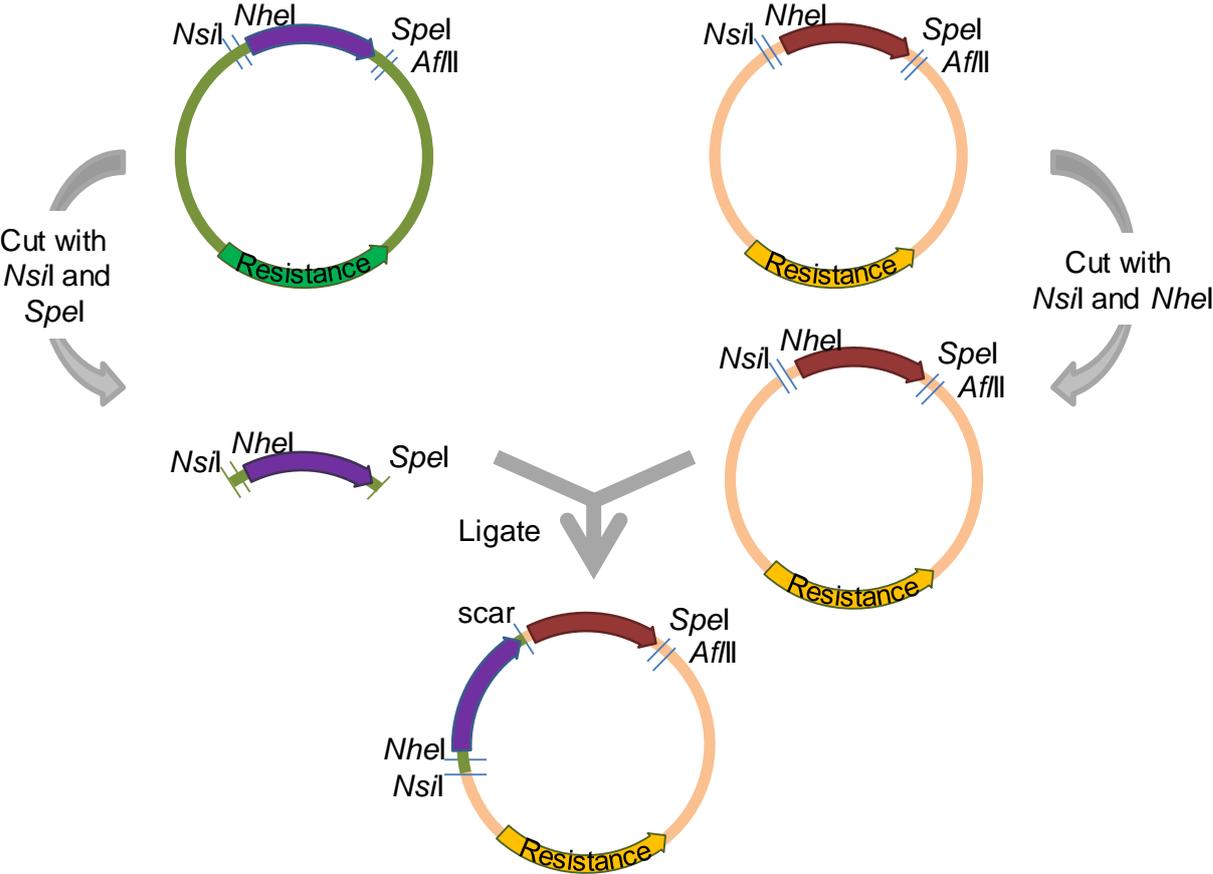
# Modular and Integrative Vectors for Synthetic Biology Applications in *Streptomyces* spp.

Céline AUBRY<sup>a</sup>, Jean-Luc PERNODET<sup>a</sup> and Sylvie LAUTRU<sup>a#</sup>

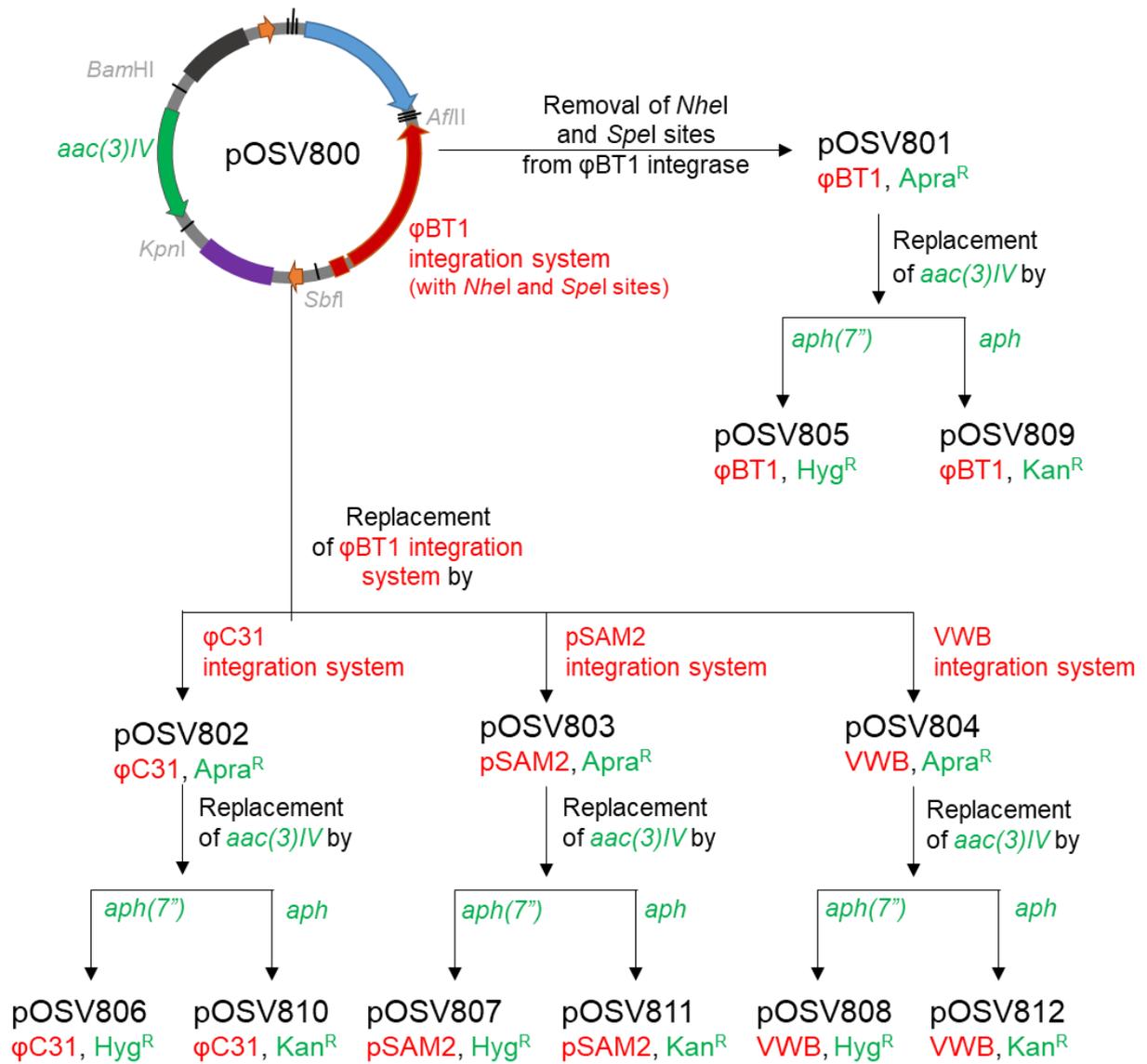
<sup>a</sup> Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

<sup>#</sup> Corresponding author: Sylvie LAUTRU, [sylvie.lautru@i2bc.paris-saclay.fr](mailto:sylvie.lautru@i2bc.paris-saclay.fr)

## Supplemental material

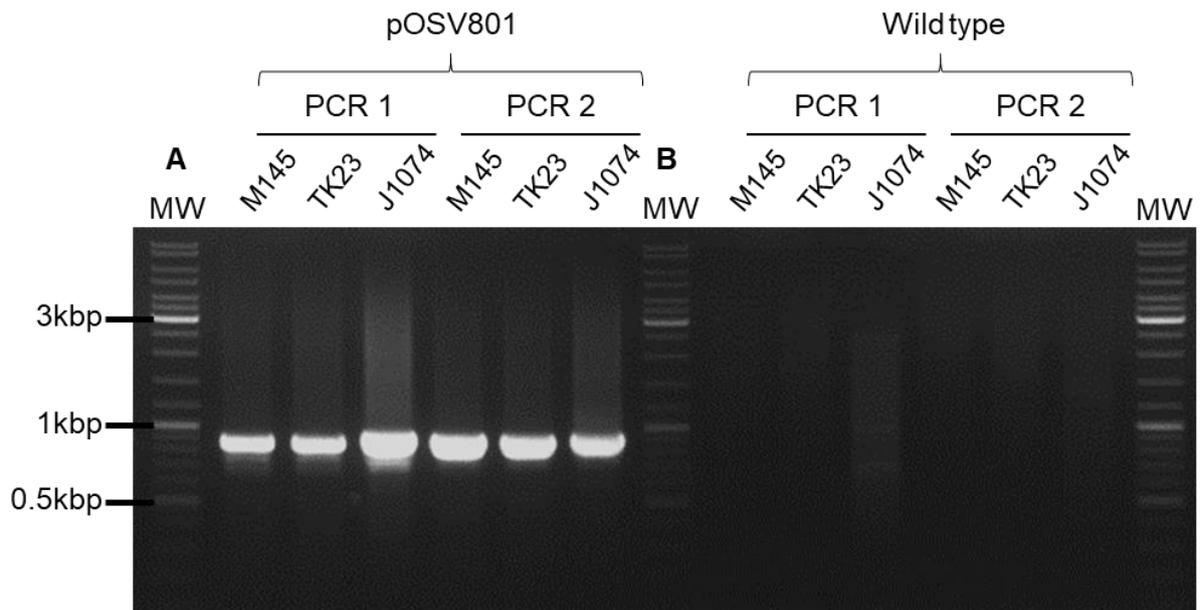


**Figure S1:** Principle of the Biobrick cloning method.



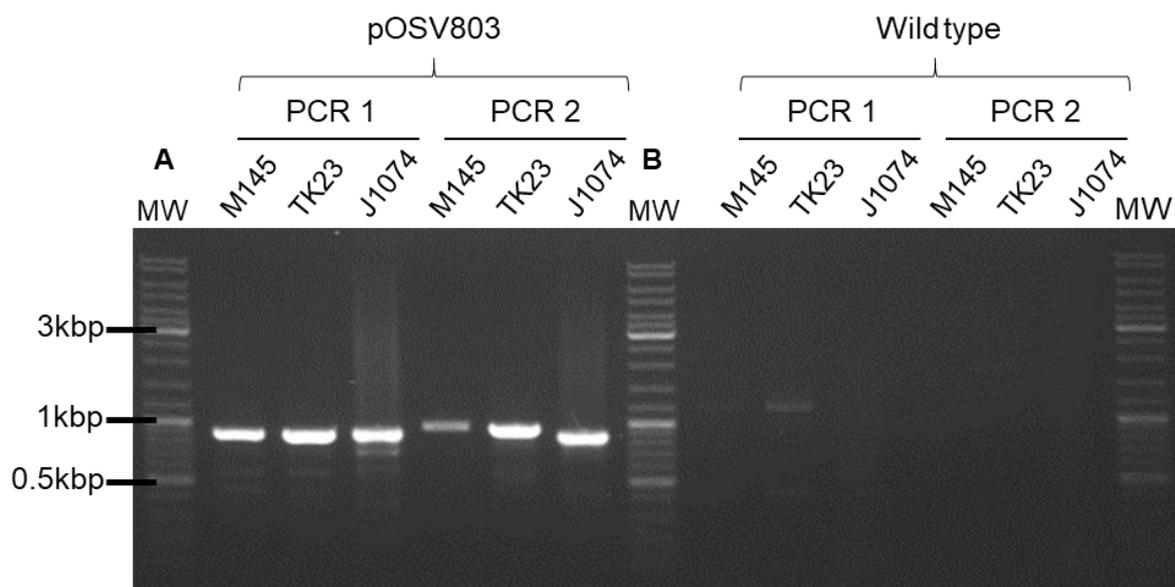
**Figure S2:** General scheme for the construction of the pOSV801-812 vectors.

In red is shown the integration systems borne by the vector and in green the antibiotic resistance conferred by the vector.



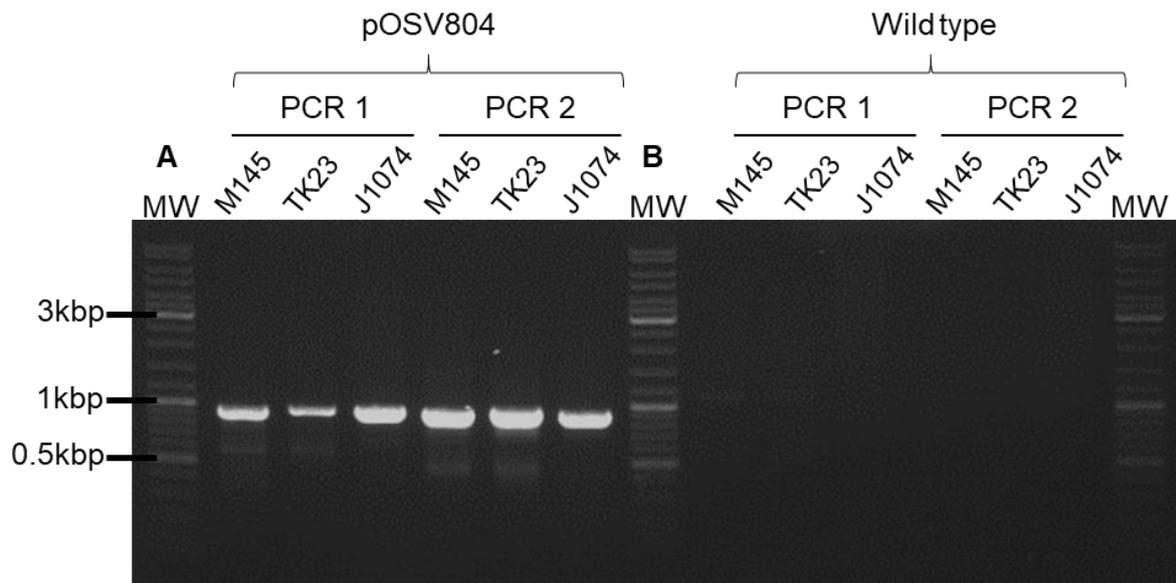
**Figure S3:** Verification of the integration of pOSV801 in *S. calicolor* M145, *S. lividans* TK23 and *S. albus* J1074 chromosomes.

A. PCR fragments obtained by PCR 1 (*attL* region; expected size: 870 bps for M145 and TK23, 899 bps for J1074) and by PCR 2 (*attR* region; expected size: 877 bps for M145 and TK23, 893 bps for J1074) on the three *Streptomyces* strains bearing pOSV801. B. Negative control PCRs using genomic DNA of the wild type *Streptomyces* strains



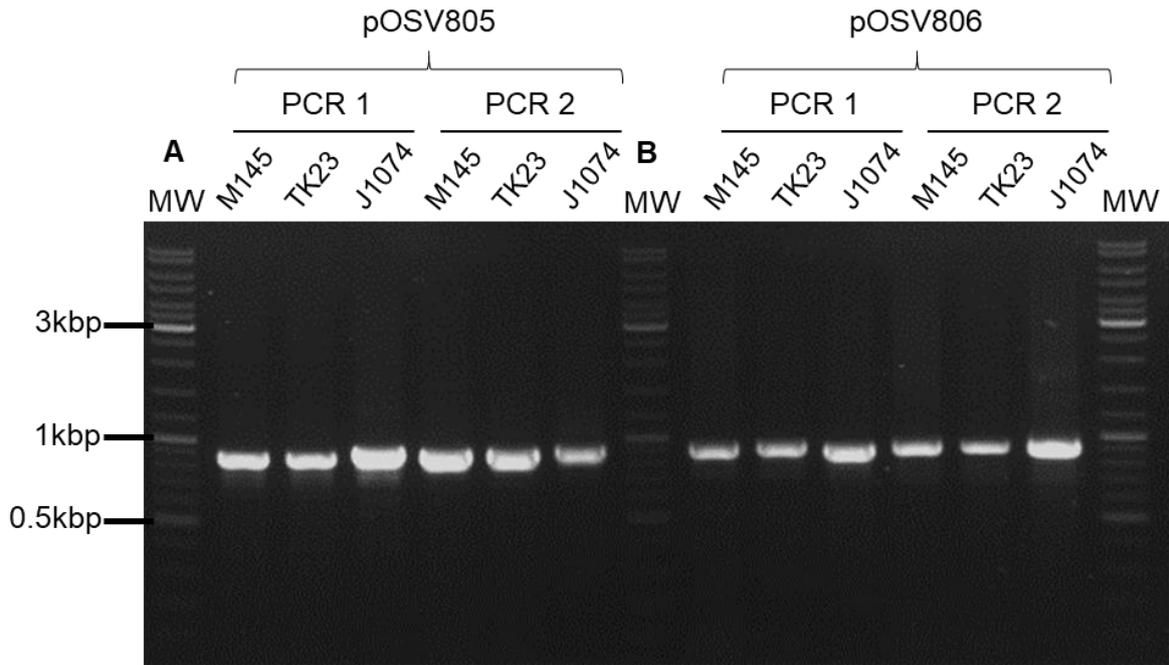
**Figure S4:** Verification of the integration of pOSV803 in *S. calicolor* M145, *S. lividans* TK23 and *S. albus* J1074 chromosomes.

A. PCR fragments obtained by PCR 1 (*attL* region; expected size: 896 bps for M145 and TK23, 911 bps for J1074) and by PCR 2 (*attR* region; expected size: 937 bps for M145 and TK23, 873 bps for J1074) on the three *Streptomyces* strains bearing pOSV803. B. Negative control PCRs using genomic DNA of the wild type *Streptomyces* strains.



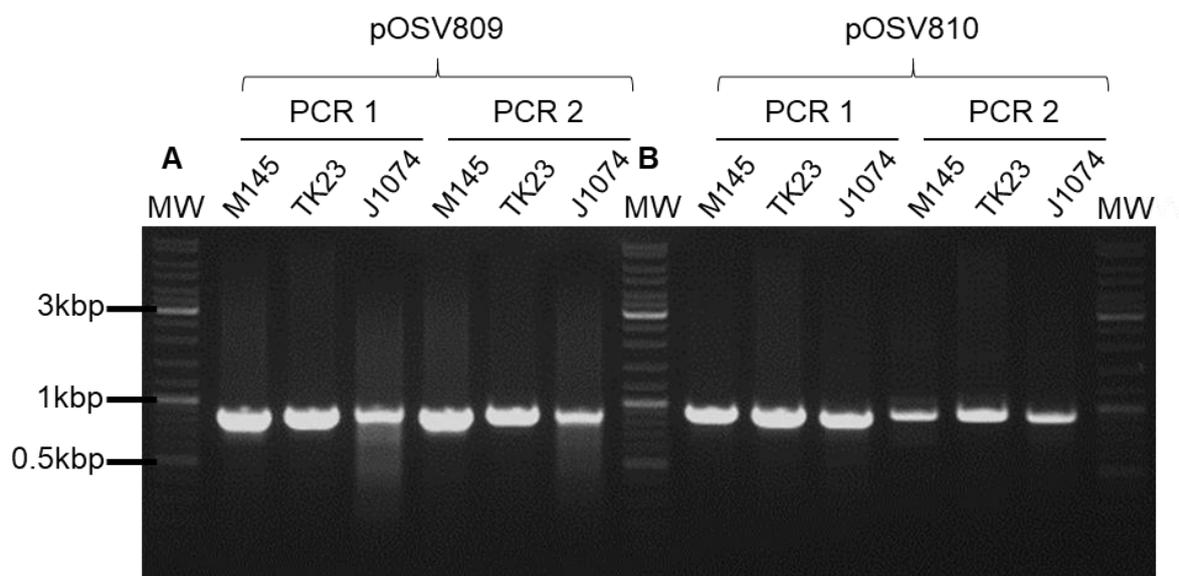
**Figure S5:** Verification of the integration of pOSV804 in *S. calicolor* M145, *S. lividans* TK23 and *S. albus* J1074 chromosomes.

A. PCR fragments obtained by PCR 1 (*attL* region; expected size: 915 bps for M145 and TK23, 925 bps for J1074) and by PCR 2 (*attR* region; expected size: 905 bps for M145, 920 bps for TK23 and 884 bps for J1074) on the three *Streptomyces* strains bearing pOSV804. B. Negative control PCRs using genomic DNA of the wild type *Streptomyces* strains.



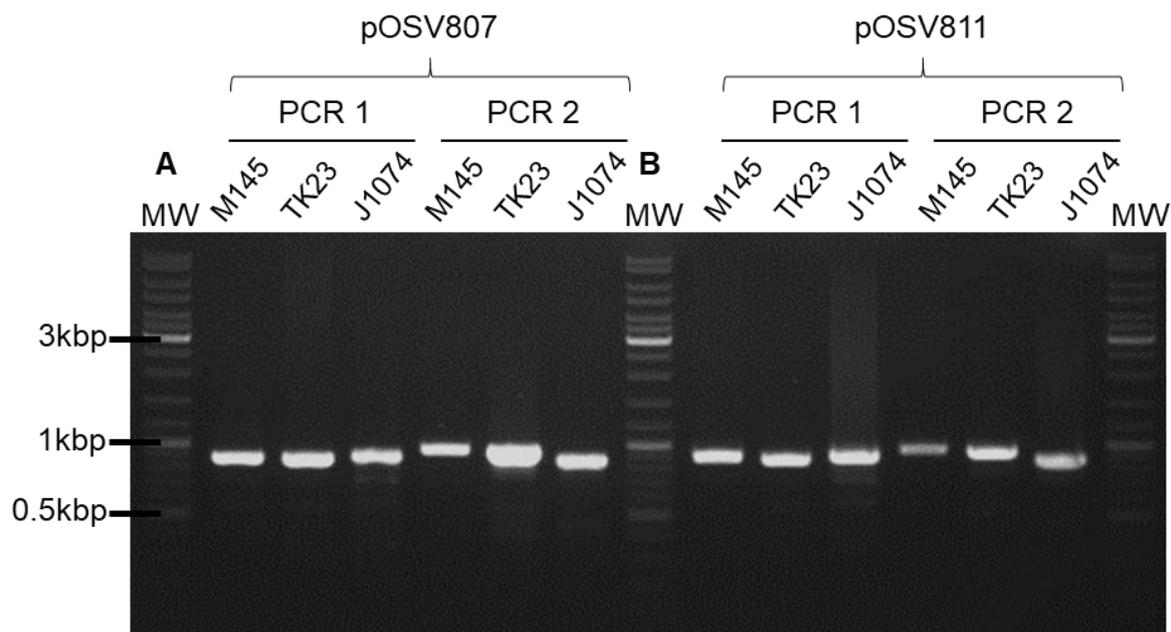
**Figure S6:** Verification of the integration of pOSV805 and pOSV806 in *S. cellicolor* M145 and *S. lividans* TK23, *S. albus* J1074 chromosomes.

A. PCR fragments obtained by PCR 1 (*attL* region; expected size: 870 bps for M145 and TK23, 899 bps for J1074) and by PCR 2 (*attR* region; expected size: 877 bps for M145 and TK23, and 893 bps for J1074) on the three *Streptomyces* strains bearing the pOSV805. B. PCR fragments obtained by PCR 1 (*attL* region; expected size: 913 bps for M145 and TK23, 888 bps for J1074) and by PCR 2 (*attR* region; expected size: 911 bps for M145 and TK23, and 907 bps for J1074) on the three *Streptomyces* strains bearing pOSV806.



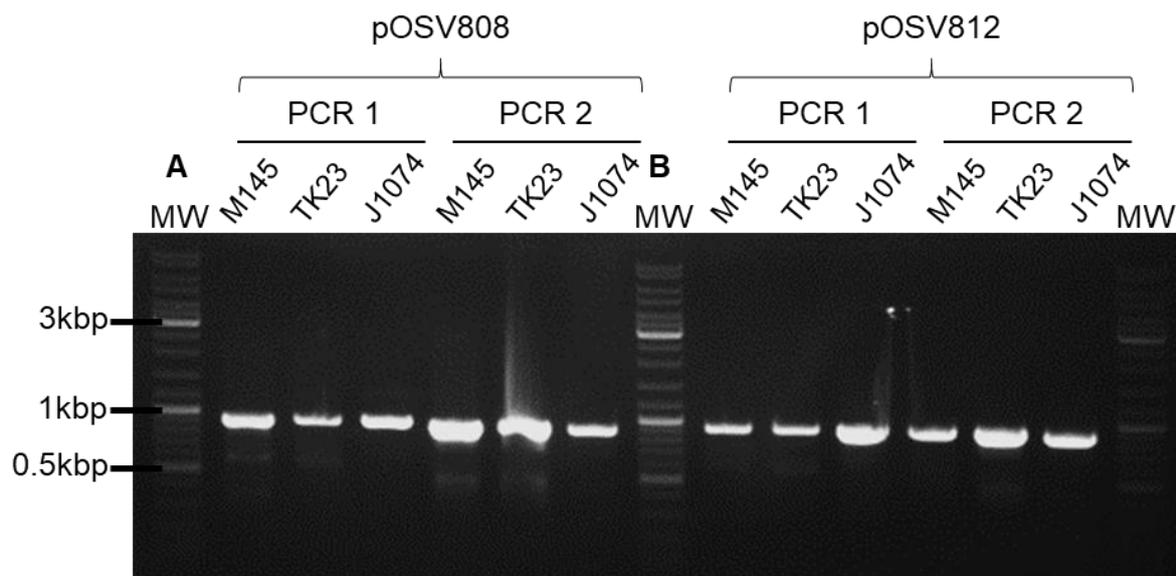
**Figure S7:** Verification of the integration of pOSV809 and pOSV810 in *S. calicolor* M145 and *S. lividans* TK23, *S. albus* J1074 chromosomes.

A. PCR fragments obtained by PCR 1 (*attL* region; expected size: 870 bps for M145 and TK23, 899 bps for J1074) and by PCR 2 (*attR* region; expected size: 877 bps for M145 and TK23, and 893 bps for J1074) on the three *Streptomyces* strains bearing pOSV809. B. PCR fragments obtained by PCR 1 (*attL* region; expected size: 913 bps for M145 and TK23, 888 bps for J1074) and by PCR 2 (*attR* region; expected size: 911 bps for M145 and TK23, and 907 bps for J1074) on the three *Streptomyces* strains bearing pOSV810.



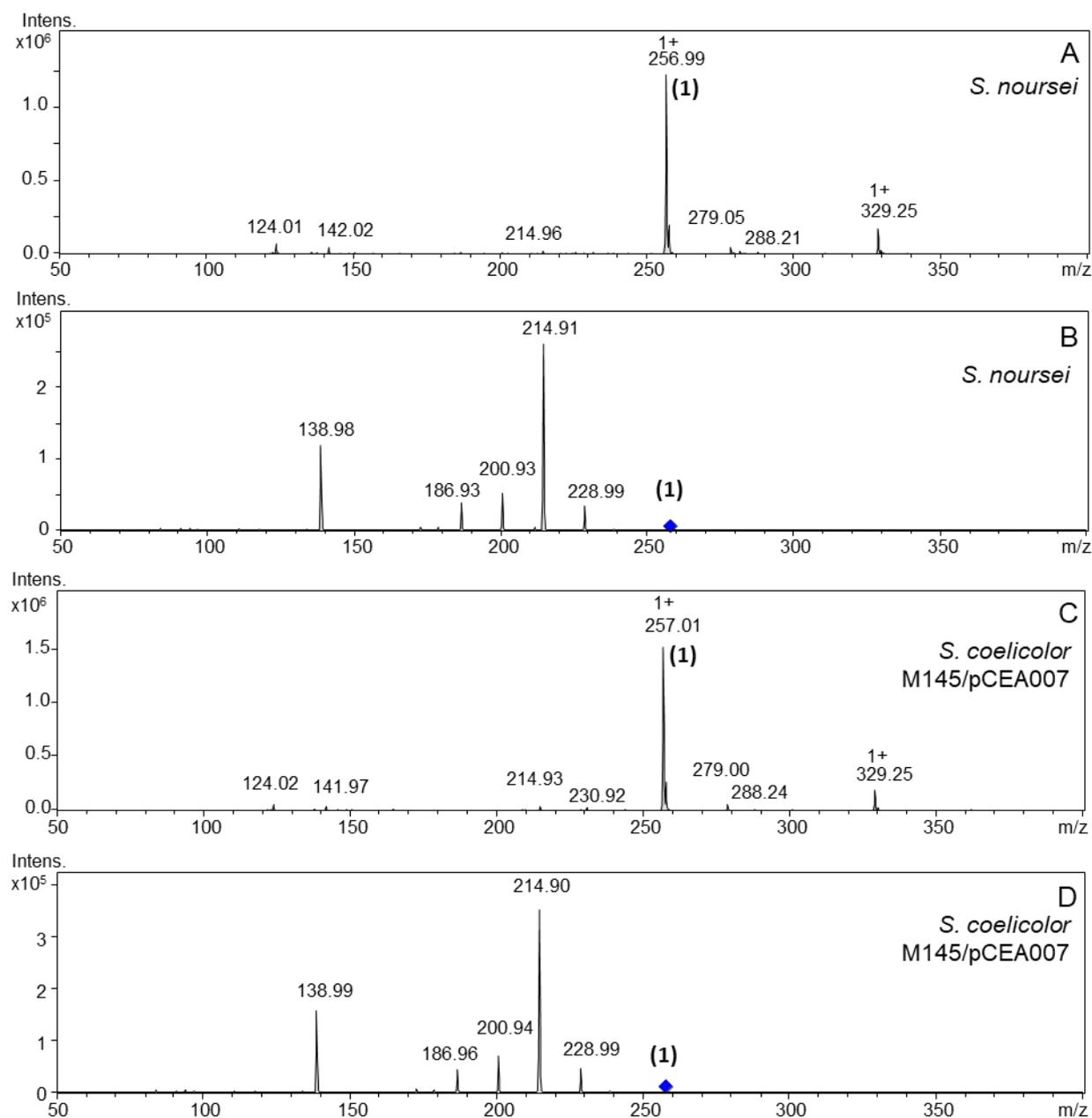
**Figure S8:** Verification of the integration of pOSV807 and pOSV811 in *S. calicolor* M145, *S. lividans* TK23 and *S. albus* J1074 chromosomes.

A. PCR fragments obtained by PCR 1 (*attL* region; expected size: 915 bps for M145 and TK23, 925 bps for J1074) and by PCR 2 (*attR* region; expected size: 905 bps for M145 and TK23, and 884 bps for J1074) on the three *Streptomyces* strains bearing pOSV807. B. PCR fragments obtained by PCR 1 (*attL* region; expected size: 915 bps for M145 and TK23, 925 bps for J1074) and by PCR 2 (*attR* region; expected size: 905 bps for M145 and TK23, and 884 bps for J1074) on the three *Streptomyces* strains bearing pOSV811.



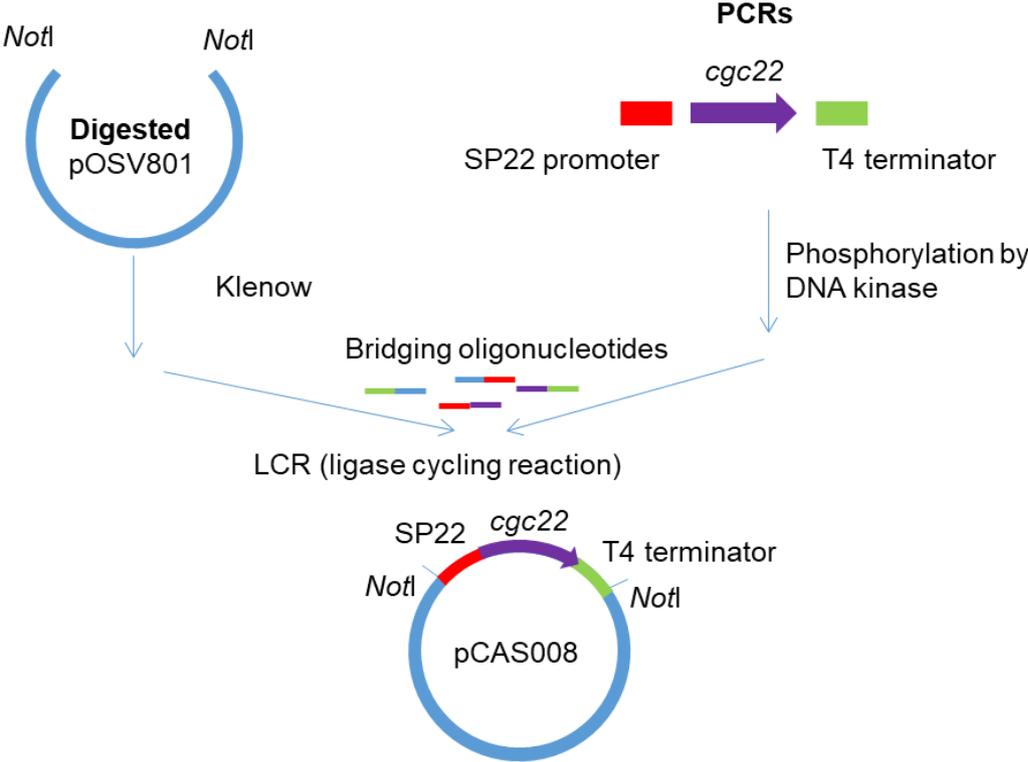
**Figure S9:** Verification of the integration of pOSV808 and pOSV812 in *S. calicolor* M145, *S. lividans* TK23, *S. albus* J1074 chromosomes.

A. PCR fragments obtained by PCR 1 (*attL* region; expected size: 896 bps for M145 and TK23, 911 bps for J1074) and by PCR 2 (*attR* region; expected size: 937 bps for M145 and TK23, and 873 bps for J1074) on the three *Streptomyces* strains bearing pOSV808. B. PCR fragments obtained by PCR 1 (*attL* region; expected size: 896 bps for M145 and TK23, 911 bps for J1074) and by PCR 2 (*attR* region; expected size: 937 bps for M145 and TK23, and 873 bps for J1074) on the three *Streptomyces* strains bearing pOSV812.



**Figure S10:** ESI-MS and ESI-MS/MS fragmentation patterns in positive mode of albonoursin in the culture supernatants.

MS spectrum of albonoursin ( $m/z$  257.1) from A. *S. noursei* supernatant and C. *S. coelicolor* M145/pCEA007 supernatant and MS/MS fragmentation patterns of (1) in B. and D.



**Figure S11:** Scheme for the LCR assembly of the pCAS008 vector.

# Chapter III - Refactoring of the congocidine biosynthetic gene cluster: from gene cassettes to gene cluster

## Chapter III introduction:

In this third chapter, I present my work on the refactoring of the congocidine biosynthetic gene cluster (21 genes) in *Streptomyces lividans* TK23, using synthetic gene cassettes and integrative vectors we constructed. Indeed, this successful refactoring of a known pyrrolamide biosynthetic pathway confirms the potential of our approach and should open the way to combinatorial biosynthesis experiments using pyrrolamide biosynthetic pathways.

This work is presented using the format of an article, and a short perspective at the end of the chapter describes the elements missing to complete this work.

# Refactoring of the congocidine biosynthetic gene cluster: from gene cassettes to gene cluster

Céline AUBRY<sup>a</sup>, Jennifer PERRIN<sup>a</sup>, Yacine Mohammed SELLAH<sup>a</sup>, Jean-Luc PERNODET<sup>a</sup> and Sylvie LAUTRU<sup>a#</sup>

<sup>a</sup> Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

## ABSTRACT

Pathway refactoring is a synthetic biology approach that consists in rewriting DNA sequence containing all the genetic information necessary for the expression and functioning of a metabolic pathway in heterologous or native host. It is often used to decouple gene expression from its native complex regulation, which, in the field of specialized metabolism, allows the expression of silent biosynthetic gene clusters. It can also be used to optimize the production yield of a metabolite or as a first step towards the generation of analogs by combinatorial biosynthesis.

We report here the refactoring of the biosynthetic gene cluster of the pyrrolamide congocidine (*cgc*). We constructed 11 basic gene cassettes, designed to constitute functional units, to express the 21 genes of the *cgc* cluster. The functionality of each cassette was verified through a combination of genetic complementation of mutant strains, HPLC analyses and bioassays. The gene cassettes were then assembled on two compatible integrative plasmids. After introduction of both constructs in *Streptomyces lividans* TK23, congocidine production was confirmed in the host strain. This work opens the way to future combinatorial biosynthesis experiments based on the pyrrolamide biosynthetic gene clusters.

**KEYWORDS** *Streptomyces*, refactoring, pyrrolamide

## INTRODUCTION

Microbial specialized metabolites have been an important source of pharmaceuticals (Newman and Cragg, 2016) and are likely to continue, as these metabolites constitute our best line of defense so far against pathogenic microorganisms. The explosion of microbial genome sequencing in the last 15 years, combined with the exploration of new ecological niches and microbial genera, has highlighted the extraordinary reservoir of specialized metabolites that remains to be explored and that may deliver the antibiotics of tomorrow.

In parallel to this genomic exploration, the search of new pharmaceutically active metabolites has led to the emergence of a new field, the synthetic biology of specialized metabolites. Here, the objective is to genetically manipulate biosynthetic gene clusters directing the biosynthesis of specialized metabolites, towards the production of new and unnatural metabolites.

Both the mining of microbial genomes and the synthetic biology of specialized metabolites rely on the availability of genetic tools. Indeed, the genomic exploration of the best-studied model microorganisms has shown that these microorganisms still have the potential to produce one to a few dozens of specialized metabolites that were not detected previously (Aigle et al., 2014; Ōmura et al., 2001). In many cases, the gap between the number of biosynthetic genes clusters identified in genomes and the number of observed specialized metabolites is linked to the absence of gene cluster expression in standard laboratory conditions. To activate the expression of such silent (or cryptic) gene clusters, empirical methods have been developed (Genilloud, 2018). However, these methods are usually not well suited when a specific gene cluster is targeted. In this case, genetic methods are generally used. Such methods include the heterologous expression in a genetically tractable host (Gomez-Escribano and Bibb, 2014) or the genetic manipulation (deletion or overexpression) of pathway-specific regulators (Yamanaka et al., 2014). Yet, pathway-specific regulators are not always present in biosynthetic gene clusters and heterologous expression does not always result in gene cluster expression and metabolite production. In these cases, a recently developed method, called pathway refactoring, is increasingly used.

Pathway refactoring is a synthetic biology approach that was first developed to decouple pathway expression from its native regulation and to facilitate the transfer of gene clusters into relatively distant microbial host (Temme et al., 2012). In the field of specialized metabolism, it has so far essentially been used to replace regulatory native elements, such as promoters, by well-characterized elements, thus removing all native regulation and allowing (constitutive) gene expression and metabolite production (Bauman et al., 2019; Eyles et al., 2018; Kim et al., 2019; Luo et al., 2013; Song et al., 2019; Tan et al., 2017). Yet, pathway refactoring is not restricted to the modification of regulatory elements. In synthetic biology, it has also been used to optimize DNA sequence for heterologous expression (Osswald et al., 2014), or to create artificial and functional transcriptional units that can then be assembled to reconstitute a functional gene cluster. This is often seen as a first step toward the genetic manipulation of the cluster and the production of new unnatural metabolites (Basitta et al., 2017; Osswald et al., 2014). It was with this objective that we undertook the refactoring of the congocidine gene cluster.

Congocidine (also called netropsin, Figure S1A) belongs to the family of pyrrolamide metabolites (distamycin, anthelvencin, pyrromycins...) characterized by the presence of 4-aminopyrrole-2-carboxylic moieties. Most pyrrolamides are minor groove binders. They display a variety of biological activities (antibacterial, antifungal, antiviral activities), but none have been exploited in medicine, mostly due to their toxicity. Congocidine, and more generally pyrrolamides, are assembled by enzymes of the non ribosomal peptide synthetase (NRPS) family. NRPSs are usually large, multimodular enzymes (Strieker et al., 2010) that are difficult to biochemically (and genetically for their associated genes) manipulate. On the contrary, NRPSs involved in the biosynthesis of pyrrolamides consist of stand-alone modules and domains (Juguet et al., 2009; Vingadassalon et al., 2015; Aubry et al., unpublished). Furthermore, pyrrolamides appear to be combinatorially assembled from a limited number of precursors and “natural combinatorial biosynthesis” has already been observed in *Streptomyces netropsis* DSM40846 producing congocidine, distamycin and disgocidine (Vingadassalon et al., 2015). For these reasons, we thought that the biosynthetic systems of pyrrolamides constituted attractive systems to carry out combinatorial biosynthetic experiments, notably at the level of the NRPS modules and domains, and study the various key elements (substrate specificity, protein interactions...) essential to the success of NRPS synthetic biology. With this goal in mind, we undertook the refactoring of the congocidine biosynthetic cluster. Our aims were (i) to control the expression of the *cgc* genes and, later on, of other pyrrolamide biosynthetic genes (remove the *cgc* native regulation), and (ii) to reorganize the genes into new transcriptional and functional units that will be reusable for combinatorial biosynthesis experiments (design of standardized gene cassettes, orthogonal and easily exchangeable).

## RESULTS AND DISCUSSION

### *Principles of the cgc gene cluster refactoring*

The refactoring of the congocidine biosynthetic gene cluster constitutes a first step towards the combinatorial biosynthesis of pyrrolamides. Thus, each gene cassette constructed for the refactoring of the *cgc* gene cluster was designed with this future use in mind. Four types of basic gene cassettes were designed: the Precursor, the Assembly, the Tailoring and the Resistance gene cassettes (Figure S1B). These basic gene cassettes are then meant to be progressively assembled into composite gene cassettes by a Biobrick-type of assembly to reconstitute the *cgc* gene cluster.

The Precursor gene cassettes include all genes necessary for the biosynthesis of a given precursor. Congocidine is assembled from three precursors, 3-aminopropionamide, guanidinoacetate and 4-acetamidopyrrole-2-carboxylate. Thus, three precursor gene cassettes were constructed. The 3-aminopropionamide gene cassette is constituted of the three genes *cgc4*, *cgc5* and *cgc6* involved in the biosynthesis of this precursor (Elie et al., unpublished). The biosynthesis of guanidinoacetate involves *cgc6* and *cgc7* (Elie et al., unpublished). As *cgc6* is already included in the 3-aminopropionamide gene cassette, the guanidinoacetate gene cassette is solely constituted of *cgc7*. The biosynthesis of 4-acetamidopyrrole-2-carboxylate requires the eight genes *cgc3*, *cgc8-cgc13* and *cgc17* (Lautru et al., 2012). These genes were included in the composite pyrrole gene cassette, together with *cgc14*. Indeed, although *cgc14* is not involved in the biosynthesis of 4-acetamidopyrrole-2-carboxylate, it codes for the enzyme responsible for the deacetylation of this molecule, once loaded on the PCP domain Cgc19. As this deacetylation is a prerequisite before any

condensation of the pyrrole precursor with another molecule, *cgc14* was included in the pyrrole gene cassette. Five Assembly gene cassettes were designed, each containing a single gene (*cgc2*, *cgc16*, *cgc19*, *cgc18* and *cgc22* respectively). The assembly genes were not combined as they should be individually exchangeable in combinatorial biosynthesis experiments. Finally, one Tailoring (*cgc15*, coding for a methyltransferase) and one Resistance (*cgc20* and *cgc21* encoding an ABC transporter) gene cassettes were designed.

Each gene cassette but one (the composite pyrrole precursor gene cassette) is constituted of a transcriptional unit, composed of a promoter, a ribosome binding site (RBS), one or several congoicidine biosynthetic genes (*cgc*) and a terminator. To overcome the native transcriptional regulation of the *cgc* genes, we opted for the use of synthetic elements to induce expression. Thus, the promoters were chosen among a set of synthetic promoters (SP), derived from the optimized and strong *kasOp\** promoter and classified by their relative strength compared to that of this promoter (Bai et al., 2015). Several studies have emphasized that the outcome of the use of genetic elements such as promoters or RBS is often influenced by genetic context (Vilanova et al., 2015; Yeung et al., 2017). Yet, because it is difficult to predict the influence of this context, we chose the six different promoters we used based on their relative strength as defined by Bai and colleagues (2015). The strength of the promoters we used varies between 0.25 (SP20) and 1.87 (SP44) fold the strength of *kasOp\**, but most (SP22-SP25) have roughly half the strength of this promoter. The weakest promoter SP20 was chosen for the expression of the resistance genes (*cgc20* and *cgc21*), as overexpression of membrane proteins Cgc20 and Cgc21 may have deleterious effects on membrane integrity (Wagner et al., 2007). The expression of all biosynthetic genes but *cgc19* is under the control of four medium strength promoters (SP22-SP25), to avoid imposing too much of a metabolic burden to the cell. Different promoters were chosen to limit sequence repetitions. As for *cgc19*, the PCP domain encoded by this gene is central in congoicidine biosynthesis, carrying all covalently tethered intermediates along the biosynthetic chain (Al-Mestarihi et al., 2015; Vingadassalon et al., 2015). For this reason, we chose a stronger promoter (SP44) for *cgc19* expression. The RBS of the gene coding for the protein of the  $\varphi$ C31 phage capsid, used by Bai *et al.* (2015) during their promoter characterization, was used in all constructions. To better insulate our gene cassettes and allow their sequential and orthogonal use, we decided to add a terminator at the end of the each cassette. While synthetic promoters have been developed recently for *Streptomyces* species, the number of characterized terminators remained really low at the onset of this study. We settled to use the T4 terminator associated to the gene *ssb* (gp32) in the T4 bacteriophage (Prentki and Krisch, 1984) in all our gene cassettes.

As previously mentioned, each basic gene cassette is constituted of a single transcriptional unit, except for the composite pyrrole precursor gene cassette, constituted of nine genes spanning nearly 12 kb. We kept the *cgc8-cgc14* genes, natively cotranscribed in one operon (Vingadassalon et al., unpublished). The two remaining genes, *cgc3* and *cgc17*, physically separated in the native gene cluster, were placed together under the control of another promoter to form a new operon.

Altogether, we designed 11 synthetic gene cassettes to refactor the congoicidine biosynthetic gene cluster (Table 1).

Table 1: List of *cgc* gene cassettes constructed in this study

Name	Composition	Plasmid name
Basic <i>cgc</i> gene cassettes		
CAS001	SP23-cgc4-6-term T4	pCAS001
CAS002	SP20-cgc15-term T4	pCAS002
CAS003	SP23-cgc7-term T4	pCAS003
CAS005	SP23-cgc3 and cgc17-term T4	pCAS005
CAS006	SP20-cgc20-21-Term T4	pCAS006
CAS007	SP25-cgc8-14-Term T4	pCAS007
CAS008	SP22-cgc22-Term T4	pCAS008
CAS009	SP24-cgc2-Term T4	pCAS009
CAS010	SP22-cgc18-Term T4	pCAS010
CAS011	SP24-cgc16-Term T4	pCAS011
CAS013	SP44-cgc19-Term T4	pCAS013
Composite <i>cgc</i> gene cassettes		
CAS016	cgc18; cgc15	pCAS016
CAS017	cgc22; cgc19	pCAS017
CAS018	cgc2; cgc16	pCAS018
CAS019	cgc4-6; cgc7	pCAS019
CAS020	cgc3 and cgc17; cgc8-14	pCAS020
CAS022	cgc18; cgc15; cgc2; cgc16	pCAS022
CAS023	cgc4-6; cgc7; cgc3 and cgc17; cgc8-14	pCAS023
CAS024	cgc22; cgc19; cgc18; cgc15; cgc2; cgc16	pCAS024
CAS026	cgc4-6; cgc7; cgc3 and cgc17; cgc8-14; cgc20-21	pCAS026

### ***Construction of the basic gene cassettes***

Each basic gene cassette (except for the pyrrole precursor gene cassette CAS007, see below) was assembled using the ligase cycling reaction (LCR) (de Kok et al., 2014). This seamless assembly is based on the use of a thermostable ligase and multiple temperature cycles of denaturation-annealing-ligation. Bridging oligonucleotides, whose sequences are complementary to the sequences of the extremities of two DNA fragments to be assembled, are used as a matrix to anneal the two fragments, which are then ligated by the thermostable ligase (Figure 1). The modular vectors pOSV801 and pOSV812, previously constructed to facilitate gene cassette constructions and assembly (Aubry et al., 2019), were used as backbones. The composite pyrrole precursor gene cassette is constituted of two operons, the *cgc8-cgc14* operon, and the *cgc3* and *cgc17* operon. The *cgc3* and *cgc17* operon (CAS005) was assembled by LCR as described above. Due its relatively large size (8 kb), the assembly of the *cgc8-cgc14* operon (CAS007) required two LCR reactions followed by a classical restriction enzyme-based cloning (Figure S2). The DNA fragment constituted of the promoter SP25-RBS was joined to the DNA fragment containing the *cgc8* to *cgc11* genes by LCR. Similarly, the *cgc12* to *cgc14* DNA fragment was assembled with the T4 terminator DNA fragment by LCR. As LCR assembly of the obtained DNA fragments together with the pOSV801 vector repeatedly failed, both LCR fragments were cloned into pCR blunt. To allow the assembly of the two fragments with the pOSV801 vector, the *cgc12* to *cgc14*-T4 terminator fragment was amplified by PCR, using the oligonucleotide onCAS074 and onCAS010bis. The onCAS074 oligonucleotide

allowed the addition at the 5' extremity of the fragment of 19 bp (end of *cgc11*) containing an *Xba*I site. The onCAS010bis oligonucleotide allowed the reconstitution of the complete BioBrick suffix at the 3' extremity of the fragment. The two LCR fragments were then assembled with the pOSV801 vector by classical restriction enzyme-based cloning. The basic gene cassettes CAS005 (*cgc3* and *cgc17*) and CAS007 (*cgc8-cgc14*) were then assembled using the Biobrick type of cloning, generating the composite Precursor gene cassette CAS20.

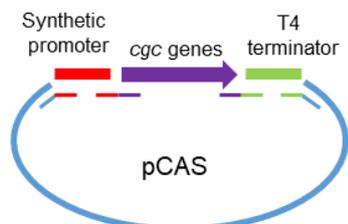


Figure 1: General principle of gene cassette construction using Ligase Cycling Reaction.

The sequences of all the basic gene cassettes constructed (Table 1) were confirmed by sequencing.

### Verification of the functionality of basic gene cassettes by genetic complementation

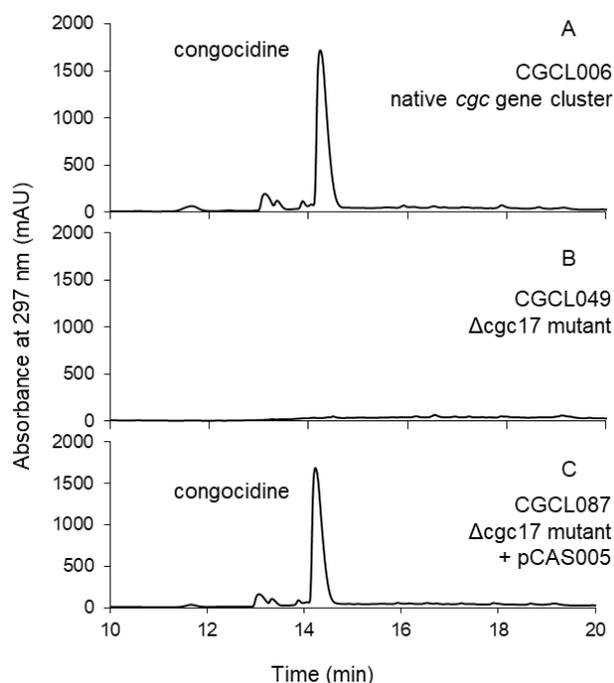


Figure 2: Verification of the functionality of the CAS005 gene cassette: genetic complementation of a *cgc17* deletion mutant.

HPLC chromatograms of culture supernatants of *S. lividans* A) CGCL006 (containing native *cgc* cluster), B) CGCL049 (*cgc* cluster with *cgc17* deleted), C) CGCL087 (CGCL049 with CAS005 containing *cgc3* and *cgc17*). Samples were analyzed on a reverse phase  $C_{18}$  column, eluted in isocratic conditions with 0.1% HCOOH in  $H_2O$  (solvent A)/ 0.1% HCOOH in  $CH_3CN$  (solvent B) (95:5) at  $1 \text{ ml} \cdot \text{min}^{-1}$  for 7 min, followed by a gradient to 40:60 A/B over 23 min. Absorbance was monitored at 297 nm.

As we were completely refactoring the congocidine biosynthetic gene cluster, we wanted to ensure that each of the basic gene cassettes we constructed was functional before assembling these

cassettes together to reconstitute the *cgc* gene cluster. For this purpose, we took advantage of the library of *cgc* gene deletion mutants constructed during the studies of congocidine biosynthesis (Juguet et al., 2009; Lautru et al., 2012). These mutant strains were genetically complemented with the gene cassette expressing the gene deleted in the mutant. Thus, to verify that the CAS005 gene cassette, containing the *cgc17* gene, was functional, the pCAS005 plasmid was introduced by intergeneric conjugation in the strain *S. lividans* CGCL049, which contains the whole native *cgc* cluster except for *cgc17*, which has been deleted (Lautru et al., 2012). Exconjugants, named CGCL087, were verified by PCR. The *S. lividans* CGCL087 strain was then grown in liquid MP5 at 28°C (Pernodet et al., 1993), together with *S. lividans* CGCL049 and *S. lividans* CGCL006 (heterologously expressing the native *cgc* gene cluster) as controls. The supernatants of 4-day cultures were analyzed by HPLC at 297 nm. The chromatograms presented in Figure 2A-C show that congocidine production is restored in the complemented mutant *S. lividans* CGCL087, thereby confirming that Cgc17 is functional when produced from the CAS005 gene cassette.

The gene cassettes CAS001 (*cgc4-cgc6*), CAS002 (*cgc15*), CAS003 (*cgc7*), CAS005 (*cgc3*), CAS008 (*cgc22*), CAS010 (*cgc18*), CAS011 (*cgc16*) and CAS013 (*cgc19*) were all verified using the same protocol (Figure S3 to S11 and (Aubry et al., 2019)), and all were proven to be functional.

#### ***Verification of the functionality of the CAS009 (cgc2) gene cassette***

The genetic complementation of the *cgc2* mutant *S. lividans* CGCL035 failed to restore congocidine production. As we suspected that this failure originated from the *S. lividans* CGCL035 strain rather than from the pCAS009 plasmid, we decided to try to genetically complement a mutant of *dst2* and *dst25* genes (*S. lividans* DSTL020), orthologs of *cgc2* in the gene clusters directing the biosynthesis of congocidine, disgocidine and distamycin in *Streptomyces netropsis* DSM40846 (Vingadassalon et al., 2015). The double mutant *S. lividans* DSTL020 does not produce any of the three pyrrolamides. As *S. lividans* DSTL020 already harbors a kanamycin resistance marker, we replaced the kanamycin resistance cassette of pCAS009 by an apramycin resistance cassette by simple restriction enzyme-based cloning, yielding pCAS014. pCAS014 was introduced in *S. lividans* DSTL020 by intergeneric conjugation. Exconjugants were verified by PCR and the strain, named DSTL028, was cultivated for 4 days in MP5 medium at 28°C, together with *S. lividans* DSTL005 (expressing the complete *dst* gene clusters) and DSTL020 strains. Culture supernatants were analyzed by HPLC and the chromatograms (Figure S12) indicated that congocidine and disgocidine production was restored. We did not observe the production of distamycin by the strain. This could be due to an absence of cross-complementation of Dst25 by Cgc2. Alternatively, this could also be due a production of distamycin too low to be observed by HPLC, as in the *S. lividans* strain heterologously expressing the *dst* gene clusters (DSTL005), the production of distamycin is already quite low.

#### ***Verification of the functionality of the CAS006 (cgc20-cgc21) gene cassette***

The functionality of the Resistance gene cassette (CAS006) was verified by testing its ability to confer resistance to congocidine. The pCAS006 gene cassette was introduced by intergeneric conjugation in *S. lividans* TK23, a strain that is naturally sensitive to congocidine. The resulting strain (CGCL088), *S. lividans* TK23 and *S. lividans* CGCL006 (containing the native *cgc* cluster) were

streaked on GYM medium with or without congocidine (40 µg/mL) and the plates were incubated at 28°C for 72h. All strains grew on GYM medium (Figure S13A). On GYM supplemented with congocidine however, the *S. lividans* TK23 strain did not grow (except for a few clones that might be spontaneously resistant) whereas *S. lividans* CGCL006 and CGCL088 strains grew well (Figure S13B). This confirmed that the CAS006 cassette is functional and confers resistance to congocidine.

### ***Verification of the functionality of the CAS007 (cgc8-cgc14) gene cassette***

To verify the functionality of the CAS007 gene cassette, we introduced it by intergeneric conjugation in the *S. lividans* strain already expressing the CAS005 (*cgc3-cgc17*) gene cassette and checked for the production of the expected product, the 4-acetamidopyrrole-2-carboxylate. Indeed, this metabolite is excreted in culture supernatants and absorbs at 297 nm (Lautru et al., 2012). The exconjugants, named CGCL094, were verified by PCR. *S. lividans* CGCL089 (containing only CAS005) and *S. lividans* CGCL094 (containing both CAS005 and CAS007) were grown in liquid MP5 at 28°C for 72h and the culture supernatants were analyzed by HPLC. The chromatograms (Figure S14) show that *S. lividans* CGCL094 produced 4-acetamidopyrrole-2-carboxylate, identified by comparison with an authentic standard. This confirmed the functionality of the CAS007 cassette and showed that combined, the two cassettes CAS005 and CAS007 are therefore sufficient to produce 4-acetamidopyrrole-2-carboxylate. It should be noted, however, that this experiment did not allow confirming the expression of Cgc14 as an active enzyme, as Cgc14 deacetylates 4-acetamidopyrrole-2-carboxylate loaded on Cgc19.

### ***Assembly of the gene cassettes by Biobrick-like assembly and reconstruction of the cgc cluster***

As each individual gene cassette was confirmed to be functional, we proceeded to the assembly of the different gene cassettes. The objective was to assemble all gene cassettes on a single plasmid. However, as we were aware that this might prove difficult, we devised the construction of two plasmids: one containing the Precursor and Resistance gene cassettes, and another one containing the Assembly and Tailoring gene cassettes. For this, we used the two compatible plasmids pOSV801 and pOSV812 (Aubry et al., 2019). These plasmids allow a Biobrick-type of assembly (Shetty et al., 2008). The six Assembly and Tailoring gene cassettes (CAS002, CAS008, CAS009, CAS010, CAS011 and CAS013) were assembled in pOSV812 as presented in Figure 3, yielding pCAS024. Similarly, the Precursor and Resistance gene cassettes (CAS001, CAS003, CAS005, CAS006 and CAS007) were assembled in pOSV801 as presented on Figure 4, yielding pCAS026. Attempts to assemble the CAS024 and CAS06 gene cassette failed repeatedly. Taken together, pCAS024 and pCAS026 harbor all the 21 genes necessary for congocidine production in a *Streptomyces* host, organized in 11 transcriptional units.

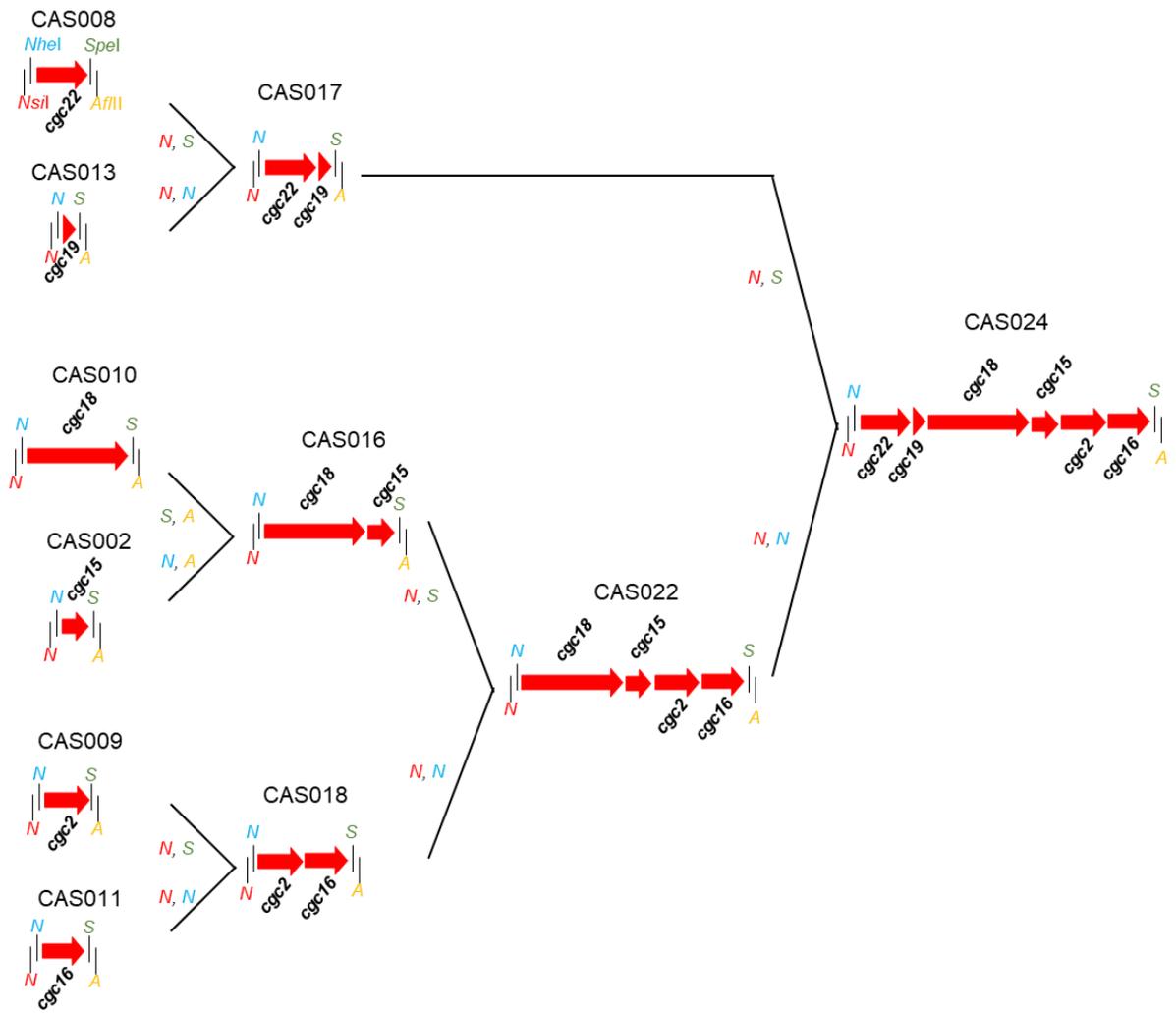


Figure 3: Scheme of the assembly of the Assembly and Tailoring *cgc* gene cassettes. Promoters and terminators are not represented on the figure. *N*: *NsiI*, *N*: *NheI*, *S*: *SpeI*, *A*: *AflIII*

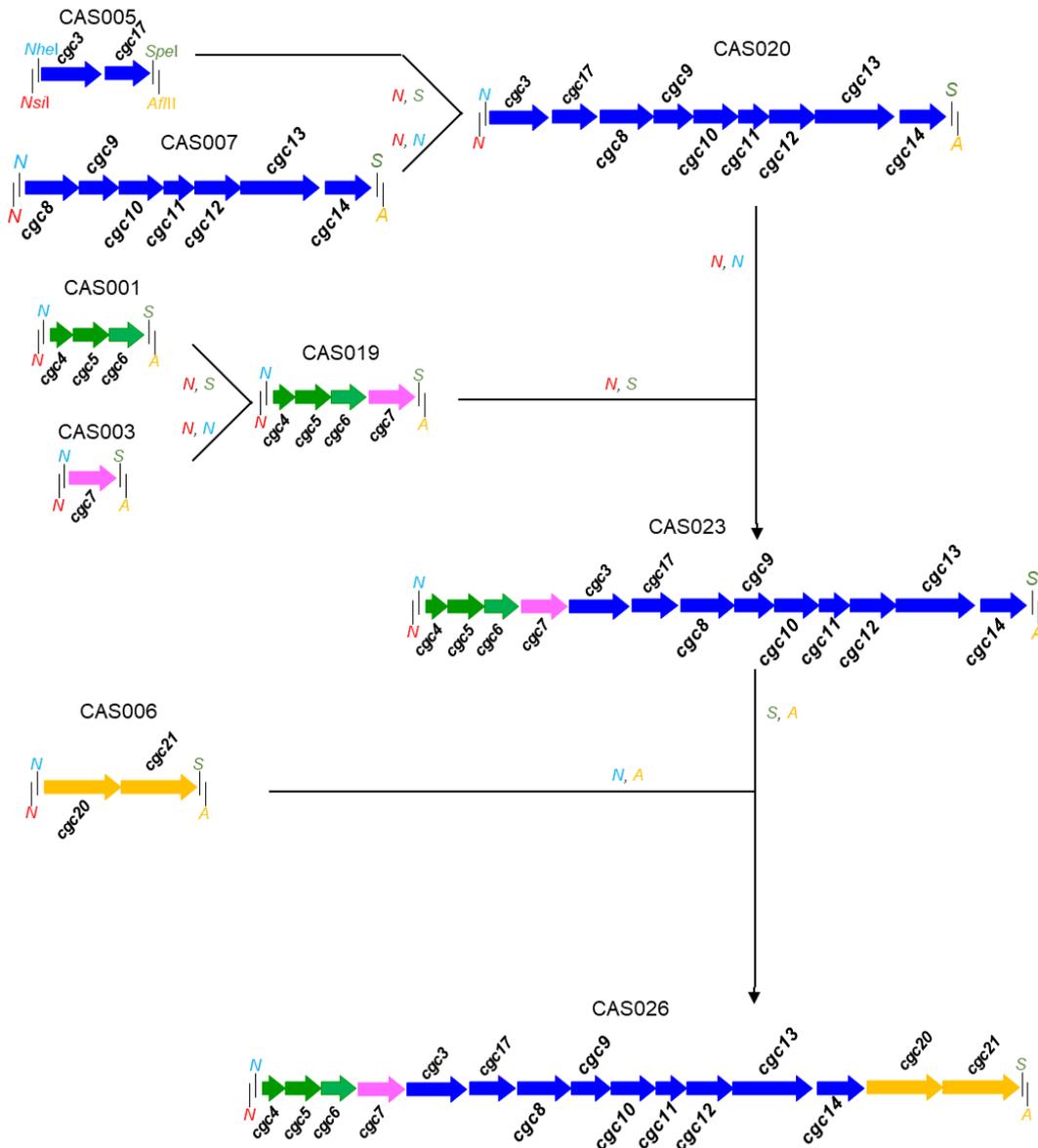


Figure 4: Scheme of the assembly of the Precursor and Resistance gene cassettes  
Promoters and terminators are not represented on the figure. *N*: *NsiI*, *N*: *NheI*, *S*: *SpeI*, *A*: *AflIII*

### ***Heterologous expression of the refactored *cgc* gene cluster in *S. lividans* TK23***

The next step consisted in the introduction by intergeneric conjugation of the pCAS024 and pCAS026 in *S. lividans* TK23. We chose this host as a chassis as all our previous heterologous expression of pyrrolamide gene clusters had been carried out in this host (Juguet et al., 2009; Lautru et al., 2012; Vingadassalon et al., 2015). The strains that are usually used for *E. coli*/*Streptomyces* intergeneric conjugations are *E. coli* ET12567/pUZ8002 and *E. coli* S17-1 (Flett et al., 1997; Simon et al., 1983). However, we noticed a high genetic instability of the pCAS024 and pCAS026 in these strains (loss of (part of) the inserts), instability that was not observed during the assembly of the gene cassettes in *E. coli* DH5 $\alpha$ . Sequencing of one of the plasmids extracted from *E. coli* ET12567/pUZ8002 transformed with pCAS026 suggests that recombination likely occurred between the multiple copies of the 126-bp T4 terminator sequences. This genetic instability and its probable cause, the repetition of the terminator sequence, underline the necessity, in the type of

approach we chose, to vary the genetic elements (promoters, terminators...), making use for example of those recently developed in the group of Andriy Luzhetskyy (Horbal et al., 2018a), for the construction of gene cassettes.

*E. coli* DH10B/pUZ8002 has also been used for *E. coli*/*Streptomyces* intergeneric conjugations (Coëffet-Le Gal et al., 2006). We thus transformed this strain with pCAS026. Genetic instability appears to be much reduced in this strain compared to *E. coli* ET12567/pUZ8002 and *E. coli* S17-1. However, the conjugation efficiency using standard conditions was also greatly reduced.

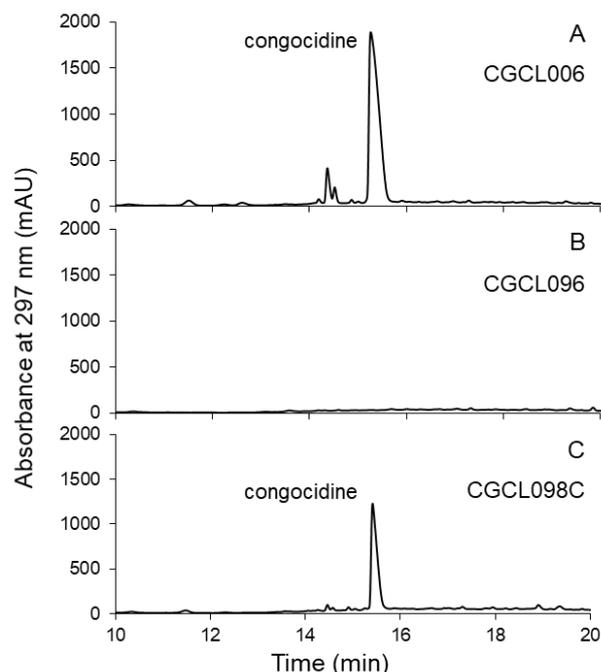


Figure 5: Production of congocidine by the refactored *cgc* gene cluster.

HPLC chromatograms of *S. lividans* A) CGCL006 (TK23 containing native *cgc* cluster), B) CGCL096 (TK23 with CAS024), C) CGCL098C (TK23 with CAS024 and CAS026, clone C) supernatants. Samples were analyzed on a reverse phase C<sub>18</sub> column, eluted in isocratic conditions with 0.1% HCOOH in H<sub>2</sub>O (solvent A)/ 0.1% HCOOH in CH<sub>3</sub>CN (solvent B) (95:5) at 1 ml.min<sup>-1</sup> for 7 min, followed by a gradient to 40:60 A/B over 23 min. Absorbance was monitored at 297 nm.

The pCAS024 plasmid was introduced in *S. lividans* TK23 by intergeneric conjugation from the *E. coli* S17-1 strain. Out of the four ex-conjugants that were carefully verified by PCR, only one, called CGCL098, appeared correct. This clone was used for the introduction of pCAS026 from *E. coli* ET12567/pUZ8002. To verify the resulting ex-conjugants, we carried out a bioassay based on the antibiotic activity of congocidine. Indeed, if the intact pCAS026 had been introduced in *S. lividans* CGCL098, then we expected the resulting strain to produce congocidine. Out of 27 clones tested, five inhibited *Micrococcus luteus* growth (Figure S15). These clones were verified by PCR and named CGCL098A-E. They were cultivated in liquid MP5 at 28°C for 4 days and their supernatant was analyzed by HPLC at 297 nm. All clones produced congocidine, as exemplified by the chromatogram of the *S. lividans* CGCL098C (Figure 5). From this preliminary experiment, we

estimate that congocidine production from the refactored cluster is roughly one third of that obtained with the native gene cluster.

## CONCLUSIONS

In this study, we refactored the congocidine biosynthetic gene cluster. For this purpose, we first designed and constructed synthetic gene cassettes constituted of transcriptional units (promoter-RBS-genes-terminator). These cassettes were also designed to constitute functional units, involved in either precursor biosynthesis, congocidine resistance, assembly and tailoring. Each of the 11 gene cassettes was functionally validated by genetic complementation, HPLC analysis or antibiotic bioassay. They were then assembled on two compatible and integrative plasmids using Biobrick-like assembly. Integration of both plasmids in the *S. lividans* host resulted in production of congocidine, confirming that the refactored cluster was functional. This successful refactoring now opens the way to the optimization of congocidine production, playing for example with regulatory elements, as already done in other studies (Horbal et al., 2018b; Hu et al., 2019; Song et al., 2019). More importantly, it now offers us a functional platform to elaborate pyrrolamide-based combinatorial biosynthesis experiments, and to bring forth, for example by exchanging NRPS genes, the knowledge on these systems that is still required for their successful engineering.

## MATERIAL AND METHODS

### ***Bacterial strains, plasmids and growth conditions***

Strains and plasmids used in this study are listed in Table S1 and S2. *E. coli* strains were grown at 37 °C in LB or SOB medium complemented with MgSO<sub>4</sub> (20 mM final), supplemented with appropriate antibiotics as needed. The Soya Flour Mannitol (SFM) medium (Kieser et al., 2000) was used for genetic manipulations of *Streptomyces* strains and spore stocks preparations. *Streptomyces* strains were grown at 28°C in MP5 (Pernodet et al., 1993) for congocidine and pyrrole production, and bioassays were performed on HT medium (Kieser et al., 2000) or GYM medium (Shima et al., 1996).

### ***DNA Preparation and manipulations***

All oligonucleotides used in this study were purchased from Eurofins and are listed in Table S3. The High fidelity DNA polymerase Phusion (Thermo Fisher Scientific) was used to amplify the fragments used for the construction of the cassettes. DreamTaq polymerase (Thermo Fisher Scientific) was used for PCR verification of plasmid integration in *Streptomyces* strains. Restriction enzymes used were from New England Biolabs or Thermo Fisher Scientific, the thermostable ligase was also ordered from New England Biolabs. DNA fragments were purified from agarose gels using the Nucleospin Gel and PCR clean-up kit from Macherey-Nagel. *Escherichia coli* transformations and *E. coli*/*Streptomyces* conjugations were performed according to standard procedures (Sambrook and Russell, 2001; Kieser et al., 2000).

### ***Construction of the gene cassettes by Ligase cycling reaction assembly***

Each basic gene cassette (CAS001-003; CAS005-006; CAS008-CAS013) was assembled in a plasmid using the Ligase Cycling Reaction assembly (LCR) as shown on Figure 1 (Chandran, 2017).

The construction of the CAS007 cassette, more complex, is described in a separated paragraph below.

The plasmids (pOSV801 or pOSV812) were digested by *NotI*/Klenow and the 5 kb fragments were purified on agarose gel. The *cgc* genes constituting the gene cassettes were amplified from the pCGC002 cosmid (Juguet et al., 2009) using the primers described in Table S3. The synthetic promoters SP (Bai et al., 2015) were ordered from Eurofins Genomics as synthetic gene fragments and amplified with the primers onCAS001bis and onCAS002. The T4 terminator sequence was amplified from the pOSV215 plasmid (Raynal et al., 2006) with the primers onCAS007 and onCAS008bis. The primers upstream of the promoter SP and downstream of the terminator were designed in order to recreate the prefix (*NsiI*, *NotI*, *NbeI*) and suffix (*SpeI*, *NotI*, *AflII*) located upstream and downstream the biobrick respectively. All fragments were then phosphorylated and ligated *via* LCR. The resulting pCAS plasmids were confirmed by sequencing.

To replace the kanamycin resistance cassette by the apramycin resistance cassette of the pCAS009 plasmid, pCAS009 was digested by *HindIII* and *KpnI*, excising the kanamycin resistance cassette. It was then ligated with the 1.2 kb *BamHI*-*KpnI*-digested apramycin resistance fragment coming from pOSV801. The plasmid pCAS014 obtained was verified by restriction enzyme digestions.

### **Construction of the CAS007 cassette**

The CAS007 cassette contains the genes *cgc8-cgc14* and spans 8 kb. To construct this cassette, we combined LCR (Chandran, 2017) with classical restriction enzyme-based cloning, as shown in Figure S2. Two LCR were performed, one assembling the promoter SP25 with the fragment containing *cgc8* to *cgc11*, the other assembling the *cgc12* to *cgc14* fragment with the T4 terminator. Each LCR product was then cloned into the pCR blunt vector (Thermo Fisher Scientific), yielding the vectors pCR-blunt-SP25-cgc8-11 and pCR-blunt-cgc12-14-T4ter. The pCR-blunt-cgc12-14-T4ter was used to PCR amplify the cgc12-14-T4ter fragment with oligonucleotides onCAS074 adding 19 base pairs corresponding to the end of *cgc11* and onCAS010bis reconstituting the complete suffix sequence. The amplified fragment was digested by *XhoI* (site introduced by the onCAS074 primer) and *AflII*. It was ligated with the *NbeI*/*XhoI*-digested SP25-cgc8-11 fragment of pCR-blunt-SP25-cgc8-11 and the *NbeI*/*AflII*-digested pOSV801, yielding pCAS007. The complete sequence of the 8 kb cassette was verified by sequencing.

### **Integration of each basic gene cassette in *S. lividans* strains**

The pCAS001-pCAS003, pCAS005, pCAS008, pCAS010-pCAS013 were introduced by intergeneric conjugation following the standard procedure (Kieser et al., 2000) in *Streptomyces lividans* mutant strains expressing the *cgc* cluster except for one gene of the tested cassette (Juguet et al., 2009), gene whose functionality was tested. The pCAS014 (CAS009) was introduced in *Streptomyces lividans* DSTL020 expressing the *dst* gene clusters except for *dst2* and *dst25* (Vingadassalon et al., 2015). The pCAS006 was introduced in *S. lividans* TK23 and the pCAS007 in *S. lividans* CGCL089 already containing the pCAS005 plasmid. *E. coli* ET12567/pUZ8002 was used as a donor strain for the pCAS plasmids conferring resistance to apramycin (Table S2) and *E. coli* S17-1 for the pCAS

plasmids that confer resistance to kanamycin. All resulting strains were verified by PCRs amplifying the sequence of the gene(s) introduced and the attL and attR regions.

### ***Assembly of all gene cassettes to reconstruct the cgc cluster***

The synthetic *cgc* gene cluster was assembled on two plasmids: one containing the Precursor and Resistance gene cassettes (Figure 4), and another one containing the Assembly and Tailoring gene cassettes (Figure 3) using a Biobrick-like assembly. One of the advantages of this type of assembly is that gene cassettes can be assembled two by two in parallel, generating composite gene cassettes that can then be assembled together. At each step, the recipient plasmid is opened either upstream (in the prefix) or downstream (in the suffix) of the existing cassette, using respectively *NsiI/NheI* or *SpeI/AflIII*. The cassette to be inserted is digested either by *NsiI/SpeI* or *NheI/AflIII* respectively, and two fragments are ligated together. Since after ligation, both the prefix and the suffix are reformed upstream and downstream the composite cassette and only a scar is left between the assembled cassettes, the same protocol can be repeated until the final plasmid is obtained. All plasmids were verified by restriction digestion before pursuing to the next assembly step. The final plasmids pCAS024 and pCAS026 were introduced in *S. lividans* TK23 by intergeneric conjugation. Clones were verified by PCR.

### ***Bioassay protocols***

To confirm the functionality of CAS006 (resistance genes *cgc20* and *cgc21*), we carried out a bioassay testing the ability of this cassette to confer congocidine resistance. The strains *S. lividans* CGCL089 (expressing CAS006), *S. lividans* CGCL006 (expressing the native *cgc* gene cluster, positive control) and *S. lividans* TK23 (susceptible to congocidine, negative control) were streaked on GYM plates with or without 40 µg/mL congocidine. Growth was observed after 3 days at 28°C.

*S. lividans* clones containing the pCAS024 and pCAS026 plasmids were screened for congocidine production using a bioassay based on the antibacterial activity of congocidine. They were patched on HT plates. After two days of growth at 28°C, the plates were overlaid with soft nutrient agar (SNA) containing *Micrococcus luteus* and left at 37°C overnight. Clones exhibiting a halo of *M. luteus* growth inhibition, therefore producing an antibiotic compound, were selected for further analyses.

### ***LC analyzes***

For congocidine and 4-acetamidopyrrole-2-carboxylate production, *S. lividans* strains were cultivated in MP5 medium for 3 to 4 days at 28°C. Supernatants were filtered using Mini-UniPrep syringeless filter devices (0.2 µm, Whatman). Before injection in the HPLC instrument, the supernatants of the cultures producing 4-acetamidopyrrole-2-carboxylate were acidified to pH 4.5, to avoid the splitting of the HPLC peak into two peaks. The samples were then analyzed on an Atlantis C<sub>18</sub> T3 column (250 mm x 4.6 mm, 5 µm, column temperature 30°C) using an Agilent 1200 HPLC instrument with a quaternary pump. Samples were eluted in isocratic conditions with 0.1% HCOOH in H<sub>2</sub>O (solvent A)/ 0.1% HCOOH in CH<sub>3</sub>CN (solvent B) (95:5) at 1 ml.min<sup>-1</sup> for 7 min, followed by a gradient to 40:60 A/B over 23 min. Congocidine was detected by monitoring absorbance at 297 nm (Juguet et al., 2009).

## ***Acknowledgements***

The research received funding from ANR-14-CE16-0003-01. The funders had no role in study design, data collection and interpretation.

## ***References***

Aigle, B., Lautru, S., Spitteller, D., Dickschat, J.S., Challis, G.L., Leblond, P., and Pernodet, J.-L. (2014). Genome mining of *Streptomyces ambofaciens*. *J. Ind. Microbiol. Biotechnol.* *41*, 251–263.

Al-Mestarihi, A.H., Garzan, A., Kim, J.M., and Garneau-Tsodikova, S. (2015). Enzymatic evidence for a revised congoicidine biosynthetic pathway. *Chembiochem* *16*, 1307–1313.

Aubry, C., Pernodet, J.-L., and Lautru, S. (2019). A set of modular and integrative vectors for synthetic biology in *Streptomyces*. *Appl. Environ. Microbiol.* Aug 1;85(16).

Bai, C., Zhang, Y., Zhao, X., Hu, Y., Xiang, S., Miao, J., Lou, C., and Zhang, L. (2015). Exploiting a precise design of universal synthetic modular regulatory elements to unlock the microbial natural products in *Streptomyces*. *Proc. Natl. Acad. Sci. U.S.A.* *112*, 12181–12186.

Basitta, P., Westrich, L., Rösch, M., Kulik, A., Gust, B., and Apel, A.K. (2017). AGOS: A plug-and-play method for the assembly of artificial gene operons into functional biosynthetic gene clusters. *ACS Synth Biol* *6*, 817–825.

Bauman, K.D., Li, J., Murata, K., Mantovani, S.M., Dahesh, S., Nizet, V., Luhavaya, H., and Moore, B.S. (2019). Refactoring the cryptic streptophenazine biosynthetic gene cluster unites phenazine, polyketide, and nonribosomal peptide biochemistry. *Cell Chem Biol* *26*, 724-736.e7.

Chandran, S. (2017). Rapid assembly of DNA via Ligase Cycling Reaction (LCR). *Methods Mol. Biol.* *1472*, 105–110.

Coëffet-Le Gal, M.-F., Thurston, L., Rich, P., Miao, V., and Baltz, R.H. (2006). Complementation of daptomycin *dptA* and *dptD* deletion mutations in *trans* and production of hybrid lipopeptide antibiotics. *Microbiology (Reading, Engl.)* *152*, 2993–3001.

Eyles, T.H., Vior, N.M., and Truman, A.W. (2018). Rapid and robust yeast-mediated pathway refactoring generates multiple new bottromycin-related metabolites. *ACS Synth Biol* *7*, 1211–1218.

Flett, F., Mersinias, V., and Smith, C.P. (1997). High efficiency intergeneric conjugal transfer of plasmid DNA from *Escherichia coli* to methyl DNA-restricting streptomycetes. *FEMS Microbiol. Lett.* *155*, 223–229.

Genilloud, O. (2018). Mining actinomycetes for novel antibiotics in the omics era: are we ready to exploit this new paradigm? *Antibiotics (Basel)* *7*.

Gomez-Escribano, J.P., and Bibb, M.J. (2014). Heterologous expression of natural product biosynthetic gene clusters in *Streptomyces coelicolor*: from genome mining to manipulation of biosynthetic pathways. *J. Ind. Microbiol. Biotechnol.* *41*, 425–431.

Horbal, L., Siegl, T., and Luzhetskyy, A. (2018a). A set of synthetic versatile genetic control elements for the efficient expression of genes in Actinobacteria. *Sci Rep* *8*, 491.

Horbal, L., Marques, F., Nadmid, S., Mendes, M.V., and Luzhetskyy, A. (2018b). Secondary metabolites overproduction through transcriptional gene cluster refactoring. *Metab. Eng.* *49*, 299–315.

Hu, F., Liu, Y., and Li, S. (2019). Rational strain improvement for surfactin production: enhancing the yield and generating novel structures. *Microb. Cell Fact.* *18*, 42.

Juguet, M., Lautru, S., Francou, F.-X., Nezbedová, S., Leblond, P., Gondry, M., and Pernodet, J.-L. (2009). An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congoicidin in *Streptomyces ambofaciens*. *Chem. Biol.* *16*, 421–431.

Kieser, T., Bibb, M., Buttner, M., and Hopwood, D.A. (2000). *Practical Streptomyces genetics*, John Innes Foundation, Norwich NR47UH, UK.

Kim, S.-H., Lu, W., Ahmadi, M.K., Montiel, D., Ternei, M.A., and Brady, S.F. (2019). Atolypenes, tricyclic bacterial sesterterpenes discovered using a multiplexed *in vitro* Cas9-TAR gene cluster refactoring approach. *ACS Synth Biol* *8*, 109–118.

de Kok, S., Stanton, L.H., Slaby, T., Durot, M., Holmes, V.F., Patel, K.G., Platt, D., Shapland, E.B., Serber, Z., Dean, J., et al. (2014). Rapid and reliable DNA assembly via ligase cycling reaction. *ACS Synth Biol* *3*, 97–106.

Lautru, S., Song, L., Demange, L., Lombès, T., Galons, H., Challis, G.L., and Pernodet, J.-L. (2012). A sweet origin for the key congoicidin precursor 4-acetamidopyrrole-2-carboxylate. *Angew. Chem. Int. Ed. Engl.* *51*, 7454–7458.

Luo, Y., Huang, H., Liang, J., Wang, M., Lu, L., Shao, Z., Cobb, R.E., and Zhao, H. (2013). Activation and characterization of a cryptic polycyclic tetramate macrolactam biosynthetic gene cluster. *Nat Commun* *4*, 2894.

Newman, D.J., and Cragg, G.M. (2016). Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* *79*, 629–661.

Ōmura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M., Takahashi, Y., Horikawa, H., Nakazawa, H., Osonoe, T., et al. (2001). Genome sequence of an industrial microorganism *Streptomyces avermitilis*: Deducing the ability of producing secondary metabolites. *Proc Natl Acad Sci U S A* *98*, 12215–12220.

Osswald, C., Zipf, G., Schmidt, G., Maier, J., Bernauer, H.S., Müller, R., and Wenzel, S.C. (2014). Modular construction of a functional artificial epothilone polyketide pathway. *ACS Synth Biol* *3*, 759–772.

Pernodet, J.L., Alegre, M.T., Blondelet-Rouault, M.H., and Guérineau, M. (1993). Resistance to spiramycin in *Streptomyces ambofaciens*, the producer organism, involves at least two different mechanisms. *J. Gen. Microbiol.* *139*, 1003–1011.

Prentki, P., and Krisch, H.M. (1984). *In vitro* insertional mutagenesis with a selectable DNA fragment. *Gene* *29*, 303–313.

Raynal, A., Karray, F., Tuphile, K., Darbon-Rongère, E., and Pernodet, J.-L. (2006). Excisable cassettes: new tools for functional analysis of *Streptomyces* genomes. *Appl. Environ. Microbiol.* *72*, 4839–4844.

Sambrook, J., and Russell, D.W. (2001). *Molecular cloning: a laboratory manual*, Third edition. CSHL Press, Cold Spring Harbor, NY.

Shetty, R.P., Endy, D., and Knight, T.F.J. (2008). Engineering BioBrick vectors from BioBrick parts. *J. Biol. Eng* 2:5.

Shima, J., Hesketh, A., Okamoto, S., Kawamoto, S., and Ochi, K. (1996). Induction of actinorhodin production by *rpsL* (encoding ribosomal protein S12) mutations that confer streptomycin resistance in *Streptomyces lividans* and *Streptomyces coelicolor* A3(2). *J. Bacteriol.* 178, 7276–7284.

Simon, R., Prierer, U., and Pühler, A. (1983). A Broad host range mobilization system for in vivo genetic engineering: transposon mutagenesis in Gram negative bacteria. *Nature Biotechnology* 1, 784–791.

Song, C., Luan, J., Cui, Q., Duan, Q., Li, Z., Gao, Y., Li, R., Li, A., Shen, Y., Li, Y., et al. (2019). Enhanced heterologous spinosad production from a 79-kb synthetic multioperon assembly. *ACS Synth Biol* 8, 137–147.

Strieker, M., Tanović, A., and Marahiel, M.A. (2010). Nonribosomal peptide synthetases: structures and dynamics. *Curr. Opin. Struct. Biol.* 20, 234–240.

Tan, G.-Y., Deng, K., Liu, X., Tao, H., Chang, Y., Chen, J., Chen, K., Sheng, Z., Deng, Z., and Liu, T. (2017). Heterologous biosynthesis of spinosad: an omics-guided large polyketide synthase gene cluster reconstitution in *Streptomyces*. *ACS Synth Biol* 6, 995–1005.

Temme, K., Zhao, D., and Voigt, C.A. (2012). Refactoring the nitrogen fixation gene cluster from *Klebsiella oxytoca*. *Proc. Natl. Acad. Sci. U.S.A.* 109, 7085–7090.

Vilanova, C., Tanner, K., Dorado-Morales, P., Villaescusa, P., Chugani, D., Frías, A., Segredo, E., Molero, X., Fritschi, M., Morales, L., et al. (2015). Standards not that standard. *J Biol Eng* 9, 17.

Vingadassalon, A., Lorieux, F., Juguet, M., Le Goff, G., Gerbaud, C., Pernodet, J.-L., and Lautru, S. (2015). Natural combinatorial biosynthesis involving two clusters for the synthesis of three pyrrolamides in *Streptomyces netropsis*. *ACS Chem. Biol.* 10, 601–610.

Wagner, S., Baars, L., Ytterberg, A.J., Klussmeier, A., Wagner, C.S., Nord, O., Nygren, P.-A., van Wijk, K.J., and de Gier, J.-W. (2007). Consequences of membrane protein overexpression in *Escherichia coli*. *Mol. Cell Proteomics* 6, 1527–1550.

Yamanaka, K., Reynolds, K.A., Kersten, R.D., Ryan, K.S., Gonzalez, D.J., Nizet, V., Dorrestein, P.C., and Moore, B.S. (2014). Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc Natl Acad Sci U S A* 111, 1957–1962.

Yeung, E., Dy, A.J., Martin, K.B., Ng, A.H., Del Vecchio, D., Beck, J.L., Collins, J.J., and Murray, R.M. (2017). Biophysical constraints arising from compositional context in synthetic gene networks. *Cell Syst* 5, 11-24.e12.

# Refactoring of the congocidine biosynthetic gene cluster: from gene cassettes to gene cluster

Céline AUBRY<sup>a</sup>, Jennifer PERRIN<sup>a</sup>, Yacine Mohammed SELLAH<sup>a</sup>, Jean-Luc PERNODET<sup>a</sup> and Sylvie LAUTRU<sup>a#</sup>

<sup>a</sup> Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

## Supplemental material

Table S1: Strains used in this study

Strain	Description	Reference
<i>Escherichia coli</i> DH5a	General cloning host	Promega
<i>E. coli</i> S17-1	Host strain for conjugation from <i>E. coli</i> to <i>Streptomyces</i> when using vectors containing the kanamycin resistance cassette	(Simon et al., 1983)
<i>E. coli</i> ET12567 pUZ8002	Host strain for conjugation from <i>E. coli</i> to <i>Streptomyces</i>	(Flett et al., 1997)
<i>E. coli</i> DH10B/pUZ8002	Host strain for conjugation from <i>E. coli</i> to <i>Streptomyces</i>	Our unpublished data
<i>S. lividans</i> TK23	<i>Streptomyces</i> host strain for heterologous expression	(Kieser et al., 2000)
CGCL006	TK23 containing pCGC002 ( <i>egc</i> cluster)	(Juguet et al., 2009)
CGCL022	TK23 containing <i>egc</i> cluster with <i>egc4</i> deleted	(Lautru et al., 2012)
CGCL028C	TK23 containing <i>egc</i> cluster with <i>egc19</i> deleted	(Juguet et al., 2009)
CGCL029	TK23 containing <i>egc</i> cluster with <i>egc18</i> deleted	(Juguet et al., 2009)
CGCL030	TK23 containing <i>egc</i> cluster with <i>egc22</i> deleted	(Juguet et al., 2009)
CGCL031	TK23 containing <i>egc</i> cluster with <i>egc15</i> deleted	(Juguet et al., 2009)
CGCL032B/C	TK23 containing <i>egc</i> cluster with <i>egc16</i> deleted	(Juguet et al., 2009)
CGCL045D	TK23 containing <i>egc</i> cluster with <i>egc3</i> deleted	(Lautru et al., 2012)
CGCL049D	TK23 containing <i>egc</i> cluster with <i>egc17</i> deleted	(Lautru et al., 2012)
CGCL051	TK23 containing <i>egc</i> cluster with <i>egc5</i> deleted	(Lautru et al., 2012)
CGCL056A	TK23 containing <i>egc</i> cluster with <i>egc6</i> deleted	(Lautru et al., 2012)
CGCL058A	TK23 containing <i>egc</i> cluster with <i>egc7</i> deleted	(Lautru et al., 2012)
CGCL076	CGCL022 complemented with pCAS001	This study
CGCL077	CGCL051 complemented with pCAS001	This study
CGCL078	CGCL056 complemented with pCAS001	This study
CGCL079	CGCL031 complemented with pCAS002	This study
CGCL080	CGCL058 complemented with pCAS003	This study
CGCL081	CGCL056 complemented with pCAS004	This study
CGCL082	CGCL058 complemented with pCAS004	This study
CGCL083	CGCL030 complemented with pCAS008	(Aubry et al., 2019)
CGCL085	CGCL029 complemented with pCAS010	This study
CGCL086	CGCL045 complemented with pCAS005	This study
CGCL087	CGCL049 complemented with pCAS005	This study
CGCL088	TK23 containing pCAS006	This study
CGCL089	TK23 containing pCAS005	This study
CGCL091	CGCL032 complemented with pCAS011	This study
CGCL093	CGCL028 complemented with pCAS013	This study
CGCL094	TK23 containing pCAS005 and pCAS007, pyrrole producer	This study
CGCL096	TK23 containing pCAS024 (plasmid with all the <i>egc</i> assembly and tailoring genes)	This study
CGCL097	TK23 containing pCAS026 (plasmid with all the <i>egc</i> precursor genes and resistance genes)	This study
CGCL098	TK23 containing both pCAS024 and pCAS026 (with all the <i>egc</i> genes)	This study
DSTL020	TK23 containing <i>dst</i> cluster with double deletion <i>dst2/dst25</i>	(Vingadassalon et al., 2015)
DSTL028	Complementation of DSTL020 with pCAS009	This study

Table S2: Plasmids used in this study

Plasmid	Description	Reference
pUZ8002	RK2 derivative with defective oriT ( <i>aph</i> )	(Flett et al., 1997)
pCR®-Blunt	<i>E. coli</i> cloning vector	Invitrogen (Thermo Fisher Scientific)
pOSV801	Plasmid constructed containing apramycin resistance and $\phi$ BT1 integrase	(Aubry et al., 2019)
pOSV812	Plasmid constructed containing kanamycin resistance and VWB integrase	(Aubry et al., 2019)
pCR-SP25- <i>cgc8-11</i>	Fragment CAS007 (SP25- <i>cgc8-11</i> ) in pCR blunt	This study
pCR- <i>cgc12-14-ter</i>	Fragment CAS007 ( <i>cgc12-14-T4 ter</i> ) in pCR blunt	This study
pCAS001	pOSV801 containing CAS001	This study
pCAS002	pOSV801 containing CAS002	This study
pCAS003	pOSV801 containing CAS003	This study
pCAS005	pOSV812 containing CAS005	This study
pCAS006	pOSV812 containing CAS006	This study
pCAS007	pOSV801 containing CAS007	This study
pCAS008	pOSV812 containing CAS008	(Aubry et al., 2019)
pCAS009	pOSV812 containing CAS009	This study
pCAS010	pOSV812 containing CAS010	This study
pCAS011	pOSV812 containing CAS011	This study
pCAS013	pOSV812 containing CAS013	This study
pCAS014	pCAS009 with modified resistance cassette ( <i>aacIII(4)</i> instead of <i>aph</i> )	This study
pCAS016	pOSV812 containing CAS016	This study
pCAS017	pOSV812 containing CAS017	This study
pCAS018	pOSV812 containing CAS018	This study
pCAS019	pOSV801 containing CAS019	This study
pCAS020	pOSV801 containing CAS020	This study
pCAS022	pOSV812 containing CAS022	This study
pCAS023	pOSV801 containing CAS023	This study
pCAS024	pOSV812 containing CAS024	This study
pCAS026	pOSV801 containing CAS026	This study

Table S3: Oligonucleotides used in this study

Oligonucleotides	Sequence	Description
CEA_vec_seq14	ATTTTCAGTGCAATTTATCTCTTC	Sequencing of beginning of the gene cassettes
CEA_vec_seq21	CACGGAATCCTGCGGATCAC	Sequencing of end of the cassettes inserted in pOSV812
JWseq6	CCCTTTTTTGGCCTTGAAAT	Sequencing of end of the cassettes inserted in pOSV801
oncas001bis	<u>GCTGCTAGCTGTT</u> CACATTCGAACCGTCTCTG	Amplification synthetic promoters forward (partial <i>NotI</i> and <i>NheI</i> sites underlined)
oncas002	ATGGACACTCCTTACTTAGAC	Amplification synthetic promoters reverse
oncas003	GTATAGGAACTTCATGCATGCGGCCGCTGCTAGCTGTTACATTTCGAACCG	Bridging oligonucleotide between plasmid (pOSV801-pOSV812) and promoter
oncas004	GAATACGACAGTCTAAGTAAGGAGTG TCCATATGTCTTCGTCCACGGCTACGAG	Bridging oligonucleotide between promoter and <i>egc15</i>
oncas005	TCGCATGGGGCGTCAAGTAAGCTGATCCGGTGGATGACCTTTTGAATG	Bridging oligonucleotide between <i>egc15</i> and T4 terminator
oncas006	CACAAAACGGTTTACAAGCATAACTAGTAGCGGCCGCTTAAGCGCTCCCTG	Bridging oligonucleotide between T4 terminator and pOSV801
oncas007	TGATCCGGTGGATGACCTTTTG	Amplification T4 terminator forward
oncas008bis	<u>GCTACTAGTTATGCTTGTAAACCGTTT</u> TG	Amplification T4 terminator reverse (partial <i>NotI</i> and <i>SpeI</i> sites underlined)
oncas010bis	<u>AAACTTAAGCGGCCGCTACTAGTTATGCTTGTAAACCGTTTTG</u>	Amplification T4 terminator reverse (complete <i>AflII</i> , <i>NotI</i> and <i>SpeI</i> suffix underlined)
oncas011	ATGTCCITCGTCCACGGCTAC	Amplification <i>egc15</i> forward
oncas012	GCTTACTTGACGCCCCATGC	Amplification <i>egc15</i> reverse
oncas013	ATGAGGGACACCACGGTGGC	Amplification <i>egc4-6</i> forward
oncas014	GCTCACGGGGACGCGGCGACC	Amplification <i>egc4-6</i> reverse
oncas015	GAATACGACAGTCTAAGTAAGGAGTG TCCATATGAGGGACACCACGGTGGCCGG	Bridging oligonucleotide between promoter and <i>egc4-6</i>
oncas016	CGCCGCGTCCCCGTGAGCTGATCCGGTGGATGACCTTTTGAATG	Bridging oligonucleotide between <i>egc4-6</i> and T4 terminator
oncas017	CGGGAGGCCGTGATGTC	Sequencing for verification of <i>egc5-6</i>
oncas018	ATGCGCCTGCCTCCCCATGAAC	Amplification <i>egc7</i> forward
oncas019	TTATCAGCCGACGACCCAGTG	Amplification <i>egc7</i> reverse
oncas020	GAATACGACAGTCTAAGTAAGGAGTG TCCATATGCGCCTGCCTCCCC	Bridging oligonucleotide between promoter and <i>egc7</i>
oncas021	CCACTGGGTCGTGCGGCTGATGATCCGGTGGATGACCTTTTGAATG	Bridging oligonucleotide between <i>egc7</i> and T4 terminator
oncas022bis	ATGAGGGCGATGCGGCAAC	Amplification <i>egc6</i> forward (GTG changed to ATG)
oncas023bis	GAATACGACAGTCTAAGTAAGGAGTG TCCATATGAGGGCGATGCGGCAACGC GAC	Bridging oligonucleotide between promoter and <i>egc6</i>

oncas024	ATGCCGCAGGTGAACGCC	Amplification <i>cgc3</i> forward (GTG changed to ATG)
oncas025	TTATCATGACATCTCCCGATCTG	Amplification <i>cgc3</i> reverse
oncas026	CCTGCCGCGAACCGGAGG	Amplification <i>cgc17</i> forward
oncas027	TCACGGGATCAGCACCACCTTG	Amplification <i>cgc17</i> reverse
oncas028	GAATACGACAGTCTAAGTAAGGAGTG TCCATATGCCGCAGGTGAACGCC	Bridging oligonucleotide between promoter and <i>cgc3</i>
oncas029	CCAAGCAGATCGGGAGATGTCATGAT AACCTGCCGCGAACCGGAG	Bridging oligonucleotide between <i>cgc3</i> and <i>cgc17</i>
oncas030	CAAGGTGGTGTGCTGATCCCGTGATGAT CCGGTGGATGACCTTTTGAATGAC	Bridging oligonucleotide between <i>cgc17</i> and T4 terminator
oncas031	ATGGCCACCGAGTCCGCCACC	Amplification <i>cgc22</i> forward
oncas032	CTACCCGCCGTCGCCGTCGC	Amplification <i>cgc22</i> reverse
oncas033	GAATACGACAGTCTAAGTAAGGAGTG TCCATATGGCCACCGAGTCCGCC	Bridging oligonucleotide between promoter and <i>cgc22</i>
oncas034	GACGGCGACGGCCGGGTAGTGATCCG GTGGATGACCTTTTGAATGAC	Bridging oligonucleotide between <i>cgc22</i> and T4 terminator
oncas035bis	ATGAGCATCTCCACCACCGCCCC	Amplification <i>cgc18</i> forward (GTG changed to ATG)
oncas036bis	TCACAGCTCGGCCTCGG	Amplification <i>cgc18</i> reverse
oncas037	GAATACGACAGTCTAAGTAAGGAGTG TCCATATGAGCATCTCCACCACCGCC	Bridging oligonucleotide between promoter and <i>cgc18</i>
oncas038	CCCGAGGCCGAGCTGTGATGATCCGG TGGATGACCTTTTGAATGAC	Bridging oligonucleotide between <i>cgc18</i> and T4 terminator
oncas039bis	ATGGCGCTACCCGTTTCGCACC	Amplification <i>cgc2</i> forward (GTG changed to ATG)
oncas040bis	TCAACGCCCGTCCGCCACC	Amplification <i>cgc2</i> reverse
oncas041	GAATACGACAGTCTAAGTAAGGAGTG TCCATATGGCGCTACCCGTTTCGC	Bridging oligonucleotide between promoter and <i>cgc2</i>
oncas042	GGCCGACGGGCGTTGATGATCCGGTG GATGACCTTTTGAATGAC	Bridging oligonucleotide between <i>cgc2</i> and T4 terminator
bridge4	ACGGTTTACAAGCATAACTAGTAGCG GCCGCTTAAGGTCGACCCGTCIG	Bridging oligonucleotide between T4 terminator and pOSV812
oncas043	ATGACCGCCGAGACCGTCC	Amplification <i>cgc20-21</i> forward
oncas044	TCACGCCTTCTCTCGAC	Amplification <i>cgc20-21</i> reverse
oncas045	GGAGAATACGACAGTCTAAGTAAGGA GTGTCCATATGACCGCCGAGACCGTCC	Bridging oligonucleotide between promoter and <i>cgc20</i>
oncas046	CCGTCGAGAGGAAGGCGTGATGATCC GGTGGATGACCTTTTGAATGACC	Bridging oligonucleotide between <i>cgc21</i> and T4 terminator
oncas047	TGCGGAACGGTGTGGATCAAC	Sequencing of <i>cgc3</i>
oncas048	CGGTGTCTGTAGCCGAACAG	Sequencing of <i>cgc17</i>
oncas049	ATGTCAATGCCAGCGAACAGG	Amplification <i>cgc8</i> forward
oncas050	CCGGTCACCGCCCTCG	Amplification <i>cgc11</i> reverse
oncas051	ATGACGGCCITCGACGTCC	Amplification <i>cgc12</i> forward
oncas052	TCAACTCATCGGTTCCGACG	Amplification <i>cgc14</i> reverse
oncas053	GAATACGACAGTCTAAGTAAGGAGTG TCCATATGTCAATGCCAGCGAACAGGC	Bridging oligonucleotide between promoter and <i>cgc8</i>
oncas055	CCCGTCCGAACCGATGAGTTGATGATC CGGTGGATGACCTTTTGAATGAC	Bridging oligonucleotide between <i>cgc14</i> and T4 terminator
oncas056	CCTACCGCGACGCCTCCGTG	Sequencing of <i>cgc20</i>
oncas057	CGCGCTGGTGGCCGATCCC	Sequencing of <i>cgc20</i>
oncas058	GAGCTGGGCCAGCCAGTCCG	Sequencing of <i>cgc21</i>
oncas059	CTGCGGCTGCTCGTCTGGG	Sequencing of <i>cgc18</i>
oncas060	CGTACGCGGCGTAGGAGACC	Sequencing of <i>cgc18</i>
oncas061	TGCGCCTGCGTGGTCTGGG	Sequencing of <i>cgc18</i>

oncas062	ATGACGAACCATGCGGACAAC	Amplification <i>cgc19</i> forward (GTG changed to ATG)
oncas063	TCAGGGGGTCTCGTTCGG	Amplification <i>cgc19</i> reverse
oncas064	ATGGAGAAGAGAGCCGGGACG	Amplification <i>cgc16</i> forward
oncas065	TCATGTGTCTCCGGTTTCG	Amplification <i>cgc16</i> reverse
oncas066	GAATACGACAGTCTAAGTAAGGAGTG TCCATATGGAGAAGAGAGCCGGGACG	Bridging oligonucleotide between promoter and <i>cgc16</i>
oncas067	CGCGAACCGGAGGACACATGATGATC CGGTGGATGACCTTTTGAATGAC	Bridging oligonucleotide between <i>cgc16</i> and T4 terminator
oncas068	GAATACGACAGTCTAAGTAAGGAGTG TCCATATGACGAACCATGCGGACAACC C	Bridging oligonucleotide between promoter and <i>cgc19</i>
oncas069	GCCGAACGAGACCCCTGATGATCCG GTGGATGACCTTTTGAATGAC	Bridging oligonucleotide between <i>cgc19</i> and T4 terminator
oncas074	<u>TCTCGAGGGCGGTGACCGGATGACGG</u> CCTTCGAC	Amplification of <i>cgc12</i> forward with <i>Xba</i> I site underlined (used with oncas010)
oncas075	ATCACCACGCCGCAGCGCTC	Sequencing of <i>cgc13</i>
oncas076	ATGCGCGTCGATGATCAC	Sequencing of <i>cgc13</i>
cmj55F	CGTCTTCTGGGCCGACTTTG	Sequencing of <i>cgc13</i>
cmj55R	GAGTCCGCGTGGATGATCTC	Sequencing of <i>cgc12-13</i>
cmj66F	GACGCCCGGATCCTGCTCTC	Sequencing of <i>cgc8</i>
cmj66R	GGACCCGCCAGGTGTCGTAG	Sequencing of <i>cgc8</i>
cmj67F	CCACCTCCTCGACTGGCTCTC	Sequencing of <i>cgc9</i>
cmj67R	CTCGACGAACTGCGGGATCAC	Sequencing of <i>cgc8-9</i>
cmj68F	GTGAAGGTCCAGCCGTTCCC	Sequencing of <i>cgc10</i>
cmj68R	GGTCCCTGGCCGATGATGTG	Sequencing of <i>cgc9-10</i>
cmj69F	CCTGTGGTCCCACCACAAGAAG	Sequencing of <i>cgc11-12</i>
cmj69R	CAGTCGCCCTCGATGACGTAG	Sequencing of <i>cgc10-11</i>
cmj70F	TGGCCCTGATCGAGGACTGC	Sequencing of <i>cgc12</i>
cmj70R	CGAGCTGGACACGTCCGATG	Sequencing of <i>cgc11-12</i>
cmj71R	GGCTGGTACGAGCCGAAGATG	Sequencing of <i>cgc14</i>

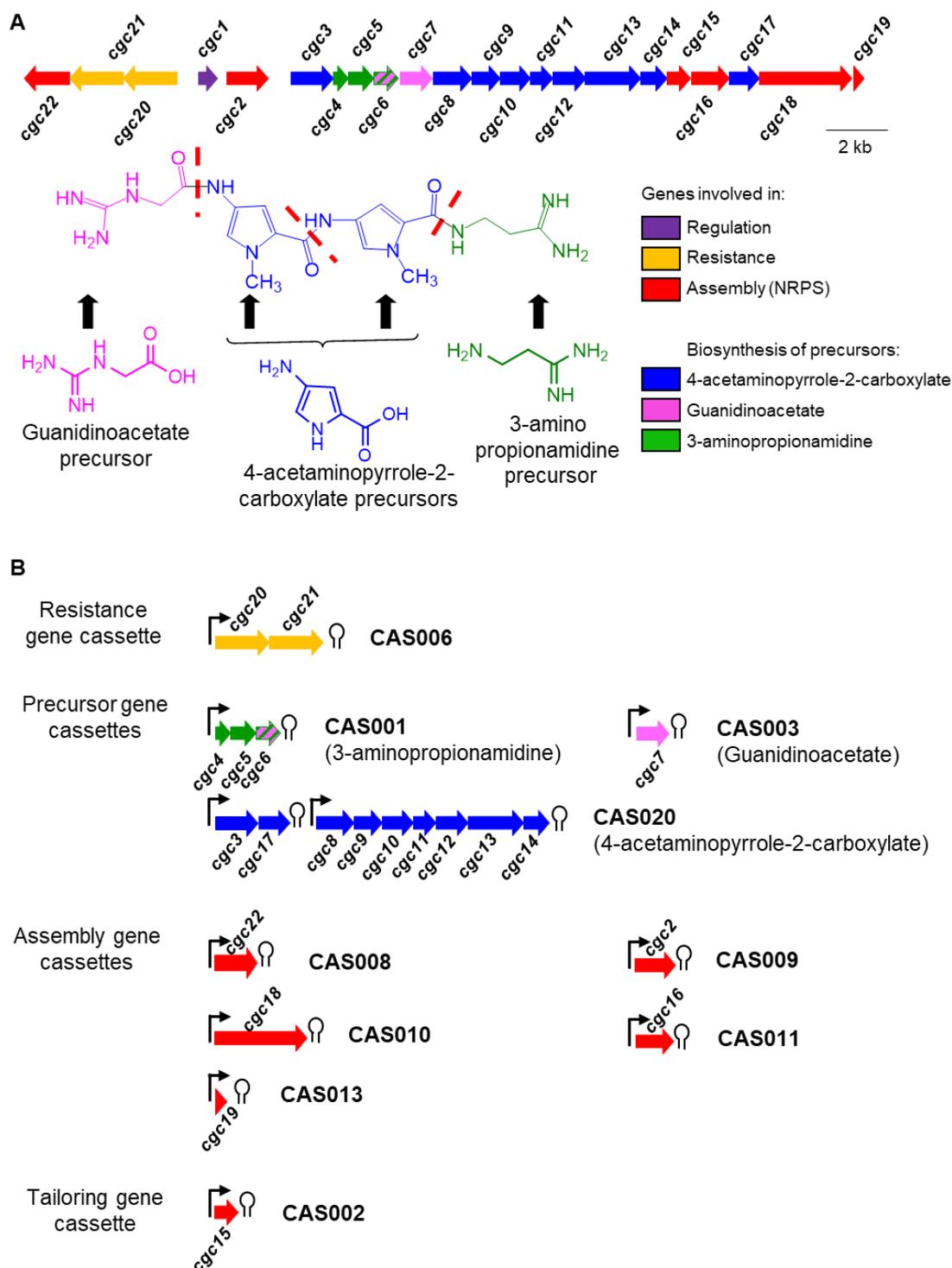


Figure S1: Congocidine biosynthetic gene cluster and gene cassette constructed

A) Native *S. ambifaciens* congocidine (*cgc*) biosynthetic gene cluster and congocidine structure. Red dashed lines separate the different monomers of congocidine

B) Synthetic gene cassettes constructed

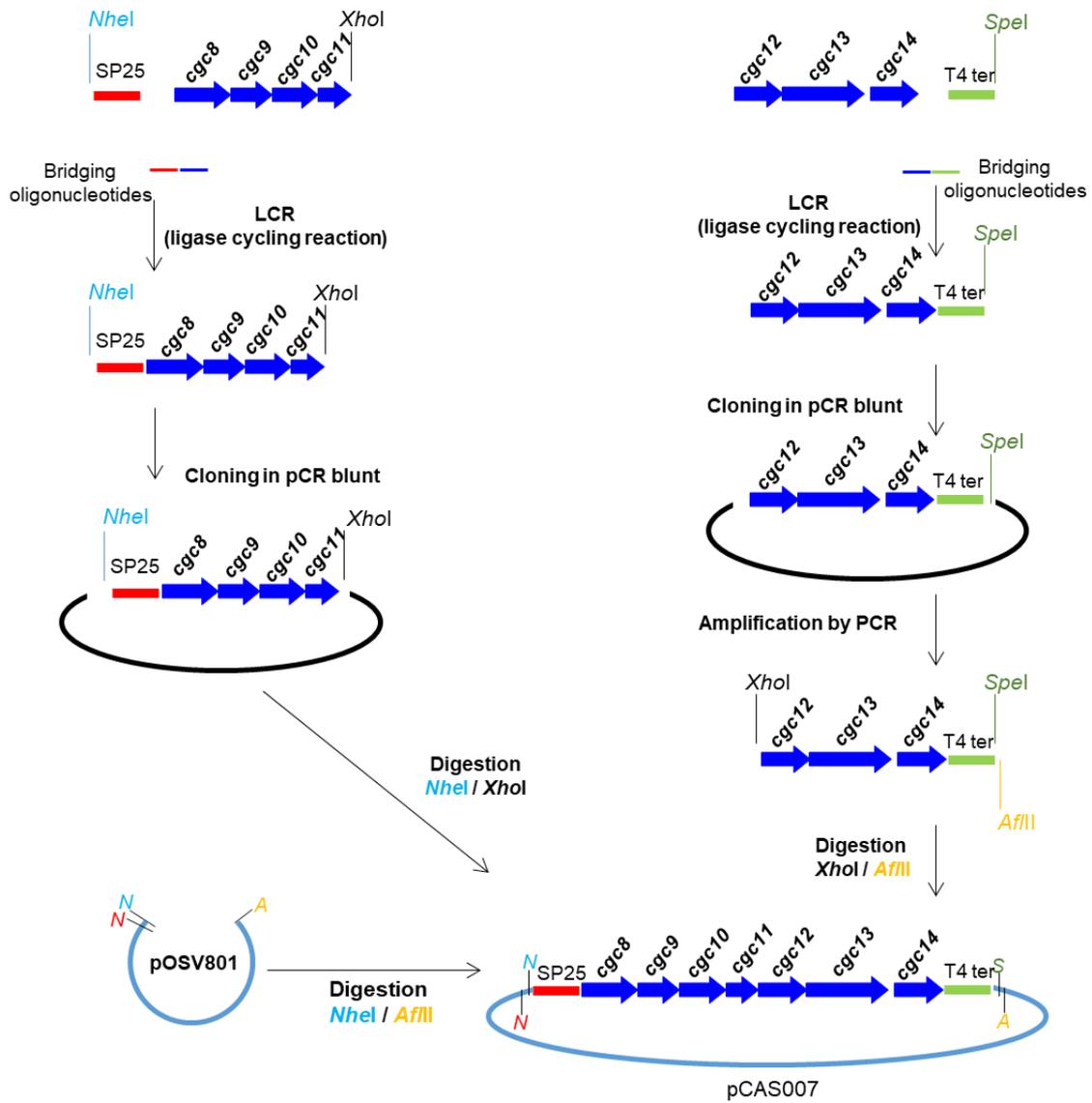


Figure S2: Scheme of the construction of the CAS007 cassette

*N*: *NsiI*, *N*: *NheI*, *S*: *SpeI*, *A*: *AflIII*, T4 ter : T4 terminator

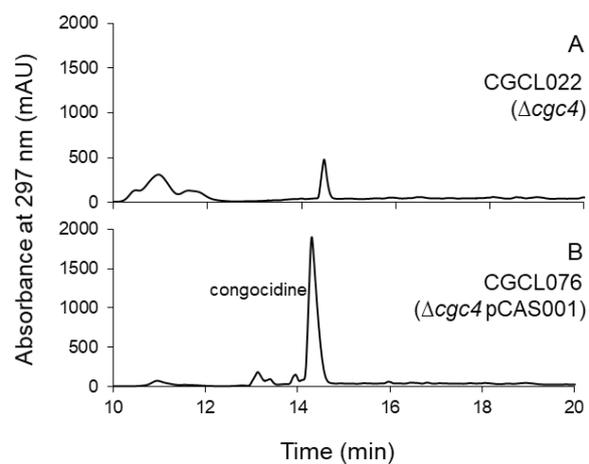


Figure S3: Verification of the functionality of the CAS001 gene cassette: genetic complementation of a *cgc4* deletion mutant.

HPLC chromatograms of culture supernatants of *S. lividans* A) CGCL022 (*cgc* cluster with *cgc4* deleted), B) CGCL076 (CGCL022 with CAS001 containing *cgc4*, *cgc5* and *cgc6*). Samples were analyzed on a reverse phase C<sub>18</sub> column, eluted in isocratic conditions with 0.1% HCOOH in H<sub>2</sub>O (solvent A)/ 0.1% HCOOH in CH<sub>3</sub>CN (solvent B) (95:5) at 1 ml.min<sup>-1</sup> for 7 min, followed by a gradient to 40:60 A/B over 23 min. Absorbance was monitored at 297 nm.

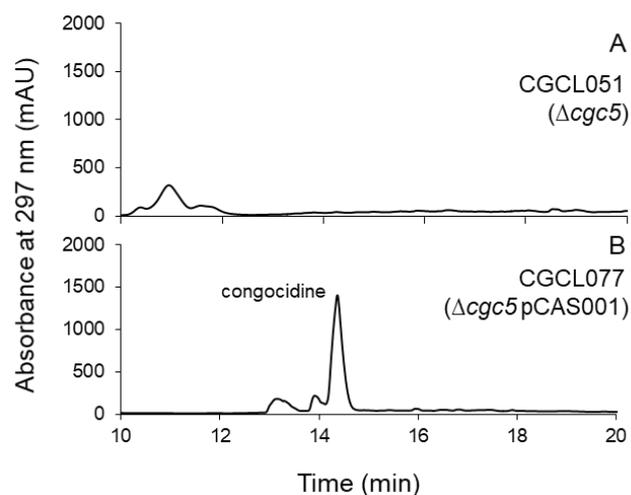


Figure S4: Verification of the functionality of the CAS001 gene cassette: genetic complementation of a *cgc5* deletion mutant.

HPLC chromatograms of culture supernatants of *S. lividans* A) CGCL051 (*cgc* cluster with *cgc5* deleted), B) CGCL077 (CGCL051 with CAS001 containing *cgc4*, *cgc5* and *cgc6*). Samples were analyzed on a reverse phase  $C_{18}$  column, eluted in isocratic conditions with 0.1% HCOOH in  $H_2O$  (solvent A)/ 0.1% HCOOH in  $CH_3CN$  (solvent B) (95:5) at  $1\text{ ml}\cdot\text{min}^{-1}$  for 7 min, followed by a gradient to 40:60 A/B over 23 min. Absorbance was monitored at 297 nm.

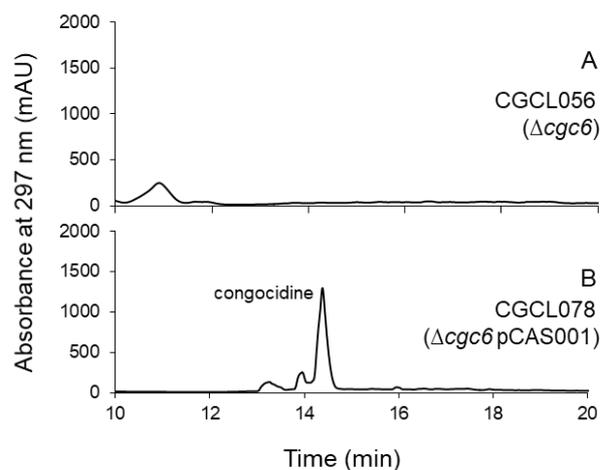


Figure S5: Verification of the functionality of the CAS001 gene cassette: genetic complementation of a *cgc6* deletion mutant.

HPLC chromatograms of culture supernatants of *S. lividans* A) CGCL056 (*cgc* cluster with *cgc6* deleted), B) CGCL078 (CGCL056 with CAS001 containing *cgc4*, *cgc5* and *cgc6*). Samples were analyzed on a reverse phase  $C_{18}$  column, eluted in isocratic conditions with 0.1% HCOOH in  $H_2O$  (solvent A)/ 0.1% HCOOH in  $CH_3CN$  (solvent B) (95:5) at  $1\text{ ml}\cdot\text{min}^{-1}$  for 7 min, followed by a gradient to 40:60 A/B over 23 min. Absorbance was monitored at 297 nm.

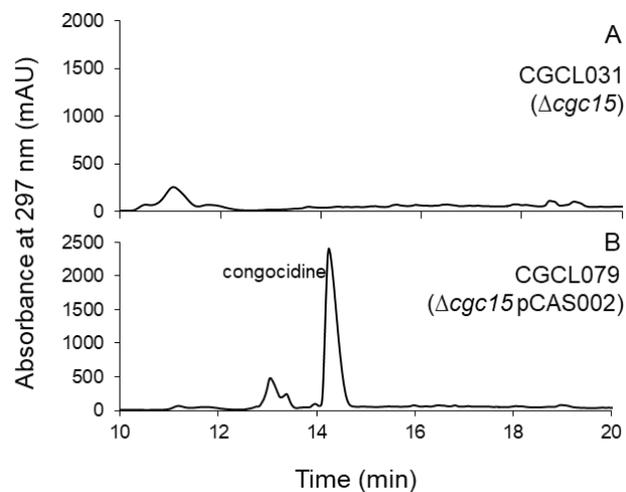


Figure S6: Verification of the functionality of the CAS002 gene cassette: genetic complementation of a *cgc15* deletion mutant.

HPLC chromatograms of culture supernatants of *S. lividans* A) CGCL031 (*cgc* cluster with *cgc15* deleted), B) CGCL079 (CGCL031 with CAS002 containing *cgc15*).

Samples were analyzed on a reverse phase C<sub>18</sub> column, eluted in isocratic conditions with 0.1% HCOOH in H<sub>2</sub>O (solvent A)/ 0.1% HCOOH in CH<sub>3</sub>CN (solvent B) (95:5) at 1 ml.min<sup>-1</sup> for 7 min, followed by a gradient to 40:60 A/B over 23 min. Absorbance was monitored at 297 nm.

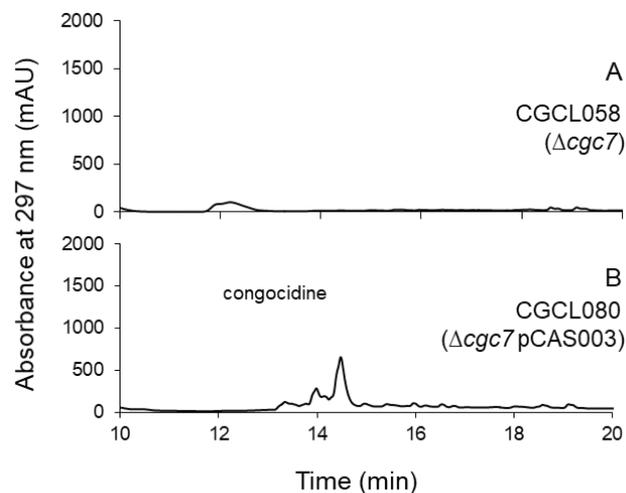


Figure S7: Verification of the functionality of the CAS003 gene cassette: genetic complementation of a *cgc7* deletion mutant.

HPLC chromatograms of culture supernatants of *S. lividans* A) CGCL058 (*cgc* cluster with *cgc7* deleted), B) CGCL080 (CGCL058 with CAS003 containing *cgc7*).

Samples were analyzed on a reverse phase  $C_{18}$  column, eluted in isocratic conditions with 0.1% HCOOH in  $H_2O$  (solvent A)/ 0.1% HCOOH in  $CH_3CN$  (solvent B) (95:5) at  $1 \text{ ml} \cdot \text{min}^{-1}$  for 7 min, followed by a gradient to 40:60 A/B over 23 min. Absorbance was monitored at 297 nm.

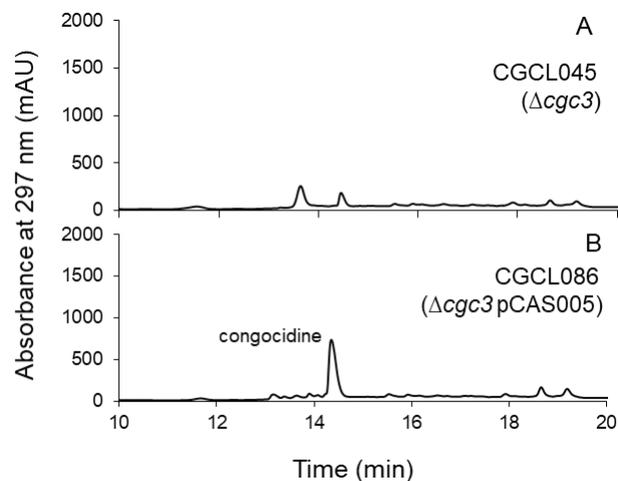


Figure S8: Verification of the functionality of the CAS005 gene cassette: genetic complementation of a *cgc3* deletion mutant.

HPLC chromatograms of culture supernatants of *S. lividans* A) CGCL045 (*cgc* cluster with *cgc3* deleted), B) CGCL086 (CGCL045 with CAS005 containing *cgc3* and *cg17*).

Samples were analyzed on a reverse phase  $C_{18}$  column, eluted in isocratic conditions with 0.1% HCOOH in  $H_2O$  (solvent A)/ 0.1% HCOOH in  $CH_3CN$  (solvent B) (95:5) at  $1 \text{ ml} \cdot \text{min}^{-1}$  for 7 min, followed by a gradient to 40:60 A/B over 23 min. Absorbance was monitored at 297 nm.

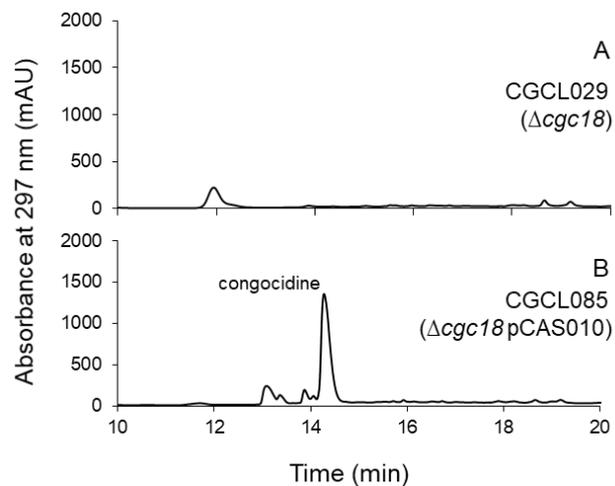


Figure S9: Verification of the functionality of the CAS010 gene cassette: genetic complementation of a *cgc18* deletion mutant.

HPLC chromatograms of culture supernatants of *S. lividans* A) CGCL029 (*cgc* cluster with *cgc18* deleted), B) CGCL085 (CGCL029 with CAS010 containing *cgc18*).

Samples were analyzed on a reverse phase  $C_{18}$  column, eluted in isocratic conditions with 0.1% HCOOH in  $H_2O$  (solvent A)/ 0.1% HCOOH in  $CH_3CN$  (solvent B) (95:5) at  $1\text{ ml}\cdot\text{min}^{-1}$  for 7 min, followed by a gradient to 40:60 A/B over 23 min. Absorbance was monitored at 297 nm.

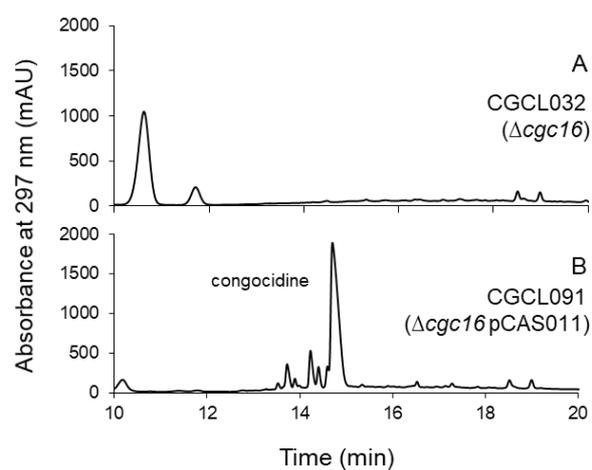


Figure S10: Verification of the functionality of the CAS011 gene cassette: genetic complementation of a *cgc16* deletion mutant.

HPLC chromatograms of culture supernatants of *S. lividans* A) CGCL032 (*cgc* cluster with *cgc16* deleted), B) CGCL091 (CGCL032 with CAS011 containing *cgc16*).

Samples were analyzed on a reverse phase  $C_{18}$  column, eluted in isocratic conditions with 0.1% HCOOH in  $H_2O$  (solvent A)/ 0.1% HCOOH in  $CH_3CN$  (solvent B) (95:5) at  $1 \text{ ml} \cdot \text{min}^{-1}$  for 7 min, followed by a gradient to 40:60 A/B over 23 min. Absorbance was monitored at 297 nm.

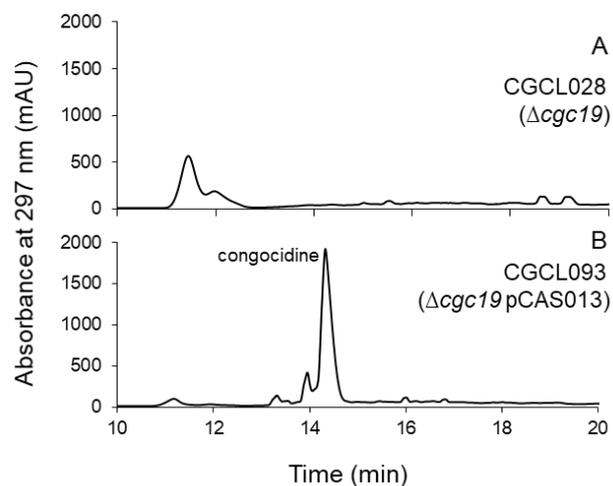


Figure S11: Verification of the functionality of the CAS013 gene cassette: genetic complementation of a *cgc19* deletion mutant.

HPLC chromatograms of culture supernatants of *S. lividans* A) CGCL028 (*cgc* cluster with *cgc19* deleted), B) CGCL093 (CGCL028 with CAS013 containing *cgc19*).

Samples were analyzed on a reverse phase C<sub>18</sub> column, eluted in isocratic conditions with 0.1% HCOOH in H<sub>2</sub>O (solvent A)/ 0.1% HCOOH in CH<sub>3</sub>CN (solvent B) (95:5) at 1 ml.min<sup>-1</sup> for 7 min, followed by a gradient to 40:60 A/B over 23 min. Absorbance was monitored at 297 nm.

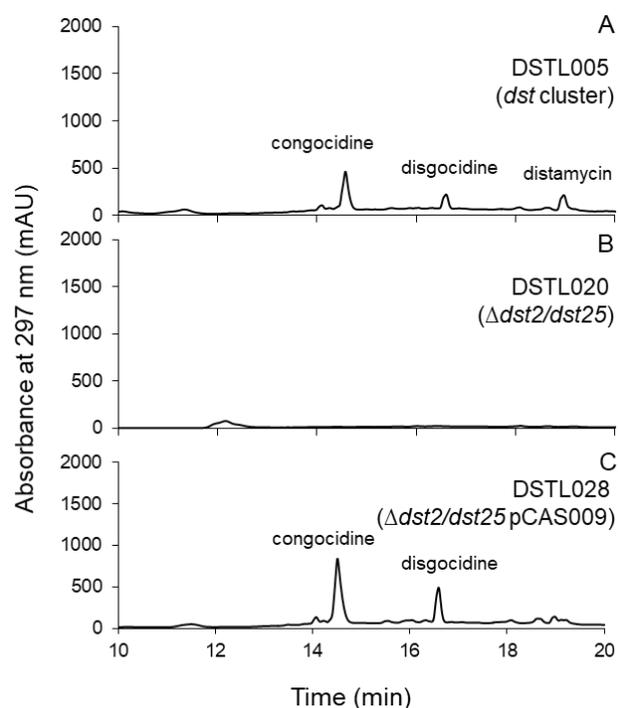


Figure S12: Verification of the functionality of the CAS009 gene cassette: genetic complementation of a *dst2/dst25* deletion mutant.

HPLC chromatograms of culture supernatants of *S. lividans* A) DSTL005 (containing both native *dst* clusters), B) DSTL020 (*dst* clusters with *dst2* and *dst25* deleted) C) DSTL028 (DSTL020 with CAS009 containing *cgc2*).

Samples were analyzed on a reverse phase C<sub>18</sub> column, eluted in isocratic conditions with 0.1% HCOOH in H<sub>2</sub>O (solvent A)/ 0.1% HCOOH in CH<sub>3</sub>CN (solvent B) (95:5) at 1 ml.min<sup>-1</sup> for 7 min, followed by a gradient to 40:60 A/B over 23 min. Absorbance was monitored at 297 nm.

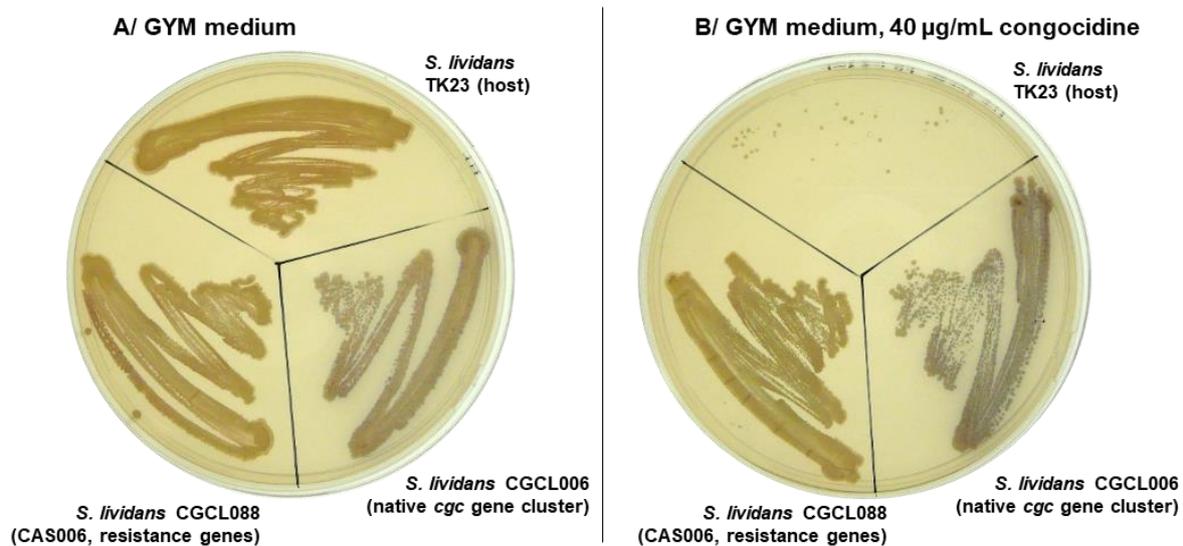


Figure S13: Verification of the functionality of CAS006. The various strains were plated on GYM medium without (A) or with (B) 40 µg/ml of congocidine.

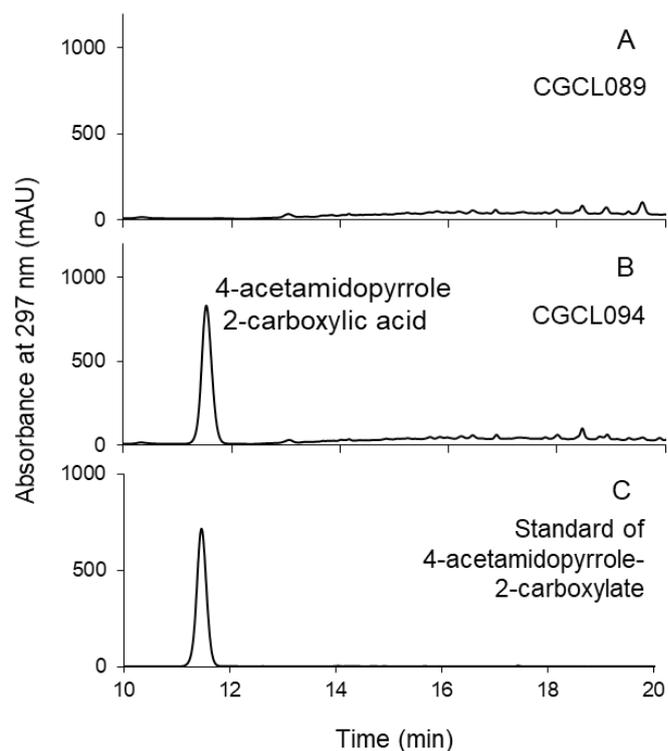


Figure S14: Verification of the functionality of the CAS007 gene cassette: production of 4-acetamidopyrrole-2-carboxylate.

HPLC chromatograms of culture supernatants of *S. lividans* A) CGCL089 (TK23 with CAS005 containing *cgc3* and *cgc17*), B) CGCL094 (TK23 with CAS005 and CAS007), C) Standard of 4-acetamidopyrrole-2-carboxylate. Samples were analyzed on a reverse phase C<sub>18</sub> column, eluted in isocratic conditions with 0.1% HCOOH in H<sub>2</sub>O (solvent A)/ 0.1% HCOOH in CH<sub>3</sub>CN (solvent B) (95:5) at 1 ml.min<sup>-1</sup> for 7 min, followed by a gradient to 40:60 A/B over 23 min. Absorbance was monitored at 297 nm.

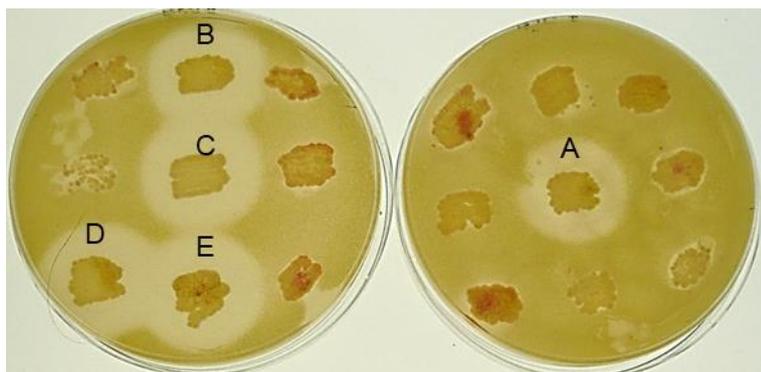


Figure S15: Screening for congoicidin producing clones.

After 2 days of growth of *S. lividans* CGCL098 on HT at 28°C, an overlay of *M. luteus* was added to the plate. The pictures were taken after overnight incubation at 37°C.

## References:

- Aubry, C., Pernodet, J.-L., and Lautru, S. (2019). A set of modular and integrative vectors for synthetic biology in *Streptomyces*. *Appl. Environ. Microbiol.* Aug 1;85(16).
- Flett, F., Mersinias, V., and Smith, C.P. (1997). High efficiency intergeneric conjugal transfer of plasmid DNA from *Escherichia coli* to methyl DNA-restricting streptomycetes. *FEMS Microbiol. Lett.* *155*, 223–229.
- Juguet, M., Lautru, S., Francou, F.-X., Nezbedová, S., Leblond, P., Gondry, M., and Pernodet, J.-L. (2009). An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congocidine in *Streptomyces ambofaciens*. *Chem. Biol.* *16*, 421–431.
- Kieser, T., Bibb, M., Buttner, M., and Hopwood, D.A. (2000). *Practical Streptomyces genetics*, John Innes Foundation, Norwich NR47UH, UK.
- Lautru, S., Song, L., Demange, L., Lombès, T., Galons, H., Challis, G.L., and Pernodet, J.-L. (2012). A sweet origin for the key congocidine precursor 4-acetamidopyrrole-2-carboxylate. *Angew. Chem. Int. Ed Engl.* *51*, 7454–7458.
- Simon, R., Prierer, U., and Pühler, A. (1983). A broad host range mobilization system for *in vivo* genetic engineering: transposon mutagenesis in gram negative bacteria. *Nat. Biotechnol.* *1*, 784–791.
- Vingadassalon, A., Lorieux, F., Juguet, M., Le Goff, G., Gerbaud, C., Pernodet, J.-L., and Lautru, S. (2015). Natural combinatorial biosynthesis involving two clusters for the synthesis of three pyrrolamides in *Streptomyces netropsis*. *ACS Chem. Biol.* *10*, 601–610.

## Chapter III perspectives:

In the third chapter, I refactored the congocidine biosynthetic gene cluster. However, due to time constraints, I could not perform all the experiments planned to analyze the production of congocidine in the *S. lividans* host. Thus, to better characterize congocidine production from the refactored gene cluster, precise kinetics and quantification of the production are required, and should be compared with the kinetics/quantification of the native gene cluster. qRT-PCR analyzes would give some insight on the transcription of the different genes and the strength of the promoters used in our genetic context. It may also help identifying possible bottlenecks in congocidine biosynthesis. Additionally, since we observed an instability of the plasmids bearing the refactored gene cluster in some *E. coli* strains, the stability of the constructions in *Streptomyces* should be assessed. It would also be possible to introduce the refactored cluster in other genetic backgrounds and to compare congocidine production in the various hosts.

In this project, we were confronted to unwanted homologous recombination in *E. coli* strains due to the repeated terminator sequences. This resulted in the instability of the two plasmids harboring the refactored cluster these strains. This observation raises concerns for future engineering experiments. The only previous report of instability in a refactoring pathway was made for the epothilone pathway (Osswald et al., 2014). The same promoter-RBS region (PTn5, 140 bps) and the same terminator (TD1, about 50 bp) were used in three gene cassettes, and the final vector containing the three cassettes was unstable. The problem was circumvented by the use of two different compatible plasmids. In our case, the use of different terminators such as the ones reported by Horbal et al. (2018a) should reduce sequence repetitions and alleviate the problem of homologous recombination we faced.

# General Conclusion

Researchers in the specialized metabolism field aim at discovering new compounds with (therapeutic) applications, and synthetic biology is one of the tools used to reach that goal. Non ribosomal peptide synthetases are modular enzymes responsible for the production of extremely diverse compounds, some of which are currently used in medicine. Were we able to modify in a plug-and-play manner these enzymes, then a huge number of metabolites with potential pharmaceutical applications could be synthesized by combinatorial biosynthesis. Currently, NRPS engineering is, however, limited by our imperfect understanding of the biosynthetic process: the substrate specificity of adenylation, condensation or thioesterase domains, and the protein/protein interactions among domains, modules or protein subunits are yet to be fully deciphered. Due to their unusual architecture (stand-alone NRPS domains or modules), and the existence of some kind of natural combinatorial biosynthesis for the synthesis of some pyrrolamides, the pyrrolamide NRPSs constitute a model to probe the limiting factors impeding the success of NRPS combinatorial biosynthesis approaches. During my PhD project, I aimed at constructing tools to permit combinatorial biosynthesis of the pyrrolamide biosynthetic genes.

Characterization of anthelvencin biosynthetic gene cluster allowed to understand the biosynthesis of anthelvencins A, B and C, and it also resulted in the addition of new pyrrolamides NRPS genes to our library. These genes can be selected for NRPS exchanges to question the factors limiting efficient metabolite production. The two genes directing respectively the biosynthesis and assembly of a novel pyrrolamide moiety (4-amino-dihydropyrrole-2-carboxylate) were also identified, and could be of use to develop pyrrolamide analogs at a later stage.

To establish a platform for combinatorial biosynthesis, we simultaneously proceeded to the construction of integrative plasmids. I built flexible modular backbones, compatible with different assembly methods and easy to modify. These plasmids are integrated in *Streptomyces* strains, and after genome integration, a system allows the excision of sequences that are identical among all vectors, and the recycling of the resistance marker. The utility of these vectors goes well beyond the unique goal of combinatorial biosynthesis of the pyrrolamide biosynthetic genes, and the plasmids were offered to the *Streptomyces* research community as tools for synthetic biology applications.

The integrative plasmids were then used as backbones for the refactoring of a pyrrolamide biosynthetic gene cluster. Refactoring the congocidine biosynthetic gene cluster followed two purposes. Firstly, it aimed at producing congocidine using a standardized gene cluster freed of the native regulation. Secondly, it was a prerequisite for combinatorial biosynthesis experiments, to prove the feasibility of the *de novo* construction of a biosynthetic gene cluster using synthetic gene cassettes. Using 11 gene cassettes harboring the 21 congocidine biosynthetic genes, we successfully refactored the congocidine biosynthetic gene cluster.

The refactored congocidine biosynthetic gene cluster can now be used as a platform to exchange NRPS genes and probe NRPS protein/protein interactions and substrate specificities. A first step could consist in exchange of domains with identical role, such as the peptidyl-carrier

protein domain of the pyrrole moiety. Since it has no catalytic role, success or failure of congoicidine production after the exchange could lead to the identification of the regions of the NRPSs involved in protein/protein interactions. Conversely, exchange of condensation domains could be very informative concerning substrate specificities. Cross complementation observed in the third chapter (*cgc2* can restore congoicidine and disgocidine production in a *dst2/dst25* mutant) suggests that substrate specificities of the pyrrolamide condensation domains are quite relaxed, but still exist (distamycin production could not be restored to a detectable level with *cgc2*).

The question of docking domains can also be tackled using our system. Indeed, no COM domains were detected in the pyrrolamide NRPSs. Thorough bioinformatics analyses of the NRPS sequence could, however, reveal unconventional docking domains, as the ones reported for rhabdopeptides and xenortide peptides (Hacker et al., 2018). Then our refactored biosynthetic gene cluster could be used to modify these potential domains through deletions or mutations and to study the impact on congoicidine production.

In the event of absence of pyrrolamide production, whether during domain exchange experiments or during docking domain modification experiments, the identification of the intermediaries bound to the PCP domain would bring very valuable information. Recently described chemical non-hydrolyzable “chain termination” probes (Ho et al., 2017), which capture the biosynthetic NRP intermediate *in vivo*, could be used in such intent.

*In vitro* studies would be complementary to the approaches previously mentioned. Purification of a C domain for example would allow to study its substrate specificities, using either chemically synthesized substrate analogs or PCP-bound substrate analogs. Such experiments should help clarify in particular the specificity of C domains at the donor site.

While I could not expect to complete combinatorial biosynthetic experiments during my project, combinatorial biosynthesis being by nature impossible to exhaust, I was a little bit disappointed not to have the time to perform at least a few genes replacements. I started my thesis confident that I would reach that step, and later on, as the project was delayed, I still thought that an extra year would allow me to do so. In the end, even the refactoring of the congoicidine gene cluster was challenging and only obtained during the last weeks of experiments.

How can we explain the gap between my experience as a young researcher, and the claims concerning synthetic biology applied to specialized metabolites research? In most definitions given in the field of specialized metabolites, synthetic biology is linked to the concepts of design and engineering. Guzmán-Trampe and colleagues (2017) present it “as an engineering approach to improve or completely create systems and organisms with specific or desirable functions”. Porcar (2019) remarks that synthetic biology, “as it is the case in any other engineering branch, would be expected to be fully rationally based, straightforward, and predictable”. Therefore, I would expect that genetically modifying a microorganism should be a reachable task, consisting of well-defined steps. Anecdotally, during a class of my second year of master in Systems and Synthetic Biology, a plant biologist even compared bacteria to “bags of enzymes”. In his opinion, the study of these unicellular organisms with no organelle was too simple to be of interest compared to that of higher eukaryotes.

I do not wish to imply here that plants are not complex and not worthy of interest, my point is to underline that we still cannot predict/control/engineer our “bags of enzymes” as we plan. The rational choice to opt for synthetic regulatory elements, as it was the case for promoters during the refactoring of congocidine gene cluster (see chapter III), is more often than not a choice of necessity, brought by our little understanding of the complex native regulation. Even synthetic genetic elements, which are meant to be well-defined and controlled, are often influenced by genetic context. Promoters, for instance, are defined by their strength of expression, but the protein production depends not only on the promoter, but also on the ribosome binding site, the gene coding sequence, the terminator, and even on the host strain (Bai et al., 2015; Horbal et al., 2018; Vilanova et al., 2015; Yeung et al., 2017). If any of those components changes, the expected results may not be transferable any more.

Unexplained failures usually do not get published, at most they can be briefly mentioned in an article reporting successful experiments. For example, concerning daptomycin engineering, Baltz (2014) reports that “in early studies at Cubist on combinatorial biosynthesis, attempts were made to transplant A domains without success (unpublished data)”. Conversely, some successes can come as surprises, though they are assumed as straightforward later on. For instance, in the 2018 Applied Natural Products Symposium taking place in Palaiseau, Professor Helge Bode made a presentation on “Peptide natural products made by microbes and men”. He shared with us a suggestion from one of his students to place the fusion site to exchange NRPSs inside a condensation domain. He admitted being highly skeptical, but still let the student proceed with the experiment. One year later, the concept of XUC unit, explained in the introduction (See Introduction 3.3.6.) was published (Bozhüyük et al., 2019). It is interesting to note that no doubt concerning the possible success of this concept is expressed in this paper.

Delays and failures are intrinsic to research in synthetic biology, although it is rarely stated in research articles. It is quite a paradox that synthetic biology is described as rational designing, or compared to efficient engineering, when we still function mainly with trials and errors (Porcar, 2019). Still, even if we do not control the systems as we claim, some experiments are remarkably successful. It was far from being obvious that substantial production of congocidine would be observed with the refactored biosynthetic pathway (see chapter III). Similarly, the use of a fusion point inside the condensation domain worked especially well (Bozhüyük et al., 2019). Do we really have to claim a complete control of the biological systems, whereas we would still be able to make incredible discoveries in the field of synthetic biology while accepting that we are fumbling in the mist?

# References

- Aigle, B., Lautru, S., Spitteller, D., Dickschat, J.S., Challis, G.L., Leblond, P., and Pernodet, J.-L. (2014). Genome mining of *Streptomyces ambofaciens*. *J. Ind. Microbiol. Biotechnol.* *41*, 251–263.
- Al-Mestarihi, A.H., Garzan, A., Kim, J.M., and Garneau-Tsodikova, S. (2015). Enzymatic evidence for a revised congocidine biosynthetic pathway. *ChemBiochem Eur. J. Chem. Biol.* *16*, 1307–1313.
- Arcamone, F., Penco, S., Orezzi, P., Nicoletta, V., and Pirelli, A. (1964). Structure and synthesis of distamycin A. *Nature* *203*, 1064–1065.
- Arcamone, F., Cassinelli, G., Fantini, G., Grein, A., Orezzi, P., Pol, C., and Spalla, C. (2000). Adriamycin, 14-hydroxydaunomycin, a new antitumor antibiotic from *S. penicetius* var. *caesius*. Reprinted from *Biotechnology and Bioengineering*, Vol. XI, Issue 6, Pages 1101-1110 (1969). *Biotechnol. Bioeng.* *67*, 704–713.
- Arnison, P.G., Bibb, M.J., Bierbaum, G., Bowers, A.A., Bugni, T.S., Bulaj, G., Camarero, J.A., Campopiano, D.J., Challis, G.L., Clardy, J., et al. (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* *30*, 108–160.
- Asai, A., Sakai, Y., Ogawa, H., Yamashita, Y., Kakita, S., Ochiai, K., Ashizawa, T., Mihara, A., Mizukami, T., and Nakano, H. (2000). Pyrromycin A and B, novel antitumor antibiotics containing pyrrole-amide repeating unit, produced by *Streptomyces* sp. *J. Antibiot. (Tokyo)* *53*, 66–69.
- Bai, C., Zhang, Y., Zhao, X., Hu, Y., Xiang, S., Miao, J., Lou, C., and Zhang, L. (2015). Exploiting a precise design of universal synthetic modular regulatory elements to unlock the microbial natural products in *Streptomyces*. *Proc. Natl. Acad. Sci. U. S. A.* *112*, 12181–12186.
- Baltz, R.H. (2010). *Streptomyces* and *Saccharopolyspora* hosts for heterologous expression of secondary metabolite gene clusters. *J. Ind. Microbiol. Biotechnol.* *37*, 759–772.
- Baltz, R.H. (2011). Function of MbtH homologs in nonribosomal peptide biosynthesis and applications in secondary metabolite discovery. *J. Ind. Microbiol. Biotechnol.* *38*, 1747–1760.
- Baltz, R.H. (2014). Combinatorial biosynthesis of cyclic lipopeptide antibiotics: a model for synthetic biology to accelerate the evolution of secondary metabolite biosynthetic pathways. *ACS Synth. Biol.* *3*, 748–758.
- Baltz, R.H. (2016). Genetic manipulation of secondary metabolite biosynthesis for improved production in *Streptomyces* and other actinomycetes. *J. Ind. Microbiol. Biotechnol.* *43*, 343–370.
- Baltz, R.H. (2018). Synthetic biology, genome mining, and combinatorial biosynthesis of NRPS-derived antibiotics: a perspective. *J. Ind. Microbiol. Biotechnol.* *45*, 635–649.
- Barrett, M.P., Gemmell, C.G., and Suckling, C.J. (2013). Minor groove binders as anti-infective agents. *Pharmacol. Ther.* *139*, 12–23.
- Bauman, K.D., Li, J., Murata, K., Mantovani, S.M., Dahesh, S., Nizet, V., Luhavaya, H., and Moore, B.S. (2019). Refactoring the cryptic streptophenazine biosynthetic gene cluster unites phenazine, polyketide, and nonribosomal peptide biochemistry. *Cell Chem. Biol.* *26*, 724-736.e7.
- Beer, R., Herbst, K., Ignatiadis, N., Kats, I., Adlung, L., Meyer, H., Niopek, D., Christiansen, T., Georgi, F., Kurzawa, N., et al. (2014). Creating functional engineered variants of the single-module non-ribosomal peptide synthetase IndC by T domain exchange. *Mol. Biosyst.* *10*, 1709–1718.

## References

- Bérdy, J. (2012). Thoughts and facts about antibiotics: where we are now and where we are heading. *J. Antibiot. (Tokyo)* *65*, 385–395.
- Bhaduri, S., Ranjan, N., and Arya, D.P. (2018). An overview of recent advances in duplex DNA recognition by small molecules. *Beilstein J. Org. Chem.* *14*, 1051–1086.
- Bibb, M.J. (2005). Regulation of secondary metabolism in streptomycetes. *Curr. Opin. Microbiol.* *8*, 208–215.
- Bibb, M., and Hesketh, A. (2009). Chapter 4. Analyzing the regulation of antibiotic production in streptomycetes. *Methods Enzymol.* *458*, 93–116.
- Binz, T.M., Maffioli, S.I., Sosio, M., Donadio, S., and Müller, R. (2010). Insights into an unusual nonribosomal peptide synthetase biosynthesis: identification and characterization of the GE81112 biosynthetic gene cluster. *J. Biol. Chem.* *285*, 32710–32719.
- Blin, K., Wolf, T., Chevrette, M.G., Lu, X., Schwalen, C.J., Kautsar, S.A., Suarez Duran, H.G., de Los Santos, E.L.C., Kim, H.U., Nave, M., et al. (2017). antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* *45*, W36–W41.
- Bode, H.B., Bethe, B., Höfs, R., and Zeeck, A. (2002). Big effects from small changes: possible ways to explore nature's chemical diversity. *Chembiochem Eur. J. Chem. Biol.* *3*, 619–627.
- Bolhuis, A., and Aldrich-Wright, J.R. (2014). DNA as a target for antimicrobials. *Bioorganic Chem.* *55*, 51–59.
- Bozhüyük, K.A.J., Fleischhacker, F., Linck, A., Wesche, F., Tietze, A., Niesert, C.-P., and Bode, H.B. (2018). *De novo* design and engineering of non-ribosomal peptide synthetases. *Nat. Chem.* *10*, 275–281.
- Bozhüyük, K.A.J., Linck, A., Tietze, A., Kranz, J., Wesche, F., Nowak, S., Fleischhacker, F., Shi, Y.-N., Grün, P., and Bode, H.B. (2019). Modification and *de novo* design of non-ribosomal peptide synthetases using specific assembly points within condensation domains. *Nat. Chem.*
- Brannon, D.R., Fukuda, D.S., Mabe, J.A., Huber, F.M., and Whitney, J.G. (1972). Detection of a cephalosporin C acetyl esterase in the carbamate cephalosporin antibiotic-producing culture, *Streptomyces clavuligerus*. *Antimicrob. Agents Chemother.* *1*, 237–241.
- Brown, A.S., Calcott, M.J., Owen, J.G., and Ackerley, D.F. (2018). Structural, functional and evolutionary perspectives on effective re-engineering of non-ribosomal peptide synthetase assembly lines. *Nat. Prod. Rep.* *35*, 1210–1228.
- Bunet, R., Song, L., Mendes, M.V., Corre, C., Hotel, L., Rouhier, N., Framboisier, X., Leblond, P., Challis, G.L., and Aigle, B. (2011). Characterization and manipulation of the pathway-specific late regulator AlpW reveals *Streptomyces ambofaciens* as a new producer of kinamycins. *J. Bacteriol.* *193*, 1142–1153.
- Bunet, R., Riclea, R., Laureti, L., Hôtel, L., Paris, C., Girardet, J.-M., Spitteller, D., Dickschat, J.S., Leblond, P., and Aigle, B. (2014). A single Sfp-type phosphopantetheinyl transferase plays a major role in the biosynthesis of PKS and NRPS derived metabolites in *Streptomyces ambofaciens* ATCC23877. *PloS One* *9*, e87607.
- Burg, R.W., Miller, B.M., Baker, E.E., Birnbaum, J., Currie, S.A., Hartman, R., Kong, Y.L., Monaghan, R.L., Olson, G., Putter, I., et al. (1979). Avermectins, new family of potent anthelmintic agents: producing organism and fermentation. *Antimicrob. Agents Chemother.* *15*, 361–367.
- Butz, D., Schmiederer, T., Hadatsch, B., Wohlleben, W., Weber, T., and Süssmuth, R.D. (2008). Module extension of a non-ribosomal peptide synthetase of the glycopeptide antibiotic balhimycin produced by *Amycolatopsis balhimycina*. *Chembiochem Eur. J. Chem. Biol.* *9*, 1195–1200.

## References

- Cai, X., Nowak, S., Wesche, F., Bischoff, I., Kaiser, M., Fürst, R., and Bode, H.B. (2017). Entomopathogenic bacteria use multiple mechanisms for bioactive peptide library design. *Nat. Chem.* *9*, 379–386.
- Cai, X., Zhao, L., and Bode, H.B. (2019). Reprogramming promiscuous nonribosomal peptide synthetases for production of specific peptides. *Org. Lett.* *21*, 2116–2120.
- Calcott, M.J., and Ackerley, D.F. (2014). Genetic manipulation of non-ribosomal peptide synthetases to generate novel bioactive peptide products. *Biotechnol. Lett.* *36*, 2407–2416.
- Calcott, M.J., and Ackerley, D.F. (2015). Portability of the thiolation domain in recombinant pyoverdine non-ribosomal peptide synthetases. *BMC Microbiol.* *15*, 162.
- Calcott, M.J., Owen, J.G., Lamont, I.L., and Ackerley, D.F. (2014). Biosynthesis of novel pyoverdines by domain substitution in a nonribosomal peptide synthetase of *Pseudomonas aeruginosa*. *Appl. Environ. Microbiol.* *80*, 5723–5731.
- Casazza, A.M., Fioretti, A., Ghione, M., Soldati, M., and Verini, M.A. (1965). Distamycin A, a new antiviral antibiotic. *Antimicrob. Agents Chemother.* *5*, 593–598.
- Challis, G.L., Ravel, J., and Townsend, C.A. (2000). Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* *7*, 211–224.
- Chen, W.-H., Li, K., Guntaka, N.S., and Bruner, S.D. (2016). Interdomain and intermodule organization in epimerization domain containing nonribosomal peptide synthetases. *ACS Chem. Biol.* *11*, 2293–2303.
- Chen, Y., Krol, J., Sterkin, V., Fan, W., Yan, X., Huang, W., Cino, J., and Julien, C. (1999). New process control strategy used in a rapamycin fermentation. *Process Biochem.* *34*, 383–389.
- Chevrette, M.G., Aicheler, F., Kohlbacher, O., Currie, C.R., and Medema, M.H. (2017). SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinforma. Oxf. Engl.* *33*, 3202–3210.
- Chiocchini, C., Linne, U., and Stachelhaus, T. (2006). In vivo biocombinatorial synthesis of lipopeptides by COM domain-mediated reprogramming of the surfactin biosynthetic complex. *Chem. Biol.* *13*, 899–908.
- Cimermancic, P., Medema, M.H., Claesen, J., Kurita, K., Wieland Brown, L.C., Mavrommatis, K., Pati, A., Godfrey, P.A., Koehrsen, M., Clardy, J., et al. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* *158*, 412–421.
- Cosar, C., Ninet, L., Pinnert-Sindico, S., and Preud'homme, J. (1952). [Trypanocide action of an antibiotic produced by a Streptomyces]. *Comptes Rendus Hebd. Seances Acad. Sci.* *234*, 1498–1499.
- Cragg, G.M., and Newman, D.J. (2013). Natural products: a continuing source of novel drug leads. *Biochim. Biophys. Acta* *1830*, 3670–3695.
- Crüseman, M., Kohlhaas, C., and Piel, J. (2013). Evolution-guided engineering of nonribosomal peptide synthetase adenylation domains. *Chem. Sci.* *4*, 1041–1045.
- Czaplewski, L., Bax, R., Clokie, M., Dawson, M., Fairhead, H., Fischetti, V.A., Foster, S., Gilmore, B.F., Hancock, R.E.W., Harper, D., et al. (2016). Alternatives to antibiotics—a pipeline portfolio review. *Lancet Infect. Dis.* *16*, 239–251.
- Darken, M.A., Berenson, H., Shirk, R.J., and Sjolander, N.O. (1960). Production of tetracycline by *Streptomyces aureofaciens* in synthetic media. *Appl. Microbiol.* *8*, 46–51.
- Davies, J. (2006). Where have all the antibiotics gone? *Can. J. Infect. Dis. Med. Microbiol.* *17*, 287–290.

## References

- Davies, J., and Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* *MMBR* *74*, 417–433.
- Dehling, E., Volkmann, G., Matern, J.C.J., Dörner, W., Alfermann, J., Diecker, J., and Mootz, H.D. (2016). Mapping of the communication-mediating interface in nonribosomal peptide synthetases using a genetically encoded photocrosslinker supports an upside-down helix-hand motif. *J. Mol. Biol.* *428*, 4345–4360.
- Demain, A.L. (2009). Antibiotics: natural products essential to human health. *Med. Res. Rev.* *29*, 821–842.
- Doekel, S., Coëffet-Le Gal, M.-F., Gu, J.-Q., Chu, M., Baltz, R.H., and Brian, P. (2008). Non-ribosomal peptide synthetase module fusions to produce derivatives of daptomycin in *Streptomyces roseosporus*. *Microbiol. Read. Engl.* *154*, 2872–2880.
- Drake, E.J., Miller, B.R., Shi, C., Tarrasch, J.T., Sundlov, J.A., Allen, C.L., Skiniotis, G., Aldrich, C.C., and Gulick, A.M. (2016). Structures of two distinct conformations of holo-non-ribosomal peptide synthetases. *Nature* *529*, 235–238.
- Dulmage, H.T. (1953). The production of neomycin by *Streptomyces fradiae* in synthetic media. *Appl. Microbiol.* *1*, 103–106.
- Ehrlich, J., Bartz, Q.R., Smith, R.M., Joslyn, D.A., and Burkholder, P.R. (1947). Chloromycetin, a new antibiotic from a soil actinomycete. *Science* *106*, 417.
- Engler, C., and Marillonnet, S. (2014). Golden Gate cloning. *Methods Mol. Biol.* Clifton NJ *1116*, 119–131.
- Eppelmann, K., Stachelhaus, T., and Marahiel, M.A. (2002). Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry* *41*, 9718–9726.
- Esquiliñ-Lebrón, K.J., Boynton, T.O., Shimkets, L.J., and Thomas, M.G. (2018). An orphan mbth-like protein interacts with multiple nonribosomal peptide synthetases in *Mycococcus xanthus* DK1622. *J. Bacteriol.* *200*, e00346-18.
- Evans, B.S., Chen, Y., Metcalf, W.W., Zhao, H., and Kelleher, N.L. (2011). Directed evolution of the nonribosomal peptide synthetase AdmK generates new andrimid derivatives *in vivo*. *Chem. Biol.* *18*, 601–607.
- Fair, R.J., and Tor, Y. (2014). Antibiotics and bacterial resistance in the 21st century. *Perspect. Med. Chem.* *6*, 25–64.
- Farag, S., Bleich, R.M., Shank, E.A., Isayev, O., Bowers, A.A., and Tropsha, A. (2019). Inter-modular linkers play a crucial role in governing the biosynthesis of non-ribosomal peptides. *Bioinformatics* *35*, 3584–3591.
- Feling, R.H., Buchanan, G.O., Mincer, T.J., Kauffman, C.A., Jensen, P.R., and Fenical, W. (2003). Salinosporamide A: a highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus *Salinospora*. *Angew. Chem. Int. Ed Engl.* *42*, 355–357.
- Ferri, M., Ranucci, E., Romagnoli, P., and Giaccone, V. (2017). Antimicrobial resistance: A global emerging threat to public health systems. *Crit. Rev. Food Sci. Nutr.* *57*, 2857–2876.
- Finlay, A.C., Hochstein, F.A., Sobin, B.A., and Murphy, F.X. (1951). Netropsin, a new antibiotic produced by a *Streptomyces*. *J. Am. Chem. Soc.* *73*, 341–343.
- Fischbach, M.A., Lai, J.R., Roche, E.D., Walsh, C.T., and Liu, D.R. (2007). Directed evolution can rapidly improve the activity of chimeric assembly-line enzymes. *Proc. Natl. Acad. Sci. U. S. A.* *104*, 11951–11956.

## References

- Fu, J., Bian, X., Hu, S., Wang, H., Huang, F., Seibert, P.M., Plaza, A., Xia, L., Müller, R., Stewart, A.F., et al. (2012). Full-length RecE enhances linear-linear homologous recombination and facilitates direct cloning for bioprospecting. *Nat. Biotechnol.* *30*, 440–446.
- Gao, L., Guo, J., Fan, Y., Ma, Z., Lu, Z., Zhang, C., Zhao, H., and Bie, X. (2018). Module and individual domain deletions of NRPS to produce plipastatin derivatives in *Bacillus subtilis*. *Microb. Cell Factories* *17*, 84.
- Gao, Y., Honzatko, R.B., and Peters, R.J. (2012). Terpenoid synthase structures: a so far incomplete view of complex catalysis. *Nat. Prod. Rep.* *29*, 1153–1175.
- Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* *6*, 343–345.
- Gomez-Escribano, J.P., and Bibb, M.J. (2011). Engineering *Streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. *Microb. Biotechnol.* *4*, 207–215.
- González, A., Jiménez, A., Vázquez, D., Davies, J.E., and Schindler, D. (1978). Studies on the mode of action of hygromycin B, an inhibitor of translocation in eukaryotes. *Biochim. Biophys. Acta* *521*, 459–469.
- Goodrich, A.C., Meyers, D.J., and Frueh, D.P. (2017). Molecular impact of covalent modifications on nonribosomal peptide synthetase carrier protein communication. *J. Biol. Chem.* *292*, 10002–10013.
- Goodsell, D.S., Kopka, M.L., and Dickerson, R.E. (1995). Refinement of netropsin bound to DNA: bias and feedback in electron density map interpretation. *Biochemistry* *34*, 4983–4993.
- Goss, R.J.M., Shankar, S., and Fayad, A.A. (2012). The generation of “unnatural” products: synthetic biology meets synthetic chemistry. *Nat. Prod. Rep.* *29*, 870–889.
- Gulick, A.M. (2009). Conformational dynamics in the Acyl-CoA synthetases, adenylation domains of non-ribosomal peptide synthetases, and firefly luciferase. *ACS Chem. Biol.* *4*, 811–827.
- Gulick, A.M. (2016). Structural insight into the necessary conformational changes of modular nonribosomal peptide synthetases. *Curr. Opin. Chem. Biol.* *35*, 89–96.
- Gust, B., Chandra, G., Jakimowicz, D., Yuqing, T., Bruton, C.J., and Chater, K.F. (2004). Lambda red-mediated genetic manipulation of antibiotic-producing *Streptomyces*. *Adv. Appl. Microbiol.* *54*, 107–128.
- Guzmán-Trampe, S., Ceapa, C.D., Manzo-Ruiz, M., and Sánchez, S. (2017). Synthetic biology era: Improving antibiotic’s world. *Biochem. Pharmacol.* *134*, 99–113.
- Hacker, C., Cai, X., Kegler, C., Zhao, L., Weickmann, A.K., Wurm, J.P., Bode, H.B., and Wöhnert, J. (2018). Structure-based redesign of docking domain interactions modulates the product spectrum of a rhabdopeptide-synthesizing NRPS. *Nat. Commun.* *9*, 4366.
- Hahn, M., and Stachelhaus, T. (2004). Selective interaction between nonribosomal peptide synthetases is facilitated by short communication-mediating domains. *Proc. Natl. Acad. Sci. U. S. A.* *101*, 15585–15590.
- Hahn, M., and Stachelhaus, T. (2006). Harnessing the potential of communication-mediating domains for the biocombinatorial synthesis of nonribosomal peptides. *Proc. Natl. Acad. Sci. U. S. A.* *103*, 275–280.
- Hao, C., Huang, S., Deng, Z., Zhao, C., and Yu, Y. (2014). Mining of the pyrrolamide antibiotics analogs in *Streptomyces netropsis* reveals the amidohydrolase-dependent “iterative strategy” underlying the pyrrole polymerization. *PloS One* *9*, e99077.
- Harden, B.J., and Frueh, D.P. (2017). Molecular cross-talk between nonribosomal peptide synthetase carrier proteins and unstructured linker regions. *Chembiochem Eur. J. Chem. Biol.* *18*, 629–632.

## References

- Harvey, A.L., Edrada-Ebel, R., and Quinn, R.J. (2015). The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* *14*, 111–129.
- Heide, L. (2009). Genetic engineering of antibiotic biosynthesis for the generation of new aminocoumarins. *Biotechnol. Adv.* *27*, 1006–1014.
- Herbst, D.A., Boll, B., Zocher, G., Stehle, T., and Heide, L. (2013). Structural basis of the interaction of MbtH-like proteins, putative regulators of nonribosomal peptide biosynthesis, with adenylating enzymes. *J. Biol. Chem.* *288*, 1991–2003.
- Ho, Y.T.C., Leng, D.J., Ghiringhelli, F., Wilkening, I., Bushell, D.P., Kostner, O., Riva, E., Havemann, J., Passarella, D., and Tosin, M. (2017). Novel chemical probes for the investigation of nonribosomal peptide assembly. *Chem. Commun. Camb. Engl.* *53*, 7088–7091.
- Horbal, L., Siegl, T., and Luzhetskyy, A. (2018). A set of synthetic versatile genetic control elements for the efficient expression of genes in Actinobacteria. *Sci. Rep.* *8*, 491.
- Hug, J.J., Bader, C.D., Remškar, M., Cirnski, K., and Müller, R. (2018). Concepts and methods to access novel antibiotics from actinomycetes. *Antibiot. Basel Switz.* *7*.
- Hur, G.H., Vickery, C.R., and Burkart, M.D. (2012). Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. *Nat. Prod. Rep.* *29*, 1074–1098.
- Izoré, T., and Cryle, M.J. (2018). The many faces and important roles of protein-protein interactions during non-ribosomal peptide synthesis. *Nat. Prod. Rep.* *35*, 1120–1139.
- Jaremko, M.J., Lee, D.J., Patel, A., Winslow, V., Opella, S.J., McCammon, J.A., and Burkart, M.D. (2017). Manipulating protein-protein interactions in nonribosomal peptide synthetase type II peptidyl carrier proteins. *Biochemistry* *56*, 5269–5273.
- Jenke-Kodama, H., and Dittmann, E. (2009). Bioinformatic perspectives on NRPS/PKS megasynthases: advances and challenges. *Nat. Prod. Rep.* *26*, 874–883.
- Jiang, W., Zhao, X., Gabrieli, T., Lou, C., Ebenstein, Y., and Zhu, T.F. (2015). Cas9-Assisted Targeting of CHromosome segments CATCH enables one-step targeted cloning of large gene clusters. *Nat. Commun.* *6*.
- Juguet, M., Lautru, S., Francou, F.-X., Nezbedová, S., Leblond, P., Gondry, M., and Pernodet, J.-L. (2009). An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congocidine in *Streptomyces ambofaciens*. *Chem. Biol.* *16*, 421–431.
- Julia, M., and Preau-Joseph, N. (1967). [Amidines and guanidines related to congocidine. I. Structure of congocidine]. *Bull. Soc. Chim. Fr.* *11*, 4348–4356.
- Kakule, T.B., Lin, Z., and Schmidt, E.W. (2014). Combinatorialization of fungal polyketide synthase-peptide synthetase hybrid proteins. *J. Am. Chem. Soc.* *136*, 17882–17890.
- Kallifidas, D., Jiang, G., Ding, Y., and Luesch, H. (2018). Rational engineering of *Streptomyces albus* J1074 for the overexpression of secondary metabolite gene clusters. *Microb. Cell Factories* *17*, 25.
- Keller, U., and Schauwecker, F. (2003). Combinatorial biosynthesis of non-ribosomal peptides. *Comb. Chem. High Throughput Screen.* *6*, 527–540.
- Kikuchi, M., Kumagai, K., Ishida, N., Ito, Y., Yamaguchi, T., Furumai, T., and Okuda, T. (1965). Isolation, purification, and properties of kikumycins A and B. *J. Antibiot. (Tokyo)* *18*, 243–250.

## References

- Kim, E., Moore, B.S., and Yoon, Y.J. (2015). Reinvigorating natural product combinatorial biosynthesis with synthetic biology. *Nat. Chem. Biol.* *11*, 649–659.
- Kittilä, T., Mollo, A., Charkoudian, L.K., and Cryle, M.J. (2016). New structural data reveal the motion of carrier proteins in nonribosomal peptide synthesis. *Angew. Chem. Int. Ed Engl.* *55*, 9834–9840.
- Knight, T. (2003). Idempotent vector design for standard assembly of biobricks.
- de Kok, S., Stanton, L.H., Slaby, T., Durot, M., Holmes, V.F., Patel, K.G., Platt, D., Shapland, E.B., Serber, Z., Dean, J., et al. (2014). Rapid and reliable DNA assembly via ligase cycling reaction. *ACS Synth. Biol.* *3*, 97–106.
- Komatsu, M., Uchiyama, T., Omura, S., Cane, D.E., and Ikeda, H. (2010). Genome-minimized *Streptomyces* host for the heterologous expression of secondary metabolism. *Proc. Natl. Acad. Sci. U. S. A.* *107*, 2646–2651.
- Kopka, M.L., Yoon, C., Goodsell, D., Pjura, P., and Dickerson, R.E. (1985). The molecular origin of DNA-drug specificity in netropsin and distamycin. *Proc. Natl. Acad. Sci. U. S. A.* *82*, 1376–1380.
- Kries, H., Niquille, D.L., and Hilvert, D. (2015). A subdomain swap strategy for reengineering nonribosomal peptides. *Chem. Biol.* *22*, 640–648.
- Kudo, F., Miyanaga, A., and Eguchi, T. (2019). Structural basis of the nonribosomal codes for nonproteinogenic amino acid selective adenylation enzymes in the biosynthesis of natural products. *J. Ind. Microbiol. Biotechnol.* *46*, 515–536.
- Laureti, L., Song, L., Huang, S., Corre, C., Leblond, P., Challis, G.L., and Aigle, B. (2011). Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in *Streptomyces ambofaciens*. *Proc. Natl. Acad. Sci. U. S. A.* *108*, 6258–6263.
- Lautru, S., and Challis, G.L. (2004). Substrate recognition by nonribosomal peptide synthetase multi-enzymes. *Microbiol. Read. Engl.* *150*, 1629–1636.
- Lautru, S., Gondry, M., Genet, R., and Pernodet, J.L. (2002). The albonoursin gene cluster of *S. noursei* biosynthesis of diketopiperazine metabolites independent of nonribosomal peptide synthetases. *Chem. Biol.* *9*, 1355–1364.
- Lautru, S., Deeth, R.J., Bailey, L.M., and Challis, G.L. (2005). Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nat. Chem. Biol.* *1*, 265–269.
- Lautru, S., Oves-Costales, D., Pernodet, J.-L., and Challis, G.L. (2007). MbtH-like protein-mediated cross-talk between non-ribosomal peptide antibiotic and siderophore biosynthetic pathways in *Streptomyces coelicolor* M145. *Microbiol. Read. Engl.* *153*, 1405–1412.
- Lautru, S., Song, L., Demange, L., Lombès, T., Galons, H., Challis, G.L., and Pernodet, J.-L. (2012). A sweet origin for the key congoicidine precursor 4-acetamidopyrrole-2-carboxylate. *Angew. Chem. Int. Ed Engl.* *51*, 7454–7458.
- Lewis, K. (2013). Platforms for antibiotic discovery. *Nat. Rev. Drug Discov.* *12*, 371–387.
- Li, M.Z., and Elledge, S.J. (2012). SLIC: a method for sequence- and ligation-independent cloning. *Methods Mol. Biol. Clifton NJ* *852*, 51–59.
- Linne, U., Doekel, S., and Marahiel, M.A. (2001). Portability of epimerization domain and role of peptidyl carrier protein on epimerization activity in nonribosomal peptide synthetases. *Biochemistry* *40*, 15824–15834.

## References

- Liu, G., Chater, K.F., Chandra, G., Niu, G., and Tan, H. (2013). Molecular regulation of antibiotic biosynthesis in *Streptomyces*. *Microbiol. Mol. Biol. Rev.* *MMBR* *77*, 112–143.
- Liu, H., Gao, L., Han, J., Ma, Z., Lu, Z., Dai, C., Zhang, C., and Bie, X. (2016). Biocombinatorial synthesis of novel lipopeptides by COM domain-mediated reprogramming of the plipastatin NRPS complex. *Front. Microbiol.* *7*, 1801.
- Lott, J.S., and Lee, T.V. (2017). Revealing the inter-module interactions of multi-modular nonribosomal peptide synthetases. *Struct. Lond. Engl.* *1993* *25*, 693–695.
- Luo, Y., Huang, H., Liang, J., Wang, M., Lu, L., Shao, Z., Cobb, R.E., and Zhao, H. (2013). Activation and characterization of a cryptic polycyclic tetramate macrolactam biosynthetic gene cluster. *Nat. Commun.* *4*, 2894.
- Luo, Y., Enghiad, B., and Zhao, H. (2016). New tools for reconstruction and heterologous expression of natural product biosynthetic gene clusters. *Nat. Prod. Rep.* *33*, 174–182.
- Lyddiard, D., Jones, G.L., and Greatrex, B.W. (2016). Keeping it simple: lessons from the golden era of antibiotic discovery. *FEMS Microbiol. Lett.* *363*.
- Marahiel, M.A. (2016). A structural model for multimodular NRPS assembly lines. *Nat. Prod. Rep.* *33*, 136–140.
- Martin, R., Sterner, O., Alvarez, M.A., de Clercq, E., Bailey, J.E., and Minas, W. (2001). Collinone, a new recombinant angular polyketide antibiotic made by an engineered *Streptomyces* strain. *J. Antibiot. (Tokyo)* *54*, 239–249.
- Masand, M., Sivakala, K.K., Menghani, E., Thinesh, T., Anandham, R., Sharma, G., Sivakumar, N., Jebakumar, S.R.D., and Jose, P.A. (2018). Biosynthetic potential of bioactive streptomycetes isolated from arid region of the Thar desert, Rajasthan (India). *Front. Microbiol.* *9*, 687.
- Matteoli, B., Bernardini, S., Iuliano, R., Parenti, S., Freer, G., Broccolo, F., Baggiani, A., Subissi, A., Arcamone, F., and Ceccherini-Nelli, L. (2008). In vitro antiviral activity of distamycin A against clinical isolates of herpes simplex virus 1 and 2 from transplanted patients. *Intervirology* *51*, 166–172.
- McErlean, M., Overbay, J., and Van Lanen, S. (2019). Refining and expanding nonribosomal peptide synthetase function and mechanism. *J. Ind. Microbiol. Biotechnol.* *46*, 493–513.
- Mchenney, M.A., Hosted, T.J., Dehoff, B.S., Rosteck, P.R., and Baltz, R.H. (1998). Molecular cloning and physical mapping of the daptomycin gene cluster from *Streptomyces roseosporus*. *J. Bacteriol.* *180*, 143–151.
- Medema, M.H., Breitling, R., Bovenberg, R., and Takano, E. (2011). Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. *Nat. Rev. Microbiol.* *9*, 131–137.
- Meyer, S., Kehr, J.-C., Mainz, A., Dehm, D., Petras, D., Süßmuth, R.D., and Dittmann, E. (2016). Biochemical dissection of the natural diversification of microcystin provides lessons for synthetic biology of NRPS. *Cell Chem. Biol.* *23*, 462–471.
- Miao, V., Coëffet-Le Gal, M.-F., Nguyen, K., Brian, P., Penn, J., Whiting, A., Steele, J., Kau, D., Martin, S., Ford, R., et al. (2006). Genetic engineering in *Streptomyces roseosporus* to produce hybrid lipopeptide antibiotics. *Chem. Biol.* *13*, 269–276.
- Miller, B.R., Sundlov, J.A., Drake, E.J., Makin, T.A., and Gulick, A.M. (2014). Analysis of the linker region joining the adenylation and carrier protein domains of the modular non-ribosomal peptide synthetases. *Proteins* *82*, 2691–2702.

## References

- Miller, B.R., Drake, E.J., Shi, C., Aldrich, C.C., and Gulick, A.M. (2016). Structures of a nonribosomal peptide synthetase module bound to mbth-like proteins support a highly dynamic domain architecture. *J. Biol. Chem.* *291*, 22559–22571.
- Mootz, H.D., Schwarzer, D., and Marahiel, M.A. (2000). Construction of hybrid peptide synthetases by module and domain fusions. *Proc. Natl. Acad. Sci. U. S. A.* *97*, 5848–5853.
- Mori, S., Green, K.D., Choi, R., Buchko, G.W., Fried, M.G., and Garneau-Tsodikova, S. (2018). Using MbtH-like proteins to alter the substrate profile of a nonribosomal peptide adenylation enzyme. *Chembiochem Eur. J. Chem. Biol.* *19*, 2186–2194.
- Neidle, S. (2001). DNA minor-groove recognition by small molecules. *Nat. Prod. Rep.* *18*, 291–309.
- Newman, D.J., and Cragg, G.M. (2016). Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* *79*, 629–661.
- Nguyen, K.T., Ritz, D., Gu, J.-Q., Alexander, D., Chu, M., Miao, V., Brian, P., and Baltz, R.H. (2006). Combinatorial biosynthesis of novel antibiotics related to daptomycin. *Proc. Natl. Acad. Sci. U. S. A.* *103*, 17462–17467.
- Niquille, D.L., Hansen, D.A., Mori, T., Fercher, D., Kries, H., and Hilvert, D. (2018). Nonribosomal biosynthesis of backbone-modified peptides. *Nat. Chem.* *10*, 282–287.
- Nothias, L.-F., Knight, R., and Dorrestein, P.C. (2016). Antibiotic discovery is a walk in the park. *Proc. Natl. Acad. Sci. U. S. A.* *113*, 14477–14479.
- Omura, S., and Crump, A. (2004). The life and times of ivermectin - a success story. *Nat. Rev. Microbiol.* *2*, 984–989.
- Omura, S., Iwai, Y., Takahashi, Y., Sadakane, N., Nakagawa, A., Oiwa, H., Hasegawa, Y., and Ikai, T. (1979). Herbimycin, a new antibiotic produced by a strain of *Streptomyces*. *J. Antibiot. (Tokyo)* *32*, 255–261.
- Omura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M., Takahashi, Y., Horikawa, H., Nakazawa, H., Osonoe, T., et al. (2001). Genome sequence of an industrial microorganism *Streptomyces avermitilis*: Deducing the ability of producing secondary metabolites. *Proc. Natl. Acad. Sci. U. S. A.* *98*, 12215–12220.
- Ongley, S.E., Bian, X., Neilan, B.A., and Müller, R. (2013). Recent advances in the heterologous expression of microbial natural product biosynthetic pathways. *Nat. Prod. Rep.* *30*, 1121–1138.
- Osswald, C., Zipf, G., Schmidt, G., Maier, J., Bernauer, H.S., Müller, R., and Wenzel, S.C. (2014). Modular construction of a functional artificial epothilone polyketide pathway. *ACS Synth. Biol.* *3*, 759–772.
- Owen, J.G., Calcott, M.J., Robins, K.J., and Ackerley, D.F. (2016). Generating functional recombinant NRPS enzymes in the laboratory setting via peptidyl carrier protein engineering. *Cell Chem. Biol.* *23*, 1395–1406.
- Perlova, O., Fu, J., Kuhlmann, S., Krug, D., Stewart, A.F., Zhang, Y., and Müller, R. (2006). Reconstitution of the myxothiazol biosynthetic gene cluster by Red/ET recombination and heterologous expression in *Mycrococcus xanthus*. *Appl. Environ. Microbiol.* *72*, 7485–7494.
- Pettit, R.K. (2011). Culturability and secondary metabolite diversity of extreme microbes: expanding contribution of deep sea and deep-sea vent microbes to natural product discovery. *Mar. Biotechnol. N. Y.* *N 13*, 1–11.
- Pickens, L.B., Tang, Y., and Chooi, Y.-H. (2011). Metabolic engineering for the production of natural products. *Annu. Rev. Chem. Biomol. Eng.* *2*, 211–236.

## References

- Porcar, M. (2019). The hidden charm of life. Life Basel Switz. 9.
- Probst, G.W., Hoehn, M.M., and Woods, B.L. (1965). Anthelvincins, new antibiotics with anthelmintic properties. Antimicrob. Agents Chemother. 5, 789–795.
- Procópio, R.E. de L., Silva, I.R. da, Martins, M.K., Azevedo, J.L. de, and Araújo, J.M. de (2012). Antibiotics produced by *Streptomyces*. Braz. J. Infect. Dis. Off. Publ. Braz. Soc. Infect. Dis. 16, 466–471.
- Rahman, A., O’Sullivan, P., and Rozas, I. (2019). Recent developments in compounds acting in the DNA minor groove. MedChemComm 10, 26–40.
- Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W., and Huson, D.H. (2005). Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). Nucleic Acids Res. 33, 5799–5808.
- Reimer, J.M., Aloise, M.N., Harrison, P.M., and Schmeing, T.M. (2016). Synthetic cycle of the initiation module of a formylating nonribosomal peptide synthetase. Nature 529, 239–242.
- Reimer, J.M., Haque, A.S., Tarry, M.J., and Schmeing, T.M. (2018). Piecing together nonribosomal peptide synthesis. Curr. Opin. Struct. Biol. 49, 104–113.
- Röttig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C., and Kohlbacher, O. (2011). NRSPredictor2--a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res. 39, W362–367.
- Rutledge, P.J., and Challis, G.L. (2015). Discovery of microbial natural products by activation of silent biosynthetic gene clusters. Nat. Rev. Microbiol. 13, 509–523.
- Samel, S.A., Schoenafinger, G., Knappe, T.A., Marahiel, M.A., and Essen, L.-O. (2007). Structural and functional insights into a peptide bond-forming bidomain from a nonribosomal peptide synthetase. Struct. Lond. Engl. 1993 15, 781–792.
- Sands, B., and Brent, R. (2016). Overview of post Cohen-Boyer methods for single segment cloning and for multisegment DNA assembly. Curr. Protoc. Mol. Biol. 113, 3.26.1–3.26.20.
- Schatz, A., and Waksman, S.A. (1944). Effect of streptomycin and other antibiotic substances upon *Mycobacterium tuberculosis* and related organisms. Proc. Soc. Exp. Biol. Med. 57, 244–248.
- Schmeing, T.M. (2016). Visualizing a natural antibiotic nanofactory. Clin. Investig. Med. Med. Clin. Exp. 39, E220–E226.
- Schneider, A., Stachelhaus, T., and Marahiel, M.A. (1998). Targeted alteration of the substrate specificity of peptide synthetases by rational module swapping. Mol. Gen. Genet. MGG 257, 308–318.
- Schomer, R.A., and Thomas, M.G. (2017). Characterization of the functional variance in MbtH-like protein interactions with a nonribosomal peptide synthetase. Biochemistry.
- Schomer, R.A., Park, H., Barkei, J.J., and Thomas, M.G. (2018). Alanine scanning of YbdZ, an MbtH-like protein, reveals essential residues for functional interactions with its nonribosomal peptide synthetase partner EntF. Biochemistry 57, 4125–4134.
- Shao, Z., Zhao, H., and Zhao, H. (2009). DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. Nucleic Acids Res. 37, e16.
- Shao, Z., Rao, G., Li, C., Abil, Z., Luo, Y., and Zhao, H. (2013). Refactoring the silent spectinabilin gene cluster using a plug-and-play scaffold. ACS Synth. Biol. 2, 662–669.

## References

- Shen, B., Du, L., Sanchez, C., Edwards, D.J., Chen, M., and Murrell, J.M. (2001). The biosynthetic gene cluster for the anticancer drug bleomycin from *Streptomyces verticillus* ATCC15003 as a model for hybrid peptide-polyketide natural product biosynthesis. *J. Ind. Microbiol. Biotechnol.* *27*, 378–385.
- Siegl, T., Tokovenko, B., Myronovskiy, M., and Luzhetskyy, A. (2013). Design, construction and characterisation of a synthetic promoter library for fine-tuned gene expression in actinomycetes. *Metab. Eng.* *19*, 98–106.
- Smanski, M.J., Bhatia, S., Zhao, D., Park, Y., B A Woodruff, L., Giannoukos, G., Ciulla, D., Busby, M., Calderon, J., Nicol, R., et al. (2014). Functional optimization of gene clusters by combinatorial design and assembly. *Nat. Biotechnol.* *32*, 1241–1249.
- Smanski, M.J., Zhou, H., Claesen, J., Shen, B., Fischbach, M.A., and Voigt, C.A. (2016). Synthetic biology to access and expand nature’s chemical diversity. *Nat. Rev. Microbiol.* *14*, 135–149.
- Stachelhaus, T., Schneider, A., and Marahiel, M.A. (1995). Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. *Science* *269*, 69–72.
- Stachelhaus, T., Mootz, H.D., and Marahiel, M.A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* *6*, 493–505.
- Strieker, M., Tanović, A., and Marahiel, M.A. (2010). Nonribosomal peptide synthetases: structures and dynamics. *Curr. Opin. Struct. Biol.* *20*, 234–240.
- Stumpp, T., Himbert, S., and Altenbuchner, J. (2005). Cloning of the netropsin resistance genes from *Streptomyces flavopersicus* NRRL 2820. *J. Basic Microbiol.* *45*, 355–362.
- Subramani, R., and Aalbersberg, W. (2012). Marine actinomycetes: an ongoing source of novel bioactive metabolites. *Microbiol. Res.* *167*, 571–580.
- Sun, H., Liu, Z., Zhao, H., and Ang, E.L. (2015). Recent advances in combinatorial biosynthesis for drug discovery. *Drug Des. Devel. Ther.* *9*, 823–833.
- Sundlov, J.A., Shi, C., Wilson, D.J., Aldrich, C.C., and Gulick, A.M. (2012). Structural and functional investigation of the intermolecular interaction between NRPS adenylation and carrier protein domains. *Chem. Biol.* *19*, 188–198.
- Süssmuth, R.D., and Mainz, A. (2017). Nonribosomal peptide synthesis—principles and prospects. *Angew. Chem. Int. Ed Engl.* *56*, 3770–3821.
- Symmank, H., Saenger, W., and Bernhard, F. (1999). Analysis of engineered multifunctional peptide synthetases. Enzymatic characterization of surfactin synthetase domains in hybrid bimodular systems. *J. Biol. Chem.* *274*, 21581–21588.
- Takaishi, T., Sugawara, Y., and Suzuki, M. (1972). Structure of Kikumycin A and B. *Tetrahedron Lett.* *13*, 1873–1876.
- Takizawa, M., Tsubotani, S., Tanida, S., Harada, S., and Hasegawa, T. (1987). A new pyrrole-amidine antibiotic TAN-868 A. *J. Antibiot. (Tokyo)* *40*, 1220–1230.
- Tanovic, A., Samel, S.A., Essen, L.-O., and Marahiel, M.A. (2008). Crystal structure of the termination module of a nonribosomal peptide synthetase. *Science* *321*, 659–663.
- Tarry, M.J., Haque, A.S., Bui, K.H., and Schmeing, T.M. (2017). X-Ray Crystallography and electron microscopy of cross- and multi-module nonribosomal peptide synthetase proteins reveal a flexible architecture. *Struct. Lond. Engl.* *1993* *25*, 783-793.e4.

## References

- Temme, K., Zhao, D., and Voigt, C.A. (2012). Refactoring the nitrogen fixation gene cluster from *Klebsiella oxytoca*. *Proc. Natl. Acad. Sci. U. S. A.* *109*, 7085–7090.
- Thirlway, J., Lewis, R., Nunns, L., Al Nakeeb, M., Styles, M., Struck, A.-W., Smith, C.P., and Micklefield, J. (2012). Introduction of a non-natural amino acid into a nonribosomal peptide antibiotic by modification of adenylation domain specificity. *Angew. Chem. Int. Ed Engl.* *51*, 7181–7184.
- Tian, Y., Li, Y.-L., and Zhao, F.-C. (2017). Secondary metabolites from polar organisms. *Mar. Drugs* *15*.
- Tufar, P., Rahighi, S., Kraas, F.I., Kirchner, D.K., Löhr, F., Henrich, E., Köpke, J., Dikic, I., Güntert, P., Marahiel, M.A., et al. (2014). Crystal structure of a PCP/Sfp complex reveals the structural basis for carrier protein posttranslational modification. *Chem. Biol.* *21*, 552–562.
- Umezawa, H., Hamada, M., Suhara, Y., Hashimoto, T., and Ikekawa, T. (1965). Kasugamycin, a new antibiotic. *Antimicrob. Agents Chemother.* *5*, 753–757.
- Vilanova, C., Tanner, K., Dorado-Morales, P., Villaescusa, P., Chugani, D., Frías, A., Segredo, E., Molero, X., Fritschi, M., Morales, L., et al. (2015). Standards not that standard. *J. Biol. Eng.* *9*, 17.
- Vingadassalon, A., Lorieux, F., Juguet, M., Le Goff, G., Gerbaud, C., Pernodet, J.-L., and Lautru, S. (2015). Natural combinatorial biosynthesis involving two clusters for the synthesis of three pyrrolamides in *Streptomyces netropsis*. *ACS Chem. Biol.* *10*, 601–610.
- Wakaki, S., Marumo, H., Tomioka, K., Shimizu, G., Kato, E., Kamada, H., Kudo, S., and Fujimoto, Y. (1958). Isolation of new fractions of antitumor mitomycins. *Antibiot. Chemother. Northfield Ill* *8*, 228–240.
- Walsh, C.T., Garneau-Tsodikova, S., and Howard-Jones, A.R. (2006). Biological formation of pyrroles: nature's logic and enzymatic machinery. *Nat. Prod. Rep.* *23*, 517–531.
- Weissman, K.J. (2015). The structural biology of biosynthetic megaenzymes. *Nat. Chem. Biol.* *11*, 660–670.
- van Wezel, G.P., McKenzie, N.L., and Nodwell, J.R. (2009). Chapter 5. Applying the genetics of secondary metabolism in model actinomycetes to the discovery of new antibiotics. *Methods Enzymol.* *458*, 117–141.
- Wildfeuer, M.E. (1964). The biosynthesis of netropsin. Thesis, University of Delaware, Newark, Delaware.
- Winn, M., Fyans, J.K., Zhuo, Y., and Micklefield, J. (2016). Recent advances in engineering nonribosomal peptide assembly lines. *Nat. Prod. Rep.* *33*, 317–347.
- Yakimov, M.M., Giuliano, L., Timmis, K.N., and Golyshin, P.N. (2000). Recombinant acylheptapeptide lichenysin: high level of production by *Bacillus subtilis* cells. *J. Mol. Microbiol. Biotechnol.* *2*, 217–224.
- Yamanaka, K., Reynolds, K.A., Kersten, R.D., Ryan, K.S., Gonzalez, D.J., Nizet, V., Dorrestein, P.C., and Moore, B.S. (2014). Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc. Natl. Acad. Sci. U. S. A.* *111*, 1957–1962.
- Yeung, E., Dy, A.J., Martin, K.B., Ng, A.H., Del Vecchio, D., Beck, J.L., Collins, J.J., and Murray, R.M. (2017). Biophysical constraints arising from compositional context in synthetic gene networks. *Cell Syst.* *5*, 11–24.e12.
- Zarins-Tutt, J.S., Barberi, T.T., Gao, H., Mearns-Spragg, A., Zhang, L., Newman, D.J., and Goss, R.J.M. (2016). Prospecting for new bacterial metabolites: a glossary of approaches for inducing, activating and upregulating the biosynthesis of bacterial cryptic or silent natural products. *Nat. Prod. Rep.* *33*, 54–72.
- Zhou, Z., Lai, J.R., and Walsh, C.T. (2007). Directed evolution of aryl carrier proteins in the enterobactin synthetase. *Proc. Natl. Acad. Sci. U. S. A.* *104*, 11621–11626.

## References

Zhu, M., Wang, L., and He, J. (2019). Chemical diversification based on substrate promiscuity of a standalone adenylation domain in a reconstituted NRPS system. *ACS Chem. Biol.* *14*, 256–265.

Zimmer, C., Reinert, K.E., Luck, G., Wähnert, U., Löber, G., and Thrum, H. (1971). Interaction of the oligopeptide antibiotics netropsin and distamycin A with nucleic acids. *J. Mol. Biol.* *58*, 329–348.

Zotchev, S., Haugan, K., Sekurova, O., Sletta, H., Ellingsen, T.E., and Valla, S. (2000). Identification of a gene cluster for antibacterial polyketide-derived antibiotic biosynthesis in the nystatin producer *Streptomyces noursei* ATCC 11455. *Microbiol. Read. Engl.* *146 (Pt 3)*, 611–619.

(2007). All natural. *Nat. Chem. Biol.* *3*, 351.

## Webography

Cancer (2018) World Health Organisation,

Retrieved 13/07/2019 from <https://www.who.int/en/news-room/fact-sheets/detail/cancer>

Global fungal diseases (2018) Centers for disease control and prevention.

Retrieved 13/07/2019 from <https://www.cdc.gov/fungal/global/index.html>

Review on Antimicrobial Resistance, Jim O'Neill (2014) Antimicrobial Resistance: Tackling a Crisis for the Future Health and Wealth of Nations.

Retrieved 13/07/2019 from <https://amr-review.org/Publications.html>

Soil-transmitted helminth infections (2019) World Health Organisation,

Retrieved 13/07/2019 from <https://www.who.int/news-room/fact-sheets/detail/soil-transmitted-helminth-infections>

WHO publishes list of bacteria for which new antibiotics are urgently needed (2017) World Health Organisation,

Retrieved 13/07/2019 from <https://www.who.int/news-room/detail/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed>

# French summary of the thesis / Résumé de la thèse en Français

Introduction :

Les métabolites spécialisés sont des petites molécules produites en particulier par des microorganismes et des plantes, non nécessaires à la croissance de l'organisme en milieu riche. De nombreux médicaments ont été développés à partir de ces métabolites spécialisés, notamment des anticancéreux et des anti-infectieux (Newman and Cragg, 2016). Cependant aujourd'hui, les bactéries pathogènes résistantes aux antibiotiques sont devenues une vraie menace de santé publique (Ferri et al., 2017), alors même que le nombre d'autorisations de mises sur le marché de nouveaux antibiotiques a fortement décliné. La recherche de nouveaux antibiotiques est donc cruciale, et les métabolites spécialisés demeurent une source potentielle d'un grand intérêt.

De nos jours, il existe deux stratégies principales visant à obtenir de nouveaux antibiotiques. La première consiste à chercher de nouveaux métabolites spécialisés, soit en explorant des nouvelles niches écologiques ou des nouveaux genres microbiens, soit en étudiant les génomes des microorganismes déjà connus (Genilloud, 2018). Des outils sont notamment développés afin d'induire l'expression de groupes de gènes cryptiques, qui ne sont pas exprimés dans des conditions standards de laboratoire. La deuxième stratégie est basée sur la biologie synthétique des métabolites spécialisés, et vise à produire des métabolites spécialisés non naturels par ingénierie des groupes de gènes de biosynthèse (Pickens et al., 2011; Smanski et al., 2016). Ces approches de modification ou de substitution d'enzymes, souvent appelées approches de biosynthèse combinatoire, sont particulièrement adaptées à l'ingénierie d'enzymes de biosynthèse modulaires telles que les synthétases de peptides non ribosomiques (NRPS) (Awakawa et al., 2018; Baltz, 2018) et les polycétides synthases (PKS) (Yuzawa et al., 2018).

Les NRPS sont de grandes enzymes multi-modulaires responsables de la biosynthèse de peptides non ribosomiques (NRP). Elles peuvent être composées de plusieurs sous-unités, chacune étant constituée de modules (Figure 1). Chaque module incorpore un monomère au peptide final. Chaque module est divisé en domaines. Il y a trois domaines principaux. Le domaine d'adénylation (A) reconnaît l'acide aminé, l'active et le lie de façon covalente au bras 4'-phosphopantéthéinyl du domaine de transport de peptide (PCP) (Keller and Schauwecker, 2003). Le domaine PCP présente aux autres domaines le substrat covalamment lié à son cofacteur. Le domaine de condensation (C) catalyse la formation d'une liaison amide entre deux acides aminés et, par conséquent, l'élongation de la chaîne peptidique. À l'extrémité de la chaîne d'assemblage, le module de terminaison contient habituellement un domaine de thioestérase (TE), qui libère le produit par hydrolyse de la liaison thioester, parfois par cyclisation intramoléculaire (McErlean et al., 2019). Il peut également exister des domaines optionnels modifiant l'acide aminé incorporé (par exemple des domaines d'épimérisation, d'oxydation, de méthylation...).

Les domaines A sont responsables de la sélection et de l'activation des monomères, et présentent donc généralement une grande spécificité pour leur substrat (Strieker et al., 2010). Cependant, les domaines C et les domaines TE présentent eux aussi une certaine spécificité de substrats (Lautru and Challis, 2004), quoique moins stricte que celle des domaines A, qui n'a pas encore été complètement élucidée.

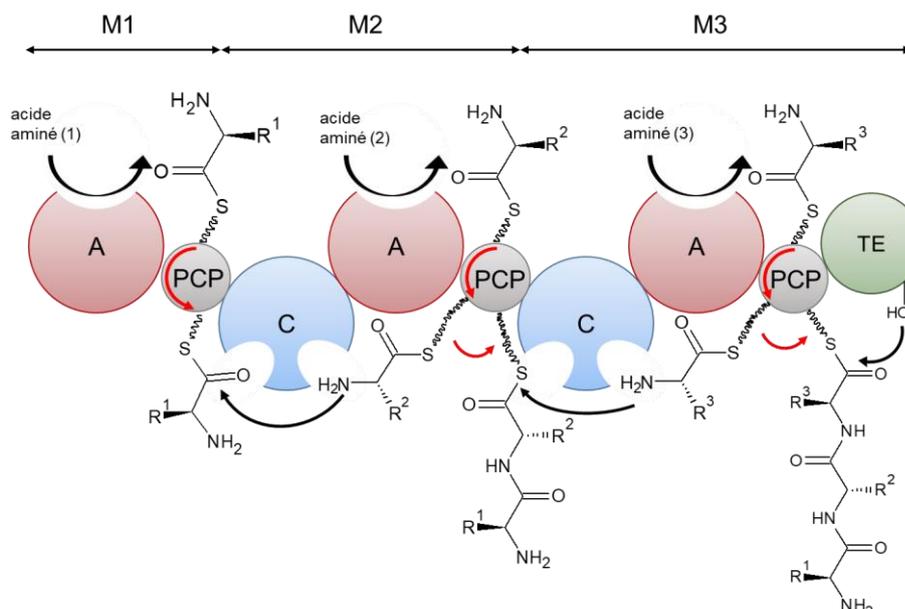


Figure 1 : Modèle de biosynthèse des NRPS

M1, M2 et M3 correspondent aux différents modules. Le module d'initiation M1 contient un domaine d'adénylation (A) et un domaine de transport de peptide (PCP). Le module d'extension M2 possède les deux mêmes domaines précédés d'un domaine de condensation (C). Le domaine de terminaison M3 a un domaine supplémentaire, le domaine thioestérase (TE) qui hydrolyse et libère le composé final.

Un autre point important pour la biosynthèse des NRP concerne les interactions entre domaines, modules et sous-unités, qui doivent être respectées pour que les partenaires interagissent correctement. Au cours du cycle catalytique, des réarrangements de domaines sont en effet nécessaires (Izoré and Cryle, 2018). Les mouvements des domaines A (partie C-terminale) et PCP sont particulièrement importants, l'adoption de différentes conformations permettant au bras 4'-phosphopantéthéinyl d'accéder à tous les sites catalytiques. Ces mouvements impliquent que les interactions protéine / protéine varient au cours du cycle catalytique, et les linkers reliant les domaines jouent donc un rôle essentiel en maintenant les interactions protéiques tout en permettant les changements de conformation. Dans certains cas, des petits domaines de communication, détectés aux extrémités des sous-unités des NRPS, permettent une interaction fonctionnelle et spécifique entre les différentes sous-unités des NRPS.

Appliquer des approches de biosynthèse combinatoire aux NRPS constitue une démarche particulièrement attrayante, du fait de la modularité de ces enzymes et de la diversité extrême de composés synthétisés. Des expériences d'ingénierie, principalement basées sur deux approches différentes, ont été menées et ont contribué à notre connaissance des NRPS. Une première approche consiste à modifier la spécificité de substrat du domaine A, par des mutations ponctuelles ou des substitutions de sous-domaines (Figure 2A et B). Ces approches minimisent la modification des interfaces, mais elles sont limitées dans la plupart des cas par la spécificité de substrat des domaines C. Une alternative permettant de limiter les problèmes de spécificité de substrat du domaine C consiste à substituer plusieurs domaines ou modules (Figure 2C et D). Des substitutions des domaines C-A ou C-A-PCP sont les plus fréquemment utilisées, même si des cas présentant d'autres substitutions ont été rapportés. Quelle que soit la stratégie adoptée, les approches de biosynthèse combinatoire ont généralement pour résultat un faible rendement. Les multiples éléments qui entrent en jeu pour le bon fonctionnement des NRPS expliquent très probablement la difficulté rencontrée pour concevoir des chaînes d'assemblage fonctionnelles.

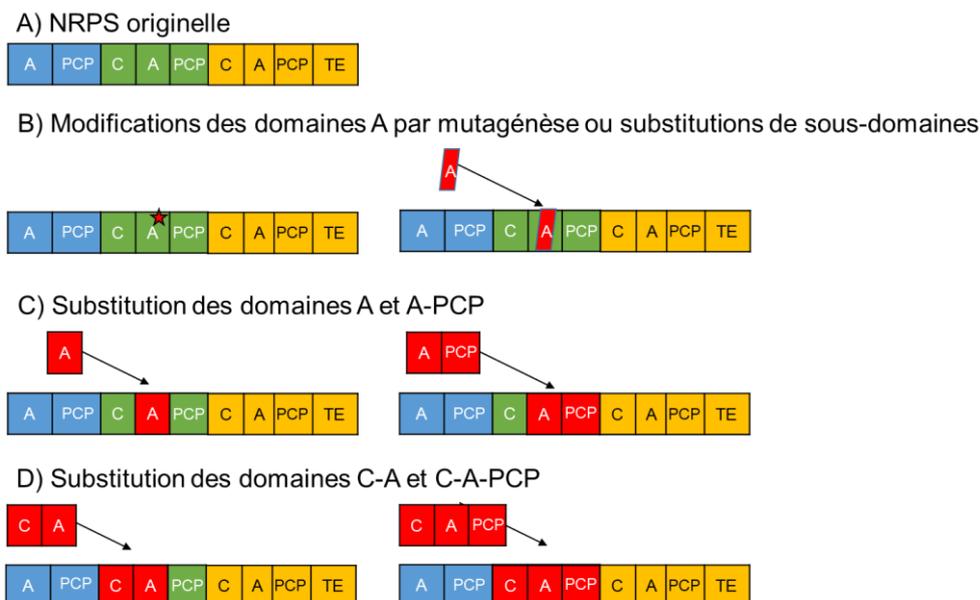


Figure 2: Possibilités de substitution de domaines de NRPS

Si les principes généraux de la biosynthèse de peptides non ribosomiques sont bien compris, un travail important est encore nécessaire pour déchiffrer les mécanismes détaillés permettant le fonctionnement coordonné des nombreux domaines enzymatiques constituant ces méga complexes. Des études structurales et biochimiques seront sans aucun doute nécessaires, mais l'utilisation de la biosynthèse combinatoire pour aborder ces questions apporte également des informations importantes. À cet égard, les NRPS qui dirigent la biosynthèse des pyrrolamides pourraient constituer un bon modèle. En effet, ces systèmes NRPS atypiques sont uniquement constitués de modules et de domaines autonomes, objets beaucoup plus petits que les sous-unités NRPS classiques et donc plus faciles à manipuler génétiquement ou biochimiquement.

Les pyrrolamides (congoïcine, distamycine, anthelvencine, pyrromycine...) constituent une famille de métabolites secondaires caractérisés par la présence de 4-aminopyrrole-2-carboxylates dans leur structure (Figure 3). La plupart des pyrrolamides se lient au petit sillon de l'ADN de façon non covalente. Ils présentent une variété d'activités biologiques (activités antibactériennes, antifongiques, antivirales), mais aucun n'a été exploité en médecine, principalement en raison de leur toxicité.

Les pyrrolamides sont assemblés par des NRPS atypiques composées de modules et de domaines autonomes, facilement manipulables (Juguet et al., 2009; Vingadassalon et al., 2015; Aubry et al., unpublished). De plus, les pyrrolamides semblent être assemblés de façon combinatoire à partir d'un nombre limité de précurseurs et de la « biosynthèse combinatoire naturelle » a déjà été observée dans la souche *Streptomyces netropsis*, productrice de congoïcine, distamycine et disgocidine (Vingadassalon et al., 2015).

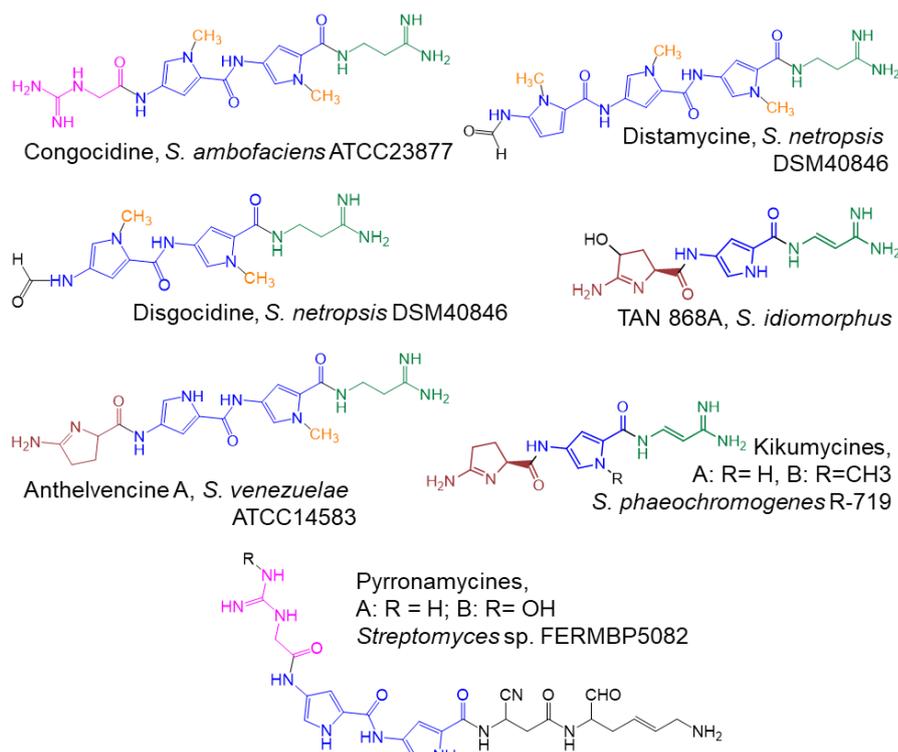


Figure 3 : Structures chimiques des membres de la famille des pyrrolamides et nom de leurs producteurs *Streptomyces*

Pour ces raisons, nous avons pensé que les systèmes de biosynthèse des pyrrolamides constituaient des systèmes attrayants pour effectuer des expériences de biosynthèse combinatoire, visant à mieux comprendre les différents éléments clés (spécificité du substrat, interactions protéiques...) essentiels au succès de la biologie synthétique des NRPS. Mon projet de doctorat a consisté à construire les outils nécessaires à la future biosynthèse combinatoire des pyrrolamides. Le projet a été divisé en trois axes, chacun développé dans un chapitre de thèse distinct :

(i) La caractérisation du groupe de gènes de biosynthèse du pyrrolamide anthelvencine.

Une condition préalable à la biosynthèse combinatoire est d'avoir à disposition des gènes provenant de différents groupes de gènes de biosynthèse. En effet, ces gènes sont les éléments de base qui fournissent les précurseurs et les enzymes qui doivent être échangés. Au début de mon projet, le laboratoire avait caractérisé les voies biosynthétiques de la congocidine (dans *Streptomyces ambofaciens* (2009) et *Streptomyces netropsis* (non publié)), et des distamycine/disgocidine/congocidine (dans *S. netropsis* (2015)). Toutefois, les gènes de biosynthèse des autres pyrrolamides n'avaient pas été identifiés. J'ai donc entrepris **la caractérisation du groupe de gènes de biosynthèse de l'anthelvencine, un pyrrolamide produit par *Streptomyces venezuelae* ATCC 14583**, qui est présentée dans le chapitre I.

(ii) La construction de vecteurs pour l'assemblage de groupes de gènes chez *Streptomyces*.

La biosynthèse combinatoire implique d'avoir des vecteurs qui permettent la manipulation génétique de nombreuses constructions génétiques. Les plasmides intégratifs historiques sont encore très utilisés aujourd'hui, mais ils ne sont pas normalisés et ne sont pas particulièrement adaptés à cet objectif. J'ai donc **développé une série de 12 vecteurs intégratifs**. Ces plasmides

modulaires ont été conçus pour faciliter la construction de cassettes de gènes. Ils ont également été construits pour permettre des intégrations multiples ou itératives dans le chromosome de *Streptomyces* et un système d'excision a été mis en place pour recycler les marqueurs de résistance et supprimer les éléments superflus après l'intégration. La construction de ces vecteurs est présentée dans le chapitre II.

(iii) La reconstruction du groupe de gènes de biosynthèse de la congocidine.

L'échange de gènes suppose l'existence d'une banque de cassettes de gènes normalisées. J'ai ainsi conçu des cassettes de gènes constituées d'un promoteur synthétique associé à un RBS, d'un ou plusieurs gène(s) de biosynthèse de pyrrolamides et d'un terminateur, cassettes qui correspondent à des « briques standard » à assembler. Une première étape logique avant de passer à la biosynthèse combinatoire consistait à reconstruire une voie de biosynthèse connue et à confirmer la production de pyrrolamides. J'ai donc entrepris **la reconstruction du groupe de gènes de biosynthèse de la congocidine en construisant et en assemblant toutes les cassettes génétiques** nécessaires à la production, et en évaluant la production de congocidine dans la souche hôte *S. lividans* TK23. Cette reconstitution est présentée dans le troisième et dernier chapitre de cette thèse.

#### I- Chapitre I : Caractérisation du groupe de gènes de biosynthèse de l'anthelvencine chez *Streptomyces venezuelae* ATCC 14583

Les anthelvencines A et B (Figure 4A) sont des métabolites spécialisés qui ont été isolés en 1965 de cultures de *Streptomyces venezuelae* ATCC 14583-14585 et qui présentent des activités antibactériennes et anthelminthiques modérées (Probst et al., 1965). Ils appartiennent à la famille des métabolites pyrrolamides, dont les membres les mieux caractérisés sont la congocidine et la distamycine.

Pour isoler le groupe de gènes dirigeant la biosynthèse de l'anthelvencine, nous avons séquencé le génome de la souche *S. venezuelae* ATCC 14583. Le groupe de gènes qui dirige la biosynthèse de l'anthelvencine a été identifié par recherche d'homologues des gènes impliqués dans la biosynthèse de la congocidine (Juguet et al., 2009). Nous avons identifié un groupe de gènes (*ant*) qui s'étend sur 26 kb et contient 22 gènes (Figure 4B). Vingt des protéines Ant présentent une identité de séquence d'acides aminés élevée avec les protéines Cgc (de 64 à 84 % d'identité de séquence) et elles ont très probablement une fonction semblable à leurs homologues Cgc. Ainsi, les numéros de gènes attribués aux gènes *ant* ont été choisis pour suivre la nomenclature *cgc* dans la mesure du possible. L'organisation génétique du groupe de gènes *ant* est remarquablement semblable à celle du groupe de gènes *cgc* (Figure 4B) (Juguet et al., 2009). Deux gènes *cgc* (*cgc7* et *cgc18*) impliqués dans la biosynthèse du précurseur guanidinoacétate de la congocidine (absent dans l'anthelvencine) et de son assemblage n'ont pas d'homologues dans le groupe de gènes *ant*. Le groupe de gènes contient en revanche deux gènes, *ant24* et *ant23*, probablement impliqués dans la biosynthèse du 5-amino-3,4-dihydro-2H-pyrrole-2-carboxylate [4] et son assemblage avec le premier précurseur du pyrrole, respectivement.

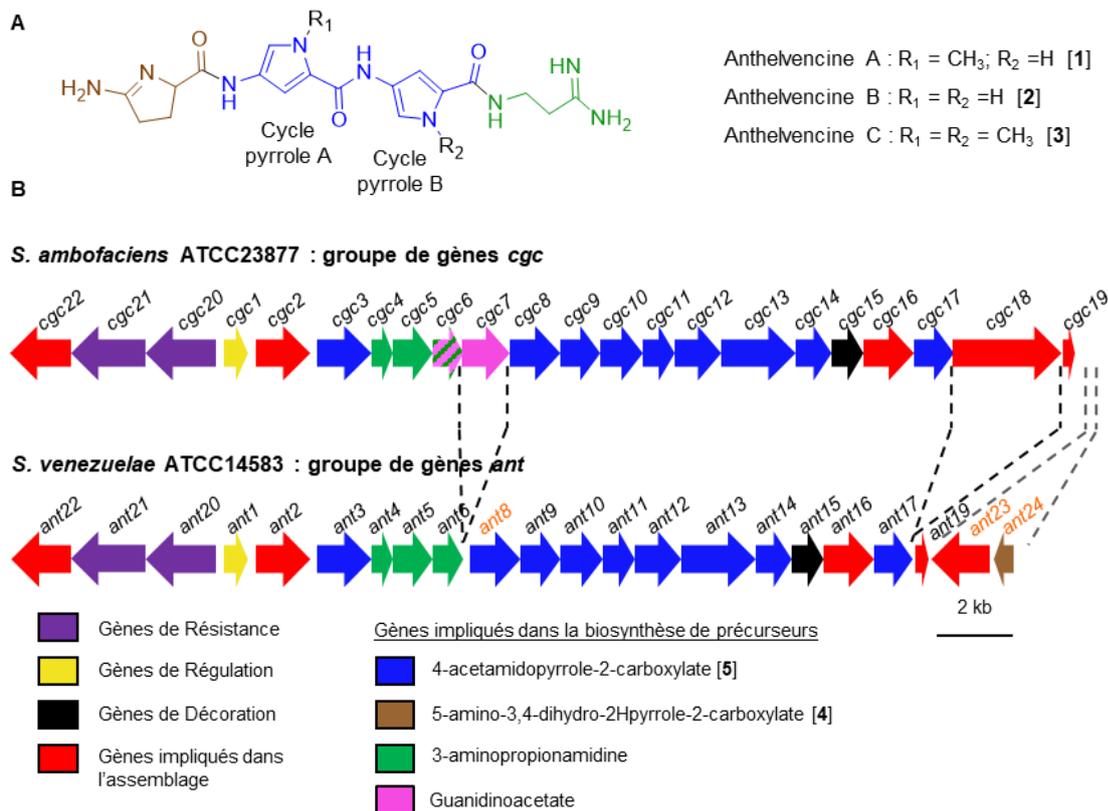


Figure 4 : Structure et groupe de gènes de l'anthelvincine

A) Structure des anthelvincines A, B et C

B) Organisation génétique du groupe de gènes de biosynthèse de la congocidine chez *S. ambofaciens* ATCC 23877 comparé à celui de l'anthelvincine chez *S. venezuelae* ATCC 14583.

Les gènes *ant* écrits en orange ont été remplacés par une cassette de résistance dans le cadre de cette étude.

Pour vérifier que le groupe de gènes *ant* est impliqué dans la biosynthèse de l'anthelvincine, nous avons inactivé *ant8* par remplacement par une cassette de résistance à l'apramycine. *ant8* est l'orthologue de *cgc8* qui est impliqué dans la biosynthèse du 4-acétaminopyrrole-2-carboxylate [5], précurseur de la congocidine (Lautru et al., 2012) et probablement précurseur de l'anthelvincine. Les surnageants de culture de la souche sauvage et du mutant ont été analysés par HPLC. Les chromatogrammes (Figure 5) montrent que quatre métabolites présents dans le surnageant de la souche de type sauvage (pics I à IV) sont absents dans le surnageant de la souche mutante ANT007 (*ant8::aac(3)IV*). Le premier métabolite (pic I, temps de rétention de 11,5 min) correspond au 4-aminopyrrole-2-carboxylate [5] (Lautru et al., 2012). Les trois pics II (temps de rétention de 13,3 min), III (temps de rétention de 14,3 min) et IV (temps de rétention de 15,5 min) ont des spectres d'absorption UV typiques des pyrrolamides (Vingadassalon et al., 2015).

Pour déterminer la nature chimique des métabolites II, III et IV, nous les avons partiellement purifiés. Une analyse en spectrométrie de masse à haute résolution et fragmentation (HR-MSMS) a confirmé que II correspondait à l'anthelvincine B. La masse exacte de III correspond à celle de l'anthelvincine A. Toutefois, le profil de fragmentation indique que la position du groupement méthyle ne se trouve pas sur le cycle pyrrole B, comme cela avait été proposé précédemment (mais jamais établi expérimentalement (Probst et al., 1965)), mais plutôt sur le cycle pyrrole A (Figure 4). Les expériences de RMN faites sur le composé III purifié n'ont jusqu'à présent pas permis de confirmer la position du groupement méthyle. La masse exacte et le

profil de fragmentation du composé IV indiquent qu'il s'agit d'une anthelvencine méthylée sur les deux groupements pyrroles, anthelvencine que nous avons nommée anthelvencine C. Nous avons essayé de purifier l'anthelvencine C pour confirmer sa structure chimique avec des analyses de RMN mais ce métabolite s'est avéré très instable, comme déjà observé par M. Lee et ses collaborateurs (Lee et al., 1988).

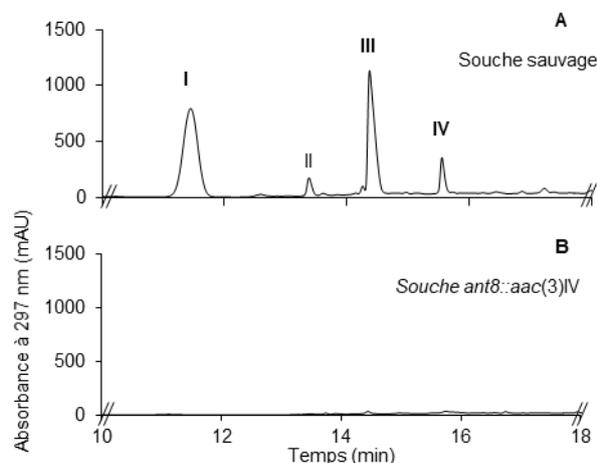


Figure 5: Analyse HPLC de surnageants de culture

- A) *S. venezuelae* ATCC14583 souche sauvage,  
 B) *S. venezuelae* ATCC14583 ANT007 (*ant8::aac(3)IV*)

Pour vérifier qu'*ant24* participe à la biosynthèse de **[4]** (5-amino-3,4-dihydro-2H-pyrrole-2-carboxylate), nous l'avons remplacé par une cassette de résistance. Le surnageant de la souche mutante obtenue a été analysé par HPLC. Aucune production d'anthelvencine n'a été observée, ce qui confirme que *ant24* est nécessaire pour la production de ces métabolites. L'ajout de **[4]** synthétisé chimiquement a permis de rétablir la production d'anthelvenicines A et C, confirmant ainsi l'implication de *ant24* dans la biosynthèse du précurseur de l'anthelvencine **[4]**. De la même manière, le remplacement de *ant23* par une cassette de résistance a eu pour conséquence l'arrêt de la production d'anthelvenicines. Pour nous assurer que le phénotype observé était dû au remplacement de *ant23* par la cassette *aac(3)IV*, nous avons génétiquement complété la souche en utilisant une expression plasmidique de *ant23* et *ant24* sous un promoteur constitutif. La production d'anthelvencine a été rétablie dans la souche complétée, confirmant ainsi que *ant23* est impliqué dans la biosynthèse de l'anthelvencine.

D'après les résultats présentés ci-dessus et les caractérisations antérieures de biosynthèse des pyrrolamides (Al-Mestarihi et al., 2015; Juguet et al., 2009; Lautru et al., 2012; Vingadassalon et al., 2015), nous proposons que les anthelvenicines soient assemblées à partir de 3-aminopropionamidine, 4-aminopyrrole-2-carboxylate et 5-amino-3,4-dihydro-2H-pyrrole-2-carboxylate (Figure 6). Comme déjà observé pour la biosynthèse d'autres pyrrolamides (congoïcine, distamycine), la synthèse de peptide non ribosomique impliquée dans l'anthelvencine est constituée uniquement de domaines autonomes (domaines C et PCP). Aucun domaine d'adénylation n'est impliqué dans l'activation des groupes carboxylés des précurseurs. Au lieu de cela, l'activation du groupe carboxylate du précurseur du pyrrole **[5]** et le lien covalent du précurseur activé au domaine PCP Ant19 est catalysé par Ant22, qui appartient à la famille des synthétases d'acyl-CoA. La formation de la première liaison amide entre **[4]** et **[5]** lié à Ant19 est

probablement catalysée par Ant23, une enzyme de la famille des enzymes de ligase ATP-grasp. Deux domaines de condensation autonomes, Ant16 et Ant2, catalysent la formation des autres liaisons amides, ajoutant respectivement un deuxième précurseur du pyrrole et la 3-aminopropionamide.

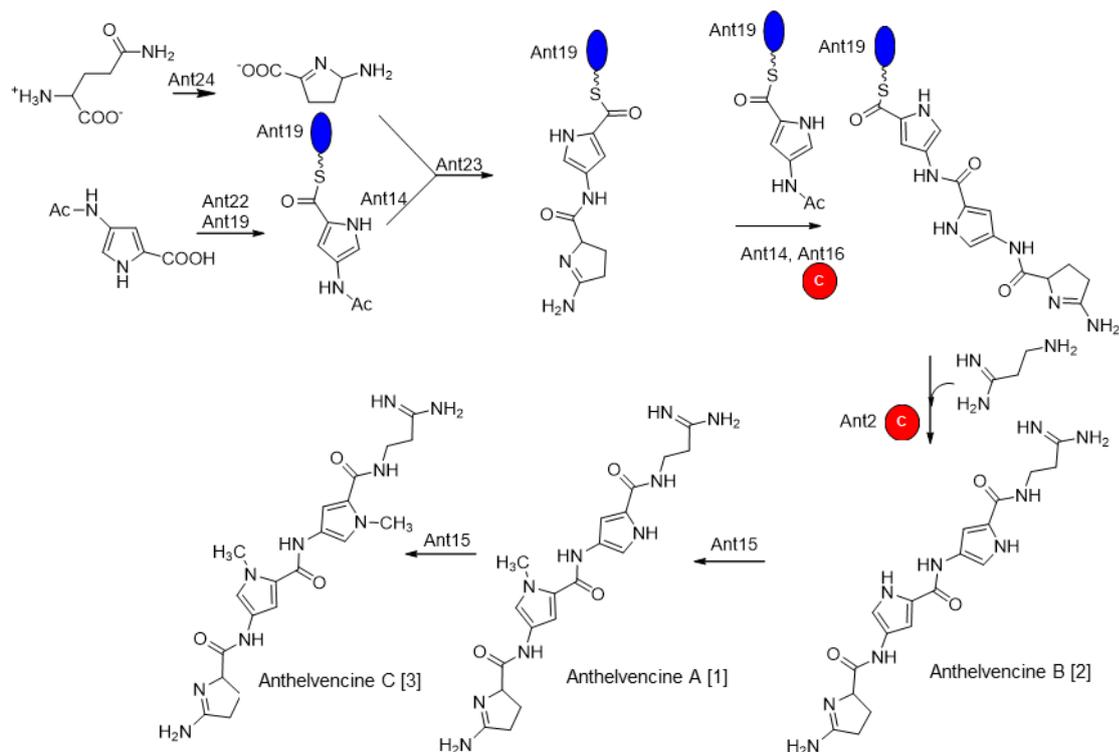


Figure 6: Voie de biosynthèse proposée pour les anthelvencines A, B and C.

En conclusion, nous avons identifié et caractérisé le groupe de gènes dirigeant la biosynthèse de l'anthelvencine dans *S. venezuelae* ATCC 14583. Nous avons montré que ce groupe dirige la biosynthèse de deux métabolites connus, l'anthelvencine A (pour laquelle nous proposons une structure révisée) et l'anthelvencine B, et d'une nouvelle anthelvencine que nous avons appelée anthelvencine C. Les nouveaux gènes pyrrolamide découverts s'ajoutent à notre bibliothèque de gènes NRPS, et seront probablement utiles plus tard pour procéder aux échanges de NRPS pendant les expériences de biosynthèse combinatoire.

## II- Chapitre II : Construction de vecteurs modulaires et intégratifs chez *Streptomyces*

Le développement de la biologie synthétique dans le domaine du métabolisme spécialisé nécessite le développement d'outils et de méthodes dédiés. En particulier, il nécessite des hôtes optimisés pour la production de métabolites spécialisés, des bibliothèques de fragments d'ADN synthétiques tels que des promoteurs, des séquences Shine-Dalgarno (RBS) ou des terminateurs, ainsi que des vecteurs et des méthodes d'assemblage de l'ADN pour l'assemblage *de novo* de groupes de gènes. Différents contextes expérimentaux sont susceptibles de nécessiter des approches de clonage différentes ou même une combinaison d'approches. Par conséquent, les vecteurs utilisés pour le clonage doivent être flexibles et facilement adaptables à diverses méthodes d'assemblage. Pourtant, dans le domaine de la biologie synthétique des métabolites spécialisés, peu de ces vecteurs

ont été construits. Nous avons donc entrepris la construction d'un ensemble de 12 vecteurs normalisés et modulaires, conçus pour permettre l'assemblage de groupes de gènes de biosynthèse à l'aide de diverses méthodes de clonage chez *Streptomyces*, producteurs prolifiques de métabolites spécialisés.

Les vecteurs ont été conçus pour répondre aux spécifications suivantes (Figure 7). Il doit être possible d'utiliser plusieurs vecteurs dans la même souche (orthogonalité). En conséquence, différentes cassettes de résistance aux antibiotiques et différents systèmes d'intégration à des sites spécifiques dans le chromosome de *Streptomyces* doivent être utilisés pour la construction des vecteurs. Les vecteurs doivent également être des vecteurs de navette entre *Escherichia coli* et *Streptomyces* afin que les constructions génétiques puissent être préparées dans *E. coli* avant d'être introduites dans les souches de *Streptomyces*. Enfin, les vecteurs doivent être modulaires et flexibles, de sorte que chaque module puisse être facilement remplacé par un autre équivalent si nécessaire.

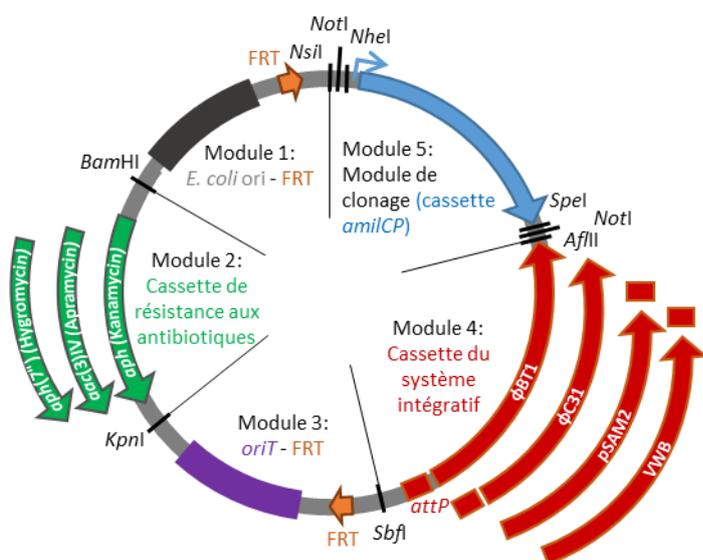


Figure 7 : Représentation schématique de l'ensemble des vecteurs modulaires et intégratifs pOSV801-pOSV812.

Les différentes cassettes de résistance aux antibiotiques et les systèmes d'intégration utilisés sont indiqués. Chaque site enzymatique de restriction indiqué est unique, sauf *NotI* (deux sites). *E. coli* ori correspond à l'origine de réplication p15A d'*E. coli*. *oriT* est l'origine de transfert. *amilCP* est le gène codant une chromoprotéine d'*Acropora millepora*, une protéine de couleur bleue. FRT correspond aux sites reconnus par la recombinaison Flp. Le promoteur du module 5 ne fonctionne que dans *E. coli*. Les sites *attP* sont utilisés par des intégrases pour intégrer le plasmide dans le génome de *Streptomyces* à un site spécifique.

Chaque vecteur est constitué de cinq modules (Figure 7). Le premier module est constitué de l'origine de réplication chez *E. coli* et d'un site FRT ciblé par la Flp pour recombinaison. Le deuxième module consiste en un marqueur de résistance aux antibiotiques. Trois gènes de résistance différents, fonctionnels chez *E. coli* et *Streptomyces*, ont été choisis. Le troisième module est constitué de l'origine de transfert RP4, et d'un deuxième site FRT. Le quatrième module est la cassette du système d'intégration (intégrases et leur site *attP* correspondant) qui permet l'intégration spécifique du site dans les chromosomes de *Streptomyces* après la conjugaison. Le dernier module est le module de clonage. Notre objectif pour ce module était de permettre le clonage et l'assemblage de gènes ou de cassettes de gènes utilisant une variété de méthodes de clonage (basées sur les

régions d'homologie ou sur l'utilisation d'enzymes de restriction), car différents projets peuvent nécessiter des approches de clonage différentes. Ce module a donc été conçu pour permettre l'assemblage itératif de gènes (ou de cassettes de gènes) en utilisant la méthode d'assemblage BioBrick (Shetty et al., 2008). Le module de clonage comprend un gène *amiC<sub>P</sub>* entre les séquences de préfixe et de suffixe des biobriques. Ce gène code une chromoprotéine, donnant une couleur bleue à la cellule. Cette cassette est destinée à être remplacée par la construction d'intérêt et offre un moyen pratique de cribler les clones contenant la nouvelle construction.

Pour vérifier que les 12 vecteurs que nous avons construits étaient tous fonctionnels, nous les avons intégrés dans le chromosome de trois souches de *Streptomyces* couramment utilisées pour l'expression hétérologue : *Streptomyces coelicolor* M145, *Streptomyces lividans* TK23 et *Streptomyces albus* J1074. Une difficulté potentielle lorsque plusieurs constructions génétiques doivent être intégrées dans les chromosomes de *Streptomyces* est le nombre limité de marqueurs de résistance aux antibiotiques qui sont fonctionnels dans une souche donnée. Pour permettre le recyclage des marqueurs de résistance, nous avons inclus dans nos vecteurs des sites FRT entourant le module 1 (origine de réplication chez *E. coli*), le module 2 (cassette de résistance aux antibiotiques) et le module 3 (origine de transfert). Ainsi, une fois un vecteur intégré dans un chromosome *Streptomyces*, ces trois modules, qui ne sont plus nécessaires, peuvent être excisés en utilisant la recombinase Flp amenée en *trans* par un plasmide réplicatif. La faisabilité de l'excision a été démontrée en prenant l'exemple d'un des vecteurs, intégré dans *S. coelicolor* M145.

Pour illustrer certaines utilisations possibles de nos vecteurs, nous avons reconstruit le groupe de gènes de l'albonoursine produite par *Streptomyces noursei*, en utilisant la méthode d'assemblage BioBrick. Nous avons également utilisé la méthode de clonage par réaction en cycle de ligase (LCR) pour assembler une unité de transcription dans l'un des vecteurs et compléter génétiquement une souche mutante.

En conclusion, nous avons construit un ensemble de plasmides dédié à l'assemblage et l'intégration d'ADN dans les chromosomes de *Streptomyces*. Nous voulions proposer une plateforme modulaire et flexible pouvant être utilisée dans différents contextes expérimentaux, de l'assemblage de petites cassettes de gènes à l'assemblage de fragments d'ADN plus grands, et qui soit compatible avec une grande variété de méthodes de clonage. Tous nos plasmides sont à la disposition de la communauté par le biais du dépôt dans les collections de plasmides (Addgene et BCCM).

### III- Chapitre III : Reconstruction du groupe de gènes de biosynthèse de la congocidine

La reconstruction d'une voie de biosynthèse est une approche de biologie synthétique qui consiste à réécrire la séquence d'ADN contenant toutes les informations génétiques nécessaires à l'expression et au fonctionnement de cette voie. Cette approche a d'abord été développée pour découpler l'expression des voies de biosynthèse de leur régulation naturelle (Temme et al., 2012), mais peut aussi être utilisée pour créer des unités de transcription artificielles qui peuvent ensuite être assemblées pour reconstituer un groupe de gènes fonctionnels. On considère souvent qu'il s'agit d'un premier pas vers la manipulation génétique du groupe de gènes de biosynthèse et la production de nouveaux métabolites non naturels (Basitta et al., 2017; Osswald et al., 2014). C'est dans ce but que nous avons entrepris la reconstruction du groupe de gènes de biosynthèse de la congocidine, un des pyrrolamides les mieux caractérisés (Figure 8A).

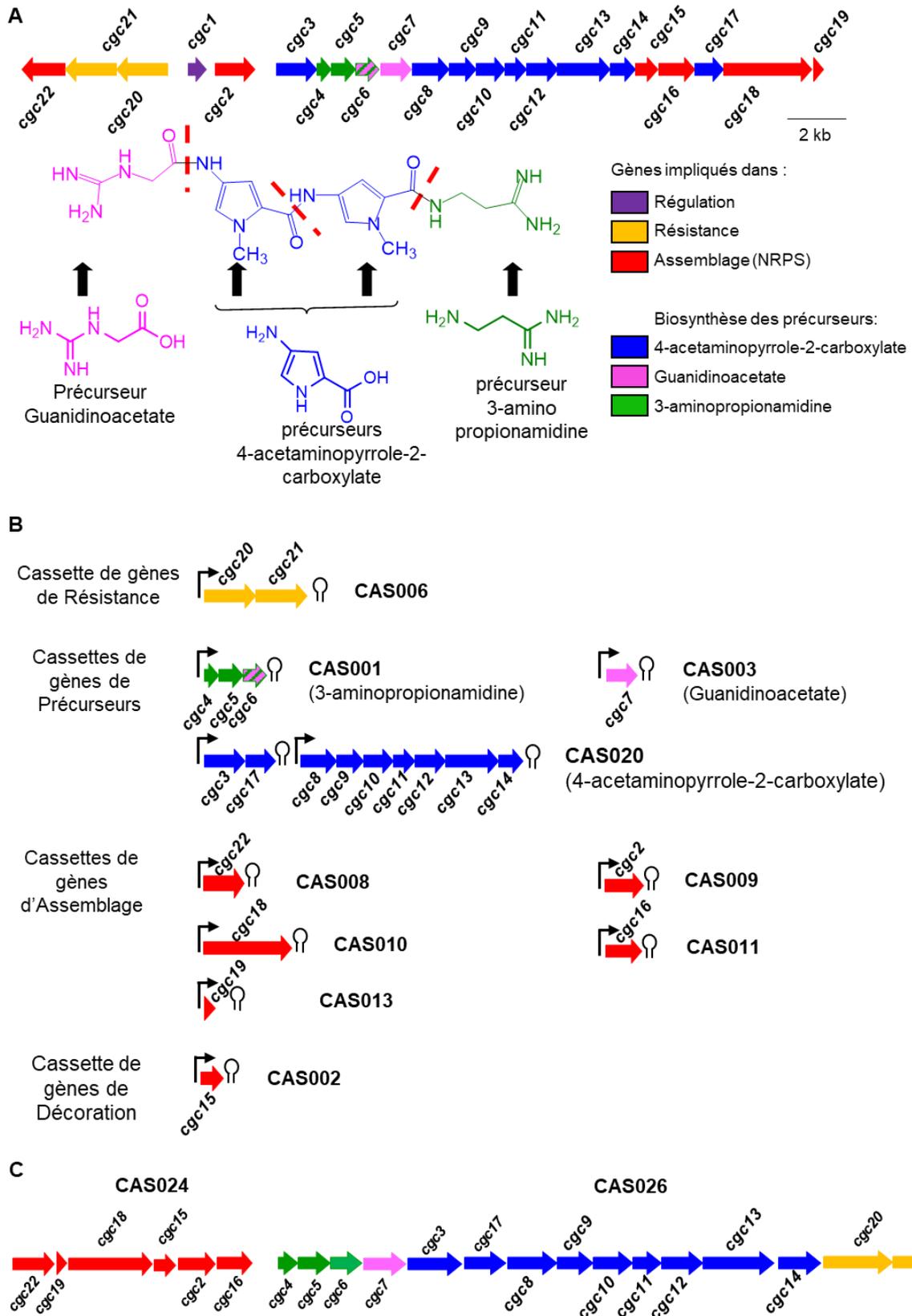


Figure 8: Groupe de gènes de biosynthèse de la congocidine et cassettes de gènes construites  
 A) Groupe de gènes de biosynthèse natif de la congocidine (*cgc*) produite par *S. ambifaciens* et structure de la congocidine. Les tirets en rouge séparent les différents monomères de la congocidine  
 B) Cassette synthétique de gènes construites  
 C) Schéma du cluster *cgc* reconstitué (par souci de clarté les promoteurs et terminateurs ne sont pas indiqués)

Nos objectifs étaient (i) de contrôler l'expression des gènes *cgc* et, plus tard, d'autres gènes de biosynthèse des pyrrolamides (supprimer la régulation transcriptionnelle naturelle) et (ii) de réorganiser les gènes en nouvelles unités de transcription fonctionnelles qui seront ré-utilisables pour des expériences de biosynthèse combinatoire (conception de cassettes génétiques normalisées, orthogonales et facilement échangeables).

Nous avons construit 11 cassettes de gènes basiques, conçues pour constituer des unités fonctionnelles, pour exprimer les 21 gènes du groupe de gènes *cgc*. Chaque cassette de gènes a été conçue en tenant compte de l'utilisation future dans des approches de biosynthèse combinatoire des pyrrolamides. Quatre types de cassettes de gènes basiques ont été construits : les cassettes de précurseurs, d'assemblage, de décoration et de résistance (Figure 8B).

Les cassettes de gènes des précurseurs comprennent tous les gènes nécessaires à la biosynthèse d'un précurseur donné. La congocidine est constituée de trois précurseurs, la 3-aminopropionamide, le guanidinoacétate et le 4-acétaminopyrrole-2-carboxylate. Ainsi, trois cassettes de gènes de précurseurs ont été construites. Cinq cassettes de gènes d'assemblage ont été construites, chacune contenant un seul gène (*cgc2*, *cgc16*, *cgc19*, *cgc18* et *cgc22* respectivement), car les gènes d'assemblage devront pouvoir être échangés individuellement dans le cadre d'expériences de biosynthèse combinatoire. Enfin, une cassette de gène de décoration (*cgc15*, codant une méthyltransférase) et une cassette de gènes de résistance (*cgc20* et *cgc21* codant un transporteur ABC) ont été construites.

Chaque cassette de gènes basique est constituée d'une unité de transcription, composée d'un promoteur synthétique, d'une séquence de Shine-Dalgarno (RBS), d'un ou de plusieurs gènes de biosynthèse de la congocidine (*cgc*) et d'un terminateur T4. Chaque cassette de gènes basique a été assemblée à l'aide de la réaction en cycle de ligase (LCR) (de Kok et al., 2014). Cet assemblage est basé sur l'utilisation d'une ligase thermostable et de plusieurs cycles de température de dénaturation-appariement-ligature. Des oligonucléotides chimères, dont les séquences sont complémentaires aux séquences des extrémités de deux fragments d'ADN à assembler, sont utilisés comme matrice pour appairer les deux fragments, qui sont ensuite ligaturés par la ligase thermostable.

La fonctionnalité de chaque cassette a été vérifiée au moyen d'une combinaison de complémentation génétique de souches mutantes, d'analyses HPLC et d'essais biologiques. Ces cassettes de gènes basiques ont ensuite été progressivement assemblées en cassettes de gènes composites par un assemblage de type Biobrick. Au final, deux plasmides intégratifs compatibles contenaient l'ensemble des cassettes nécessaires pour reconstituer le groupe de gènes *cgc*.

L'étape suivante a consisté en l'introduction des deux plasmides dans *S. lividans* TK23 par conjugaison inter-générique. Nous avons remarqué une grande instabilité génétique des deux plasmides chez les souches conjuguantes de *E. coli* (perte d'une partie des inserts), instabilité qui n'a pas été observée lors de l'assemblage des cassettes génétiques dans *E. coli* DH5 $\alpha$ . Une analyse de séquence a montré que cette instabilité était probablement due à de la recombinaison homologue entre les multiples copies des séquences terminatrices T4. Pour sélectionner les exconjugants contenant les plasmides non recombinés, nous avons effectué un essai biologique basé sur l'activité antibiotique de la congocidine. En effet, si les plasmides intacts ont été introduits dans *S. lividans*, alors la souche devrait produire de la congocidine. Les clones inhibant la croissance de *Micrococcus*

*lutens* ont été cultivés et leurs surnageants de culture ont été analysés par HPLC. Tous les clones ont produit de la congocidine, comme en témoigne le chromatogramme d'un clone présenté sur la Figure 9.

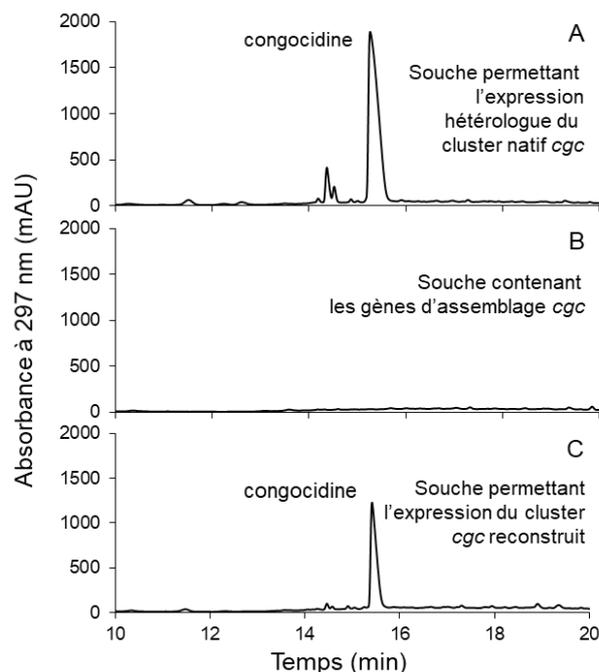


Figure 9 : Production de congocidine par le groupe de gènes *cgc* reconstruit.

Chromatogrammes HPLC des surnageants de *S. lividans* :

A) CGCL006 (TK23 contenant le groupe de gènes natif *cgc*),

B) CGCL096 (TK23 avec CAS024, contenant tous les gènes d'assemblage),

C) CGCL096 (TK23 avec CAS024 (gènes d'assemblage) et CAS026 (gènes de résistance et de biosynthèse des précurseurs)

En conclusion, dans cette étude, nous avons reconstitué le groupe de gènes de biosynthèse de la congocidine et avons confirmé que le groupe de gènes reconstruit était fonctionnel. Cette reconstruction réussie ouvre maintenant la voie à l'optimisation de la production de congocidine. Plus important encore, elle nous offre une plate-forme fonctionnelle pour élaborer des expériences de biosynthèse combinatoire basées sur les pyrrolamides, et d'accroître, par exemple en échangeant des gènes de NRPS, les connaissances qui sont encore requises afin de maîtriser leur ingénierie.

Conclusion :

En raison de leurs propriétés (domaines ou modules NRPS autonomes, gènes homologues parmi les différents groupes de gènes de biosynthèse, existence d'une biosynthèse combinatoire naturelle), nous avons choisi la famille des pyrrolamides comme modèle pour sonder les facteurs limitants qui nuisent au succès des approches de biosynthèse combinatoire de la NRPS. Au cours de mon projet de doctorat, j'ai cherché à construire des outils pour permettre la biosynthèse combinatoire des gènes de biosynthèse des pyrrolamides. La caractérisation du groupe de gènes de biosynthèse de l'anthelvencine a notamment permis d'ajouter de nouveaux gènes de NRPS à notre banque de gènes. Afin d'établir une plate-forme facilitant la biosynthèse combinatoire, j'ai construit des plasmides intégratifs flexibles et compatibles avec différentes techniques d'assemblage. J'ai ensuite utilisé ces plasmides pour entreprendre la reconstruction du groupe de gènes de biosynthèse

de la congocidine, afin de prouver la faisabilité de cette approche basée sur la construction de cassettes synthétiques de gènes dans une démarche de biosynthèse combinatoire.

La voie de biosynthèse de la congocidine reconstruite peut maintenant servir de plate-forme pour échanger des gènes NRPS et sonder les interactions protéines/protéines des NRPS et les spécificités des substrats des différents domaines. Une première étape pourrait consister en l'échange de domaines ayant un rôle identique, comme le domaine PCP transportant les intermédiaires au cours de la biosynthèse des pyrrolamides. Comme ce domaine n'a pas de rôle catalytique, le succès ou l'échec de la production de congocidine après l'échange pourrait conduire à l'identification des régions des NRPS impliquées dans les interactions protéines/protéines. Inversement, certains des domaines de condensation ont des rôles similaires dans des voies de biosynthèse distinctes. La substitution de ces domaines de condensation par des homologues plus ou moins proches pourrait être très instructive en ce qui concerne les spécificités des substrats.

#### Bibliographie :

Al-Mestarihi, A.H., Garzan, A., Kim, J.M., and Garneau-Tsodikova, S. (2015). Enzymatic evidence for a revised congocidine biosynthetic pathway. *Chembiochem Eur. J. Chem. Biol.* *16*, 1307–1313.

Awakawa, T., Fujioka, T., Zhang, L., Hoshino, S., Hu, Z., Hashimoto, J., Kozono, I., Ikeda, H., Shin-Ya, K., Liu, W., et al. (2018). Reprogramming of the antimycin NRPS-PKS assembly lines inspired by gene evolution. *Nat. Commun.* *9*, 3534.

Baltz, R.H. (2018). Synthetic biology, genome mining, and combinatorial biosynthesis of NRPS-derived antibiotics: a perspective. *J. Ind. Microbiol. Biotechnol.* *45*, 635–649.

Basitta, P., Westrich, L., Rösch, M., Kulik, A., Gust, B., and Apel, A.K. (2017). AGOS: A plug-and-play method for the assembly of artificial gene operons into functional biosynthetic gene clusters. *ACS Synth. Biol.* *6*, 817–825.

Demain, A.L. (2009). Antibiotics: natural products essential to human health. *Med. Res. Rev.* *29*, 821–842.

Ferri, M., Ranucci, E., Romagnoli, P., and Giaccone, V. (2017). Antimicrobial resistance: a global emerging threat to public health systems. *Crit. Rev. Food Sci. Nutr.* *57*, 2857–2876.

Genilloud, O. (2018). Mining actinomycetes for novel antibiotics in the omics era: are we ready to exploit this new paradigm? *Antibiot. Basel Switz.* *7*.

Izoré, T., and Cryle, M.J. (2018). The many faces and important roles of protein-protein interactions during non-ribosomal peptide synthesis. *Nat. Prod. Rep.* *35*, 1120–1139.

Juguet, M., Lautru, S., Francou, F.-X., Nezbedová, S., Leblond, P., Gondry, M., and Pernodet, J.-L. (2009). An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congocidine in *Streptomyces ambofaciens*. *Chem. Biol.* *16*, 421–431.

Keller, U., and Schauwecker, F. (2003). Combinatorial biosynthesis of non-ribosomal peptides. *Comb. Chem. High Throughput Screen.* *6*, 527–540.

de Kok, S., Stanton, L.H., Slaby, T., Durot, M., Holmes, V.F., Patel, K.G., Platt, D., Shapland, E.B., Serber, Z., Dean, J., et al. (2014). Rapid and reliable DNA assembly via ligase cycling reaction. *ACS Synth. Biol.* *3*, 97–106.

- Lautru, S., and Challis, G.L. (2004). Substrate recognition by nonribosomal peptide synthetase multi-enzymes. *Microbiol. Read. Engl.* *150*, 1629–1636.
- Lautru, S., Song, L., Demange, L., Lombès, T., Galons, H., Challis, G.L., and Pernodet, J.-L. (2012). A sweet origin for the key congocidine precursor 4-acetamidopyrrole-2-carboxylate. *Angew. Chem. Int. Ed. Engl.* *51*, 7454–7458.
- Lee, M., Coulter, D.M., and Lown, J.W. (1988). Total synthesis and absolute configuration of the antibiotic oligopeptide (4S)-(+)-anthelvencin A and its 4R(-) enantiomer. *J. Org. Chem.* *53*, 1855–1859.
- McErlean, M., Overbay, J., and Van Lanen, S. (2019). Refining and expanding nonribosomal peptide synthetase function and mechanism. *J. Ind. Microbiol. Biotechnol.* *46*, 493–513.
- Newman, D.J., and Cragg, G.M. (2016). Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* *79*, 629–661.
- Osswald, C., Zipf, G., Schmidt, G., Maier, J., Bernauer, H.S., Müller, R., and Wenzel, S.C. (2014). Modular construction of a functional artificial epothilone polyketide pathway. *ACS Synth. Biol.* *3*, 759–772.
- Pickens, L.B., Tang, Y., and Chooi, Y.-H. (2011). Metabolic engineering for the production of natural products. *Annu. Rev. Chem. Biomol. Eng.* *2*, 211–236.
- Probst, G.W., Hoehn, M.M., and Woods, B.L. (1965). Anthelvencins, new antibiotics with anthelmintic properties. *Antimicrob. Agents Chemother.* *5*, 789–795.
- Shetty, R.P., Endy, D., and Knight, T.F.J. (2008). Engineering BioBrick vectors from BioBrick parts. *J. Biol. Eng.* *2*:5.
- Smanski, M.J., Zhou, H., Claesen, J., Shen, B., Fischbach, M.A., and Voigt, C.A. (2016). Synthetic biology to access and expand nature’s chemical diversity. *Nat. Rev. Microbiol.* *14*, 135–149.
- Strieker, M., Tanović, A., and Marahiel, M.A. (2010). Nonribosomal peptide synthetases: structures and dynamics. *Curr. Opin. Struct. Biol.* *20*, 234–240.
- Temme, K., Zhao, D., and Voigt, C.A. (2012). Refactoring the nitrogen fixation gene cluster from *Klebsiella oxytoca*. *Proc. Natl. Acad. Sci. U. S. A.* *109*, 7085–7090.
- Vingadassalon, A., Lorieux, F., Juguet, M., Le Goff, G., Gerbaud, C., Pernodet, J.-L., and Lautru, S. (2015). Natural combinatorial biosynthesis involving two clusters for the synthesis of three pyrrolamides in *Streptomyces netropsis*. *ACS Chem. Biol.* *10*, 601–610.
- Yuzawa, S., Backman, T.W.H., Keasling, J.D., and Katz, L. (2018). Synthetic biology of polyketide synthases. *J. Ind. Microbiol. Biotechnol.* *45*, 621–633.

**Titre :** Vers la biosynthèse combinatoire d'antibiotiques pyrrolamides chez *Streptomyces*

**Mots clés :** métabolisme spécialisé, biologie synthétique, *Streptomyces*, pyrrolamide

**Résumé:** Depuis plus de 80 ans, le métabolisme spécialisé nous fournit de nombreuses molécules utilisées en médecine, en particulier comme anti-infectieux. Aujourd'hui, avec l'augmentation mondiale de la résistance aux antimicrobiens, de nouveaux antibiotiques sont indispensables. Une des réponses à cette pénurie grave pourrait provenir de la biologie synthétique. Dans le domaine du métabolisme spécialisé, la biologie synthétique est utilisée en particulier pour la biosynthèse de métabolites non naturels. Parmi les métabolites spécialisés, les peptides non ribosomiques constituent une cible attrayante, car ils nous ont déjà fourni des molécules à haute valeur clinique (ex. les antibiotiques vancomycine et daptomycine). De plus, la plupart sont synthétisés par des enzymes multimodulaires appelées synthétases de peptides non ribosomiques (NRPS), et sont diversifiés davantage par des enzymes de décoration. Ainsi, ces voies de biosynthèse se prêtent particulièrement à la biosynthèse combinatoire, consistant à combiner des gènes de biosynthèse provenant de divers groupes de gènes ou, dans le cas des NRPS, à combiner des modules ou domaines pour créer de nouvelles enzymes. Cependant, si plusieurs études ont établi la faisabilité de telles approches, de nombreux obstacles subsistent avant que les approches combinatoires de biosynthèse soient totalement efficaces pour la synthèse de nouveaux métabolites.

Les travaux présentés ici s'inscrivent dans le cadre d'un projet visant à comprendre les facteurs limitant les approches de biosynthèse combinatoire basées sur les NRPS, en utilisant une approche de biologie synthétique. Nous avons choisi de travailler avec les NRPS responsables de la biosynthèse des pyrrolamides. En effet, ces NRPS sont constituées uniquement de modules et de domaines autonomes, et donc particulièrement adaptés aux manipulations génétiques et biochimiques. La caractérisation du groupe de gènes de biosynthèse du pyrrolamide anthelvencine constitue la première partie de cette thèse et nous a fourni de nouveaux gènes pour notre étude. La deuxième partie a consisté à construire des vecteurs intégratifs modulaires, outils essentiels pour la construction et l'assemblage de cassettes génétiques. La dernière partie présente la reconstruction du groupe de gènes du pyrrolamide congocidine, basée sur la construction et l'assemblage de cassettes de gènes synthétiques. Dans l'ensemble, ces travaux ouvrent la voie à de futures expériences de biosynthèse combinatoire, expériences qui devraient contribuer à une meilleure compréhension du fonctionnement précis des NRPS.

**Title:** Towards combinatorial biosynthesis of pyrrolamide antibiotics in *Streptomyces*

**Keywords:** specialized metabolism, synthetic biology, *Streptomyces*, pyrrolamide

**Abstract:** For more than 80 years, specialized metabolism has provided us with many molecules used in medicine, especially as anti-infectives. Yet today, with the rise of antimicrobial resistance worldwide, new antibiotics are crucially needed. One of the answers to this serious shortage could arise from synthetic biology. In the field of specialized metabolism, synthetic biology is used in particular to biosynthesize unnatural metabolites. Among specialized metabolites, non-ribosomal peptides constitute an attractive target as they have already provided us with clinically valuable molecules (e.g. the vancomycin and daptomycin antibiotics). In addition, most are synthesized by multimodular enzymes called non-ribosomal peptide synthetases (NRPS) and further diversified by tailoring enzymes. Thus, such biosynthetic pathways are particularly amenable to combinatorial biosynthesis, which consists in combining biosynthetic genes coming from various gene clusters or, in the case of NRPSs, combining modules or domains to create a new enzyme. Yet, if several studies have established the feasibility of such approaches, many obstacles remain before combinatorial biosynthesis approaches are fully effective for the synthesis of new metabolites.

The work presented here is part of a project aiming at understanding the limiting factors impeding NRPS-based combinatorial biosynthesis approaches, using a synthetic biology approach. We chose to work with the NRPSs involved in the biosynthesis of pyrrolamides. Indeed, these NRPSs are solely constituted of stand-alone modules and domains, and thus, particularly amenable to genetic and biochemical manipulations. The characterization of the biosynthetic gene cluster of the pyrrolamide anthelvencin constitutes the first part of this thesis, and provided us with new genes for our study. The second part involved the construction of modular integrative vectors, essential tools for the construction and assembly of gene cassettes. The final part presents the successful refactoring of the congocidine pyrrolamide gene cluster, based on the construction and assembly of synthetic gene cassettes. Altogether, this work paves the way for future combinatorial biosynthesis experiments that should help decipher the detailed functioning of NRPSs.

