University of Strathclyde Glasgow

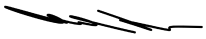
BIOGEOGRAPHICAL AND PHYLOGENETIC BOUNDARIES OF THE SPECIALISED METABOLISM OF *Pseudonocardia* FROM MARINE ORIGIN

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

Jonathan Parra Villalobos

Declaration

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

Signed: 

Date: 27th January 2022

Toledo en vuelo (2021) by Mayela Villalobos (my mom)

Dedicated to all the faces, scents, colours and sounds that I call home.

Dedicado a todos los rostros, aromas, colores y sonidos a los que llamo hogar.

Acknowledgments

Costa Rica

Un especial agradecimiento a Yalieth, quien me apoyó de muchas formas para que pudiera cumplir mi sueño. Aunque el faro decidió apagarse y el barco se perdió en la tormenta, agradezco la luz que por un instante iluminó el regreso a casa.

Mi mayor agradecimiento será para mis padres, Guillermo y Mayela, quienes me construyeron como persona. Gracias por enseñarme el camino a seguir y darme la libertad de construir mi propio destino. A ellos agradeceré esté y todos mis triunfos. También quiero agradecer a mis hermanos Marco, Pablo y Jesús, quienes junto con mis padres, han sido fieles testigos y actores principales en mi vida.

Agradezco a la Universidad de Costa Rica (UCR) y al Programa de Innovación y Capital Humano para la Competitividad (PINN) del Ministerio de Ciencia, Innovación, Tecnología y Telecomunicaciones de Costa Rica (MICITT) por el apoyo financiero que permitió realizar este proyecto.

Scotland and the rest of the world

First and foremost, I have to thank Dr Katherine Duncan. Kate believed in me from the beginning, more than I did. She taught me a lot about science, but she also taught me great lessons about myself. Thanks to Kate I learned that I may have limitations, but it is up to me to decide if these limitations are going to define me. I will always be grateful.

I would like to thank all the members of the Duncan group. Laia, Ally, Darren and Lily were my family in Scotland. I love you all. I have to confess that I would not have survived in Scotland without Ally. Thanks a lot, Ally. I have also to thank Laia for being such a supportive and good friend. Of course, I have to thank my dear friend Darren, who always cheered me up. Last but not least, thanks to Lily, one of the sweetest and most talented people I have ever met. I cannot forget to thank Diana and Connie, two wonderful people as well as talented scientists. Finally, I also would like to thank the former postdocs of the Duncan group Sylvia and Alejandro.

I also express my gratitude to all the collaborators of this project. I would like to thank Eszter Denes and Professor Carol Munro for their collaboration in the antifungal tests. Special thanks to Dr Scott Jarmusch, who kindly helped me to conclude this project with his amazing MSI data. Also, I would like to thank all the PIs of the Strathclyde micro-group for their valuable opinions and advice. In particular, I would like to thank Dr Paul Herron, who demonstrated a genuine interest in this project.

I would like to thank each and every member of the Strathclyde micro-group. I was very fortunate to work with such a talented and kind group of people. Special thanks to Andy (my first friend in Glasgow), Ainsley, Becca, Liam, Emily, Lis, Jonny, Jordan, Anna, Gordon (I really enjoyed our philosophical debates), David, Charlie (my coffee pal), Adriana, Eilidh, Elmira, and Robyn. I have to especially thank Tiago, who helped me a lot in the lab and was an extraordinary friend outside of it. Lastly, I would like to thank Alemão; Glasgow would not have been the same without him.

During the last year of my PhD, I built a supportive network of PhD students and early career researchers from all over the world who helped me a lot during the final stage of this project. So, I would like to thank Jana, Lina, Agustina, Camila and Kristiina for all their support.

Abstract

Multidrug-resistant pathogens have become a global threat. In this context, filamentous actinomycetes have been proven to be an exceptional source of antimicrobial metabolites. In particular, rare actinomycetes isolated from marine environments have been proposed as a potential source of yet untapped specialised metabolites. In this study, two novel species, *Pseudonocardia abyssalis* sp. nov. and *Pseudonocardia oceani* sp. nov., isolated from deep Southern Ocean sediments are described, both in terms of their phenotypic and genomic characterization. Furthermore, the genomic architecture, with a focus on Biosynthetic Gene Clusters (BGC), across eight strains belonging to the two novel species was investigated. As a result, a total of 13 Gene Cluster Families (GCF) were identified, of which six GCFs comprise BGCs from both species, and one was annotated specifically of each species. Following genome analysis, a comparative mass-spectrometry based metabolomics analysis was carried out, including phylogenetically closely-related non-marine species, as well as other *Pseudonocardia* strains isolated from different marine environments. Then, genomics and metabolomics data were correlated through NPLinker, an unsupervised method for integrating paired omics data. As a result, it was demonstrated that the BGC evolution and distribution across strains is mainly shaped by phylogeny over any biogeographical pattern.

Furthermore, metabolomics and genome mining tools were used to assess the role of the specialised metabolites on interactions between members of the *Pseudonocardia* genus isolated from different marine environments. To this end, using challenge bioassays, antagonist interaction between nine *Pseudonocardia* spp. isolated from marine sediments and two non-marine species was examined and correlated to the strains' phylogeny and their metabolomics profile. To understand the spatial dynamics, matrix-assisted laser desorption/ionization (MALDI)-mass spectrometry imaging (MSI) was used to examine *in situ* species chemical interactions between the Southern Ocean strains and their closest relatives. This analysis showed that phylogeny is the main predictor of antagonistic interactions among free-living *Pseudonocardia* strains. Moreover, a group of parent ions produced by *P. abyssalis* were identified as possible metabolites involved in the inter-species interactions.

Overall, this work showcases the power of a combined genomics-metabolomics approach to investigate rare actinomycetes from understudied locations and have uncovered a wealth of both biosynthetic and chemical diversity for further investigation.

Previously Published Work

The results of chapter 3 have been published in the following research article:

Parra, J., Soldatou, S., Rooney, L. M., & Duncan, K. R. (2021). *Pseudonocardia abyssalis* sp. nov. and *Pseudonocardia oceani* sp. nov., two novel actinomycetes isolated from the deep Southern Ocean. *International Journal of Systematic and Evolutionary Microbiology*, *71*(9), 005032. https://doi.org/10.1099/ijsem.0.005032

List of Tables

List of Figures

Table of Contents

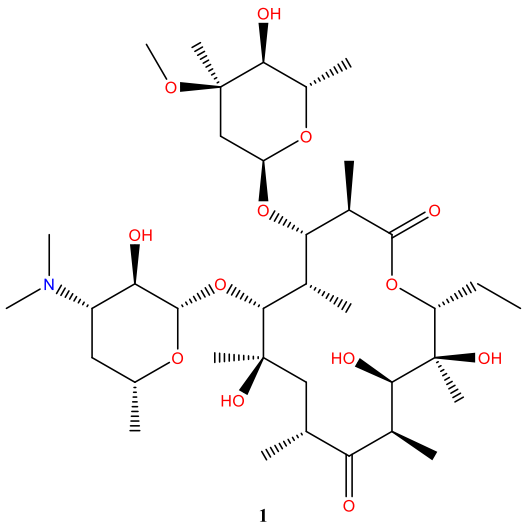
Chapter 1 Introduction

1.1. Antibiotics

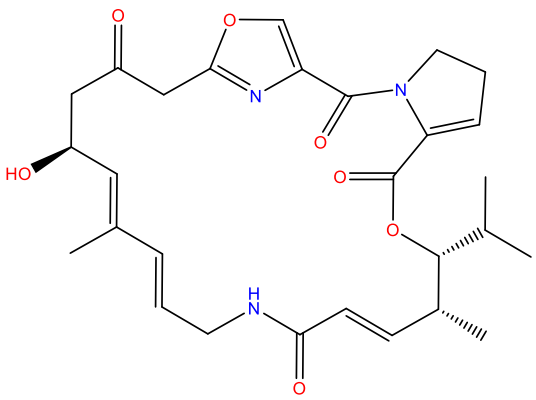
1.1.1. Antibiotics

The term ‘antibiotic’ was first proposed by Waksman (1941) to describe the antagonistic effect of *Streptomyces antibioticus* living in mixture with other microorganisms [1]. Later, this term was defined as *a chemical substance, produced by microorganisms, which has the capacity to inhibit the growth of and even to destroy bacteria and other microorganisms* [2]. Nowadays the word ‘antibiotic’ is used with a broader meaning to include designed molecules. Moreover the terms antibacterial and antifungal are used to describe specific actions against bacteria and fungi, respectively [3].

The antibacterial antibiotics are usually classified based on the cellular component or pathway they affect (Fig. 1-1). Additionally, they are classified according their growth inhibitory (bacteriostatic) or killing (bactericidal) activities [3, 4]. Most bactericidal antibiotics inhibit DNA, RNA, cell wall or protein synthesis. For example, macrolides like erythromycin A (**1**), bind to the 50 S ribosomal subunit, resulting in dissociation of peptidyl-tRNA from the ribosome. This mechanism is shared by other antibiotics such as lincosamides and streptogramins (**2**, streptogramin A) [5].



1



2

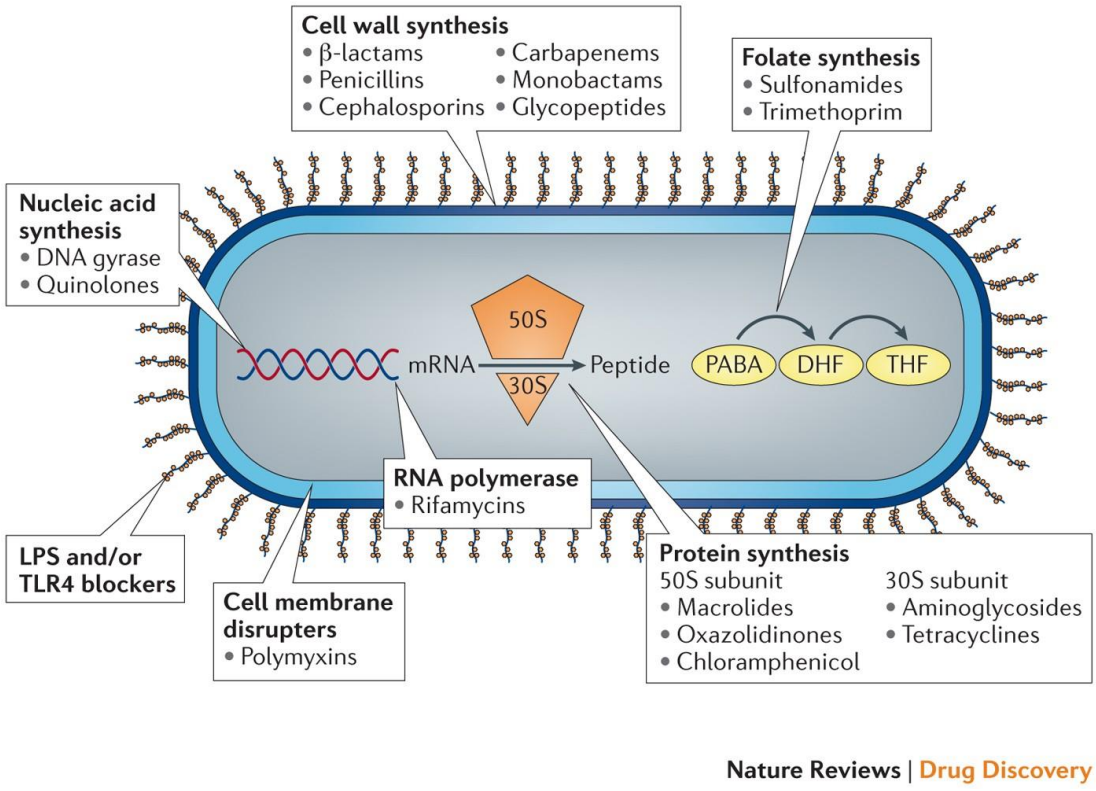
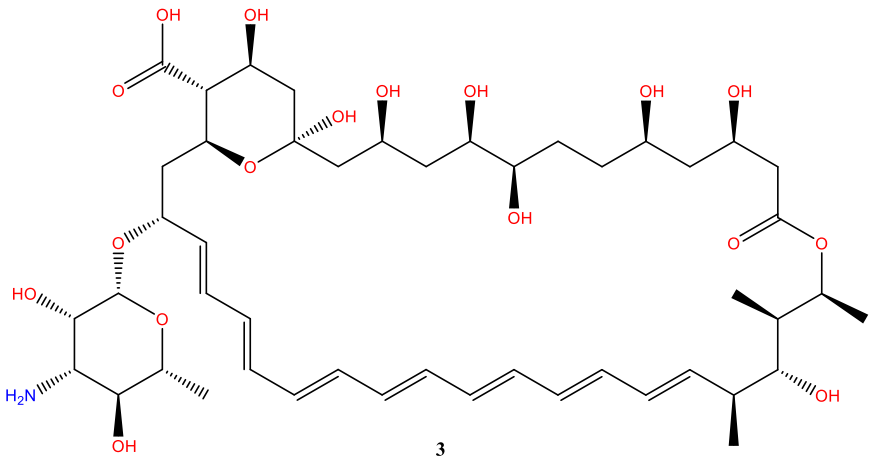


Figure 1-1 Main groups of antibacterial antibiotics in clinical use, organised according to their mechanism of action. Reproduced from Brown 2015 [6].

Similarly, antifungal antibiotics inhibit the spread of fungi by killing fungal cells or spores, or preventing their growth. Most of them do this through the alteration of the fungal cell wall and fungal cell membrane (Fig. 1-2) [7, 8]. For example, polyenes like amphotericin B (**3**) work by directly binding to ergosterol in the fungal cell membrane, forming pores that cause leakage of intracellular components and, ultimately, result in cell death [9].



3

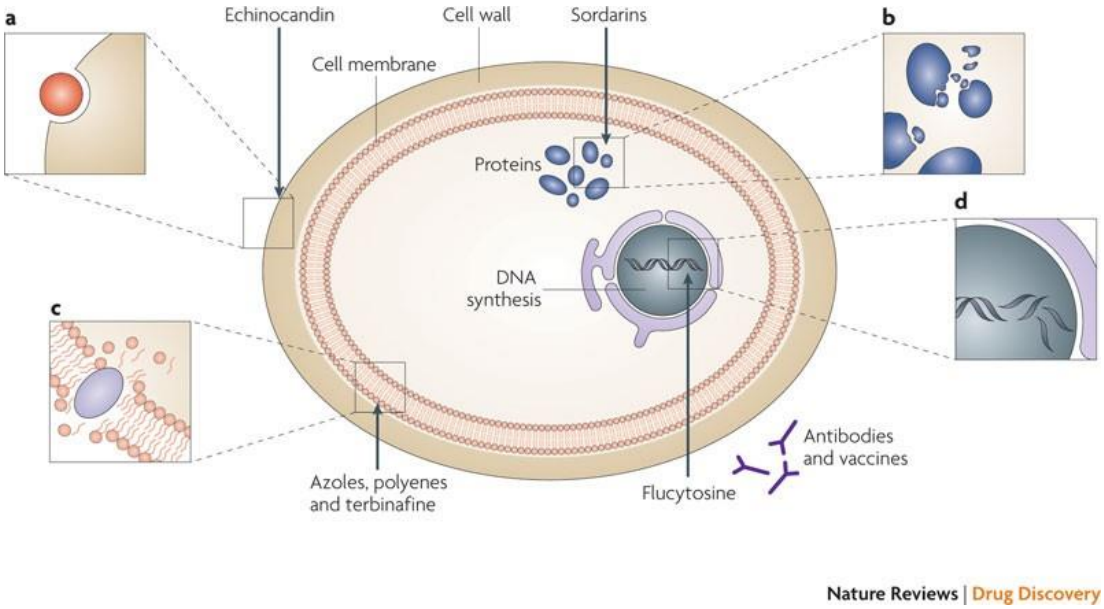


Figure 1-2 Main groups of antifungal antibiotics in clinical use, organised according to their mechanism of action. Reproduced from Ostrosky-Zeichner *et al*. 2010 [8].

Antibiotics represented a milestone in the history of medicine, allowing the effective treatment of diseases such as tuberculosis, bacterial pneumonia, and sepsis [10]. For example, the clinical use of penicillin reduced the case fatality rates for bacterial pneumonia and bloodstream infections from 80% to 17% between 1935 and 1952 [11]. However, during its discovery, Fleming (1929) noticed that the growth of a group of *E. coli* and other related bacteria were not inhibited by penicillin [12]. Later, Chain and Abraham (1940) made the first report that these bacteria produce an enzyme with the ability to destroy penicillin, which explained their resistance. Globally, as a result of increasing resistance to antibiotics, the costs of treatment and mortality are increasing with low- and middle-income countries especially impacted [13].

1.1.2. The antimicrobial resistance problem

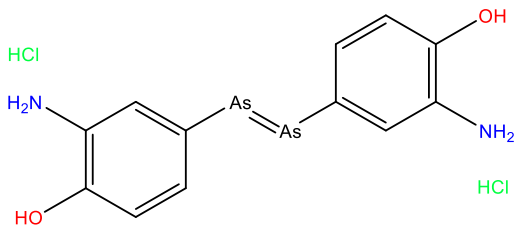
Antimicrobial resistance (AMR) is widely recognized as one of the major global health challenges of the 21st century [14, 15]. Globally, AMR caused more than 1.27 million deaths in 2019 [16], a figure that could rise to 10 million by 2050 if no actions are taken now to slow down the rise of drug resistance [17]. In this context, the World Health Organization (WHO) set a priority list for research and development of new antibiotics for antibiotic-resistant

bacteria that is led by multidrug-resistant and extensively-resistant *Mycobacterium tuberculosis*, followed by carbapenem-resistant *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and members of the Enterobacteriaceae family (*Escherichia coli*, *Salmonella typhimurium*, *Klebsiella pneumoniae*, *Enterobacter* spp.) considered as a critical priority [18]. Besides bacterial pathogens, climate change has facilitated the emergence of multidrug-resistant fungal pathogens like *Candida auris* [19, 20]. Although *C. albicans* remains the most frequently isolated *Candida* species in the clinical setting, the appearance of azole-resistant *C. auris* isolates has become a global public health threat due to its challenging identification and treatment [21, 22].

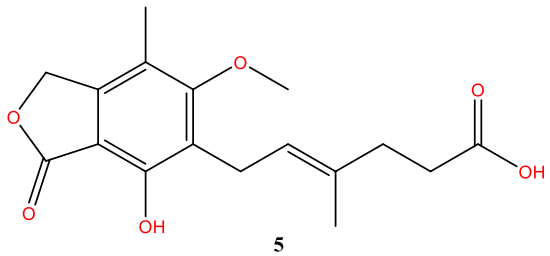
The emergence of multi-drug resistant bacterial and fungal strains could put at risk the achievements in health measures obtained during the so-called "antibiotics era" [23]. Consequently, the discovery of novel therapeutic alternatives against these emerging multi-drug resistant pathogens is essential to tackle the antimicrobial resistance crisis.

1.1.3. Antibiotic discovery

Arsphenamine (Salvarsan) (**4**), introduced in 1910 by Paul Ehrlich and Sahachirō Hata [24], was the first antimicrobial compound used in the clinic, for the treatment of syphilis and trypanosomiasis [10]. However, the first antibiotic discovered from nature was mycophenolic acid (**5**). Isolated from *Penicillium brevicompactum* by Bartolomeo Gosio (1896), mycophenolic acid showed activity against *Bacillus anthracis* [25, 26].

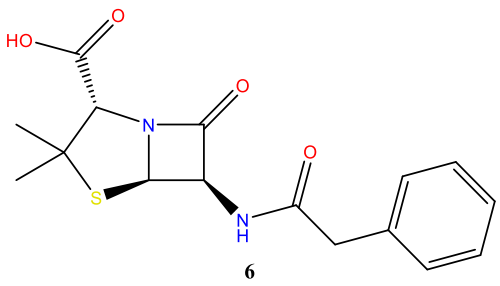


4

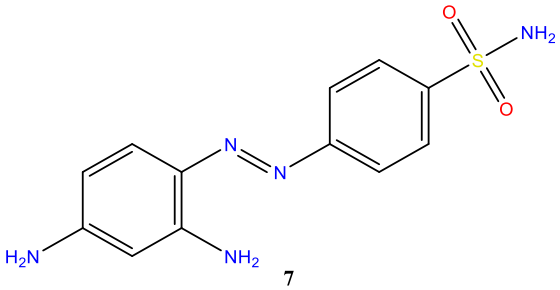


5

These early discoveries of Salvarsan and mycophenolic acid preceded the discoveries of penicillin (**6**, penicillin G), described by Alexander Fleming (1929) from *Penicillium rubens* [12], and sulfamidochrysoidin (Prontosil) (**7**) synthesised by Gerhard Domagk (1932) [27]. With the later purification of penicillin and its clinical application by Chain *et al*. (1940) [28, 29], the modern antibiotic era began [23].

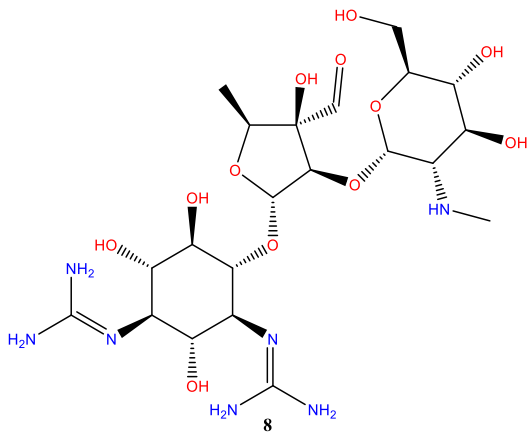


6

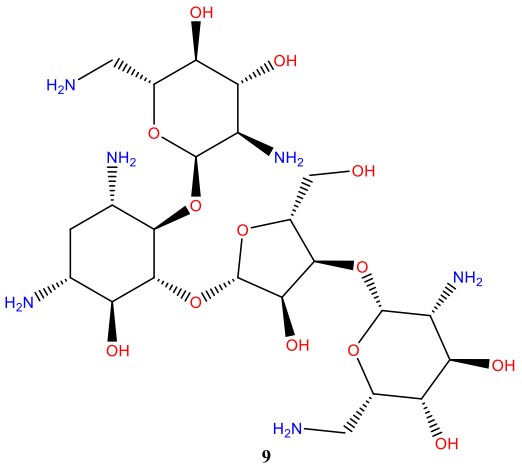


7

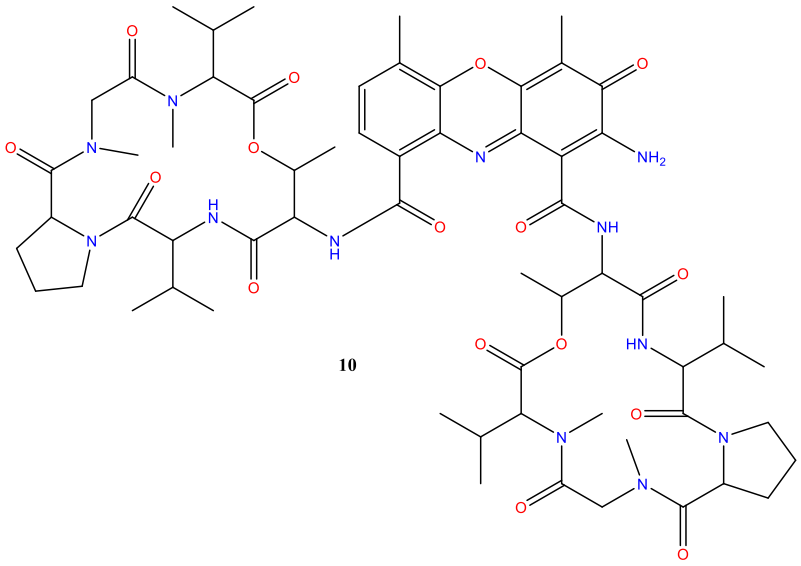
After the discovery of penicillin, Selman Waksman developed a platform that systematised the search of antimicrobial activity from soil bacteria, particularly focussed on the *Streptomyces* genus [30, 31]. Following this strategy, several antibiotics were discovered, including streptomycin (**8**) [32], neomycin (**9**) [33], and actinomycin D (Dactinomycin) (**10**) [34]. Due to its success, the Waksman platform was widely adopted by the pharmaceutical industry, leading to the discovery of several classes of antibiotics over the next twenty years [30, 35]. As result, this period is known as the "golden age" of antibiotic discovery [31, 36, 37] (Fig. 1-3).



8



9



10

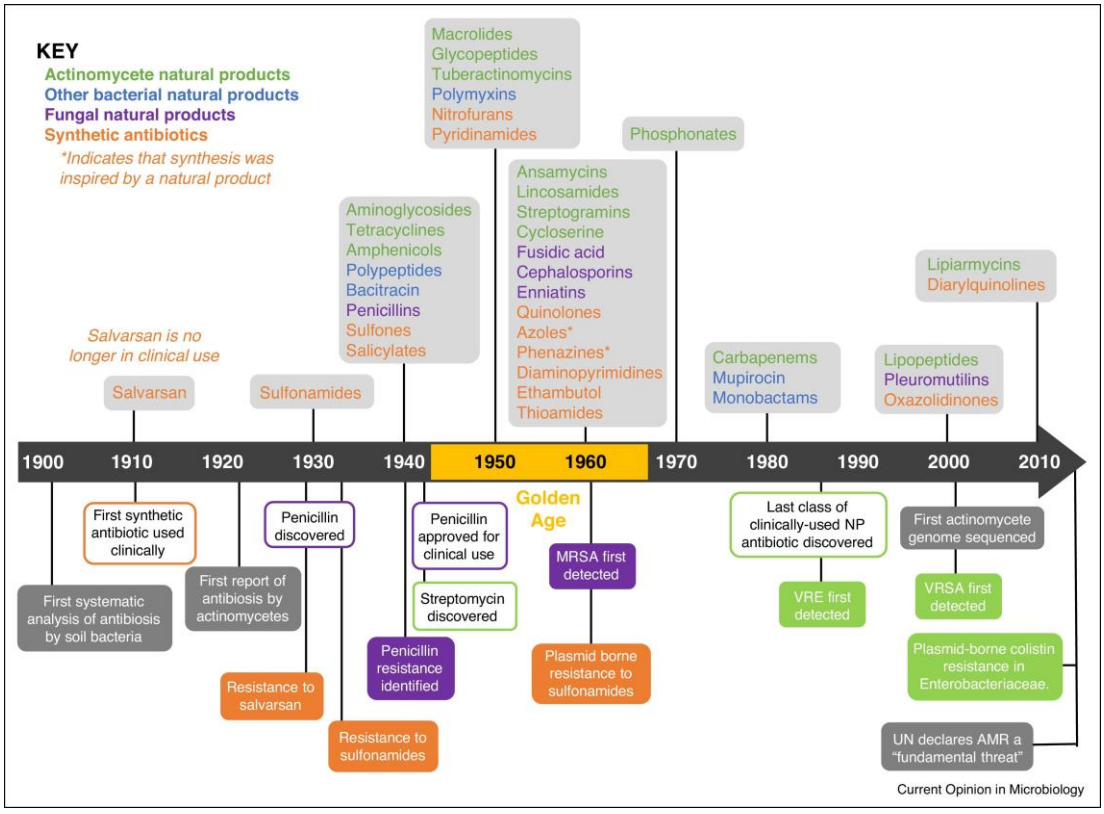
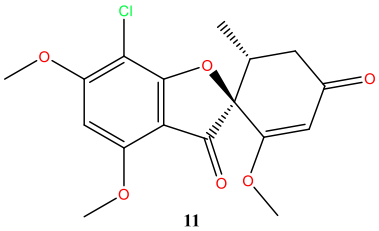
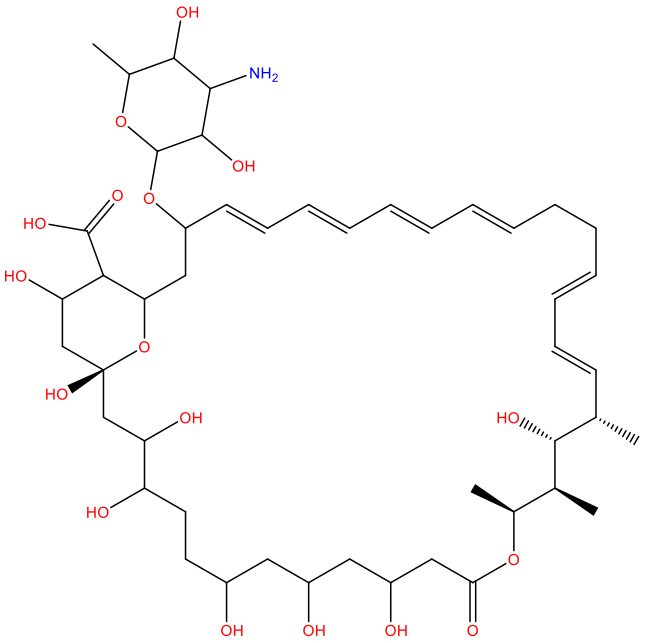


Figure 1-3 Antibiotic discovery timeline showing the new classes of antibiotic discovered in each decade (top) and milestones related to antibiotic discovery and antimicrobial resistance (bottom). Antibiotics from actinomycetes are coloured in green, while blue is used for other bacteria, purple for fungi, and orange for those of synthetic origin. Reproduced from Hutchings *et al*. 2019 [31].

The golden age of antibiotic discovery also resulted in the discovery of important antifungal compounds [38]. The first antifungal described was griseofulvin (**11**), isolated from *Penicillium griseofulvum* by Oxford *et al*. (1939) [39]. However, this compound was not commercially available until 1959 after its bioactivity was clinically proven [40, 41]. Early in the 1950s the first polyene macrolide antifungal, nystatin (**12**), was discovered from *Streptomyces noursei* by Rachel Fuller-Brown and Elizabeth Lee-Hazen [42]. Later, in 1953, amphotericin B (**3**) was isolated from a *Streptomyces nodosus* culture at the Squibb Institute for Medical Research [43].



11



12

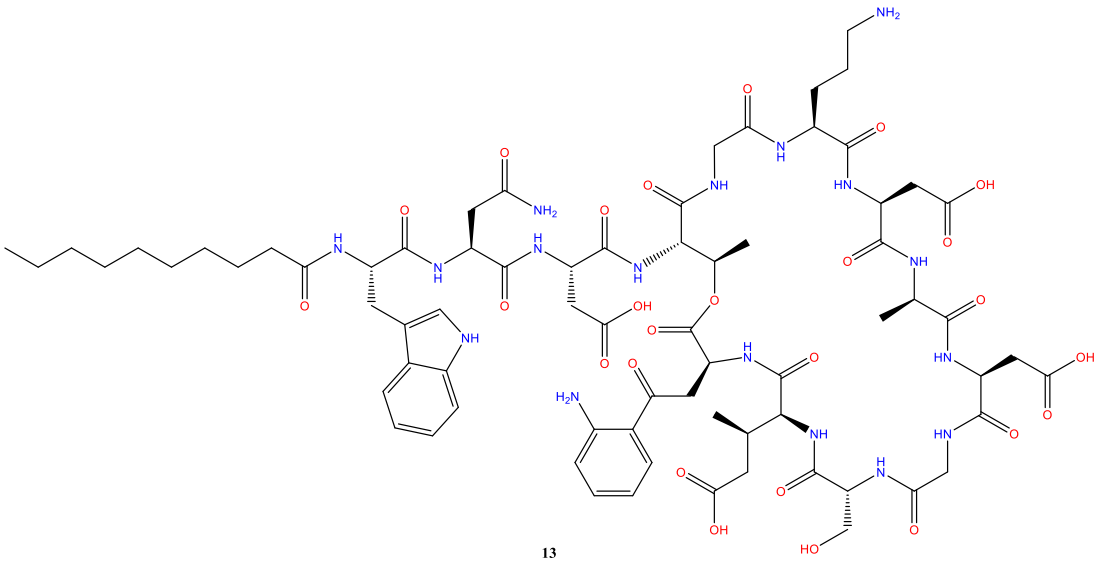
The end of the golden age was marked by a drop of the rate of new chemical discoveries [36, 37]. By the mid-1960s, new antibiotic scaffolds were becoming harder to discover using the Waksman platform. Then, this period was followed by the medicinal chemistry era, where antibiotic scaffold discovery was the focus for synthetic modification [44]. For example, after the synthesis of miconazole in 1969 [45], the discovery of this antifungal was dominated by synthetic azole compounds as they represented a less toxic treatment for invasive fungal infections [46]. However synthetic compounds from combinatorial libraries also failed to provide new antibiotic scaffolds [44]. Then, the focus once again returned to natural products, in particular understanding the biological and ecological function of metabolites using genomics and synthetic biology tools for new chemistry [36].

1.2. Natural products

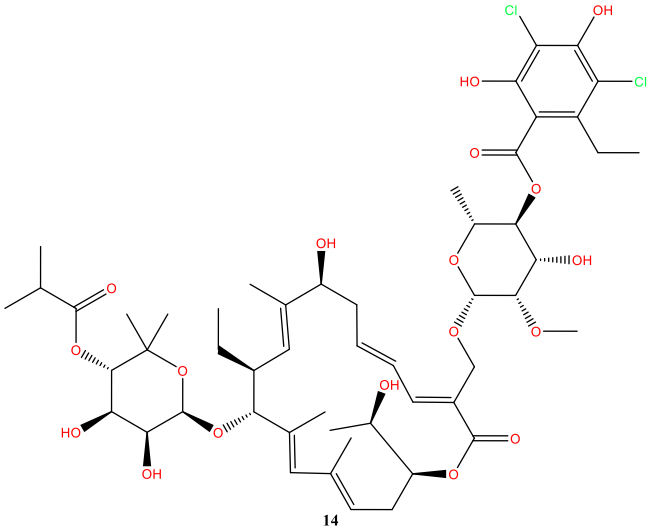
1.2.1. Natural products

Natural products are broadly defined as chemical compounds produced by living organisms [47]. Natural products like penicillin and streptomycin have been a successful source of antibiotics used in the clinic. For example, 32% of the small molecules approved as new drugs by the USA Food and Drug Administration (FDA) between 1981 and 2019 were natural products or their derivatives; accounting for approximately 70% of approved small molecule antibiotics [48]. Despite this historical importance, some authors argued that the continuous examination of natural products, in particular from actinomycetes, as a source of antibiotics is supported by past successes [49]. In addition, it has been suggested that efforts in antibiotic discovery based on natural products tend to be redundant, re-discovering targets, mechanisms of action and molecular scaffolds [50]. As a result, most antibacterial compounds in clinical development are derivatives of old antibiotic classes such as β-lactams and tetracyclines [51].

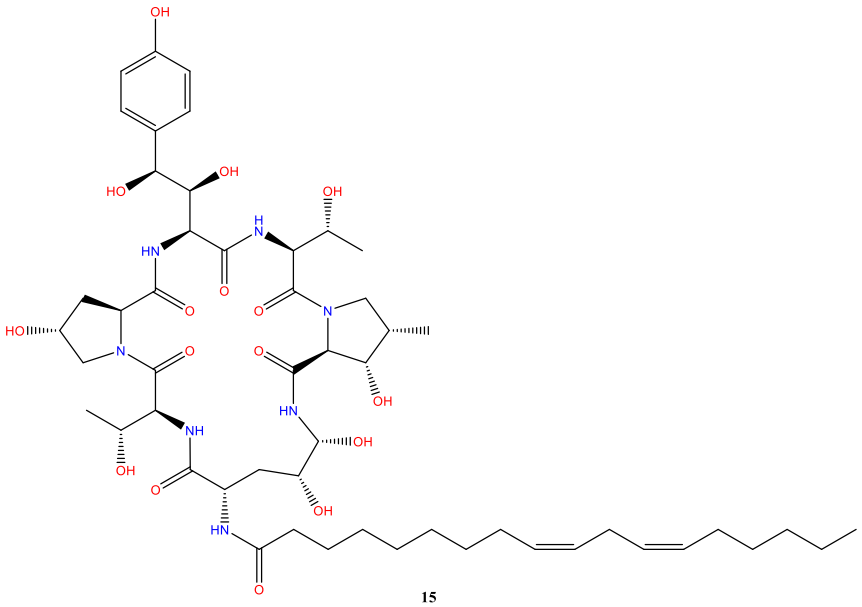
Due to the capability of testing a large number of compounds, combinatorial chemical libraries overcame fermentation screening platforms, partially explaining why the pharmaceutical industry abandoned natural product discovery projects in the 1990s [50, 52]. However, the emergence of 'omics technologies, as will be discussed later, has revealed an unexplored chemical potential [53, 54]. In addition, novel antibiotic classes from natural products have been discovered after the golden age. For example, during the period 2000-2019, 38 new antibacterial drugs were approved, of which two correspond to natural products and 16 were derived of natural products [55]. Of these new antibacterials, five were first-in-class, of which, two were natural products, and one natural product derivative [55]. These natural products that were new antibiotic classes correspond to daptomycin (**13**), discovered in the late 1980s from *Streptomyces roseosporus* [56, 57]; and fidaxomicin (tiacumicin B) (**14**), isolated from *Dactylosporangium aurantiacum* in 1986 [58]. Similarly, echinocandins like Caspofungin, the newest class of antifungals [8], are natural product derivatives based on lipopeptides like echinocandin B (**15**) isolated from *Aspergillus nidulans* in 1974 .



13



14



15

Several of the antibiotics mentioned so far, such as streptomycin, actinomycin D, and amphotericin B, showcase the relevant role of actinomycetes in antibiotic discovery. In fact, actinomycetes are unsurpassed for the production of bioactive natural products; accounting for 40% of the bioactive microbial metabolites discovered from the 1940s to 2010. Moreover, more than 70% of the natural products produced by actinomycetes has been shown to have antimicrobial activity [59]. Furthermore, actinomycetes isolated from marine environments have been described as a promising source of new antibiotics [60, 61]. For example, during the period 2010-2015, 50 bioactive compounds were obtained from marine-derived bacteria, and 69% of them were obtained from actinomycetes; additionally, these compounds showed high chemical diversity, with the identification of 10 new molecular scaffolds [62].

1.2.2. Marine natural products

The oceans cover 71% of the planet surface and are considered the largest ecosystem on Earth. In particular, deep oceans are unexplored environments, with less than 5% explored with remote instruments and less than 0.01% of the deep sea-floor sampled, despite evidence that they are one of the most biodiverse ecosystems [63]. For example, by the end of 2016, approximately 28,500 compounds were reported from marine sources, and six of them have been approved as drugs [64, 65]. Some examples of clinically-used marine-derived drugs are eribulin [66], a semisynthetic derivate of the polyether macrolide halichondrin B isolated from the marine sponge *Halichondria okada* [67], and trabectedin (ecteinascidin 743), an alkaloid originally isolated from the sea squirt *Ecteinascidia turbinata* [68]. Both have been approved for the treatment of metastatic breast cancer, and soft tissue sarcoma and ovarian cancer, respectively. These examples show the importance of marine natural products for drug discovery [69].

Early marine natural products discovery programs were focused on marine invertebrates like sponges and tunicates. However, symbiotic microorganisms, such as actinomycetes and cyanobacteria, have been recognized as the real source of several of these bioactive metabolites [70]. For example, an analysis of clinical and preclinical marine natural products discovered by the end of 2010 showed that despite being collected from marine macroorganims, 80% of the clinical trial and approved pharmaceutical agents discovered are predicted to be produced by bacteria and cyanobacteria (Fig. 1-4) [70].

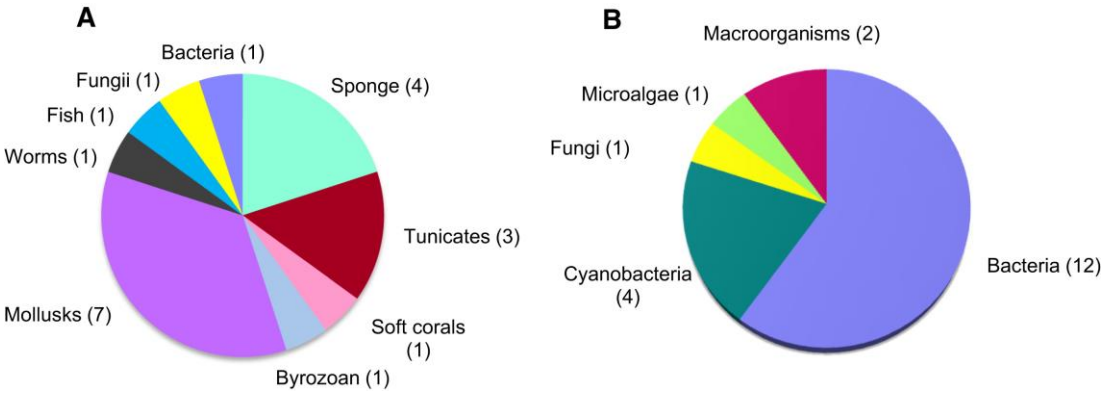


Figure 1-4 Pie charts showing the collected sources (A) and predicted biosynthetic sources (B) of 20 marine derived or inspired drugs and clinical trial agents by the end of 2010. Reproduced from Gerwick and Moore 2012 [70].

An example of a marine natural product produced by an a microorganism associated with a macroorganism is the cytotoxic polyketide polyether okadaic acid, which was isolated originally from the sponge *Halichondria okadai* [71]. However, the dinoflagellate *Prorocentrum lima* was later identified as the metabolite producer [72]. Another example are the diketopiperazines *cyclo*-(L-Pro-L-Leu), *cyclo*-(L-Pro-L-Val), and *cyclo*-(Pro-Ala) reported as metabolites of the marine sponge *Tedania ignis* [73], but later identified as bacterial metabolites of the actinomycete *Micrococcus* sp. [74]. Moreover, the close relationship between the structure of trabectedin and other bacterial metabolites like saframycin A from *Streptomyces lavendulae* [75], saframycin Mx1 from *Myxococcus xanthus* [76], and safracin B from *Pseudomonas fluorescens* [77] suggests trabectedin has a microbial origin. In fact, a metagenomic analysis of the *Ecteinascidia turbinate* microbiome proposed the bacterial symbiont *Candidatus Endoecteinascidia frumentensi* as the real trabectedin producer. This was because biosynthetic genes congruent with the molecule were identified in its metagenome-assembled genome (MAG). However, as the putative microbial producer has not been cultured, confirmation has not been possible [78, 79]. Similarly, metagenomics analysis revealed that misakinolide A, originally isolated from the sponge *Theonella swinhoei* [80], is produced by an uncultivated 'Entotheonella' symbiont. In this case, the heterologous expression of the biosynthetic genes, as well as the co-localization of 'Entotheonella' and misakinolide A using a combination of CARD-FISH (catalysed reporter deposition-fluorescence in situ hybridization) and MALDI-IMS (matrix-assisted laser desorption ionization-mass spectrometry imaging) confirmed the biosynthetic origin of misakinolide A [81]. The description of the microbial origin of several marine natural products has

contributed to the search for new bioactive metabolites in marine environments being increasingly directed towards microorganisms. For example, a recent review showed that from 2015 to 2019, most of the new reported sources of marine natural products in the literature correspond to microorganisms, led by fungi and followed by bacteria (Fig. 1-5) [82].

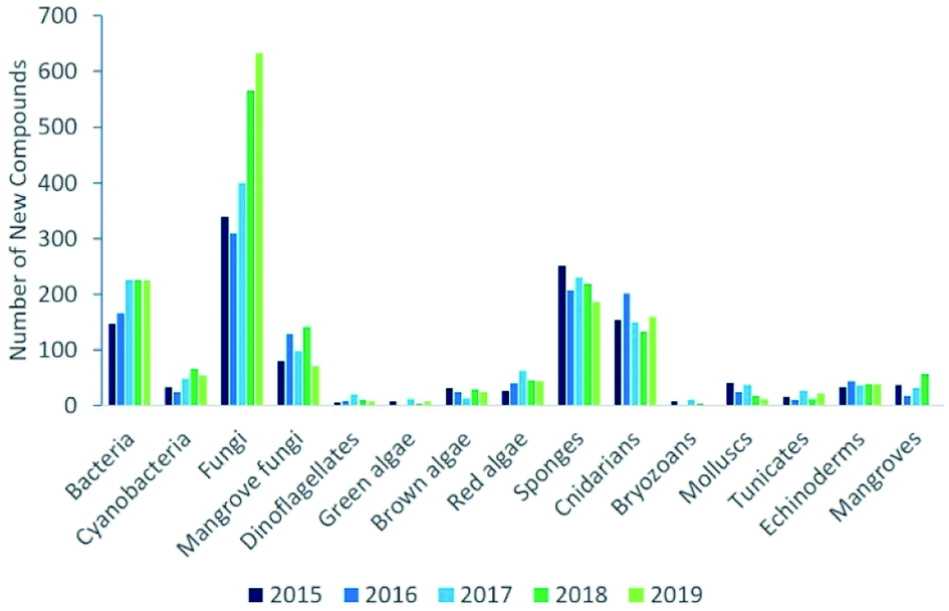
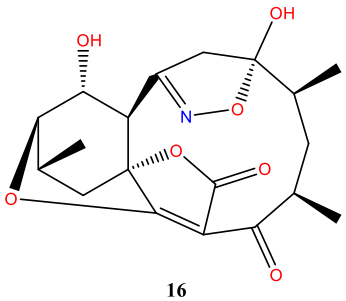


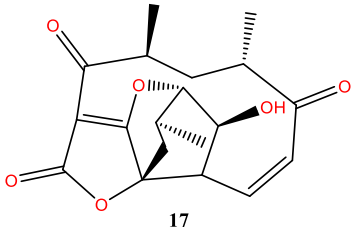
Figure 1-5 Sources of new marine natural products reported over the period 2015–2019. Reproduced from Carroll *et al.* 2021 [82].

1.2.3. Marine natural products from actinomycetes

Among microorganisms of marine origin, actinomycetes have been proposed as a promising source of chemically diverse natural products [83–85]. Deep-sea actinomycetes in particular have gained attention for antibiotic discovery [86, 87]. For example, the abyssomicins (**16**, abyssomicin B; **17**, abyssomicin C) were isolated from *Verrucosispora maris* AB-18-032 isolated from a sediment sample collected from the Sea of Japan at a depth of 289 m. These metabolites through the inhibition of the *p*-aminobenzoate biosynthesis showed antibiotic activity against multi-drug resistant clinical isolates of *Staphylococcus aureus* [88, 89].

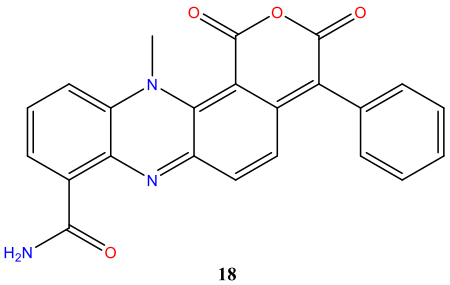


16

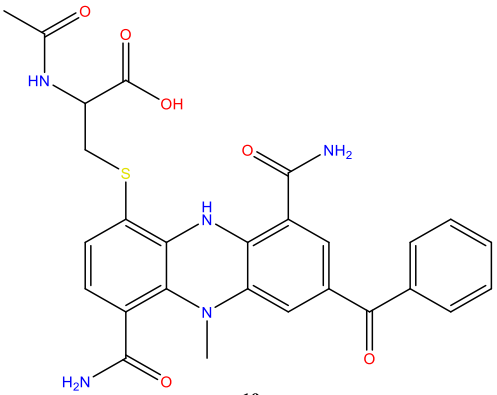


17

Another example is the dermacozines (dermacozine F, **18**; dermarcozine J, **19**) produced by *Dermacoccus abyssi* isolated from Mariana Trench sediments collected at a depth of 10 898 m. Despite no antibiotic activity reported, some members exhibited moderate cytotoxic activity against leukaemia cells [90, 91].

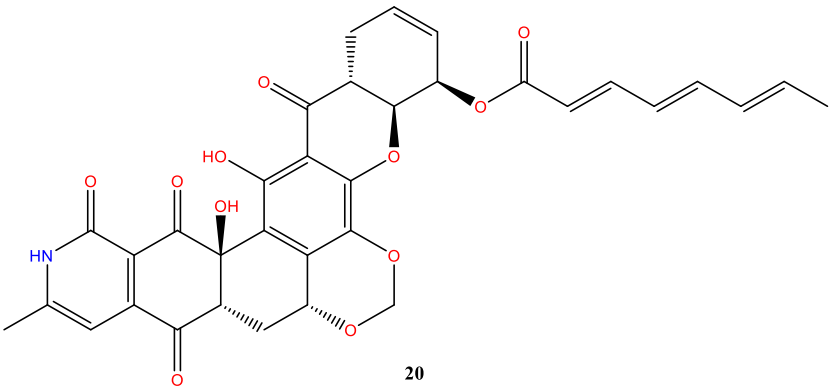


18



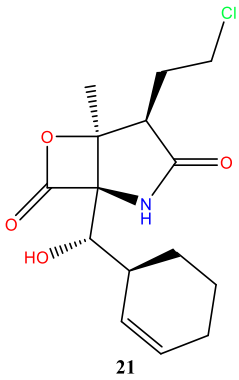
19

Actinomycetes from marine origin have also been a source of antifungal compounds. For example, turbinmicin (**20**) was shown to inhibit the fungal vesicle-mediated trafficking pathway, and demonstrates broad-spectrum antifungal activity, including against multi-drug resistant *C. auris* [92]. Tubinmicin is produced by *Micromonospora* sp. isolated from the ascidian *Ecteinascidia turbinate*, after a metabolomics-guided screening of 1482 actinomycetes from marine invertebrates [92].



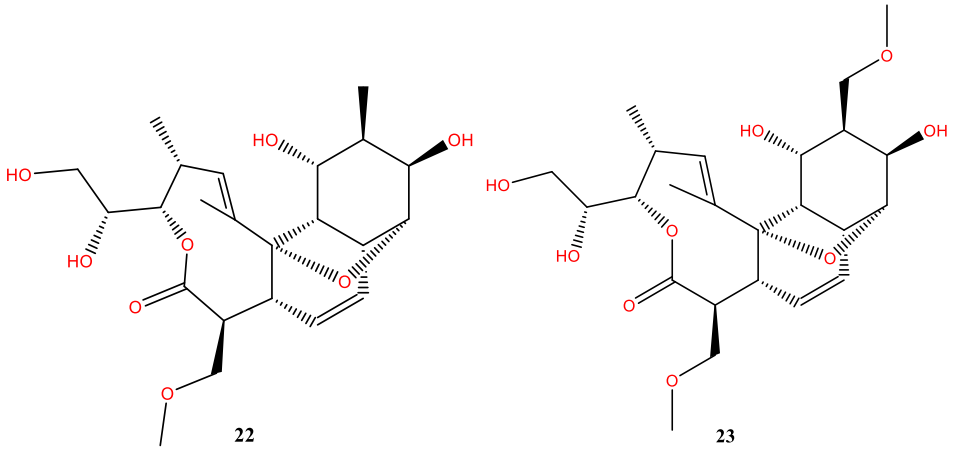
20

However, the best-known example of a bioactive metabolite isolated from marine-derived actinomycetes is that of salinosporamide A (Marizomib) (**21**) originally isolated from the obligate marine bacterium *Salinispora tropica* CNB-392. This metabolite displays potent inhibitory activity against the chymotrypsin-like proteolytic activity of the 20S proteasome [93]. Now, salinosporamide A is being studied as a treatment for glioblastoma in phase three clinical trials [65, 94].



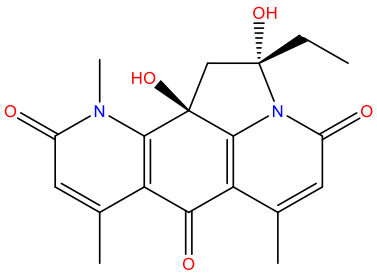
21

In the case of the *Pseudonocardia* genus, few bioactive metabolites from marine origin strains have been reported. For example, from *P. carboxydivorans* M-227 isolated from a deep-water sample collected from the Cantabrian Sea at a depth of 3000m, two bioactive metabolites called branimycins (branimycin B, **22**; branimycin C, **23**) were described. These compounds displayed antibiotic activities against several Gram-positive bacteria, including *Corynebacterium urealyticum*, *Clostridium perfringens*, and *Micrococcus luteus*, and against the Gram-negative bacterium *Neisseria meningitidis*. Additionally, branimycin B showed moderate antibacterial activity against *Bacteroides fragilis*, *Haemophilus influenzae*, and *Escherichia coli*, whilst branimycin C was bioactive against *Enterococcus faecalis* and methicillin-resistant *Staphylococcus aureus* [95].



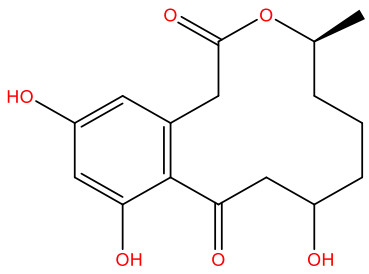
22 **23**

Furthermore, a new group of bioactive metabolites called pseudonocardians A-C (pseudonocardian B, **24**) were isolated from *P. antitumoralis* SCSIO 01299 from a sediment sample collected at the depth of 3258 m in the South China Sea. Besides exhibiting potent cytotoxic activity, these metabolites were bioactive against *Staphylococcus aureus*, *Enterococcus faecalis* and *Bacillus thuringensis* [96].



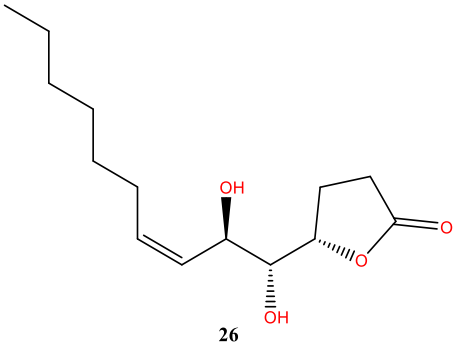
24

Another success story includes five curvularin macrolides, from *Pseudonocardia* sp. HS7 (also identified as *P. antitumoralis* according to 16S rRNA gene sequencing), isolated from the cloacal aperture of a sea cucumber. Two of these metabolites, (11R,15R)-11-hydroxycurvularin (**25**) and curvularin-7-O-α-D-glucopyranoside, exhibit inhibitory activity against *Escherichia coli* [97].



25

Finally, *Pseudonocardia* sp. YIM M13669 was selected from a screening assay for having antibacterial activity against *Mycobacterium smegmatis* MC^2 155, a non-pathogenic *Mycobacterium* used as a model system to study *M. tuberculosis*. This *Pseudonocardia* strain from South China Sea marine sediment (depth of 2448 m) was shown to produce seven γ-butyrolactone metabolites, named pseudonocardides A-G (pseudonocardide A, **26**). However, none of the pseudonocardides showed inhibitory activity against *M. smegmatis*, indicating that other metabolites may be responsible for the activity originally observed [98].



26

These examples illustrate the potential of actinomycetes, in particular non-*Streptomyces* strains, as a source of chemically diverse natural products. In fact, the examination of natural products recently isolated from actinomycetes from diverse origins supports that this group of bacteria is still one of the most important sources of chemical diversity [99].

1.3. Actinomycetes

1.3.1. Overview

Actinomycetes are filamentous bacteria within the phylum Actinobacteria. They are Gram-positive bacteria, with high G+C DNA content, representing one of the largest bacterial phyla with a ubiquitous distribution in both aquatic and terrestrial ecosystems. Most Actinobacteria are free-living organisms with a mycelial lifestyle and complex morphological differentiation (Fig. 1-6) [100].

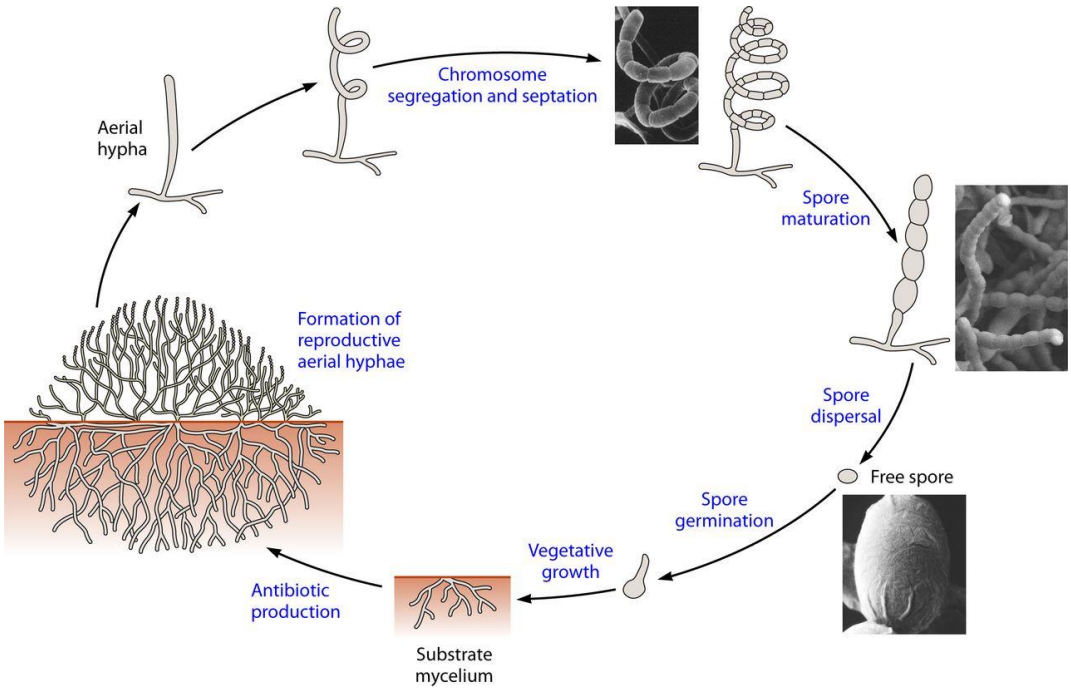


Figure 1-6 Schematic representation of the life cycle of sporulating actinomycetes. Reproduced from Barka *et al*. 2015 [100].

According to Bergey's Manual of Systematics of Archaea and Bacteria [101], the phylum Actinobacteria encompasses five classes, 19 orders, 50 families, and 221 genera. The constituent classes are Acidimicrobiia, Actinobacteria (class.), Coriobacteriia, Rubrobacteria, and Thermoleophilia [101]. Although the International Code of Nomenclature of Prokaryotes (ICNP) has recently updated the name of the phylum Actinobacteria to Actinomycetota [102], for the purposes of this work they will be referred as Actinobacteria.

Most Actinobacteria are saprophytic, soil-dwelling organisms that spend the majority of their life cycles as semi-dormant spores, particularly under nutrient-restricted conditions. However, this group contains members which are adapted to diverse environments [100]. For example, members of the genera *Mycobacterium* [103] and *Nocardia* [104] are human pathogens, while members of the genus *Bifidobacterium* are ubiquitous inhabitants of the mammalian gut microbiota [105]. Similarly, some members of the *Leifsonia* genus have been isolated from extreme environments, like the psychrophilic strains isolated from a cyanobacterial mat sample collected in Antarctica [106], while other members are plant-commensals [106] or pathogens [107].

Filamentous actinomycetes are prolific producers of antimicrobial metabolites, accounting for 64% of the known classes of naturally occurring antibiotics [31]. However, the exploration of this chemical potential has been focused in the *Streptomyces* genus. For example, the Natural Products Atlas (NPAtlas), a database of known microbial natural product structures, reported a total of 7,501 natural products related to Actinobacteria, of which 5,605 (approximately 75%) were associated to Streptomycetales [108, 109]. Moreover, rare actinomycetes such as *Salinispora*, *Micromonospora* and *Nocardopsis*, have been and still are the source of novel bioactive compounds, suggesting an untapped chemical potential [110].

1.3.2. Rare actinomycetes

Rare actinomycetes are defined as actinomycetes within the order Actinomycetales but not belonging to the genus *Streptomyces* [111] and often named such due to their less frequent isolation under standard lab conditions [112]. Examples of rare actinomycetes include the genera *Actinomadura*, *Actinoplanes*, *Amycolatopsis*, *Dactylosporangium*, *Kibdelosporangium*, *Microbispora*, *Micromonospora*, *Planobispora*, *Streptosporangium* and *Planomonospora* [112], as well as *Pseudonocardia*, which is the main focus of this project.

Despite being isolated less frequently, rare actinomycetes are widely distributed in terrestrial and aquatic ecosystems [113]. In particular, culture-dependent works, which involves strain isolation, propose marine environments as rich sources of rare actinomycetes. For example, between 2007 and mid-2013, a total of 80 new species of rare actinomycetes were reported from marine environments, of which 20 were novel genera and 3 novel families. Furthermore, the families *Micromonosporaceae*, *Pseudonocardiaceae*, *Nocardioidaceae*, as well as the suborder *Micrococcineae* were the dominant taxa [114]. Isolation efforts using deep-sea sediments has also allowed the identification of taxonomically diverse actinomycetes. For instance, the analysis of diversity of cultured actinomycetes from Mariana Trench sediment (10,898 m, the deepest region on Earth), revealed the presence of *Streptomyces* alongside rare actinomycetes such as *Dermacoccus*, *Kocuria*, *Micromonospora*, *Tsukamurella* and *Williamsia* [115].

1.3.3. *Pseudonocardia* genus.

Pseudonocardia is a genus of the Actinobacteria family Pseudonocardiaceae, described for the first time in 1957 [116]. Their phenotypic details will be introduced in Chapter 3.

Members of this genus have been isolated from various habitats. Some examples are *P. petroleophila* isolated from soil collected in Germany [117], *P. nigra* from a rock sample collected in the Atacama Desert [118], *P. hydrocarbonoxydans* from aerial contamination of a silica gel plate [119], and *P. antarctica* from a moraine sample collected from the McMurdo Dry Valleys in Antarctica [120].

The *Pseudonocardia* genus has also been studied for its ability to metabolise organic environmental contaminants [121–123]. For example, *P. dioxanivorans*, isolated from contaminated industrial sludge can grow on 1,4-dioxane, a potential human carcinogen, as well as on tetrahydrofuran, a by-product of the polymers industry [124–126]. These kinds of substances are concerning groundwater pollutants, making *Pseudonocardia* a useful tool for bio-remediation purposes [124–126]. Some other species, including *P. tetrahydrofuranoxydans* [127], and *P. benzenivorans* [128], isolated from contaminated soils have been shown to have the ability to degrade organic solvents.

Several endophytic *Pseudonocardia* species have been described in association with the medicinal plant *Artemisia annua*. Particularly, from surface-sterilized roots, *P. bannaensis* [129], *P. artemisiae* [130], *P. xishanensis* [131], and *P. kunmingensis* [132] were isolated, while *P. seranimata* [133] was isolated from its leaves. These plant-associated *Pseudonocardia* strains have been reported to have regulated plant growth, as well as the production of artemisinin [134, 135]. Other endophytic *Pseudonocardia* spp. include *P. tropica* from *Maytenus austroyunnanensi* [136], *P. acaciae* from the roots of *Acacia auriculiformis* [137], *P. endophytica* from *Lobelia clavata* [138], *P. eucalypti* from *Eucalyptus microcarpa* [139], and *P. broussonetiae* from the roots of *Broussonetia papyrifera* [140].

In addition, the fungus-growing ants (tribe Attini) such as members of the genus *Acromyrmex*, which grow the fungus *Leucoagaricus* spp. as a food source, have developed symbiotic interactions with antifungal producing *Pseudonocardia* strains for protecting their fungal gardens against pathogens such as *Escovopsis* spp. [141, 142]. Furthermore, a recent study suggested that *Pseudonocardia* can also protect fungus-growing ants from the entomopathogenic fungus *Metarhizium anisopliae* [143]. These fungus-growing ants have developed complex cuticular structures for maintaining *Pseudonocardia* symbionts in an example of convergent evolution between actinomycetes and insects [144].

Additionally, members of the genus have also been isolated from marine environments. For example, the selective isolation of actinomycetes from marine sediments collected from the Japan Trench (Pacific Ocean), the Canary Basin (Atlantic Ocean), and selected fjords in Norway, revealed the widespread presence of *Pseudonocardia*, as strains of this genus were identified in all locations [145]. Similar studies performed with sediments collected in the Republic of Palau [146], the Arctic Ocean [147], and the bay of Valparaíso in Chile [148] also retrieved *Pseudonocardia* strains. In this context, several *Pseudonocardia* strains isolated from marine environments have been taxonomically described. For example, *P. sediminis* was isolated from a marine sediment sample from the South China Sea collected at a depth of 652 m [149]. Moreover, *P. antitumoralis* was also isolated from a South China Sea sediment sample, but from a different location at a depth of 3258 m [150], while *P. profundimaris* was isolated from marine sediment of the western Pacific Ocean at the depth of 7118 m [151]. Other examples include *P. mangrovi* isolated from a soil sample collected from a mangrove forest [152], and *P. ammonioxydans* isolated from coastal sediments [153]. In general, these examples illustrate how *Pseudonocardia* spp. are ubiquitous inhabitants of marine sediments.

Although the complex ecological relationship such as that observed with fungus-growing ants has not yet been described in marine environments, there are reports of *Pseudonocardia* strains living in association with macroorganisms. For example, two *Pseudonocardia* strains, phylogenetically related to *P. antarctica* according to the 16S rRNA gene sequence analysis, were cultured from Antarctic deep-sea sponges [154]. Similarly, a *Pseudonocardia* strain phylogenetically related to *P. carboxydivorans* was also isolated from the marine sponge *Hymeniacidon perleve* collected in the South China Sea [155]. Moreover, the species *P. kongjuensis* which was originally isolated from a gold mine cave [156] was later identified as part of the microbiome of marine sponges *Haliclona* spp. [157] and *Pseudonocardia* strains were isolated from the scleractinian corals *Galaxea fascicularis* and *Acropora millepora* [158]. Finally, three *Pseudonocardia* strains, phylogenetically related to *P. adelaidensis*, were isolated from Antarctic macroalgae. In the latter case, isolates exhibited an obligate requirement for seawater for growth, which has not been reported for the genus before [159].

1.4. Actinomycetes' specialised metabolism

1.4.1. Specialised metabolites

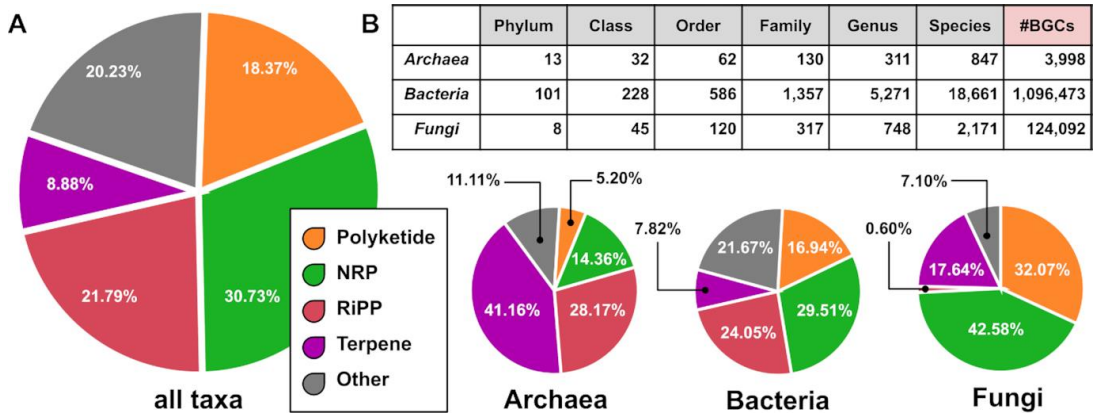
Specialised metabolites, are small organic molecules produced by living organisms presumably as a response to their adaptation to the environment [160]. Specialised metabolites are in principle natural products, but with a limited clade-specific or niche-specific distribution, known or presumed to have a specialised role in ecology or physiology [161]. Although these metabolites are believed to play an essential role in intra- and extra-cellular interactions between both cells and organisms, their specific ecological functions usually remain undescribed [162]. For instance, to kill or slow down the growth of another microorganism would reduce competition for nutrients. This would represent a selective advantage for survival, as such, it is widely accepted that antibiotics play a key role in microbial ecology. However, specialised metabolites may also be involved in more complex ecological functions, such as communication between cells [163, 164]. For example, extracellular concentrations of γ-butyrolactone SCB1 produced by *Streptomyces coelicolor* have been shown to induce actinorhodin and undecylprodigiosin production, suggesting a role as quorum-sensing signalling metabolites [165]. Moreover, experiments using a *Myxococcus xanthus Δta1* mutant, blocked myxovirescin [166] production, which revealed that this antibiotic plays a role in the predatory behaviour of these bacteria. Exogenous antibiotic supplementation restored the predation [167].

1.4.2. Specialised metabolites biosynthesis

The cellular machinery responsible for the biosynthesis of specialised metabolites is encoded by clustered groups of genes called biosynthetic gene clusters (BGCs). In general, the BGCs encode the enzymes, regulatory proteins and transporters that are necessary to produce, process and export a specialised metabolite. Hence, a BGC can be defined as a group of genes that together encode the proteins responsible for the biosynthesis of specialised metabolites [168, 169]. In bacteria, these genes tend to be physically clustered on the genome. However, sub-clusters located in different genomic loci can work together to produce a metabolite [161, 170]. For example, biosynthetic genes involved in the production of the glycosylate carotenoid sioxanthin are encoded in four separate regions of the *Salinispora tropica* CNB-

440 genome, consisting of two sub-clusters and two individual genes [171]. Similarly, the biosynthesis Lyngbyatoxin A by *Streptomyces* involves four genes in a sub-cluster (*tleABC*) and an individual gene (*tleD*) located in a different region [172]. Moreover, cluster situated regulators (CSRs) and global regulation genes can co-regulate BGCs and affect their expression despite being located in different genomic regions [173].

Specialised metabolites are often classified according to the types of enzymes involved in their biosynthesis. For example non-ribosomal peptide-synthetase (NRPS) BGCs are defined by a minimal module made by an adenylation (A) domain, a thiolation (T) domain and a condensation (C) domain [174]. Similarly, polyketide synthases (PKS) BGCs are characterised by a minimal module made by an acetyltransferase (AT), an acyl carrier protein (ACP) domain, and a ketosynthase (KS) domain [174]. In terms of microbial BGC diversity, a recent analysis of 1,225,071 BGCs was retrieved from 188,622 microbial genomes. This, combined with 20,584 MAGs and 1,910 experimentally characterized BGCs from the MIBiG (Minimum Information about a Biosynthetic Gene Cluster) database [175], was used to create the BiG-FAM database. Analysis of this dataset revealed that NRPS were the most common class of BGC found in microbial genomes, followed by ribosomally synthesized and post-translationally modified peptides (RiPPs) and PKS (Fig. 1-7). Similar analysis in 1,110 *Streptomyces* genomes [176], 43 *Amycolatopsis* genomes [177], 103 *Nocardia* genomes [178], and 75 *Salinispora* genomes [179] also showed that these three BGC classes were the most commonly found in actinomycete genomes.



	Phylum	Class	Order	Family	Genus	Species	#BGCs
Archaea	13	32	62	130	311	847	3,998
Bacteria	101	228	586	1,357	5,271	18,661	1,096,473
Fungi	8	45	120	317	748	2,171	124,092

Figure 1-7 Distribution of the collection of 1,225,071 BGCs contained in the BiG-FAM database. A. Pie charts showing the distribution of BGC classes across three microbial kingdoms. B. Taxa covered by BiG-FAM, showing the total number of BGC in each microbial kingdom. Reproduced from Kautsar *et al*. 2021 [180].

Cases of convergent evolution of BGCs, where two biosynthetic pathways evolved independently to produce the same metabolite, have been reported. For example, the biosynthesis of fosfomycin by *Pseudomonas syringae* follows a very different pathway from that of *Streptomyces fradiae* and *Streptomyces wedmorensis* [181]. On the other hand, a single pathway could be related to the biosynthesis of two or more metabolites [182]. Even taking these cases into account, in principle each specialised metabolite should be related to at least one biosynthetic gene. However, the number of characterised specialised metabolites far exceed the number of known BGCs. For example, the NP Atlas comprised a total of 32,552 metabolites by the end of 2021 [108], while MIBiG, a curated repository of known BGCs, contained 2,021 BGCs by early 2020 [175]. These finding suggests that many BGCs remain 'cryptic', which means that the corresponding metabolite has been observed, but where its associated BGC has not been experimentally identified [183]. In fact, it has been predicted that around 90% of BGCs in microorganisms are cryptic [184].

1.4.3. Genome mining of BGCs

Genome mining seeks to facilitate the discovery of specialised metabolites and their biosynthetic genes [161]. Classical genome mining techniques are based on the prediction of genes encoding enzymes putatively involved in specialised metabolite biosynthesis [185]. However, genome mining approaches have also provided insight into the function that metabolites perform in physiology and ecology [161]. Further information on the methods used for genome mining will be introduced in chapter 4.

Genome mining has revealed that the biosynthetic potential of microorganisms is much greater than that observed under standard laboratory conditions. This situation has reinvigorated specialised metabolite research [186].

1.4.4. Biosynthetic potential of actinomycetes

As mentioned above, by late 1960s, new antibiotic scaffolds from actinomycetes were becoming harder to discover and rediscovery was more frequent [59]. The high rate of rediscovery could be related to exhaustive isolation and misidentification of strains. For example, occasionally the detection of frequently recovered specialised metabolites come

from screening multiple isolates of common strains of *Streptomyces* spp. [187]. Although the apparent biosynthetic potential of actinomycetes might be biased by the relatively high number of sampled strains and sequenced genomes available, recent large-scale genome mining analysis confirmed that there is still a hidden biosynthetic potential encoded in actinomycetes' genomes. For example, using the BiG-SLiCE approach, Kautsar *et al*. analysed 1.2 million BGC from more than 200,000 publicly available microbial genomes, including MAGs, by grouping them into gene cluster families (GCFs). The diversity of the BGCs were assessed through the identification of cross-species patterns. This showed four *Streptomyces* spp. as the second richest GCF-containing species, just behind *Ralstonia solanacearum* and other fungal species [188]. In addition, a large-scale metagenome analysis of human microbiomes, which in theory is free of isolation biases, showed that actinomycete MAGs possessed the highest average number of BGCs per genome from the analysed dataset [189].

In the particular case of actinomycetes, similar genome mining studies have confirmed their biosynthetic potential. For example, 11,422 BGCs were classified into 4,122 GCFs across 830 genomes and extrapolation of this dataset estimated thousands of unknown BGC [190]. Furthermore, 178 strains were selected for metabolomics analysis which revealed 77% of their BGCs were silent as they contained BGCs of known metabolites that were not identified by mass spectrometry [190], confirming the gap between the number of known metabolites and their cognate BGCs. Furthermore, a study focused on chemotherapeutic gene clusters across 1,110 publicly available *Streptomyces* genomes confirmed a high diversity and abundance of BGCs across the genus [176] with differences between members of the same species observed. For example, of the 48 BGCs identified in *Streptomyces albus*, 18 were found across all strains, while 16 were observed only in one strain. This suggests a high strain-level BGC diversity [176]. These genome mining analyses have aided novel metabolite discoveries. For example, high-throughput genome mining of a collection of over 10,000 actinomycetes allowed the description of a group of phosphonate BGCs within the genome of 278 sequenced strains. The characterisation of this pathway led to the discovery of argolaphos, a broad-spectrum antibacterial phosphonopeptides, which represent a new type of sulphur-containing phosphonate natural product [191].

Despite the biosynthetic potential of actinomycetes, fewer than 10% of the sequenced microbial genomes available in public repositories belong to this group [168]. Since most of the genome sequencing projects are related to bacterial pathogens, the genomic information

of Actinobacteria in general, and rare actinomycetes in particular, is limited. For example, in the GOLD (Genomes OnLine Database) of the Joint Genome Institute (JGI), from a total of 196,937 registered bacterial genome sequencing projects, only 11.7% belong to Actinobacteria, 4.9% for Mycobacteriaceae, 1.7% for Streptomycetaceae, and only 0.3% for Pseudonocardiaceae family [192, 193]. As such comparative genome mining of non-*Streptomyces* is limited by the number of publicly available genomes. However, some studies suggested that rare actinomycetes harbour a slightly lower biosynthetic potential than *Streptomyces*. Pseudonocardiales in particular have a relative high incidence of BGCs (19.8 per-genome average) among actinomycetes, which is only surpassed by Streptomycetales (21.6 per-genome average) [190]. For example, a genomic study of 21 rare actinomycetes and 24 *Streptomyces* isolated from marine environments revealed that the genus *Streptomyces* had the highest number of BGC. A total of 1925 BGCs were found in *Streptomyces*, in comparison to 1386 for rare actinomycetes. However, the true diversity of these genes as a function of the Shannon Index normalized by number of BGCs, is higher for rare actinomycetes (0.2770) than for S*treptomyces* (0.1787), implying a higher novelty in the BGCs of marine-derived rare actinomycetes. In particular, *Corynebacterium* (0.9296), *Gordonia* (0.7579), *Nocardiopsis* (0.6128), *Saccharomonospora* (0.6057) and *Pseudonocardia* (0.6040) genera [194]. The Shannon Index is an entropy, giving the uncertainty in the outcome of a sampling process. When it is calculated using logarithms to the base two, it is the minimum number of yes/no questions required, on average, to determine the identity of a sampled species [195].

1.5. Evolution of specialised metabolism in actinomycetes

1.5.1. Ecological variables influencing specialised metabolism in actinomycetes

The consensus is that specialised metabolites work mainly as signalling or communication molecules that modulate interactions between organisms [162, 163, 196]. So, the study of interactions between microbes and other macro- and microorganisms could not only lead to the discovery of novel antibiotics, but it would allow further understanding of the function of specialised metabolites in the environment [197]. The fact that actinomycetes are found in diverse ecosystems underpinned by complex ecological interactions is a consequence of

their success as producers of specialised metabolites [198, 199]. The previously mentioned symbiosis between fungus-growing ants and *Pseudonocardia* is an excellent example of an ecological relationship which influences the evolution of specialised metabolisms by actinomycetes [200]. Polyene macrolides alongside other antifungal metabolites have been particularly well-studied from *Pseudonocardia* strains in a mutualistic relationship with fungus-growing ants [141, 201]. Some further details of antifungal metabolites from the *Pseudonocardia* genus will be addressed in chapter 5. Similarly, ecological relationships among beewolf digger wasps (*Philanthus* spp., Hymenoptera, Crabronidae) and *Streptomyces* spp., which are incorporated into the cocoon's silk for protecting larvae of bacterial and fungal infections have been found [202]. Another example is *Streptomyces* endophytes isolated from *Arabidopsis thaliana* that produce filipin-like polyenes with activity against the plant-pathogenic fungus *Gaeumannomyces graminis* [203].

An interesting example of metabolite ecological role is that of the earthy-smelling metabolites geosmin and 2-methylisoborneol [204, 205]. Genes responsible for the biosynthesis of these metabolites are well-conserved across the *Streptomyces* genus [206], but their ecological role had not been fully understood. Recently, experiments performed between *S. coelicolor* and the springtail *Folsomia candida*, a soil-dwelling arthropod, showed that these organisms were attracted to traps baited with *S. coelicolor* colonies [207]. This effect was explained by proving that geosmin and 2-methylisoborneol induced electrophysiological responses in the antennae of the springtails. Furthermore, the authors exhibit that the genes for biosynthesis of these metabolites are under the direct control of sporulation-specific transcription factors [207]. Overall, these results suggest that the production of geosmin and 2-methylisoborneol are an integral part of the sporulation process, completing the *Streptomyces* life cycle by facilitating dispersal of spores by soil arthropods [207].

The application of ecological concepts to the study of specialised metabolites leads to testable hypotheses regarding the roles of these molecules in shaping ecosystems [208]. For example, a genome mining study to characterise differences in the specialised metabolism focused on two related *Pseudonocardia* phylotypes associated with *Acromyrmex echinatior* fungus-growing ants. This work showed that some BGCs were specific to each phylotype [209]. Although some BGCs were conserved across both clades, each phylotype encoded different polyene antifungals. The antifungal metabolites encoded in these BGCs were found

to be fundamental traits for their ecological relationship with ants [209]. Similarly, two *Streptomyces* strains isolated from the marine sponge *Haliclona simulans* showed a correlation between phylotypes and their BGC profile [210]. These strains were shown to be closely related to terrestrial strains at a genomic level (91-99% similarity in six housekeeping genes), yet they potentially produce different compounds. This reinforced that ecological variables have a critical effect on shaping their specialised metabolism [210]. In this sense, the ecological context is undoubtedly related to the evolution of both microorganisms' lineages and their specialised metabolism. For example, the core-genome phylogenetic study of 120 *Streptomyces* spp. strains, of which 69 were related to insect microbiomes, showed that most insect-associated *Streptomyces* cluster together in discrete lineages [211]. Moreover, the evolutionary analysis of the BGCs annotated showed that their distribution within a *Streptomyces* genome is influenced by both phylogeny and ecology [211].

Beyond traditional screening based methods to look for bioactive compounds, understanding when, where, and why these metabolites are produced could facilitate exploring the untapped chemical diversity of microorganisms [212]. Although genome mining studies have allowed the identification of new biosynthetic pathways and metabolites, a further task is the functional characterization of these compounds since the interaction between bacteria and their environment will certainly involve specialised metabolites. Therefore, integrative approaches such as ecology-based mining (mining metagenomics data and using the ecological context to prioritize environmental BGCs) and functionality-based mining (use of protein domains for genome mining based on predicted function), are likely to accelerate the discovery of antimicrobial agents [213]. Beyond the discovery of novel chemistry, Medema (2018) proposes that the function of the specialised metabolites is what matters most and that the integration of cutting-edge omics technologies combined with foundational chemical and ecological concepts will provide many new insights into the chemical language of life [160].

Although the influence of the ecological context on specialised metabolite production in actinomycetes is a well-accepted concept [196, 199], the effect of geography in the evolution of its specialised metabolism is not completely clear. Whether specialised metabolism follow a sympatric or allopatric process of evolution is still an open debate [214, 215].

1.5.2. Biogeographic variables of the specialised metabolism in actinomycetes

Biogeography is defined as the science that attempts to describe and understand the spatial patterns of biological diversity [216]. As the study of biogeography enables an understanding of the distribution patterns of biodiversity across space and time, it allows surveying how the physiological, genetic, morphologic and behavioural characteristics of individuals and populations vary across its geographic ranges. Furthermore, the study of biogeography relates to the distribution of the species with their evolutionary history [216]. In the late 19th century and early 20th century, some microbiologists identified that the microbial distribution was determined by the environment, with Martinus Beijerinck as one of the most important proponents of this hypothesis [217, 218]. However, the study of microbial biogeography formally started with the work of Lourens Gerhard Marinus Baas Becking, who was influenced by the work of Beijerinck in environmental microbiology, and formulated in 1934 the hypothesis "*everything is everywhere, but the environment selects*", which has been the basis of most of the ecological theories in microbial distribution formulated later [217, 218]. This hypothesis is based on the idea that due to their abundance and size, microorganisms are not restricted by geographic barriers over geological time scales [219]. However, soil variables have been shown to shape their distribution [220]. Hence, it could be proposed that the local environmental factors, as well as the response of microorganisms to environmental changes, could describe the biogeographic patterns observed in some microorganisms, while regional factors such as dispersal could have less influence on them [221, 222]. In general, microbial ecologists agree that the development of new tools and strategies would enable all the factors involved in the distribution of microbial communities and their relation with the ecological process, such as dispersal, selection and speciation, to be considered [223–225].

Although most of the preceding studies on microbial diversity have been focused on taxonomy, a few of them applied a trait-based approach to assess the relationship between biodiversity and ecosystem functioning [226]. The study of the distribution of traits can be used to understand more complex ecological questions beyond the classical biogeographic distribution patterns. So, the principal aim of trait-based biogeography is to analyze not only the presence of some organisms in a specific area but the expression of some specific feature

as a response to ecological variables of this area [227]. For microorganisms, the best-studied trait for bacteria is antibiotic production, as it represents a key competitive strategy for soil microbial survival and performance [228]. In this sense, biogeographic patterns of BGC diversity have been studied to analyse the relationship between the geographic distribution of microbial communities and their specialised metabolism profiles. For example, the metagenomic characterisation of Australian soil microbial populations from 397 locations with PCR-amplification of AD (related to nonribosomal peptide biosynthesis) and KS (related to polyketide biosynthesis) domains revealed domain composition was correlated with pH and latitude of the collection sites [229]. A similar analysis was performed scaling-up to soil samples collected from 185 soil/sediments samples collected across five continents. The global abundance and comparative distribution of AD/KS sequences suggested a correlation between geographic distance and biome-type, and biosynthetic diversity [230]. Additionally, a culture-dependent study of 24 *Streptomyces* strains collected from soils of ecologically similar grassland sites spanning 6000 km across the continental United States revealed a phylogenetic distribution into two monophyletic groups (northern and southern clades) congruent with the geographic ranges delimited by latitude [231]. Moreover, the genome annotation retrieved a core group of nine BGCs ubiquitous in all samples, while two were specific in the Northern clade and six in the Southern clade, suggesting an influence of latitude in the BGCs distribution [231].

In the case of actinomycetes isolated from marine environments, the genus *Salinispora* has been studied as a model of microbial biogeography. For example, phylogenetic analysis based on 16S rRNA and *gyr*B (DNA gyrase subunit B) gene sequences of 152 *Salinispora* spp. strains isolated from different marine environments revealed that while *S. arenicola* showed a cosmopolitan distribution, *S. tropica* and *S. pacifica* had a restricted distribution [232]. However, the low genetic divergence between *S. tropica* and *S. pacifica* suggested that ecological differentiation rather than geographical isolation as a speciation force [232]. Moreover, the analysis of ketosynthase (KS) domains from type I polyketide synthase (PKS) genes of 30 *S. arenicola* strains isolated from six locations revealed three subpopulations for which the BGC were endemic to their locations [233]. Although the effect of biogeographic patterns over the specialised metabolisms is difficult to assess, some evidence showed that the co-occurrence of related *Salinispora* species suggests that diversification within this genus is the result of ecological differentiation more than of geographical isolation. Particularly, early reports on KS domains sequences of the salinisporamide A BGC (*salA*)

proposed that salinosporamide is only produced by the closely related *S. tropica* and *S. pacifica* with no evidence of inter-species recombination between the *salA* KS sequences, supporting the hypothesis of geographic isolation of these two lineages [234]. However, the later isolation of salinosporamide-producing *S. arenicola* strains isolated from temperate sediments of the Pacific Ocean in the Japan coasts (the highest latitude reported for this genus) suggested that the distribution of specialised metabolites genes in this genus may be wider than previously thought [235]. For *Salinispora*, the evidence suggests that phylogeny explains the distribution of species-specific metabolites regardless of the strains' origin. Moreover, the evidence of species-level vertical inheritance supports that specialised metabolites represent functional traits that help define *Salinispora* species [179, 236].

Biogeographic patterns in the production of specialised metabolites have also been studied for the *Pseudonocardia* genus. For fungus-growing ant-associated *Pseudonocardia* strains, it has been reported that their population dynamic and biosynthetic potential are structured by biogeography, where the distribution of BGCs is affected by the geographic isolation of the microbial communities. Particularly, the study of *Pseudonocardia* communities from a geographically isolated location such as an island showed a large number of monophyletic gene trees that distinguish these lineages from the mainland lineages. Likewise, the distribution of GCFs also followed isolation location as a number of GCFs were identified only in the island communities [237]. The Geographic Mosaic Theory of Coevolution (GMC) has been proposed for describing the role of biogeography in the evolutionary history of antimicrobial production by these *Pseudonocardia* communities [238]. The GMC concept was suggested previously to explain why the *Pseudonocardia* populations have a highly structured and limited dispersal over a relatively small scale, despite the well-known dispersal capacity of bacteria [239]. This theory postulates a tripartite hypothesis which suggests that (*a*) interactions may show natural selection mosaics, such that in different communities different traits are favoured; (*b*) the landscape may produce a subset of communities in which a strong reciprocal selection is occurring (hot spots); and (*c*) there is a continual geographic remixing of the range of coevolving traits as a consequence of gene flow, random genetic drift, and local extinction of populations [240, 241]. Recently, the genomic and metabolomics analysis of 42 ant-associated *Pseudonocardia* strains isolated from four different geographical locations in Brazil revealed that the novel non-ribosomal peptide attinimicin was produced by nearly two-thirds of all Brazilian strains, while it was not

detected in other previously studied strains. This represents the first specialised metabolite from ant-associated *Pseudonocardia* spp. with broad geographic distribution [242].

The GMC offers a framework to explain the biogeographic patterns of BGC distribution observed in symbiotic actinomycetes strains, but it does not explain the dynamics underlying the evolution of these BGCs. Some models suggest that even though closely related microorganisms in the same geographical location share many BGCs, genes among these BGCs are under diversifying selection that gives the advantage to produce diverse metabolites [243]. For example, the genomic region related to the siderophore pyoverdine was identified as one of the most divergent loci across *P. aeruginosa* genomes. A diversifying selection process could help to explain the divergence in pyoverdine genes and the diversity of pyoverdine molecular structures [244]. A similar evolutionary process has been proposed to explain the diversity in the prochlorosins, a group of lanthipeptides produced by marine cyanobacteria, for which the plasticity in the production of diverse structures represents a selective advantage [245]. So unless the processes governing the evolution of BGC are better understood, conclusions about the effect of geographic isolation or ecological niches occupancy on specialised metabolism are hard to assess.

1.5.3. Evolutionary dynamics on the biosynthetic gene cluster evolution

Understanding the evolutionary forces that shape chemical diversity in specialised metabolites can provide valuable information to streamline the search for chemical novelty. However, the evolutionary dynamics underlying BGC evolution remain understudied. Chevrette *et al*. (2020) proposed the Dynamic Chemical Matrix Evolutionary (DCME) hypothesis to explain the mechanisms behind specialised metabolism evolution [170]. Briefly, the DCME model suggests BGCs evolve from promiscuous enzymes (enzyme's ability to accept more than one substrate) of central or primary metabolism. This process begins with the expansion of an enzymatic family due to gene duplication and/or horizontal gene transfer (HGT). Then, the extra gene copy may be retained in the genome if it provides an advantage, serving as raw material for new metabolic pathways in a process called expansion-and-recruitment [246, 247]. Through positive selection, genes coding for these promiscuous enzymes might cluster in the genome, promoting increased interactions among promiscuous enzymes and increasing the probability of the formation of new metabolites. In this stage, the process of gene clustering, and consequent metabolic integration, promotes chemical diversity rather than specialisation of the emerging pathways. All new metabolites

experience different possible scenarios within the fitness landscape that determines whether BGCs are conserved or purged. For metabolites that provide a fitness advantage, selection drives cluster formation and maintains chemical diversity. Further positive selection and subsequent clustering produce a BGC, which could be subject to additional evolution events like HGT and genetic sweeps [170]. The reason why emerging genes clusters group into discrete BGCs is an open question. A potential explanation is that their interdependence in producing a single phenotype provides a fitness advantage [248].

An example of using evolutionary concepts for the discovery of novel bioactive metabolites is the analysis of resistance genes. For instance, using antibiotic resistance as a discriminating criterion and phylogeny-based BGCs screening, it was possible to discover the glycopeptides, pekiskomycin [249] and corbomycin [250]. Using the phylogeny of individual components of the glycopeptides BGCs, and reconciling them with a dated species tree to identify milestones in their evolution, it was shown that these BGCs arose 150–400 million years ago. Furthermore, the precursors of glycopeptide biosynthesis are far older than other components, confirming that these BGCs are the product of the gene recruitment from pre-existing pathways [251]. The same evolutionary analysis retrieved similar conclusions for streptomycin in the *Streptomyces* genus, which origin was dated to approximately 30-50 million years ago [248].

1.6. Metabolomics analysis of specialised metabolites

1.6.1. Metabolomics

Metabolomics is defined as the comprehensive study of small molecules in biological systems, which provides a direct measure of detectable metabolite production by organisms [252]. Although analytic techniques like Nuclear Magnetic Resonance (NMR) have been used for metabolomics analysis, Mass Spectrometry (MS) remains the most used instrumental technique [253]. MS metabolomics techniques allow the analysis of complex samples, which represents a highly valuable tool for the analysis of specialised metabolites. In fact, recent advances in MS metabolomics have focused on capturing the chemo-diversity of complex biological samples [254]. For example, when applied to a collection of 1,000 marine microorganisms, rapid dereplication of known molecules enabled the novel metabolites maridric acid A and B to be identified [255].

1.6.2. Mass spectrometry techniques

Mass spectrometry is an analytical technique based on measuring the mass-to-charge ratios (*m/z*) of molecular ions. In short, a mass spectrometer is made of an ionisation source and a mass analyser. First, the ion source ionises the analyte molecules (called 'metabolites' in the case of biological samples), and the resulting gas-phase ions are transported in an electric field to the mass analyser which calculates the *m/z* values [256]. In tandem mass spectrometry, also known as mass spectrometry/mass spectrometry (MS/MS), there are two spectrometers arranged one behind the other. After metabolites are ionised, the first mass detector generates a spectrum of the parent ion (MS1). Then, the kinetic energy in the collision chamber causes the ions to fragment, and these fragments are subsequently analysed in the second mass spectrometer (MS2). The fragmentation pattern of metabolites can be interpreted to determine the identity of the molecules in the sample [256]. The principle behind MS/MS metabolomics is that data acquisition under similar experimental conditions will yield highly similar MS/MS spectra, enabling identification of MS/MS spectra in libraries of reference compounds [257]. Recent advancements in mass spectrometry data deposition and mining constitute powerful tools to boost specialised metabolism research. Techniques for analysing mass spectrometry data, as well as approaches to link the metabolomics and genomics data will be introduced in chapter 4.

Among the mass spectrometry techniques, mass spectrometry imaging (MSI) is a convenient analytical technique to analyse the chemical ecology of microbial cultures, as it allows visualisation of surface and secreted metabolites directly from microbial colonies [258]. For example, the MSI analysis of metabolites produced by *S. coelicolor* interacting with five other actinomycetes allowed the role of a new group of siderophores called desferrioxamines in iron gathering within microbial interactions to be described [259]. Moreover, using MSI to analyse the metabolite production of *Bacillus subtilis* and *Streptomyces* sp. co-cultures, it was possible to identify the enzymatic hydrolysis of the cyclic lipopeptide surfactin as a mechanism of resistance by *Streptomyces* [260].

In summary, alongside genome mining, metabolomics techniques are powerful tools to facilitate natural products discovery, as well as provide data to analyse their production and understand their biological and ecological roles [252].

1.7. Project scope

This work applies genomics and metabolomics tools to assess the influence of evolution, biogeography and ecology on the production of specialised metabolites by a group of *Pseudonocardia* strains isolated from the deep Southern Ocean [261].

To this end, the strains were morphologically and taxonomically described to define them as species, and this taxonomic unit was used to study their evolutionary history and specialised metabolism. Preliminary evidence based on 16S rRNA gene sequences suggested that these strains represented three monophyletic groups that could be novel species [261]. In chapter 3, the evolutionary history of these three lineages, as well as the formal description of two of them as new species was addressed.

In chapter 4, comparative metabolomics and genome mining of these strains, phylogenetically related strains and strains isolated from diverse marine locations was performed. The assessment of BGC and metabolite diversity rooted in phylogeny enabled the evolutionary and biogeographical variables that affect specialised metabolism to be investigated. Prior work on fungus-growing related *Pseudonocardia* strains analysed the influence of phylogeny and biogeography on metabolism [237, 238]. However, comprehensive studies on the specialised metabolism of free-living *Pseudonocardia* are limited. As the three Southern Ocean *Pseudonocardia* lineages were collected from similar environments and geographical locations, it was proposed that they represent a model to study the variables that shape specialised metabolism of free-living *Pseudonocardia*. Finally, the production of bioactive metabolites by the Southern Ocean *Pseudonocardia* strains was explored through the analysis of their bioactivity in different culture conditions. Bioactivity against *S. aureus* by the Southern Ocean *Pseudonocardia* was reported previously [262]. However, preliminary tests also showed that the bioactivity of these strains against other pathogens was limited, while the optimal growing conditions in terms of bioactivity have not been described yet.

Then, in chapter 5, the antifungal activity of the Southern Ocean *Pseudonocardia*, as well as the effect of chemical elicitors on their metabolomics profile was studied. In particular a combination of co-culture and mass spectrometry imaging assays were performed to

investigate the effect of inter-species co-occurrence on the production of bioactive metabolites.

1.8. Project aims

- To describe evolutionarily and taxonomically *Pseudonocardia* spp. isolated from the Southern Ocean.
- To compare the BGC and metabolomics profiles of *Pseudonocardia* spp. strains retrieved from different marine environments.
- To evaluate the effect of co-culturing and chemical elicitors on the metabolomic profiles of *Pseudonocardia* spp. isolated from the Southern Ocean and their phylogenetically closest relatives.

Chapter 2 Methods

2.1. General microbiology and molecular methods

2.1.1. Strain collection

Pseudonocardia strains (Table 2-1) were selected from a bacterial culture collection from Antarctica and the Arctic isolated and previously described by Millán-Aguiñaga *et al*. (2019) [261]. Strains were stored in glycerol 20% at -80 °C. Initial pre-cultures were established by inoculating a colony from an International *Streptomyces* Project Medium 2 (ISP2) [263] agar plate into 7 mL of ISP2 broth supplemented with 18 g/L of Instant Ocean® Sea Salt (IO) (Instant Ocean, USA) in a culture tube (30 mL), and followed by 7-days incubation (30 °C, shaking at 150 rpm). These pre-cultures were used to inoculate (20 mL) ISP2 agar plates (room temperature, 4 weeks) for routine culturing.

Table 2-1. Isolation and taxonomic data of the *Pseudonocardia* strains selected for analysis.

Organism	Strain ID	16S rRNA gene accession	RefSeq assembly accession	Isolation media *	Collection date	Depth	Latitude and longitude
P. abyssalis	KRD168	MH725312.1	GCF_019263705.2	SW	2002-03-17	4539 m	62.95933 S 27.88716 W
P. abyssalis	KRD169	MH725311.1	GCF_019263715.1	SW	2002-03-17	4539 m	62.95933 S 27.88716 W
P. oceani	KRD176	MH725304.1	GCF_019263645.1	A1	2002-03-09	4060 m	65.33133 S 48.09300 W
P. oceani	KRD182	MH725298.1	GCF_019263685.1	SW	2002-03-09	4060 m	65.33133 S 48.09300 W
P. oceani	KRD184	MH725296.1	GCF_019263655.1	SW	2002-03-09	4060 m	65.33133 S 48.09300 W
P. oceani	KRD185	MH725295.1	GCF_019263585.2	SW	2002-03-09	4060 m	65.33133 S 48.09300 W
P. oceani	KRD188	MH725292.1	GCF_019263625.1	A1	2002-03-09	4060 m	65.33133 S 48.09300 W
Pseudonocardia sp.	KRD291	MH725273.1	GCF_019263595.1	SC	2002-03-09	4060 m	65.33133 S 48.09300 W

*A1: starch, 10 g l−1; yeast extract, 4 g l−1; peptone, 2 g l−1; agar, 14 g l−1. SW: agar, 14 g l−1. SC: starch, 10 g l−1; KNO3, 2 g l−1; K2HPO4, 2 g l−1; casein, 0.3 g l−1; MgSO4·7H2O, 0.05 g l−1; CaCO3, 0.02 g l−1; FeSO4·7H2O, 0.01 g l−1; agar, 18 g l−1.

2.1.2. Antibacterial test

Antibacterial activity was evaluated using a disc-diffusion test [264]. The pathogens *Escherichia coli* (ATCC25922), *Staphylococcus aureus* (ATCC433N), *Klebsiella pneumoniae* (ATCC603), *Acinetobacter baumannii* (ATCC19606), *Pseudomonas aeruginosa* (A14), and *Enterococcus faecalis* (ATCC54299) were grown separately overnight in 5 mL of nutrient broth (37°C, 200 rpm), then mixed with 5 mL of liquefied soft nutrient agar to a final OD_{600} of 0.01 (0.1 for *E. faecalis*) and poured on to a plate containing 10 mL of solid nutrient agar. Once solidified, a sterile paper disc (6 mm) with 20 µL of each extract (5 mg/mL) was placed onto the previously inoculated plates. Plates were incubated overnight (37° C) and the inhibition zones were recorded (mm).

2.1.3. Antifungal test

Antifungal activity was evaluated using a modified version of the method proposed by Moore *et al*. [265]. First, the *Pseudonocardia* strains were cultured separately in 5 mL of ISP2 [263] broth (3 malt extract, 3 g yeast extract, 10 g dextrose, 5 g peptone, 1 L distilled water; pH = 7.2) for one week (30 °C, 250 rpm) with a glass bead to avoid clumping. Next, 15 µL of the pre-culture was inoculated in the form of a drop on to the centre of an agar plate containing ISP2 agar (3 g malt extract, 3 g yeast extract, 10 g dextrose, 5 g peptone, 18 g agar, 1 L distilled water; pH = 7.2). After culturing at room temperature for 4 weeks, the plate was inoculated with a spore suspension of the fungal strain. To prepare the spore suspension for *Rhizopus oryzae* (unknown strain designation) and *Aspergillus niger* DSM 1957, a Mannitol - Soya flour (MS) agar (20 g soya flour, 20 g mannitol, 20 g agar, 1 L tap water) plate was inoculated with the fungal spore stock (not quantified) and incubated at room temperature for 48 h. Next, 5 mL of sterile water was added on the plate, and a cell spreader was used to gently break the mycelia. After suspending the spores, a cell spreader was used to move the mycelia to an edge of the plate, and the plate was tilted to move the spore suspension to the opposite edge to enable a spore aliquot to be taken avoiding mycelia chunks. In the case of *Candida albicans* DSM 1386, a liquid pre-culture was made inoculating a colony from a plate to a universal tube containing 5 mL of nutrient broth (5 g peptone, 2 g yeast extract, 5 g sodium chloride, 1 L distilled water). Then, the spore suspension was transferred to a 15 mL Falcon tube, and the OD_{600} was measured. Finally, a known amount of spore suspension was inoculated to 5 mL of melted soft nutrient agar (5 g peptone, 2 g yeast extract, 5 g sodium chloride, 7 g agar,

1 L distilled water) to a concentration of OD_{600} = 0.01, mixing gently by inversion, and pouring over the test plate that contained the *Pseudonocardia* colony. The plates were incubated at room temperature for 48 h, taking zone of inhibition measurements (mm) at 12, 24 and 48 h.

The antifungal assays against clinical isolates (Table 2-2) were carried out at the Institute of Medical Sciences at University of Aberdeen by MSci Eszter Denes under the supervision of Dr Carol Munro. To this end, 100 µl of pathogen solution (2 x 103 cfu/mL for *A. fumigatus* and 2 x 104 cfu/mL for *Candida* strains) was gently mixed with 5 mL soft agar (yeast extract-peptone-dextrose for *Candida* spp. and malt extract for *A. fumigatus*) and this was poured on top of the agar of the sporulating *Pseudonocardia* cultures. The plates were incubated upright (at 30 °C for *Candida* spp. and at 37 °C for *A. fumigatus*), and zone of inhibition was recorded (mm) after 48 h.

Table 2-2 Clinically isolated fungal strains used in antifungal assays showing their origin and antigungal susceptibility

Organism	Strain ID	Isolation	Azole susceptibility
Aspergillus fumigatus	AF293	Lung biopsy [266]	Susceptible
Candida auris	UACa11	Blood culture [267]	Resistant
Candida albicans	SC5314	From a patient with disseminated Candidiasis [268]	Susceptible

2.1.4. 16S rRNA gene phylogeny

The almost complete 16S rRNA gene sequences of KRD168 and KRD185 were obtained by PCR amplification using the 3-IDT® (Integrated DNA Technology) primers FC27 (5'-AGAGTTTGATCCTGGCTCAG-3') and RC1492 (5'-TACGGCTACCTTGTTACGACTT-3') [269]. Sequences were compared to sequences within the NCBI's 16S ribosomal RNA database using the Basic Local Alignment Search Tool (BLAST) [270] (BLAST+ v.2.12.0) and EzBioCloud [271] (v.20210707). A multiple-alignment of all the sequences was achieved using Clustal X [272] (v.2.1), then a Neighbour-Joining tree [273] using a Tamura-Nei model [274] was built using Mega X [275] (v.10.0.4) with 1000 bootstrap replication [276]. Full list of parameters are in Table S1. Interactive tree of life (iTOL) [277] (v.5.7) was used for phylogenetic tree visualisation.

2.2. Whole-genome sequencing

2.2.1. DNA extraction

A colony from each strain cultured on ISP2 agar was used to inoculate ISP2 broth (7 mL) supplemented with 18 g/L of Instant Ocean® sea salt in a culture tube and these were incubated for three days (25 °C, 150 rpm). The culture was then transferred to microcentrifugation tubes (1.5 mL) and centrifuged (4000 rpm, 10 minutes) to pellet the cells, discarding the supernatant. A modified organic DNA extraction protocol [278, 279] was performed for bacterial DNA isolation. Briefly, cells were suspended in a mixture of 111.0 µL Tris 50 mM (BDH, UK), 111.0 µL EDTA (ethylenediaminetetraacetic acid) 20 mM, 75 µL Lysozyme 20 mg/mL (Sigma, UK), and 3 µL RNAse 20 mg/mL (Invitrogen™ PureLink™ RNAse A, UK), and incubated (30 min, 37 °C). Subsequently, 50 µL SDS (sodium dodecyl sulfate) 10% (Sigma, UK) and 85 µL NaCl 1 M (Fisher Chemicals, UK) was added, then the DNA extracted with 400 µL phenol-chloroform-isoamyl alcohol (25:24:1) (Sigma, UK), purified using 0.5 mL isopropanol (Fisher Chemicals, UK) and washed with cold ethanol (70%, 1 mL, Fisher Chemicals, UK). Isolated DNA was re-suspended in nuclease-free water. The quality of the extracted DNA was assessed using a NanoDrop 2000 spectrophotometer (Thermo Scientific) and quantified using a Qubit® 2.0 fluorometer (Invitrogen).

Table 2-3 DNA quality of bacterial strains (KRD168 and KRD185) sent for whole genome sequencing.

	KRD168	KRD185
Conc. Nanodrop (ng/µL)	311.3	259.2
Nucleic acid 260/280 ratio	1.82	1.97
Conc. Qubit (ng/µL)	100.0	81.6

2.2.2. Genome sequencing

Genomic DNA from KRD169, KRD168, KRD176, KRD182, KRD184, KRD185 and KRD188 (Table 2-1) was sequenced by MicrobesNG (Birmingham, UK) using Illumina whole-genome sequencing in a previous work [262]. Genomic DNA libraries were prepared using a Nextera

XT Library Prep Kit (Illumina, San Diego, USA) following the manufacturer's protocol with the following modifications: 2 ng of DNA instead of 1 ng, and PCR elongation time was increased to 1 min. DNA quantification and library preparation were carried out on a Hamilton Microlab STAR automated liquid handling system. Pooled libraries were quantified using the Kapa Biosystems Library Quantification Kit for Illumina on a Roche light cycler 96 qPCR machine. Libraries were sequenced on the Illumina HiSeq™ using a 250 base pair (bp) paired-end protocol. Reads were adapter trimmed using Trimmomatic [280] (v.0.30) with a sliding window quality cut-off of Q15 and default settings.

Genomic DNA from the strains *P. abyssalis* KRD168 and *P. oceani* KRD185 (Table 1-1) were also sent to NU-OMICS (Northumbria University, Newcastle, UK) for sequencing using the PacBio sequencer platform. Briefly, genomic DNA was size sheared using Covaris g-Tube to approximately 9 Kbp, and size distribution visualised using Agilent Bioanalyzer. ExoVII was used to remove ssDNA and obtain blunt ended dsDNA. Blunt hairpin SMRTbell adapters were ligated to the polished fragmented gDNA, and ExoIII treatment was carried out to remove failed ligation products. Adapter-ligated SMRTbell templates were purified using PacBio AMPure. PacBio Polymerase binding reaction to SMRTbell template was carried out using the Sequel Binding kit 2.1 and sequenced using the PacBio Sequel instrument. Samples were sequenced using 10 h movie capture. BAM files of CCS (circular consensus sequencing) reads were used for assembly in Hierarchical Genome Assembly Process (HGAP) [281] (v.4).

2.2.3. *De novo* genome assembly

Several approaches to construct de novo assemblies were applied. PacBio long-reads were assembled using Flye (v.2.8.1), Canu [282] (v.2.1.1), and Raven [283] (v.1.2.2), while the HGAP [281] (v.4) was included in the SMRT analysis. Illumina short-reads were assembled using Unicyler [284] (v.0.4.8.0), with SPAdes [285] (v.3.14.1) and Pilon [286] (v.1.23) as dependencies. Full list of parameters is available in table S1. For quality control, reads were aligned against assemblies using minimap2 [287] (v.2.17), and the BAM files produced were analysed with Qualimap [288] (v.2.2.2) to assess the genome coverage. In addition, assembly quality was evaluated using QUAST [289] (v.5.0.2), while their completeness was evaluated using BUSCO [290] (v.3.0.2). Genomic islands were annotated using IslandViewer 4 [291] and prophages using PHASTER [292] (v.2015).

The *P. abyssalis* KRD168 and *P. oceani* KRD185 final assemblies were achieved using hybrid assembly approaches by including long and short read data. For KRD168, long-reads were assembled using Flye [293] (v.2.8.1). Then, the short reads were mapped over the assembly using Bowtie2 [294] (v.2.4.2). The produced BAM files were used for improving the genome assembly using Pilon [286] (v.1.23) by correcting bases, fixing misassemblies and filling gaps. In the case of KRD185, a consensus assembly was created to reduce the number of contigs. To this end, long reads were first assembled using HGAP [281] (v.4), Flye (v.2.8.1), Canu [282] (v.2.1.1), and Raven [283] (v.1.2.2). Then, a consensus assembly was obtained with Flye using the 'subassemblies' mode. Finally, the consensus long-reads assembly was integrated in Unicycler [284] (v.0.4.8), with SPAdes [285] (v.3.14.1) and Pilon (v.1.23) as dependencies, to create a hybrid assembly including short read data.

Before biosynthetic gene annotation, the KRD169, KRD176, KRD182, KRD184, and KRD188 draft genome scaffolding was accomplished using MeDuSa [295] (v.1.6) using default settings, and with KRD168 and KRD185 as species reference genomes. For KRD291, the *Pseudonocardia sediminis* genome (SHKL00000000.1) was used as a reference genome for scaffolding.

Genome sequences were submitted to the NCBI genome database (Table 2-1), and the annotation was added by the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) (v.4.13).

2.3. Whole-genome analysis

2.3.1. Comparative genomics

To assess the genomic distance between KRD168, KRD185, and other publicly available *Pseudonocardia* genomes (Table S2), digital DNA–DNA Hybridization (dDDH) values were calculated with the Genome-to-Genome Distance Calculator (GGDC) [296] (v.2.1) using the formula 2. Additionally, pairwise whole-genome average nucleotide identity (ANI) values were calculated using the alignment-free method FastANI [297] (v.1.3), as well as the method based on MUMmer aligment (ANIm) using PYANI [298] (v.0.2.11). Default settings were used in both ANI calculation methods. A genome-scale phylogenetic tree based on the multi-locus sequence analysis was performed using autoMLST [299] (v.20191016) with a *de novo*

workflow and performing IQ-TREE Ultrafast Bootstrap analysis (1000 replicates). The list of protein-coding genes used can be found in Table S3.

2.3.2. Pan-genome analysis

From the publicly available *Pseudonocardia* genomes, genomes with redundant confirmed taxonomy were removed for the pan-genome analysis. For selected genomes (Table S2), protein-coding regions and gene annotations were assigned using Prokka [300] (v. 1.14.6). The orthologous genes shared across these genomes (e-value cut-off set to 1E-6) were then identified with PIRATE (Pangenome Iterative Refinement and Threshold Evaluation) [301] (v.1.0.4). The Maximum Likelihood tree based on the core-genome (>95% genomes) alignment retrieved from PIRATE was reconstructed using RAxML (Randomized Axelerated Maximum Likelihood) [302] (v.8.2.4) under GTR (general time-reversible) model of nucleotide substitution with a gamma model of rate heterogeneity. Full list of parameters is available in table S1. Results were visualised using Phandango [303] (v.1.3.0).

2.3.3. Biosynthetic Gene Cluster annotation and analysis

BGC annotation was performed using antiSMASH [304, 305] (v.5.1.07). Using the output, a sequence similarity network was generated by calculating pairwise distances between gene sequences [calculated as the weighted combination of the Jaccard, GK (Goodman-Kruskal gamma function) and DDS (Domain Duplicate Score) indices] and clustering them in Gene Cluster Families (GCFs) using BiG-SCAPE [306] (v.1.0.0) under hybrid and glocal alignment mode settings at a distance cut-off of 0.5. All available BGCs from MIBiG [175] (v.2.0) were integrated into the BiG-SCAPE analysis. The BGC profile of the Southern Ocean *Pseudonocardia* strains was visualised using Circos [307] (v.0.63-9), while the BiG-SCAPE output networks were visualised using Cytoscape [308] (v.3.7.2). Synteny and genes sequence similarity into each GCFs were analysed and visualised using Clinker [309] (v.0.0.2).

2.3.4. Recombination analysis of whole genome and BGCs

The core-genome gene flow among *Pseudonocardia* strains was analysed inferring recombination events using ClonalFrameML [310] (v.1.12). The core-genome alignment retrieved from PIRATE [301], and the phylogenetic tree reconstructed from it with RAxML [302] were used as inputs, using the same settings mentioned above. The relative

contribution of recombination to mutation (r/m) was calculated using the relative rate of recombination to mutation (R/θ), the mean import length of recombining DNA (v), and the mean divergence of imported recombining DNA (δ) values. Recombinant regions were masked from the alignment using the script maskrc-svg.py (https://github.com/kwongj/maskrc-svg) (v.0.5).

To calculate the influence of recombination over GCFs, the BGC sequences in each GCF were aligned using Mugsy [311] (v.1.2.3). Then, a phylogeny for each GCF was reconstructed with RAxML to infer the relative rate of recombination to mutation with ClonalFrameML as mentioned above. The full list of parameters is available in table S1.

2.4. Physiology and chemotaxonomic characterisation of *P. abyssalis* KRD168 and *P. oceani* KRD185

2.4.1. Morphologic characterisation

Colony morphology of *P. abyssalis* KRD168 and *P. oceani* KRD185, as well as *P. petroleophila* DSM 43193 and *P. hydrocarbonoxydans* DSM 43281, were examined when cultured on ISP2, ISP3, ISP4 and ISP5 media (Table 2-3) [263]. Salinity and acidity tolerance analysis of the strains was carried out on modified ISP2 and ISP5 with pH 4-10 (Table S4) and 0-20 % NaCl (m/v).

Table 2-4 Composition of media used during the colony morphology description based on the International *Streptomyces* Project [263].

Medium	Recipe
ISP2	yeast 4 g, malt 10 g, dextrose 4 g, H_2O 1 L, pH = 7.2
ISP3	oatmeal 20 g, cooked and filtered, H_2O 1 L, pH = 7.2
ISP4	soluble starch 10 g, K_2HPO_4 1 g, $MgSO_4 \cdot 7H_2O$ 1 g, NaCl 1 g, $(NH_4)_2SO_4$ 2 g, $CaCO_3$ 2 g, H_2O 1 L, pH = 7.2
ISP5	L-asparagine 1 g, glycerol 10 g, K_2HPO_4, H_2O 1 L, pH = 7.2

2.4.2. Microscopic characterisation

Strains *P. abyssalis* KRD168, *P. oceani* KRD185, and *P. petroleophila* DSM 43193 were cultured on ISP2 agar for 28 days with an inclined (45° angle) coverslip, and the morphology of the growing colony on the coverslip was observed using a Nikon Eclipse TE2000-S Inverted microscope fitted with a 100x/1.3 numerical aperture objective lens (Nikon, Japan). Illumination was sourced from a mercury arc lamp with appropriate emission filters for FITC/PI imaging. Fluorescence microscopy was carried out using fluorescein-conjugated Wheat Germ Agglutinin (FITC-WGA) and Propidium Iodide (PI) to describe structures in the aerial hyphae and apical tip growth [312]. Phase-contrast and fluorescence images were acquired sequentially using an ORCA-100 CCD camera (Hamamatsu, Japan). Image processing and analysis was performed using FIJI [313] (ImageJ v.1.53e).

2.4.3. Chemotaxonomic characterisation

Fatty acid, metabolic activity, respiratory quinone, polar lipid, and whole-cell sugar analyses were carried out by the Identification Service at the Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany. For these analyses, strains were cultivated in GYM (glucose-yeast extract-malt extract) medium (28 °C for 8-14 days) prior to transport the Leibniz Institute. The fatty acid analysis was performed using the Sherlock MIS (MIDI Inc, Newark, USA) system after conversion into fatty acid methyl esters [314, 315]. The fatty acids annotations and quantification were calculated by the MIS Standard Software (Microbial ID) using the Aerobic Bacteria Library (TSBA6 v6.10). Metabolic activity was determined using the API 20E and API ZYM systems (BioMérieux). Polar lipids and respiratory quinones were extracted from freeze-dried cells cultured in YEME (yeast extract-malt extract) broth and analysed by chromatography [316]. Whole-cell sugars analysis and detection of isomers of 2,6-diaminopimelic acid (Dpm) and 2,6-diamino-3-hydroxypimelic acid (OH-Dpm) were performed by thin-layer chromatography (TLC) [317, 318].

2.5. LC-MS/MS analysis

2.5.1. Metabolite extract preparation

For each of the sixteen selected *Pseudonocardia* strains (Table S7), an ISP2 broth pre-culture (100 µL) was inoculated on a Petri dish containing 20 mL of ISP2 agar supplemented with Instant Ocean® (18 g/L) and *N*-acetyl glucosamine (1 mM). After 28 days, samples were dried by lyophilisation (Labconco FreeZone 2.5L Freeze Dry System) and extracted twice with 20 mL of ethyl acetate (Fisher Chemicals, UK) in 250 mL Erlenmeyer flasks with agitation (150 rpm) overnight. The extract was dried by nitrogen flow, weighed and stored at 4 °C.

2.5.2. LC-MS/MS analysis of metabolite extracts

The dried extracts were reconstituted in acetonitrile Optima™ LC/MS (Fisher Scientific, UK) to a concentration of 1 mg/mL. Analysis using Liquid Chromatography – High-Resolution Mass Spectrometry (LC-MS) using a Dionex UltiMate™ 3000 coupled to Q-Exactive™ (Thermo-Scientific, Germany) mass spectrometer with an electrospray ionization (ESI) source was carried out with a mass range of 70–1000 *m/z* in positive and negative ionization modes with a spray voltage of 4.2 kV and capillary temperature at 310 °C. A Phenomenex Kinetex 1.7 µm C18 100A (100 x 2.1 mm) column was used for chromatographic separation at 40 °C. Mobile phase A consisted of H_2O with 0.1% formic acid (Fluka® Analytical, Switzerland) and mobile phase B consisted of acetonitrile Optima™ LC/MS (Fisher Scientific, UK) with 0.1% formic acid. A gradient was used starting from 5% B and increasing to 50% from 0 to 8 min. Then to 99% from 8 to 10 mins, and it was held constant for 2 min. Finally, B was brought back to initial conditions and held for 5 min for a total run time of 15 min. A flow rate of 300 µL/min and an injection volume of 10 µL was used.

2.5.3. Molecular features analysis

The LC-MS data were acquired using Xcalibur (v.2.2) (Thermo Fisher Scientific, USA), and the obtained raw data were converted to mzXML file format using Proteowizard MSConvert [319]. Then, data were pre-processed using MZmine 2 [320] (v.2.53). Briefly, mass peaks were detected in negative mode with a centroid mass detector and $1.0x10^6$ as noise level cut-off for MS1 and $1.0x10^3$ for MS2. Chromatograms were built by the ADAP module [321] using a minimal highest intensity of $3.0x10^6$ and a *m/z* tolerance of 0.02. Then, chromatograms were denconvoluted using the Algorithm Baseline Local Minimum search with a minimum retention time range of 0.1 min, a minimum relative height of 30%, and a minimum absolute height of $1.0x10^6$. Next, a peak list was created by Join Aligner using an

m/z tolerance of 0.02, a retention time tolerance of 0.1 min, and weights for retention time of 20% and for *m/z* of 80%. After a gap-filling process and filtering peaks detected in only one sample, the peak list was exported and analysed using MetaboAnalyst 4.0 [322]. Data were normalized by Pareto scaling and analysed using univariate analysis (Volcano plots) and multivariate analysis (Partial Least-Squares Discriminant Analysis, PLS-DA).

2.5.4. Classical molecular networking

A molecular network [323] was created using the online workflow (https://ccms-ucsd.github.io/GNPSDocumentation/networking/) on the Global Natural Product Social Molecular Networking (GNPS) website (http://gnps.ucsd.edu). The data was filtered by removing all MS/MS fragment ions within +/- 17 Da of the precursor m/z. MS/MS spectra were window filtered by choosing only the top 6 fragment ions in the +/- 50Da window throughout the spectrum. The precursor ion mass tolerance was set to 2.0 Da and a MS/MS fragment ion tolerance of 0.02 Da. A network was then created where edges were filtered to have a cosine score (normalized dot-product, which is a mathematical measure of spectral similarity between two fragmentation spectra. A cosine score of 1 represents identical spectra while a cosine score of 0 denotes no similarity at all) above 0.7 and more than 6 matched peaks, according to the recommendations of Watrous *et al.* (2012) [257]. Further, edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. Finally, the maximum size of a molecular family was set to 100, and the lowest scoring edges were removed from molecular families until the molecular family size was below this threshold. The spectra in the network were then searched against GNPS' spectral libraries. The library spectra were filtered in the same manner as the input data. All matches kept between network spectra and library spectra were required to have a cosine score above 0.7 and at least 6 matched peaks. The molecular networks were visualized using Cytoscape software [308].

2.5.5. Feature-Based Molecular Networking

A molecular network was created using the Feature-Based Molecular Networking (FBMN) workflow [324] on GNPS [323]. The mass spectrometry data were first processed with MZmine 2 [320] (v.2.53), as mentioned above, and the results were exported to GNPS for FBMN analysis. The data was filtered by removing all MS/MS fragment ions within +/- 17 Da

of the precursor m/z. MS/MS spectra were window filtered by choosing only the top 6 fragment ions in the +/- 50 Da window throughout the spectrum. The precursor ion mass tolerance was set to 2.0 Da and the MS/MS fragment ion tolerance to 0.02 Da. A molecular network was then created where edges were filtered to have a cosine score above 0.7 and more than 6 matched peaks. Further, edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. Finally, the maximum size of a molecular family was set to 100, and the lowest scoring edges were removed from molecular families until the molecular family size was below this threshold. The analogue search mode was used by searching against MS/MS spectra with a maximum difference of 100.0 Da in the precursor ion value. The library spectra were filtered in the same manner as the input data. All matches kept between network spectra and library spectra were required to have a score above 0.7 and at least 6 matched peaks, according to the recommendations of Watrous *et al.* (2012) [257]. DEREPLICATOR (v. 1.3.15-GNPS) was used to annotate MS/MS spectra [325] using default settings and the molecular networks were then visualized using Cytoscape software [308].

For the Ion Identity Molecular Networking (IIMN) [326], feature finding and ion identity networking were performed using a modified version of MZmine v.2.37 available at https://mzmine.github.io/iin_fbmn_(v.corr.17.7). Molecular feature detection was carried out with the same parameters as FBMN.

2.5.6. MolNetEnhancer for chemical class annotation of molecular networks

To enhance chemical structural information within the molecular network, information from *in silico* structure annotations from GNPS Library Search, Dereplicator+ [325] (v.1.3.15-GNPS), and MS2LDA [327] (v.1.3.15-GNPS), were incorporated into the network using the GNPS MolNetEnhancer workflow [328] (https://ccms-ucsd.github.io/GNPSDocumentation/molnetenhancer/) on the GNPS website (http://gnps.ucsd.edu), using default settings. Chemical class annotations were performed using the ClassyFire chemical ontology [329] (v.1.0).

2.6. Challenge assay

The interactions between Southern Ocean strains (*P. abyssalis* KRD168, *P. oceani* KRD185, and *Pseudonocardia* sp. KRD291) and eight phylogenetically related strains (*P. sediminis* DSM 45779, *P. petroleophila* DSM 43193, *P. hydrocarbonoxydans* DSM 43281, *P. ammonioxydans* DSM 44958, *P. profundimaris* KCTC 39641, *P. antitumoralis* DSM 45322, *Pseudonocardia* sp. VO44-3, and *Pseudonocardia* sp. CNS-139) from varied environments were studied in a challenge assay following the method proposed by Poulsen *et al*. [330]. Briefly, strains were pre-cultured in ISP2 broth (30 °C, shaking at 150 rpm) for a period of 14 days. First, the 'resident' strain was cultured in ISP2 agar supplemented with Instant Ocean® sea salt after adding 10 µL of the pre-culture in the centre of a Petri dish containing 20 mL of agar. After 28 days, the 'intruder' strain was inoculated around the resident strain colony from a 14-day liquid pre-culture using a sterile cotton bud. The zone of inhibition was measured after 14 days. The eleven *Pseudonocardia* strains were tested both as a resident and as an intruder strain in triplicate. Besides inhibition, morphological changes in colonies were also recorded.

2.7. MALDI-Mass Spectrometry Imaging (MSI)

The *in-situ* production of metabolites was observed using Matrix-Assisted Laser Desorption / Ionization (MALDI) – Mass Spectrometry Imaging (MSI) with Trapped Ion Mobility Spectrometry (TIMS) - Time of Flight (ToF) mass detector. This analysis was carried out in the Centre for Microbial Secondary metabolites – CeMiSt at the Technical University of Denmark (DTU) by Dr Scott A. Jarmusch.

2.7.1. Sample preparation

Co-cultures interactions were cultured at the DTU on ISP2 agar plates (8 mL in 9 cm petri dishes) for 40 days at room temperature. The samples were excised and mounted on IntelliSlides conductive tin oxide glass slides (Bruker Daltonik GmbH). Each slide was pre-treated with 0.25 mL of 20 mg/mL 2,5-dihydrobenzoic acid (DHB) matrix solution (2 passes) using a HTX Imaging TM-Sprayer (HTX Technologies, USA), to prevent sample flaking [331]. Images were subsequently taken using TissueScout (Bruker Daltonik GmbH) and a Braun FS120 scanner, followed by overnight drying in a desiccator. The following day, samples were covered with matrix by spraying 2.5 mL of 20 mg/mL 2,5-dihydroxybenzoic acid (DHB) matrix solution (15 passes) in a nitrogen atmosphere. The matrix solution was dissolved in

acetonitrile/methanol/H_2O (70:25:5, v/v/v). Samples mounted on slides were dried in a desiccator overnight prior to MSI measurement.

2.7.2. MALDI-ToF MSI analyses

Samples were analysed using a timsTOF flex (Bruker Daltonik GmbH) mass spectrometer. Samples were run in positive mode with a 100 µm raster width and a mass range of 100-2000 Da. Additionally, MS settings include: laser imaging 100 µm, power boost 3.0%, scan range 26 µm in the XY interval, and laser power 85%; tune: funnel 1 RF 300 Vpp, funnel 2 RF 300 Vpp, multipole RF 300 Vpp, isCID 0 eV, deflection delta 70 V, MALDI plate offset 100 V, quadrupole ion energy 5 eV, quadrupole loss mass 100 m/z, collision energy 10 eV, focus preTOF transfer time 75 µs and pre-pulse storage 8 µs. Data was analysed using SCiLS software (v.2022b) (Bruker Daltonik GmbH).

Chapter 3 Genomic analysis and formal description of *P. abyssalis* sp. nov. and *P. oceani* sp. nov.

3.1. Introduction

3.1.1. Whole-genome sequencing

Next-generation sequencing (NGS) describes highly parallel or high-output sequencing methods that produce data at or beyond the genome-scale [332]. The first sequenced genome of an Actinobacteria strain was the human pathogen *Mycobacterium tuberculosis* H37Rv, published in 1998 [333]. Following this, the genomes of the model microorganism *Streptomyces coelicolor* [334] and the industrially-relevant species *Streptomyces avermitilis* [335] were sequenced. These genomes were sequenced using shotgun Sanger technology and despite being state-of-the-art at the time, the process was extremely laborious and too expensive to provide sufficient coverage for a routine whole-genome shotgun [336]. Second-generation sequencing, such as 454-pyrosequencing [337], allowed higher genome coverage to be obtained; however, the difficulties of Actinobacteria genome sequencing often hindered high-quality assembly of their genomes [336]. For example, the high GC content of Actinobacteria genomes affects the sequencing process in itself, as well as genome assembly, a consequence of low and high coverage in GC-poor and rich regions respectively [338]. In addition, the extraction of high-quality molecular weight DNA can be a further challenge due to the difficulty in lysing membranes [336]. NGS technologies, such as Illumina [339] and Pacific Biosciences (PacBio) [340], have incorporated innovations to tackle these complexities to obtain high-quality genome sequences, by producing longer reads and higher coverages [336]. Furthermore, since the DNA sequence capacity has increased and sequencing costs have decreased with time, NGS technologies are now considered a routine tool in biological research [341].

The use of hybrid mode assembly, by adding short-reads, has been proposed as a successful strategy for correcting single-molecule sequencing reads (long-reads) in *de novo* assemblies [342]. PacBio instruments are widely used for Single-Molecule Real-Time Sequencing (SMRT) sequencing. For example, PacBio Sequel produces from 3.5 to 7 Gbp per run with a read length from 8 to 12 Kbp [341]. Despite the error rate of PacBio sequencing [up to 13%, based on basic local alignment with successive refinement (BLASR)], these errors are randomly

distributed and can be corrected in the assembly process if there is adequate coverage [343]. Thus, PacBio long-reads have been shown to be a useful tool for filling gaps in short-reads based assemblies [344]. Furthermore, high-quality, complete or almost complete genomes of bacteria aid the identification of specialised metabolite Biosynthetic Gene Clusters (BGCs) [52]. For example, the hybrid strategy of using PacBio long-read and Illumina short-read sequencing technology was used for assembling high-quality genome sequences of 22 *Streptomyces* species and eight *Streptomyces venezuelae* strains, obtaining more than 99.1 % completeness (completeness of gene space was estimated using the BUSCO v3) for 29 of the 30 assemblies, 97.4% completeness for *S. clavuligerus*, and allowing the identification of 922 BGCs [345].

According to NCBI [346] and IMG [347] databases, at the time of writing (October 2021), genome sequences for 77 *Pseudonocardia* strains are publicly available, of which 14 are MAGs. Only nine of these sequenced genomes are complete-level genomes (every chromosome in the assembly is gapless). Most of these complete genomes belong to *Pseudonocardia* strains isolated in symbiosis with fungus-growing ants [348–350]. Regarding strains isolated from marine environments, only five genomes are publicly available, none of them at a complete level and only two published in a scientifc journal [194]. In terms of the sequencing technology, Illumina short-reads have been used most frequently for sequencing the *Pseudonocardia* genus resulting in 49 genomes, while PacBio long-read technology was used to generate 22 genomes. Moreover, only two of the 77 publicly available genomes were assembled by applying a hybrid approach, where long and short reads were combined during the assembly process; in both cases, high contiguity (N50 > 1 Mb) and high coverage (> 30x) were obtained. Particularly, the *Pseudonocardia dioxanivorans* genome was sequenced with Illumina and 454-pyrosequencing [126], while PacBio and Illumina reads were combined for the *Pseudonocardia autotrophica* genome sequencing [351]; the latter is the only published example of a hybrid assembly using two next-generation sequencing platforms for the *Pseudonocardia* genus.

Furthermore, whole-genome sequencing (WGS) has been used to correct some incongruences in the phylogenetic systematics of the phylum Actinobacteria which were the result of poorly resolved 16S rRNA gene trees. Hence, phylogenomic analysis has been proposed as a valuable and powerful taxonomic marker in Actinobacteria systematics, showing that whole-genome phylogenies have a higher taxonomic resolution than the

phylogenies resolved using only a few genes [352]. An example of the taxonomic resolution at the genome level is the *Salinispora* genus. Members of this genus share more than 99% 16S rRNA gene sequence identity which makes it difficult to resolve the inter-species phylogeny based on this gene, despite their phenotypic differences. However, the comparative genomic analysis that included genome-based metrics, such as Average Nucleotide Identity (ANI), as well as core-genome phylogeny, allowed the species delimitation which was confirmed by a phenotypic characterization [353, 354].

3.1.2. Species concept and operational definitions

Despite species being the fundamental taxonomic unit, there is not a consensus on the biological origin and precise boundaries of the species concept in prokaryotes. In contrast to sexual organisms where genetic cohesion can be generally defined by sexual compatibility and population structure, in asexual organisms like bacteria, the species definition remains vague [355]. According to the International Committee on Systematics of Prokaryotes (ICSP), the current species definition in prokaryotes is pragmatic, operational and universally applicable, which stablishes that *a species is a category that circumscribes a (preferably) genomically coherent group of individual isolates/strains sharing a high degree of similarity in (many) independent features, comparatively tested under highly standardized conditions* [356, 357].

There are several philosophical questions around the species concept that are far beyond this work, including the ontological nature of the concept itself. Perhaps, one of the most fundamental questions is whether the species concept refers to a unit of classification or a unit of evolution. Historically, the species concept was used first as a conceptual tool for the hierarchical classification of life proposed by Linnaeus. Then, Darwin provided a rationale of species as a result of an evolutionary process. Nevertheless, the contemporary views of species and speciation process are focused on how species evolved and how biologically different they are, are still attached to the historical nature of classification [358]. Some models to define the concept to prokaryote species are described below:

Biological species concept (BSC). Proposed by Mayr (1942), it establishes that sexual organisms, through meiotic recombination, maintain the genetic cohesion of species. Hence, species can be defined as *groups of interbreeding natural populations that are reproductively isolated from other such groups* [358, 359]. The BSC defines species as a taxonomic rank

based on the interbreeding capacity rather than phenotypic similarities. Then, in principle, the concept can be only applied to organisms that form sexual populations, with a limited application for the prokaryote species definition [360]. However, Bacteria and Archaea can transfer and exchange genes through homologous recombination, and it has been demonstrated that barriers in this exchange could define the BSC in prokaryotes with the same efficacy as in sexually reproducing eukaryotes [361].

Pragmatic approaches. To facilitate communication in the scientific community, pragmatic approaches have been developed for clustering groups of organisms based on sequence thresholds. As these approaches are independent of a epistemological definition of the "species" concept, they are also defined as "Operational Taxonomic Units" (OTUs) [355]. DNA-DNA Hybridisation (DDH) is now the most common criterion used to classify two bacterial strains as the same species [362, 363]. In general, bacterial strains are assigned to the same species if their DNA-DNA relatedness is ≥70% in a DDH experiment under standardised conditions and their ΔT_m (melting temperature) is ≤ 5°C apart [364, 365].

The 16S rRNA subunit is a gene shared by all bacteria and archaea, and it has been used for decades as a putative marker for species delimitation [355]. Comparative studies showed that a threshold of >97% similarity of the 16S rRNA gene sequence corresponds to the threshold of ≥70% in DDH experiments [366]. Although the 16S rRNA gene sequencing has become a routine assay for species classification, its taxonomic resolution is often insufficient for the correct delineation of bacterial species [357, 367].

Even though DDH used to be considered the gold standard for species delimitation, it has been criticized for being a complex and time-consuming technique [362, 368]. As a consequence of the increasing availability of genome sequences, some computational techniques have been developed to substitute wet-lab DDH determination. For example, the Genome-BLAST Distance Phylogeny (GBDP) approach uses a BLAST-based genome comparison to create a distance matrix that is used to reconstruct the phylogeny of the compared genomes [369]. The GBDP can be used to produce a set of high-scoring segment pairs or inter-genomic matches which are used to calculate the digital DDH values (dDDH). The dDDH values mimic the wet-lab DDH values and are subject to the same interpretation [369, 370]. The Average Nucleotide Identity (ANI) of all conserved genes between the strains is another widely used alternative for the DDH standard. ANI represents the mean nucleotide identity of all orthologous genes shared between two genomes [297], and their calculation

generally involves the genome sequences fragmentation, followed by nucleotide sequence search, alignment, and identity calculation [371]. The evaluation of these values showed that ANI values of 95% corresponded to the traditional 70% DDH standard of the species definition [297, 362, 372].

Phylogenetic Species Concept (PSC). After a few authors suggested a phylogenetic concept of species, Cracraft (1983) proposed that a species is *the smallest diagnosable cluster of individual organisms within which there is a parental pattern of ancestry and descent* [358, 373]. Like pragmatic approaches, the PSC takes sequence divergence into account, but also requires that species form a monophyletic group [355]. Thus, bacterial species can be defined as a group of strains that, besides sharing 70% of DDH as explained before, should accomplish the PSC with only a few exceptional cases [364].

Stable Ecotype Model (SEM). Cohan (2001) proposed the concept of SEM as a framework that applies an ecological and evolutionary theory to define units of bacterial diversity. An 'ecotype' is defined as a *group of bacteria that are ecologically similar to one another, so similar that genetic diversity within the ecotype is limited by a cohesive force, either periodic selection or genetic drift, or both* [374, 375]. When a new allele appears generating a new adaptive variant into the ecotype population, the new variant competes with their neighbour for the niche, which might lead to a diversity purge after periodic selection events due to the spread of adaptive alleles within and across populations. Neutral mutations and genetic drift also occur along with periodic selection events, generating separate clusters of ecological and genetic diversity [376]. In this model, allele diversity within an ecotype is temporary, persisting only until the next periodic selection event [375]. In the 'stable ecotype' model, the formation of new ecotypes is rare, and the periodic selection events purge diversity within ecotypes. Under these conditions, there is a one-to-one correspondence between ecotypes and sequence clusters [375, 377]. The SEM does not attempt to define the species concept by itself but rather describes evolutionary and ecological principles to explain microbial population structures and speciation [363]. Therefore, a prokaryote species could be defined in the SEM as a clade whose members are ecologically similar to one another, so that genetic diversity within the clade is limited by cohesive forces due to either periodic selection or genetic drift, or to both [375, 377].

3.1.3. Speciation

Speciation is the process by which an ancestral species diverge into descedent lineages [378]. Historically, the concept of speciation has been closely related to the BSC, and it is conceived as the evolution of barriers to gene exchange. Thus, for sexually reproducing eukaryotes, the speciation process is mainly explained as reproductive isolation in allopatry (physical separation that prevent the gene flow), while the reproductive isolation in sympatry (in the presence of gene flow) is more controversial and restricted to specific conditions [378]. However, for prokaryotes, as the SEM suggested, it is possible that two lineages evolve cohesively in sympatry and yet diverge if they occupy different ecological niches [360, 379].

In eukaryotes, sympatric speciation is favoured when relatively few loci lead to reproductive isolation and niche adaptation, otherwise, the homogenizing force of recombination can slow down the adaptation rate [380]. By applying this principle to simulated microbial populations, Friedman *et al.* (2013) showed that at early stages recombination accelerated adaptation to the new niche by combining multiple adaptive alleles, while later in the absence of gene-flow barriers, recombination homogenizes the incipient species. The proposed solution to this trade-off is reducing the number of loci required for speciation or reducing the recombination rate over time [380]. In this context, Shapiro (2012) proposed a model for microbial speciation as a function of genes flow, which results in units of genetic similarity and natural selection, which defines units of ecological fitness [376]. In this model (Fig. 3-1A), speciation is described as a gradual process where a new niche becomes accessible to a lineage when an allele emerges as a consequence of a mutation or it is acquired in an ancestral population via HGT, generating a new variant. If the recombination rate is relatively higher than the selection rate ($r/s >> 1$), the new variant is spread by recombination. If the new and ancestral niches remain fully sympatric, the process stops here without speciation. However, if the new niche is also related to barriers to recombination, genetic speciation will occur in neutral loci throughout the genome (genetic hitchhiking). These ecological barriers will be reinforced later by genetic barriers, reducing the recombination between the two new lineages [381–383]. When the selection rate is relatively higher than the recombination rate ($r/s << 1$), the model corresponds closely to the SEM. In this case, the new variant is spread by clonal expansion, and after several steps of periodic selection, the lineages are permanently separated both ecologically and genetically [376, 382].

As mentioned above, selection over a few loci could lead the speciation process without affecting the entire genome. Comparative genomics of putative sympatric species showed that some parts of the genome are highly divergent between lineages, while the rest of the genome remains undifferentiated [384]. In a model of high recombination and low selection (Fig. 3-1B.I), the new variant will share the same niche, resulting in a competitive exclusion that produces the extinction of one of the lineages. However, if the new variant is under selection (Fig. 3-1B.II), the speciation will be successful and the new variant will contain genes under divergent natural selection [383]. In this scenario ($r/s >> 1$), individual genes will sweep to fixation without affecting genetic diversity in other sites of the genome. On the other hand, if the selection rate is relatively higher than the recombination rate ($r/s << 1$), the complete genome will sweep to fixation, resulting in genome-wide divergence from the other population (Fig. 3-1B.IV). This last scenario is described in the SEM. Finally, population divergence by geographic barriers will produce allopatric speciation in a low recombination and selection rates scenario (Fig. 3-1B.III) [383].

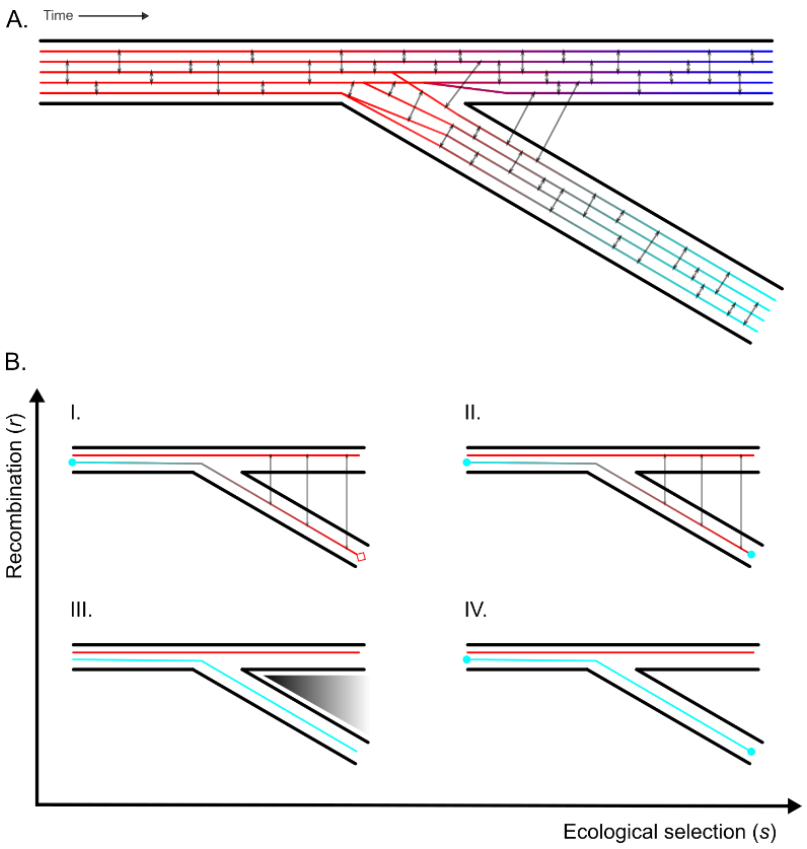


Figure 3-1 Model of speciation under different recombination (r) and selection (s) balances. A. Representation of an ancestral population which diverges into two incipient species. B. Representation of a single population of chromosomes splits into two nascent species. The ancient lineages are shown as red lines, while new variants are represented as cyan and blue lines. Recombination events between lineages are shown as grey arrows. Loci responsible for ecological differentiation are represented as a dot on the lines. The square on the line denotes lineage extinction, while the grey shaded area represents a geographic barrier.

3.1.4. Polyphasic taxonomy

Although the prokaryote species concept is still under scientific debate, bacterial species are currently described using a polyphasic approach that integrates genotypic and phenotypic properties. Polyphasic taxonomy looks for a consensus framework for groups' delimitation, attempting to describe a coherent species definition without questioning if the concept corresponds to a biological reality. In this regard, polyphasic classification is entirely empirical and is not influenced by conceptual preconceptions [385]. For practical proposes, a species is defined as *a group of strains, including the type strain, sharing at least 70% total genome DDH and less than 5∘C ΔT_m, and the phenotypic features should agree with this genotypic definition and should satisfy the PSC* [364, 386]. Although there is no universal strategy employed in the polyphasic description, a typical study should include genotypic information

like RNA sequencing (e.g. 16S rRNA gene), determination of the DNA base ratio (G+C content), and DDH studies. Moreover, a classical phenotypic analysis includes information about chemotaxonomic markers such as cellular fatty acids, mycolic acids, polar lipids, isoprenoid quinones, cell walls compounds and exopolysaccharides, as well as expressed features like morphology, physiology and enzymology [385].

3.1.5. *Pseudonocardia* genus

The genus *Pseudonocardia* was described for the first time by Hassen [116] and subsequently amended [387–390]. Members of the genus *Pseudonocardia* are aerobic, Gram-positive, non-motile, catalase-positive bacteria. Strains form branched substrate hyphae may fragment into rod-shaped elements. While aerial hyphae, if formed, can be sterile, they may be fragmented into chains of oval or square elements, or differentiate into chains of spores. Substrate and aerial hyphae show cell division in several different directions simultaneously with a tendency to form swellings. Spores are usually smooth and may be formed on the substrate or aerial hyphae. The major menaquinone is MK-8(H_4) and the predominant fatty acid is *iso*-branched hexadecanoic acid, while mycolic acids are absent [391].

According to the List of Prokaryotic names with Standing in Nomenclature (LPSN), 63 species have been described for the *Pseudonocardia* genus with validly published names at the time of writing (May 2021) [392, 393]. In this chapter, through whole-genome sequencing and phenotypic characterisation, two novel species, *Pseudonocardia abyssalis* sp. nov. and *Pseudonocardia oceani* sp. nov., are described.

3.2. Results

The studied strains are part of a collection of 25 *Pseudonocardia* sp. strains previously isolated from sediments collected in the deep Southern Ocean, for which phylogenetic analysis of the 16 rRNA gene suggested that they might belong to undescribed species. In particular, twenty-four of these strains were distributed in two closely related clades (98.7% sequence similarity, bifurcation with 80 % bootstrap support), one of which comprised seventeen strains (of which strain KRD185 was used as a representative) and the other one comprised eight (of which strain KRD168 was used as a representative), while another

remaining strain (KRD291) represented a third clade related to *P. sediminis* [261]. In a subsequent project, eight of these strains were selected for genome sequencing using Illumina short-reads technology [262]. However, due to the low contiguity (> 250 contigs) of the obtained assemblies, it was not possible to perform an accurate genomic analysis. In this work, a hybrid assembly approach was used to obtain the complete whole-genome sequence of two of the *Pseudonocardia* sp. strains by adding PacBio long-read data.

3.2.1. Genome assembly

After sequencing the genomic DNA of the KRD168 and KRD185 using Single-Molecule Real-Time Sequencing (SMRT) long-reads technology, several assemblers were tested to select the best possible methodology (Table 3-1 and 3-2). For both genome assemblies, the Hierarchical Genome Assembly Process (HGAP) and Flye performed better in terms of contiguity (number of contigs) and accuracy (indels per 100 Kbp calculated by aligning assemblies with the short-reads assembly), while Redbean (Wtdbg2) resulted in lower quality assemblies. Although Canu performed similarly to HGAP and Flye in terms of accuracy, the assembly contiguity was lower in both cases. Raven gave assemblies with good contiguity, in particular for KRD168 with similar results to HGAP and Flye. However, its accuracy and completeness were lower than the other *de novo* assemblers.

In terms of chromosome assembly, Canu and Flye, recovered a circular contig with a size of approximately 6.8 Mb for KRD168. However, Canu failed to obtain a smaller circular contig that may represent a plasmid. On the other hand, HGAP retrieved two contigs that potentially represent the chromosome and a plasmid in agreement with Flye, but the nature of the assembly methodology did not allow the analysis of their circular character. In the case of KRD185, none of the assembly methodologies recovered contigs that represent gapless circular chromosome or plasmid sequences.

Table 3-1 Contiguity and quality evaluation of long-read assembly methods for KRD168.

	HGAP	Flye	Canu	Raven	Wtdbg2
Contigs	2	2	5	2	25
Largest contig (bp)	6,273,370	6,273,349	6,287,230	6,316,582	6,288,535
Length (bp)	6,306,129	6,306,109	6,335,444	6,349,149	6,546,436
N50 (bp)	6,273,370	6,273,349	6,287,230	6,316,582	6,288,535
Mismatches per 100 Kbp*	0.30	0.31	0.33	48.86	67.81
Indels per 100 Kbp*	0.28	0.59	0.9	278.85	182.22
BUSCO (% complete single-copy genes)	100.0%	100.0%	100.0%	43.2%	69.9%

* Mismatched statistics calculated by aligning assemblies with the short-reads assembly.

Table 3-2 Contiguity and quality evaluation of long-read assembly methods for KRD185.

	HGAP	Flye	Canu	Raven	Wtdbg2
Contigs	16	14	69	22	83
Largest contig (bp)	2,883,039	4,296,828	1,416,584	1,211,147	277,314
Length (bp)	6,572,691	6,861,690	7,135,256	6,772,303	6,613,860
N50 (bp)	9,20,330	4,296,828	5,10,607	5,32,109	185,216
Mismatches per 100 Kbp*	12.15	21.19	12.8	20.83	1,107.81
Indels per 100 Kbp*	1.30	2.64	4.10	90.85	1,043.88
BUSCO (% complete single-copy genes)	98.6%	99.0%	96.9%	59.6%	19.2%

* Mismatched statistics calculated by aligning assemblies with the short-reads assembly.

With the parameters used (Table S1), HGAP and Flye gave assemblies with similar quality values for KRD168. However, as the Flye method uses repeat graphs, it was possible to confirm the circular nature of the contigs. So, this assembler was selected to create the final assembly version (Table 3). The short reads obtained by Illumina sequencing from a previous project were then mapped over the long-reads assembly for improving the sequence by correcting bases, fixing misassemblies and filling gaps. As a result, a complete genome sequence was obtained with 1150.6x genome coverage, made by a chromosome sequence of 6273229 bp and a plasmid of 32760 bp, and with a G+C content of 73.44 %.

With the parameters used (Table S1), none of the assemblers produced a gapless sequence for KRD185. Using hybrid Unicycler, an assembly of 27 contigs was made, with a total length of 6763863 bp with an N50 value of 1390031 bp, which represents a lower quality than the long-read-only assemblies. To increase the contiguity of the KRD185 hybrid assembly, the long-reads assembly step was performed using the Flye assembly instead of the Miniasm+Racon pipeline normally used as the default setting by Unicycler. In addition, this

long-reads assembly was a consensus assembly made of the HGAP, Canu and Raven assemblies, as well as the regular Flye assembly, by using the subassembly function of Flye (Table S5). As a result of this strategy, an assembly made by a linear contig of 6661555 bp that represents the chromosome, and two circular contigs of 99100 bp and 61150 bp that presumably correspond to two plasmids, were obtained. The genome coverage for KRD185 was 217.1x, and the G+C content was 73.98 %.

As expected, the coverages of circular contigs recovered that presumably represents plasmids in the KRD168 (pPab, 32760 bp, PacBio depth = 1426.1x, Illumina depth = 324.1X) and KRD185 (pPoc2, 61150 bp, PacBio depth = 819.6x, Illumina depth = 308.7x) assemblies were higher than for their chromosomes (Fig. S1). However, another circular contig (pPoc1, 99100 bp) was recovered in the KRD185 assembly, for which the values of coverage of short-reads was higher than the chromosome (Illumina depth = 47.7x), but the long-reads coverage was lower (PacBio depth = 99.1x). The contigs' multiplicity in the assembly graph assigned by Unicycler was 2x for pPoc1, and 23x for pPoc2, which implies that pPoc2 is probably present in a higher number of copies than Poc1, explaining in part the depth differences between long and short reads. Despite the discrepancies in the reads coverage, the presence of plasmid partition system genes, such as genes encoding for *ParA* family proteins, supported the plasmid nature of the circular contig. A Pulsed-Field Gel Electrophoresis (PFGE) experiment was performed for wet-lab extrachromosomal elements profiling. Unfortunately, no confirmatory bands associated with the plasmids were recovered (Fig. S2) due to technical issues in the *in-situ* DNA extraction in the agorose gel. Therefore, wet-lab confirmation of plasmid size awaits to be addressed in future projects.

According to NCBI Prokaryotic Genome Annotation Pipeline (PGAP), the KRD168 genome (Fig. 3-2) is predicted to have 5972 protein-coding genes, 45 tRNAs and two sets of complete rRNA operons (5S, 16S, 23S). Moreover, nine Biosynthetic Gene Clusters (BGCs) were annotated with antiSMASH. In the case of the KRD185 genome (Fig. 3-2), it is predicted to have 6567 protein-coding genes, along with 46 tRNAs and two sets of complete rRNA operons (5S, 16S, 23S). For this genome, ten BGCs were annotated with antiSMASH. The circular plasmids recovered from both genomes contained regions rich in genomic islands, as well as plasmid partition system genes, confirming their extrachromosomal nature.

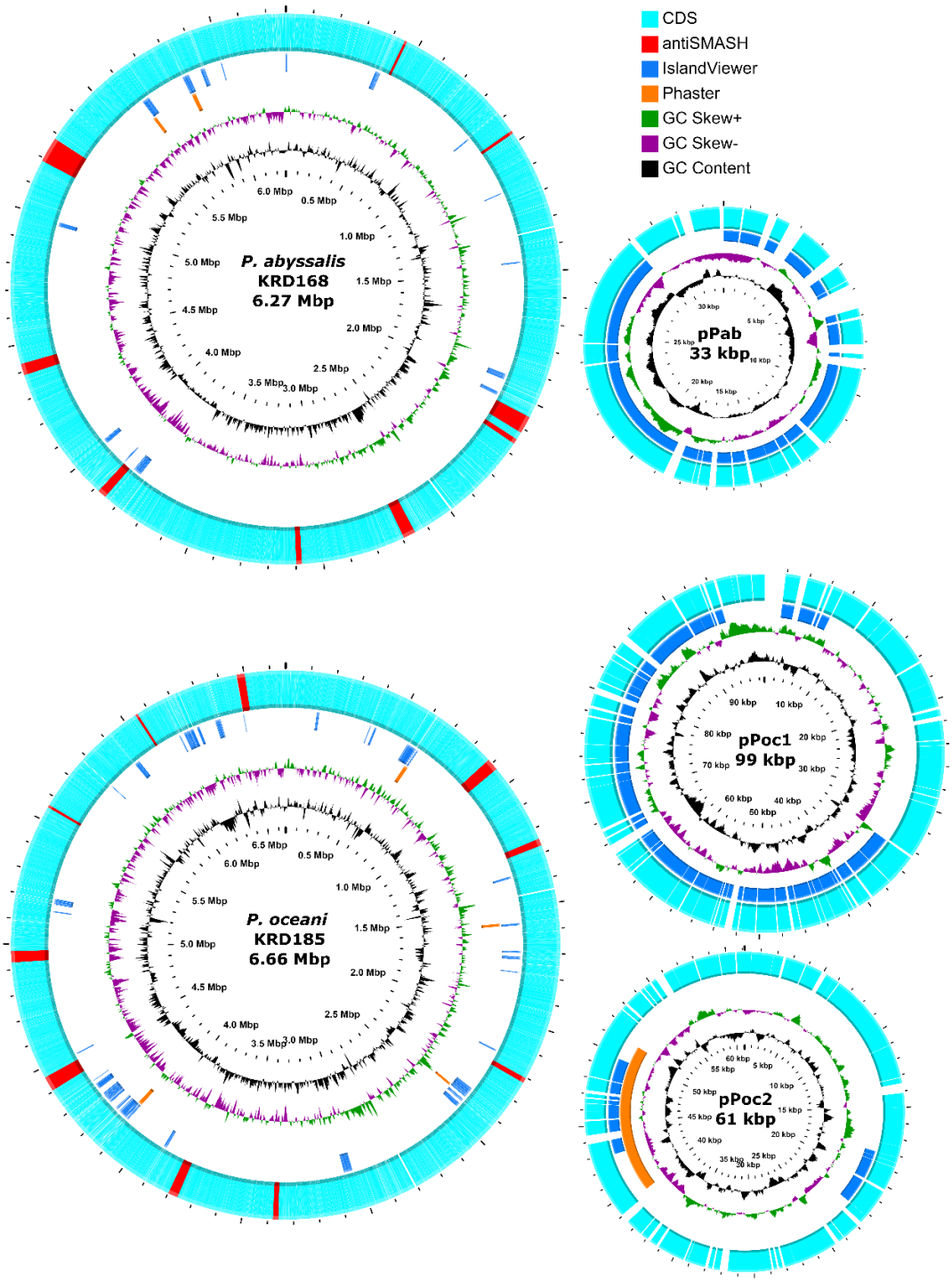


Figure 3-2 Genomic architecture overview of *P. abyssalis* KRD168 and *P.oceani* KRD185. Above. Representation of KRD168 chromosome (6.27 Mbp) and pPab plasmid (33 kbp). Below. Representation of KRD185 chromosome (6.66 Mbp), and pPoc1 (99 kbp) and pPoc2 (61 kbp) plasmids. The genomic location of annotated BGCs using antiSMASH is shown in red and the protein coding sequence (CDS) regions in cyan. Location of predicted genomic islands and detected prophages are indicated in ocean blue and orange, respectively. The next two inner circles indicate the changes in %GC content and the GC skew, respectively.

The genomic DNA of the strains KRD169, KRD176, KRD182, KRD184, KRD188, and KRD291 were sequenced by Illumina technology in a previous project. For the present project, those previously obtained raw reads were used to produce new assemblies using Unicycler (Table 3-3). As expected, assemblies with high coverage tended to show higher contiguity. The KRD291 assembly was an exception to this trend, as 225 contigs and an N50 of 18,497 bp were recovered with a coverage of 36.5x, while KRD182, KRD184, and KRD188 showed lower contiguousness with higher genome coverages. The effect of adding long-reads data on genome quality is observed by comparing the assembly contiguity values of KRD168 and KRD185 with the other genomes. The genome completeness estimated using the BUSCO (Benchmarking Universal Single-Copy Orthologues) assessment also showed the effect of the long-reads information by reducing the number of missing and fragmented single-copy orthologue genes from the Actinobacteria Dataset (actinobacteria_class_odb10). Although short-read based assemblies showed a completeness in BUSCO notation higher than 95.5%, all showed missing and fragmented single-copy orthologue genes. On the other hand, KRD168 and KRD185 showed 100% of complete single-copy orthologue genes (Fig. 3-3).

Table 3-3 Quality values and coverage for the assembly of the Southern Ocean *Pseudonocardia* genomes.

	Coverage	Contigs	Largest contig (bp)	Length (bp)	N50 (bp)
KRD168	1150.6x PacBio 215.7x Illumina	2	6,273,229	6,305,989	6,273,229
KRD169	34.0x	666	73,142	5,975,120	12,978
KRD176	124.3x	388	97,008	6,522,193	25,005
KRD182	40.9x	655	88,829	6,514,281	14,887
KRD184	53.6x	612	69,122	6,426,545	16,103
KRD185	217.1 PacBio 29.6x Illumina	3	6,661,555	6,821,805	6,661,555
KRD188	44.3x	735	64,082	6,339,047	12,830
KRD291	36.5x	525	139,490	6,571,831	18,497

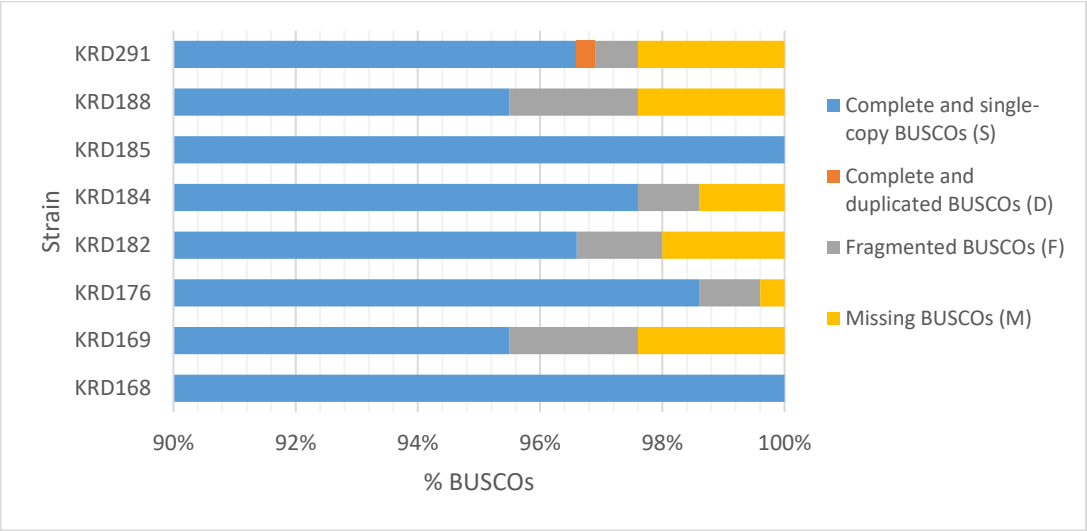


Figure 3-3 Assembly completeness assessments by Benchmarking Universal Single-Copy Orthologues (BUSCO) annotation. Bar charts show the proportion of genes present in the eight assembled Southern Ocean *Pseudonocardia* genomes. The % of complete and single copy genes (S) are shown in blue, while complete and duplicated genes (D), fragmented genes (F) and missing genes in the assemblies are shown in orange, grey and yellow, respectively.

As the main objective of the present work as a whole is to study the specialised metabolism of the *Pseudonocardia* spp. isolated from the Southern Ocean, obtaining a high-quality gapless whole-genome sequence is relevant for the analysis of Biosynthetic Gene Clusters (BGCs). The assemblies' contiguity also affected the annotation of BGCs in antiSMASH, as from the short-read based assemblies all of these were located at the edge of contigs, making it possible that the coding region (CDS) were split across different contigs. For instance, KRD168 showed more BGCs in the short-reads based assembly, including one Type I PKS (Polyketide Synthase), which is not present in the long-read assemblies (Figure 3-4). On the other hand, KRD185 had fewer annotated BGCs in the short-read assembly, which is associated with no detection of the ectoine and bacteriocin BGCs (Figure 3-4). The annotation of Non-Ribosomal Peptide Synthetase (NRPS) BGCs was affected by fragmentation the most. These differences are likely due to gene prediction errors that decreased the ability of antiSMASH to detect complete BGCs (e.g. ectoine in KRD185), as well as BGC fragmentation (eg. NRPS). These results reinforce the importance of long-read data for BGC analysis. The observed differences in the BGCs between strains will be analysed in detail in Chapter 4.

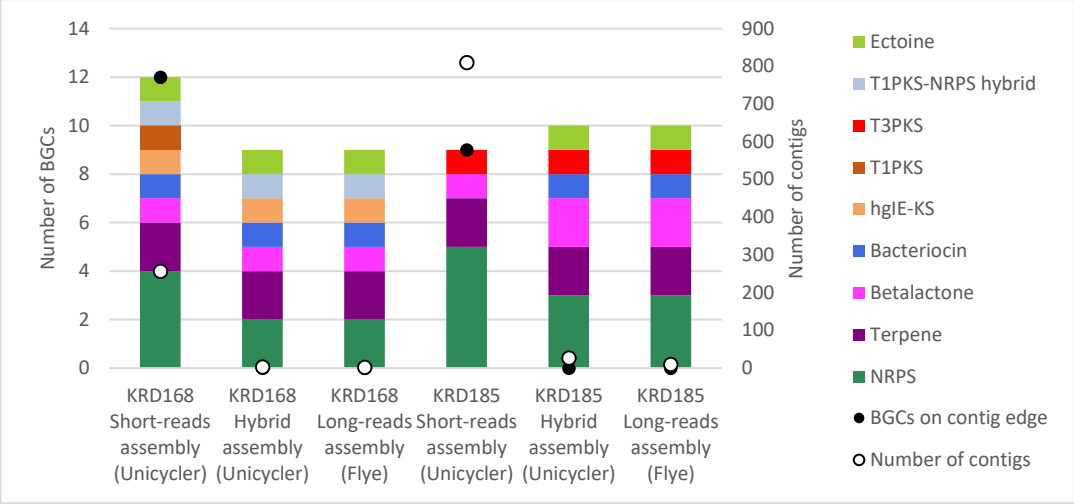


Figure 3-4 Genome assembly contiguity effect on BGC annotation in short-reads based assemblies (Unicycler), long-reads based assemblies (Flye), and hybrid assemblies (Unicycler). The colour represents antiSMASH natural product class for each BGC.

3.2.2. Comparative genomics

After obtaining the whole-genome sequence of the *Pseudonocardia* strains isolated from the deep Southern Ocean, the genomic distance between these strains was quantified through Average Nucleotide Identity (ANI) (Fig. 3-5) using the FastANI method. This analysis also included other 22 publicly available genomes from representative *Pseudonocardia* strains in the National Centre for Biotechnology Information (NCBI) database (Table S2). Following a pragmatic approach, a threshold ANI value of 95% to delimit species, it was shown that KRD176, KRD182, KRD184, KRD185, and KRD188, belong to the same species, which is closely related to *P. broussonetiae* Gen 01 (ANI ≈ 93.5%). A similar situation was identified with KRD168 and KRD169, for which an ANI value of 99.7% suggested that they belong to the same species, which is closely related to *P. petroleophila* CGMCC 4.1532 (ANI ≈ 91.7%). According to the ANI values these two identified species for the KRD strains are different to any other previously reported *Pseudonocardia* spp., suggesting two potentially novel species that will be named henceforth *Pseudonocardia oceani* (type strain: KRD185) and *Pseudonocardia abyssalis* (type strain: KRD168) respectively. In the case of KRD291, there was no ANI value higher than 82.5%, except for *P. sediminis* DSM 45779 (ANI = 88.9%) Although both genomes of KRD291 and *P. sediminis* have assemblies with low completeness, it is likely that KRD291 also represents a new species. However, further analysis would be required to confirm this.

Finally, the ANI values also revealed that the *P. abyssalis* and *P. oceani* genomes showed a high similarity (ANI ≈ 87.7%), suggesting that both lineages are closely related to each other.

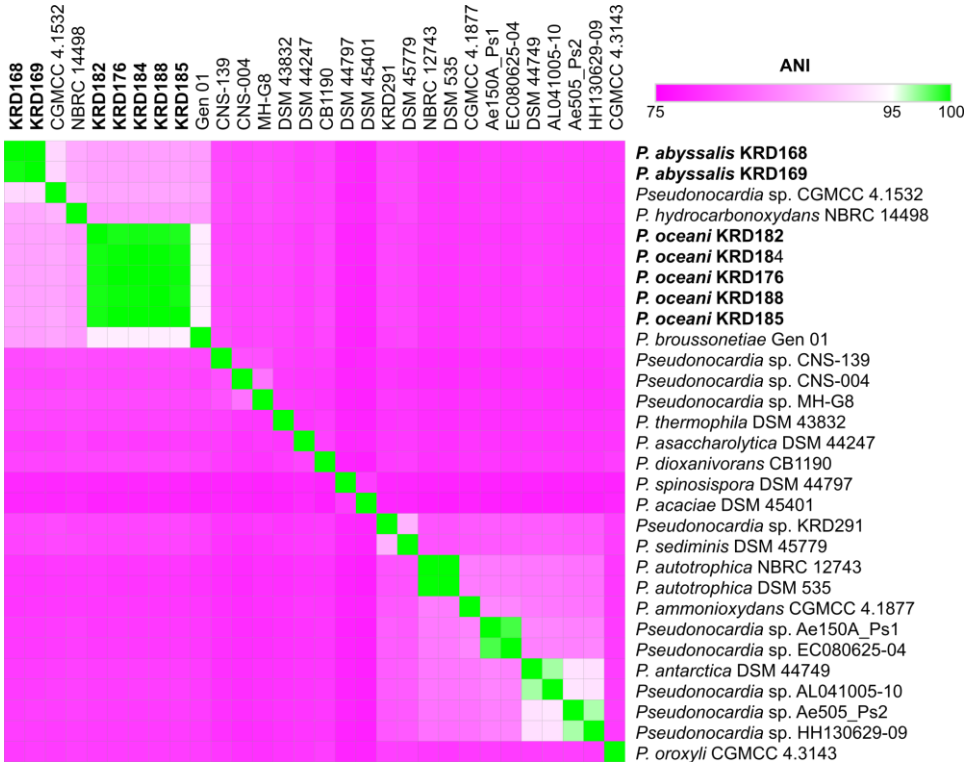


Figure 3-5 Average Nucleotide Identity (ANI) heatmap showing the sequenced Southern Ocean *Pseudonocardia* strains alongside with representative genomes of type and closest strains from the NCBI database. The ANI value is coloured from 75% to 100% (magenta, 75%; white, 95%; green, 100%). The Southern Ocean *Pseudonocardia* strains are highlighted.

To obtain more evidence that *P. abyssalis* and *P. oceani* represented novel species, the digital DNA–DNA hybridization (dDDH) values between *P. abyssalis* KRD168, *P. oceani* KRD185, and other publicly available *Pseudonocardia* genomes (Table S2) were calculated using the Genome-to-Genome Distance Calculator (GGDC). The calculated values (Fig. 3-6) showed that *P. abyssalis* KRD168 and *P. petroleophila* CGMCC 4.1532 have a dDDH of 43.5%, and an ANI of 91.8%, whilst KRD185 and *P. broussonetiae* Gen01 showed dDDH and ANI values of 51.7% and 93.7%, respectively. In both cases, the ANI values are below the recommended inter-species boundary value of 95%. Furthermore, the dDDH values also support the inter-species delimitation as they were below the threshold value of 70 %.

In the case of KRD291, the dDDH value with the closest related strain, *P. sediminis* DSM 45779, was 34%, which supported the possibility of third novel species in agreement with the ANI values. However, as mentioned above, additional data would be required to complete the KRD291 genome for a more comprehensive analysis.

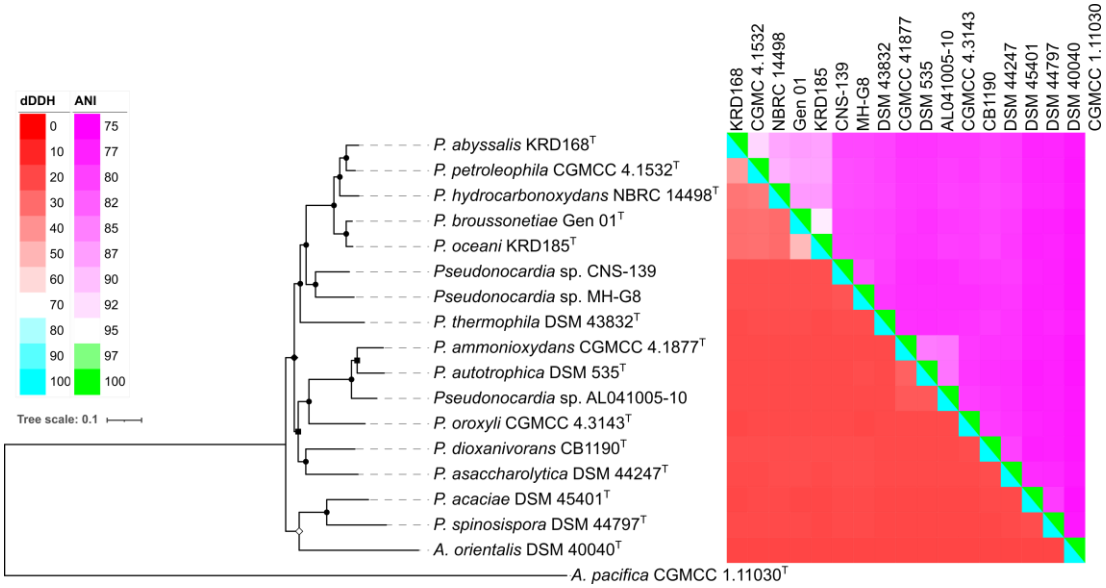


Figure 3-6 Heatmap of the genomic distance between *P. abyssalis* KRD168, *P. oceani* KRD185, and other *Pseudonocardia* spp. strains, calculated as dDDH (red, 0%; white, 70%; cyan, 100%) and ANI (magenta, 75%; white, 95%; green, 100%). The maximum-likelihood multi-locus tree based on 93 gene sequences extracted from the whole-genome sequence is shown. *Albimonas pacifica* CGMCC 111030^T was used as an outgroup. Bootstrap support is indicated as symbols in the branches (●= 100, ◆= 99, ■= 98, ◇= 95). Bar, 0.1 substitutions per nucleotide position.

As FastANI is an alignment-free method, pairwise genomic distances were confirmed by calculating ANIm, which uses MUMmer to align the genome sequences. This analysis included all the publically available genomes in the NCBI genome database at the time of writing (June 2022), including metagenome-assembled genomes (MAGs). In short, the ANIm analysis (Fig. 3-7) supported the species delimitation for the Southern Ocean *Pseudonocardia* strains at the 95% identity threshold. Moreover *P.abyssalis* and *P. oceani* showed and ANIm of 87,9%.

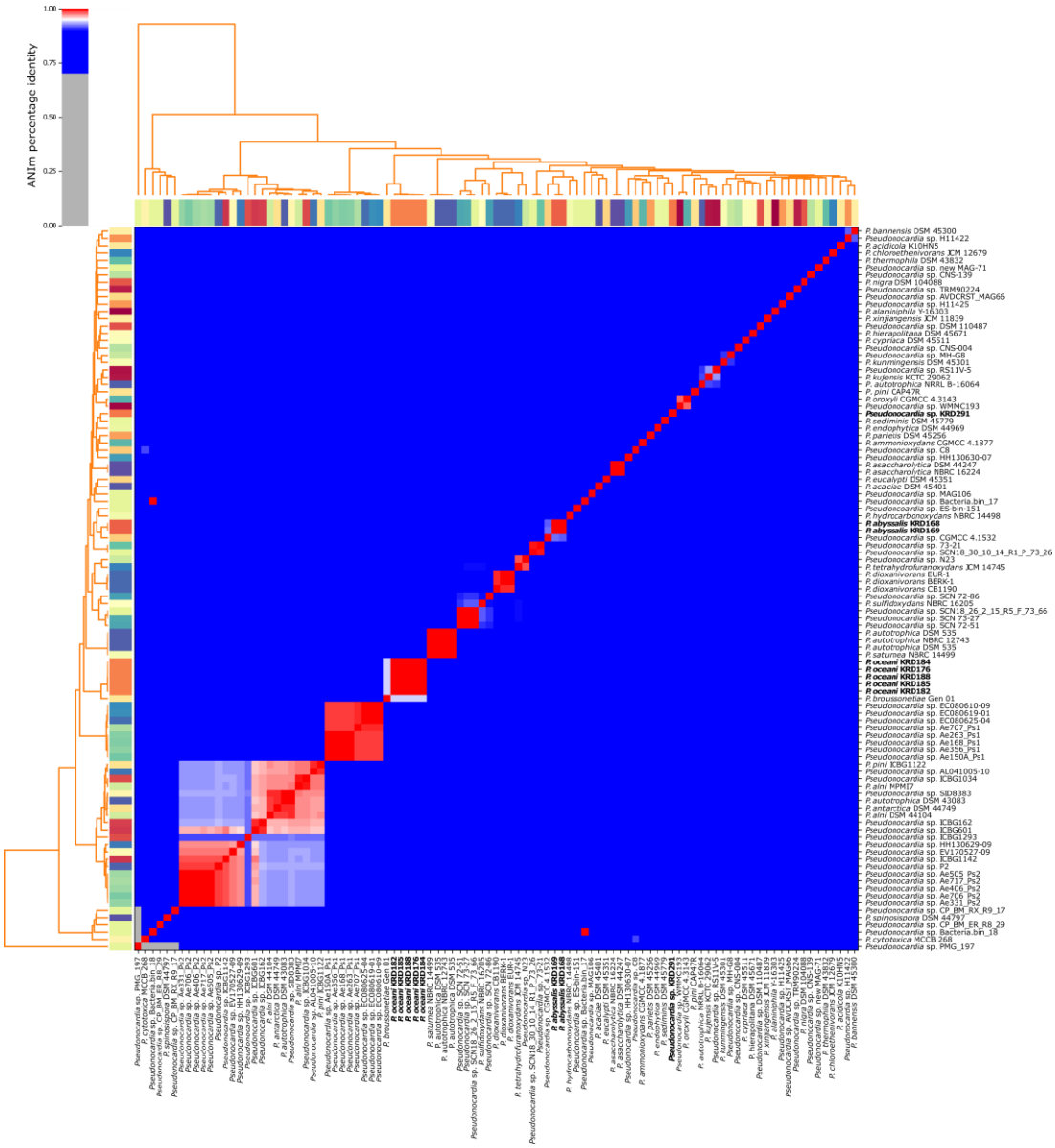


Figure 3-7 Heatmap of ANIm percentage identity for 99 genomes of *Pseudonocardia* spp. Species-level assignments and identifiers are shown as row and column labels. The ANIm percentage of identity is coloured from 0.75 to 1. Cells coulored in red correspond to 95% ANIm sequence similarity, and therefore the corresponding organisms belong to the same species. Cells coulored in blue correspond to ANIm comparisons indicating that the corresponding organisms do not belong to the same species. Hierarchical clustering of the analysis results is represented by dendrograms. The Southern Ocean Pseudonocardia strains are highlighted.

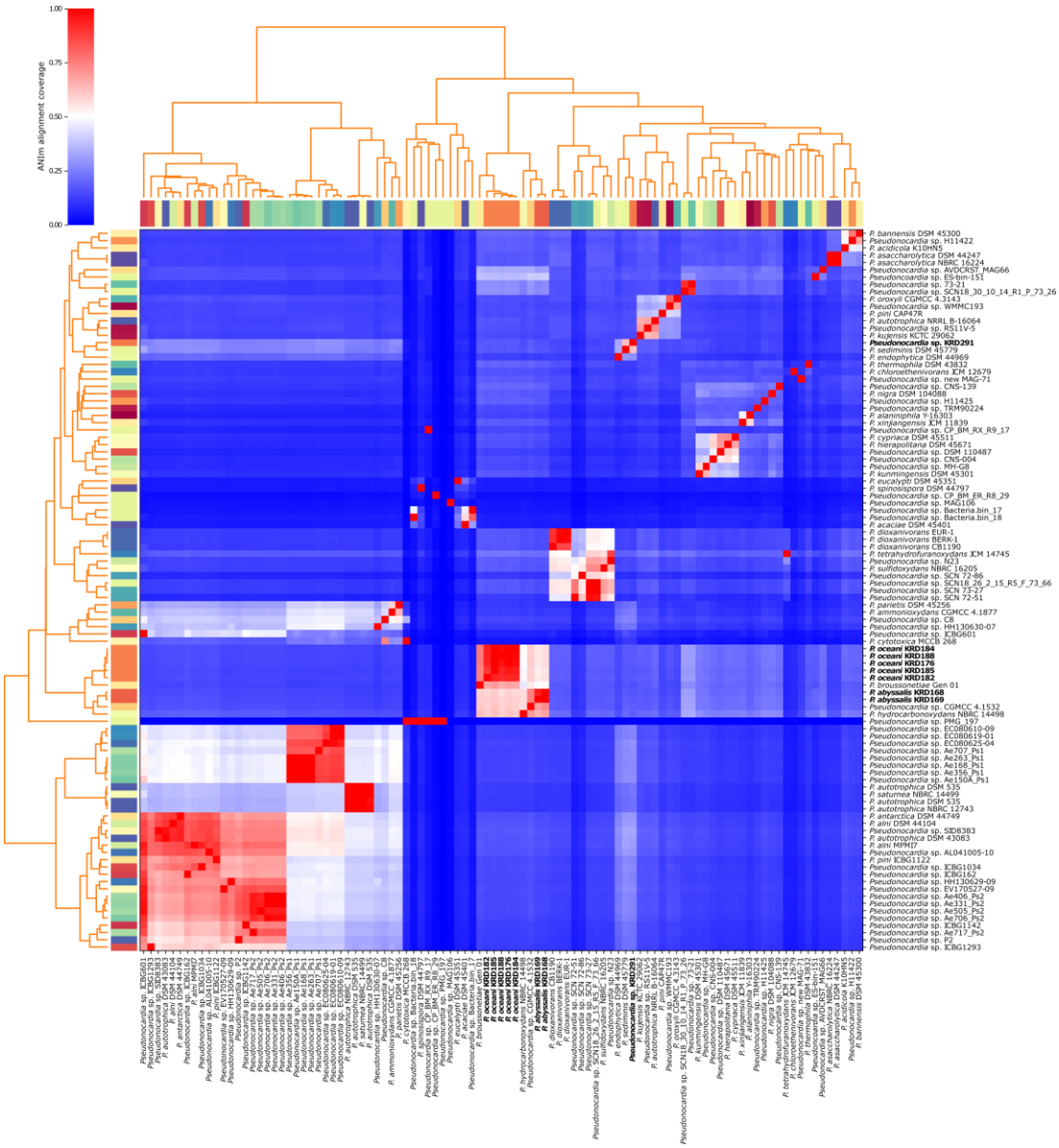


Figure 3-8 Heatmap of ANIm coverage for 99 genomes of *Pseudonocardia* spp. Species-level assignments and identifiers are shown as row and column labels. Red cells correspond to coverage of 50% or, while blue cells correspond to coverage of 50% or less. Hierarchical clustering of the analysis results is represented by dendrograms. The Southern Ocean *Pseudonocardia* strains are highlighted.

The pairwise coverage of each genome by aligned regions was also analysed. The alignment coverage plot (Fig. 3-8) shows a cluster of related species comprised of *P. abyssalis*, *P. oceani*, *P. broussonetiae*, *P. petroleophila*, and *P. hydrocarbonoxydans*, which were part of a same clade in the maximum-likelihood multi-locus analysis (Fig. 3-6). The members of this cluster align to each other over more than 50% of their total genome length. For example, the ANIm coverage between *P. abyssalis* KRD168 and *P. oceani* KRD185 was 59.8%. Similar values were

retrieved for *P. abyssalis* KRD168 and *P. petroleophila* CGMCC 4.1532 (74.1%), *P. oceani* KRD185 and *P. petroleophila* CGMCC 4.1532 (58.4%). In the case of *P. sediminis* DSM 45779 and *Pseudonocardia* sp. KRD291, ANIm coverage was 62.1%. Although there is no a widely accepted interpretation of species boundary by alignment coverage an aligned genome length above 50% gives good evidence that the aligned genomes possibly correspond to the same taxon. Therefore, this analysis confirms the genomic correlation among the members of this clade.

3.2.3. Phylogenomic analysis

Analysis using the 16S rRNA gene showed that the nearly complete (93.1%) gene sequence of *P. abyssalis* KRD168 was closely related to *P. petroleophila* ATCC 15777 (99.6% sequence similarity), *P. hydrocarbonoxydans* NRRL B-16171 (99.4% sequence similarity), and *P. seranimata* YIM 63233 (99.3% sequence similarity). Similarly, the nearly complete (95.0%) 16S rRNA gene sequence of *P. oceni* KRD185 was closely related to *P. broussonetiae* Gen01 (99.6% sequence similarity), *P. petroleophila* ATCC 15777 (98.9% sequence similarity), *P. hydrocarbonoxydans* NRRL B-16171 (98.4% sequence similarity), and *P. seranimata* YIM 63233 (98.4% sequence similarity). The neighbour-joining phylogenetic tree based on this gene (Fig. 3-7) reconstructed with MegaX (Table S1) showed that both isolates formed a monophyletic group with *P. petroleophila* ATCC 15777, *P. seranimata* YIM 63233, and *P. broussonetiae* Gen01. In particular, *P. abyssalis* KRD168, *P. seranimata* YIM 63233, and *P. petroleophila* ATCC 15777 are part of the same branch, while and *P. oceani* KRD185 and *P. broussonetiae* Gen01 formed a second branch. This topology was also supported by the maximum-likelihood (ML) and maximum-parsimony (MP) trees (Fig. S3).

The genome-scale phylogenetic tree (Fig. 3-6) based on the multi-locus sequence analysis of 93 loci (Table S3) and reconstructed with autoMLST suggested a common ancestor for *P. abyssalis* and *P. oceani*, as well as *P. petroleophila* and *P. broussonetiae*, observed in the 16S rRNA gene phylogeny. Interestingly, despite the 16S rRNA gene phylogeny suggesting a more distant relationship, multi-locus sequence analysis showed that *P. hydrocarbonoxydans* is closely related to *P. abyssalis* and *P. oceani*. The core-genome phylogeny (Fig. S3) identified and classified the orthologous gene families in *Pseudonocardia* spp. pan-genome (Fig. 3-8A) and this supported the evolutionary relationships observed in the multi-locus tree, showing the *P. abyssalis* and *P. oceani* are part of a well-differentiated monophyletic group alongside

P. petroleophila, *P. hydrocarbonoxydans*, and *P. broussonetiae*. The marine-derived *Pseudonocardia* strains CNS-139 and MG-H8 showed a close evolutionary relationship, while *P. thermophila* was shown to be the closest species.

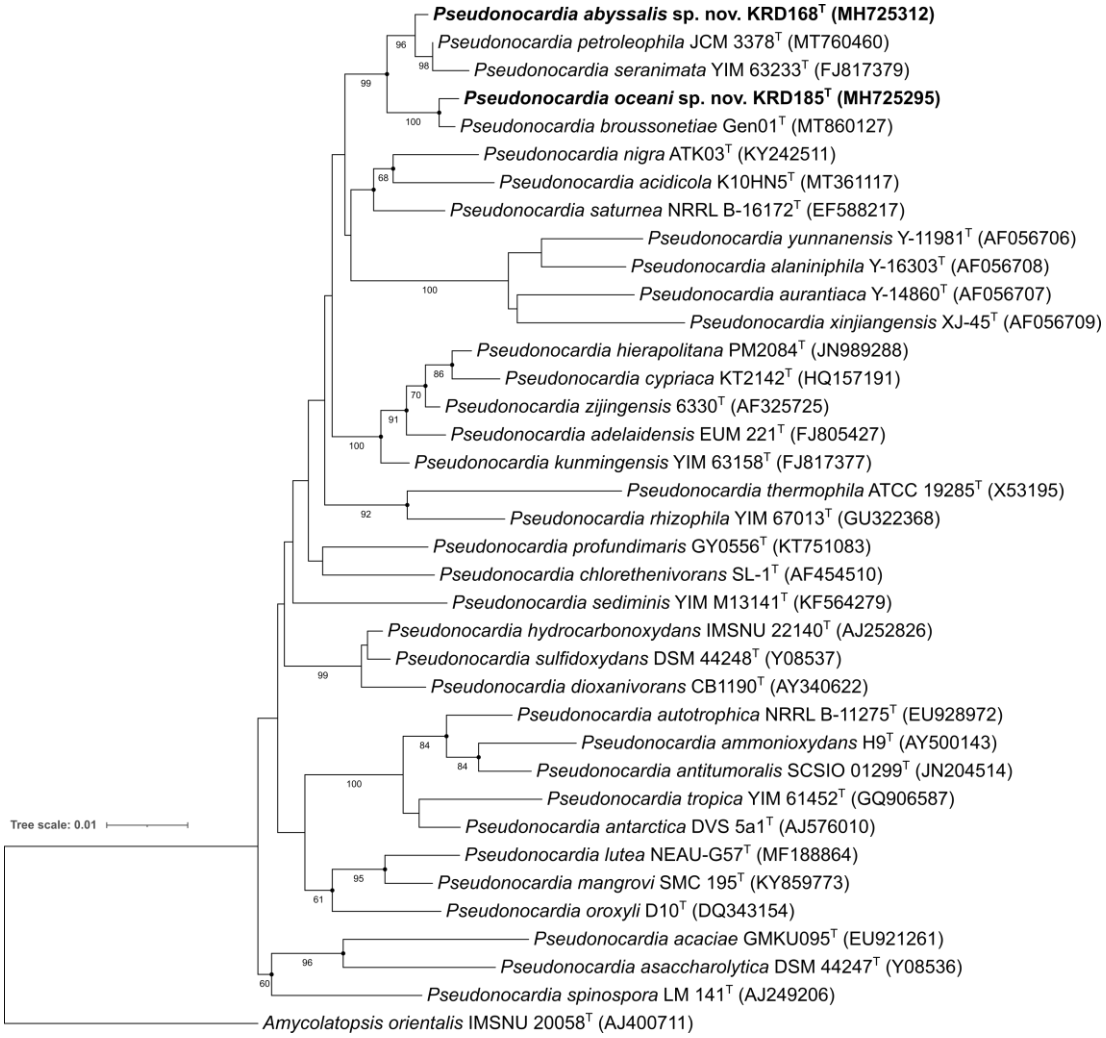


Figure 3-9 Neighbour-joining phylogenetic tree based on 16S rRNA gene sequences showing the phylogenetic relationships of *P. abyssalis* KRD-168 and *P. oceani* KRD-185, and representative members of the genus *Pseudonocardia*. Dots on branches indicate those also recovered in the maximum-likelihood and maximum-parsimony trees. Bootstrap values of >50% are shown at branching points. *Amycolatopsis orientalis* IMSNU 20058^T was used as outgroup. The scale bar represents 0.01 substitutions per nucleotide position.

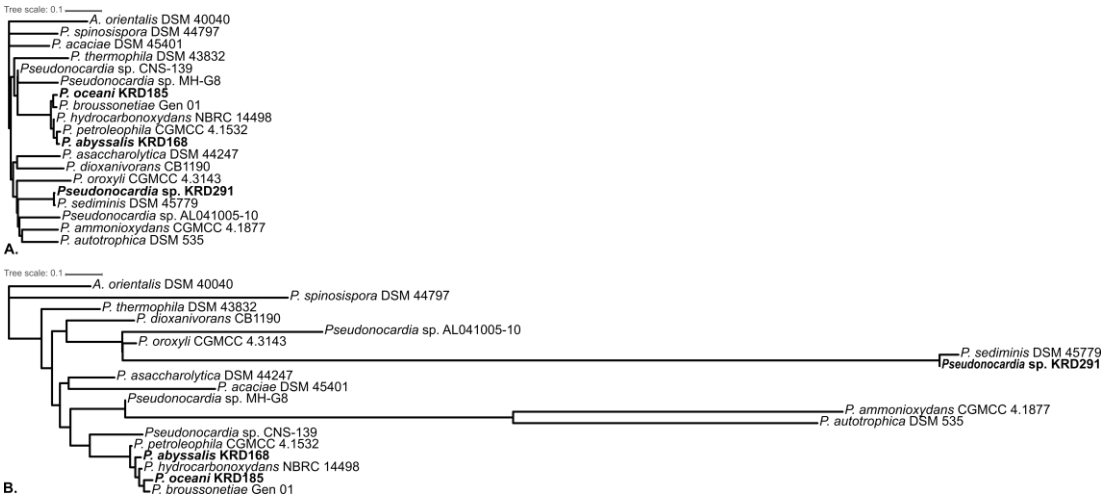


Figure 3-10 Maximum likelihood tree based on the core-genome alignment identified by PIRATE (A) and core-genome alignment after removing recombination events identified by ClonalFrameML (B). The Southern Ocean *Pseudonocardia* strains are highlighted. Bar, 0.1 substitutions per nucleotide position.

To assess the effect of homologous recombination (Table S6) on the *Pseudonocardia* spp. lineages, the maximum likelihood clonal frame phylogeny (Fig. 3-8B) was reconstructed after masking the core-genome section with evidence of recombination identified by ClonalFameML. The monophyletic group made by *P. abyssalis*¸ *P. oceani*, *P. petroleophila*, *P. hydrocarbonoxydans*, and *P. broussonetiae* remained well-defined from other species, suggesting gene-flow barriers between this clade and other lineages. Therefore, it is possible to deduce that the *P. abyssalis*¸ *P. oceani*, *P. petroleophila*, *P. hydrocarbonoxydans*, and *P. broussonetiae* lineages ramified from other *Pseudonocardia* populations in ancient times, generating a clear genome divergence, while the homologous recombination is still a cohesive force among the clade. The estimated ratio of recombination to mutation (r/m) calculated for each genome into the clade (*P. abyssalis* = 2.619334, *P. oceani* = 1.932329, *P. petroleophila* = 1.495581, *P. hydrocarbonoxydans* = 4.840895, *P. broussonetiae* = 1.261635) were higher than the value for clade's branch (0.018631), which implies that the detected recombination events happened mostly between the members of the clade.

In the case of KRD291 and *P. sediminis*, the distances between their clade, and other *Pseudonocardia* population was remarkably increased. These results imply a clear gene-exchange barrier between these two lineages and other species.. Similar to the *P. abyssalis* and *P. oceani* clade, the estimated r/m values (Table S6) denote a high level of recombination between the clade's members, but a low level between clade itself and other lineages.

In summary, despite the high sequence similarity (>99.5%) observed using the 16S rRNA gene between *P. abyssalis* KRD168 and *P. petroleophila* ATCC 15777, and between *P. oceani* KRD185 and *P. broussonetiae* Gen01, the whole-genome sequence relatedness, as well as the genome-level evolutionary analysis, evidenced that *P. abyssalis* and *P. oceani* are two novel species.

3.2.4. Chemotaxonomy and phenotypic characterization

To confirm that *P. abyssalis* KRD168 and *P. oceani* KRD185 represent two novel species as suggested by the genomic analysis, a phenotypic characterisation and a subsequent comparison between them and their closest phylogenetic relative (*P. petroleophila*, *P. hydrocarbonoxydans*, and *P. broussonetiae*) was carried out. The main differential characteristics are summarised in Table 3-4.

In terms of their morphology, strains grew on all media tested, with no diffusible pigments produced. Despite their marine origin, *P. abyssalis* KRD168 and *P. oceani* KRD185 did not display a particularly high halotolerance in comparison to their non-marine closest relatives as they both grew well at a NaCl concentration of 3%, which was 1% less than *P. petroleophila* DSM 43193 and 1% more than *P. hydrocarbonoxydans* DSM 43281. Moreover, salt was not a requirement for growth. Their morphology observed under the microscope (Fig. 3-11A) revealed budding substrate and aerial hyphae with swelling and side branches. A more filamentous structure was observed for *P. oceani* KRD185 than for *P. abyssalis* KRD168, showing longer mycelia and a higher number of ramifications. Spores were rod-like, mostly 0.6-1.4 µm wide and 1.2-1.9 µm long for *P. abyssalis* KRD168, and 0.7-1.6 µm wide and 1.4-3.9 µm long for *P. oceani* KRD185. Furthermore, fluorescence microscopy (Fig. 3-11B) was performed to describe structures in the aerial hyphae and apical tip growth. In brief, nascent peptidoglycan was observed at the tips of the hyphae, denoting elongation of the cell wall. In addition, active growth and sporulation were still observed after 28 days.

Table 3-4 Differential phenotypic characteristics of strains: 1, *P. abyssalis* KRD168; 2, *P. oceani* KRD185; 3, *P. petroleophila* DSM 43193 from Zhao *et al.* [133]; 4, *P. hydrocarbonoxydans* DSM 43281 from Zhang *et al.* [151]; and 5, *P. broussonetiae* Gen 01 from Mo *et al.* [140].

	1	2	3	4	5
Morphology in					
ISP2	Yellow	Orange	Yellow	Yellow	Yellow
ISP5	White	Pale yellow	White	White	Yellow
Growth at					
NaCl (%)	0-3	0-3	0-4	0-2	0-8
pH	6-10	6-10	6-10	6-9	5-8
API® 20E					
β-galactosidase	-	-	+	+	
Arginine dihydrolase	-	-	+	+	
Lysine decarboxylase	-	-	+	+	
Ornithine decarboxylase	-	-	+	+	
Citrate utilization	-	-	+	+	
Urease	+	-	+	+	
Acetoin production	+	-	+	+	
Gelatine hydrolysis	+	-	+	+	
API® zym					
Alkaline phosphatase	+	+	+	+	-
Esterase (C4)	+	+	+	-	+
Cystine arylamidase	-	+	+	+	-
Trypsin	-	-	+	+	-
α-chymotrypsin	-	-	+	+	-
Naphthol-AS-BI-phosphohydrolase	+	+	-	+	+
α-galactosidase	-	-	+	+	-
β-galactosidase	+	-	+	+	-
α-glucosidase	-	+	+	+	-
Fatty acids (>1%)					
12: 0	n.d.	n.d.	n.d.	n.d.	1.2
14: 0	n.d.	n.d.	n.d.	n.d.	2.9
14: 0 iso	7.6	1.9	2.0	1.5	5.6
15: 1 ω6*c*	n.d.	n.d.	n.d.	1.9	2.6
15: 0 iso	11.0	7.7	25.2	17.9	2.1
16: 1 iso H	3.2	8.9	2.6	5.8	10.7
16: 0 iso	36.3	38.5	43.2	31.7	36.1
16: 1 2-OH	n.d.	n.d.	2.4	n.d.	n.d.
16: 0	4.7	3.6	2.0	3.6	9.8
17: 0 iso	4.2	2.0	10.2	3.2	n.d.
17: 0 anteiso	1.1	2.4	n.d.	1.4	n.d.
17:1 ω8*c*	8.0	5.1	n.d.	8.3	4.0
17: 0	3.4	1.6	n.d.	1.7	1.4
17: 0 10 methyl	2.4	2.7	1.5	1.9	n.d.
18: 1 ω9*c*	1.2	n.d.	n.d.	n.d.	n.d.

[+] = enzymatic activity present; [-] = enzymatic activity absent; n.d. = not detected

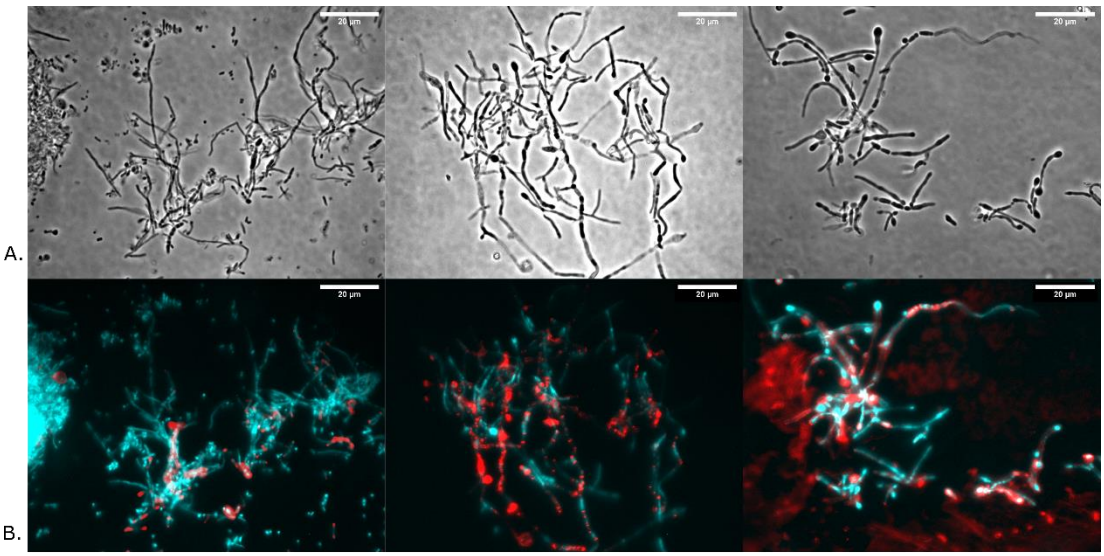


Figure 3-11 Light microscopy (A) of 1, *P. abyssalis* KRD168; 2, *P. oceani* KRD185; and 3, *P. petroleophila* DSM 43193; growing in ISP2 for 28 days. Fluorescence microscopy (B) of DNA (cyan) and nascent peptidoglycan (red) is shown. Scale bar, 20 µm.

The observed morphology of both strains, as well as the presence of *meso*-diaminopimelic acid in the cell-wall, MK-8(H_4) as the major menaquinone, and *iso*-branched hexadecanoic acid as predominant fatty acid, confirmed the typical physiology of a *Pseudonocardia* species. The polar lipid profile (Fig. S5) of both strains was characterised by the presence of phosphatidylethanolamine and diphosphatidylglycerol. In addition, two unidentified phospholipid and four unidentified glycolipids were detected in *P. abyssalis* KRD168, whilst *P. oceani* KRD185 also contains four unidentified phospholipid, two unidentified glycolipids, and an unidentified glycophospholipid. Moreover, some chemotaxonomic characteristics, such as the metabolic profile and the cellular fatty acid composition (Table 3-4), differentiated *P. abyssalis* KRD168 and *P. oceani* KRD185 from *P. petroleophila* DSM 43193 and *P. broussonetiae* Gen 01, respectively. In particular, *P. abyssalis* KRD168 was negative to the α-galactosidase and α-glucosidase activity reactions, while positive activity was reported for *P. petroleophila* DSM 43193. Similarly, *P. oceani* KRD185 was positive to the α-glucosidase activity reaction, while *P. broussonetiae* Gen 01 reported no activity.

The fatty acid analysis (Table 3-4) also revealed differences beween *P. abyssalis* KRD168 and *P. oceani* KRD185 from *P. petroleophila* DSM 43193 and *P. broussonetiae* Gen 01, respectively. Particularly, the relatively low proportion of *iso*-branched pentadecanoic acid (11.0% vs 25.2%) and the relatively high proportion of *iso*-branched tetradecanoic acid (7.6%

vs 2.0%) distinguished *P. abyssalis* KRD168 from *P. petroleophila* DSM 43193. Likewise, the relatively low proportion of *iso*-branched tetradecanoic acid hexadecanoic acid (1.9% vs 5.6%) and the relatively high proportion of *iso*-branched pentadecanoic acid (7.7% vs 2.1%) differentiated *P. oceani* KRD185 from and *P. broussonetiae* Gen 01.

3.2.5. Species description

Based on the phenotypic and chemotaxonomic analysis, strains *P. abyssalis* KRD168 and *P. oceani* KRD185 exhibit the characteristic markers for the *Pseudonodardia* genus. Furthermore, the phenotypic characterization and the analysed genomic relatedness of *P. abyssalis* KRD168 and *P. oceani* KRD185 differentiate them from their closest phylogenetic neighbours. In conclusion, these strains, alongside *P. abyssalis* KRD169, *P. oceani* KRD176, *P. oceani* KRD182, *P. oceani* KRD184, and *P. oceani* KRD188, are proposed to belong to two novel species with the following description:

Description of *Pseudonocardia abyssalis* sp. nov.

Pseudonocardia abyssalis (a.bys'sa.lis. L. n. abyssus, an abyss, deep sea; L. fem. suff. -alis, suffix denoting pertaining to; N.L. masc. adj. *abyssalis*, pertaining to the abyssal depths of the ocean).

Aerobic, Gram-positive, non-motile actinomycetes. Forms yellow aerial and substrate mycelia on ISP2, while white substrate and aerial mycelia are produced on ISP3, ISP4 and ISP5. Substrate mycelium is fragmented and rod-shaped spore chains are formed on aerial hyphae and substrate mycelium. The following enzymatic reactions are positive: Urease, acetoin production, gelatine hydrolysis, alkaline phosphatase, esterase, esterase-lipase, leucine-arylamidase, valin-arylamidase. The major menaquinone is MK-8(H_4) (91.5%), while MK-8(H_6) (8.5%) is also present. The cell wall contains *meso*-diaminopimelic acid as diagnostic amino acid, and the major whole-cell sugar is glucose with minor amounts of arabinose and xylose. The polar lipids profile includes phosphatidylethanolamine and diphosphatidylglycerol. The predominant fatty acid is iso-C16:0, while iso-C15:0, C17:1 ω8c, and iso-C14:0 are also major components.

The type strain, KRD168^T (=DSM 111918^T =NCIMB 15270^T), was isolated from sediments collected from the Southern Ocean at a depth of 4539 m. The genome of the type strain is characterized by a size of 6.31 Mbp and a G+C content of 73.44 %.

Description of *Pseudonocardia oceani* sp. nov.

Pseudonocardia oceani (o.ce.a'ni. L. gen. n. *oceani* of the ocean).

Aerobic, Gram-positive, non-motile actinomycetes. Forms yellow substrate mycelium and orange aerial mycelium with white spores formed from the aerial hyphae on ISP2. On ISP3, ISP4 and ISP5 white substrate and aerial mycelia are produced. Substrate mycelium is fragmented and rod-shaped spore chains are formed on aerial hyphae and substrate mycelium. The following enzymatic reactions are positive: alkaline phosphatase, esterase, esterase-lipase, leucine-arylamidase, valin-arylamidase, cystin-arylamidase, acid phosphatase, naphtol-AS-BI-phosphohydrolase, α-glucosidase, and β-glucosidase. The major menaquinone is MK-8(H_4) (96.2%), while MK-8(H_2) (2.9%) and MK-8(H_6) (1.5%) are also present. The cell wall contains *meso*-diaminopimelic acid as diagnostic amino acid, and the major whole-cell sugar is glucose with minor amounts of arabinose and xylose. The polar lipids profile includes phosphatidylethanolamine and diphosphatidylglycerol. The predominant fatty acid is iso-C16:0, while iso-C16:1 H, iso-C15:0, and C17:1 ω8c are also major components.

The type strain, KRD185^T (=DSM 111919^T = NCIMB 15269^T), was isolated from sediments collected from the Southern Ocean at a depth of 4060 m. The genome of the type strain is characterized by a size of 6.82 Mbp and a G+C content of 73.98 %.

3.3. Discussion

In general, no single genome assembly method performed the best on all metrics measured. However, Flye has previously been shown to be a well-balanced assembler for prokaryote genomes [394, 395]. The high coverage of the *P. abyssalis* KRD168 assembly allowed a complete-level genome with a low number of misassemblies using only the PacBio long-reads data. On the other hand, no stand-alone method was able to produce a complete genome

for KRD185, possibly due to its lower read coverage in comparison to KRD168. The consensus assembly made by Flye, Canu, HGAP and Raven increased the KRD185 assembly contiguity but failed to recover circular contigs. Due to the size selection step in the PacBio library preparation, plasmid reads may be underrepresented in the raw data. As a consequence, the only-long-reads assemblies may fail to recover plasmid sequences [394, 396]. This phenomenon was clearly observed in coverage of pPoc1 between the Illumina and PacBio reads. In this sense, combining PacBio and Illumina reads using Unicycler has proven to be a highly capable method to recover plasmids [284, 394]. As mentioned above, Unicycler was used on the long-read data to simplify the graphs produced with short-reads. The default setting for the hybrid assembly produced a long-reads assembly using Miniasm and Racon. This assembly is then used to build bridges to resolve the short-reads assembly graphs produced by SPAdes [284]. The proposed strategy for assembling the KRD185 genome used Flye's consensus long-reads assembly rather than Miniasm, this successfully reduced the number of contigs. The combination of short and long reads not only improved the contiguity and completeness of the genome assembly, but also the recovery of circular plasmids. For example, none of the assemblies using only one set of reads were able to recover circular contigs for KRD185. Small plasmids are often present in multiple copies, while bigger conjugative plasmids are usually present in one copy. To resolve whether small contigs correspond to plasmids, Unicycler applies a greedy algorithm that uses depth and graph connection to determine multiplicity values [284]. As the largest suitable contig (putative chromosome) is given a value of one, the multiplicity in the circular contigs retrieved in the KRD185 consensus hybrid assembly confirmed their extrachromosomal identity. Unfortunately, as the sequence of the KRD185 chromosome was not circular, it is not possible to be sure that the whole-genome sequence for this strain represents a complete-level genome. However, the obtained sequence is a high-quality representation of the KRD185 genome with a high level of completeness according to the BUSCO approach.

The completeness and contiguity of the sequenced strains reduced the fragmentation of open reading frames (ORFs). This fragmentation can produce false negatives during gene annotation and eliminating it can avoid splitting BGCs that may be wrongly annotated as two different clusters. For example, in a study with *Streptomyces* draft genomes, it was found that fragmentation is particularly common on protein classes with repeating multi-modular structures such as PKS, NRPS and serine/threonine kinases [397]. The effect of BGC fragmentation has been also reported for the *Pseudonocardia* genus. For instance, it was

shown that a higher number of BGCs were annotated using Oxford Nanopore MinION long-reads than Illumina MiSeq short-reads in three *Pseudonocardia* spp. assemblies and their BGCs were less fragmented. This was also observed for KRD *Pseudonocardia* strains, as the differences in the BGC profiles were mostly related to NRPS and PKS clusters, while terpene clusters were more stable [398]. High-quality assemblies are important for the analysis of the specialised metabolisms of the Southern Ocean strains in Chapter 4.

Based on ANI, the eight sequenced *Pseudonocardia* strains are distributed in three populations, confirming the 16S rRNA gene analysis previously performed by Millán-Aguiñaga *et al.* [261]. Furthermore, the genomic-level species delimitation based on dDDH [362] and ANI [362, 399] suggested that the three lineages represent three new species. However, the interpretation of the calculated ANI values depends on the individual evolutionary history of each lineage, as well as the applied algorithm. So, the new-species hypothesis for *P. abyssalis* KRD168 and *P. oceani* KRD185 strains should be confirmed based on evolutionary history as part of the polyphasic taxonomy [400]. On the other hand, the genomic-level species delimitation between *Pseudonocardia* sp. KRD291 and *P. sediminis* have to be carefully examined. In terms of the quality of a genome sequence for taxonomic purposes in prokaryotes, such as identification and classification, it has been recommended a minimum standard of a sequencing depth of coverage ≥ 50X to obtain reliable results [401]. As the *Pseudonocardia* sp. KRD291 genome does not meet these recommendations, conclusions about genomic differences between *Pseudonocardia* sp. KRD291 and *P. sediminis* have to be reviewed in the future, either by adding genomic information or by carrying out morphological and biochemical tests.

The lack of taxonomic resolution of the 16S rRNA gene to discriminate *P. abyssalis* and *P. oceani* was proved after the genotypic and phenotypic characterization of both strains. Similarly to the species delimitation for members of the genus *Salinispora* [353, 354], the comparative genomic analysis demonstrated that *P. abyssalis*, *P. oceani*, *P. petroleophila*, *P. hydrocarbonoxydans*, and *P. broussonetiae* represent different species despite the high similarity in their 16S rRNA gene sequence. Unfortunately, the whole-genome sequence of *P. seranimata* [133], another closely related species according to the 16S rRNA gene sequence, was not publicly able to include it in the genomic analysis. The high similarity of 16S rRNA gene sequences between members of this phylogenetic group was reported previously by Mo *et al*. in the taxonomic description of *P. broussonetiae* [140]. For this

species, it was shown that despite sharing 98.9% of the 16S rRNA gene sequence similarity to *P. petroleophila*, the ANI was 84.6 %, and the morphological, physiological and biochemical characteristics confirmed they were two different species [140]. As reported in the literature, a near identical 16S rRNA gene sequences does not imply whether different isolates are sufficiently related to be assigned to the same species. Moreover, a single phylogenetic marker cannot reflect the product of the genomic difference of stochastic variations and horizontal gene transfer [368]. Still, the analysis of a bigger set of genes using both multi-locus and core-genome analysis showed a species distribution congruent with the ANI values, showing that highly reliable genetic markers have a strong correlation with whole-genome relatedness. Particularly, the maximum-likelihood (ML) analysis used to build a phylogenetic tree and calculate the ML-based distances between all pairs of 43 genomes of four bacterial groups (*E. coli*, *Salmonella* spp., *Shewanella* spp., and *Burkholderia* spp.) showed a correlation between the averages for the ML distances and the ANI distances, where almost all core genes, regardless of their functions and positions in the genome, offer robust phylogenetic reconstruction [402]. Therefore, the results retrieved from the Southern Ocean *Pseudonocardia* spp. support reports for the genus *Streptomyces* which propose that the 16S rRNA gene sequence identity does not resolve species-level variations [403].

Although the phylogenomic evidence suggests *P. abyssalis* and *P. oceani* represent two different lineages (ANI > 95%), these species, as well as with other members of the clade of which they are part, showed a close genomic relatedness (ANI = 87-94%). In this context, it is possible to hypothesise a high rate of homologous recombination between the *P. abyssalis* and *P. oceani* lineages which explain the genomic similarity observed between them. In fact, the recombination over mutation ratio (r/m) values confirmed a relative high gene-exchange rate between the *P. abyssalis* and *P. oceani* clade, as well as between *Pseudonocardia* sp. KRD291 and *P. sediminis*, which undoubtedly can be related to the observed ANI values. Overall, phylogeny topology was not strongly affected by homologous recombination, confirming that the core genome phylogenies are usually highly robust to recombination events. Only reconstructed core-genome trees of species evolving under strong selective pressures are impacted by recombination due to biased gene exchange [404]. In this sense, the obtained results contrast with what was previously observed in *Pseudonocardia* populations in symbiosis with fungus-growing ants for which the recombination events were relatively infrequent [239]. Hence, the ant-related strains showed population dynamics with a correlation between genomic and geographic distances, while the Southern Ocean strains

did not exhibit evidence of genetic barriers with species isolated from distant geographic sites. This situation is consistent with other free-living microorganisms from marine environments. For example, the metagenomics analysis of the SAR11 clade of Alphaproteobacteria, the most abundant group of planktonic bacteria in the ocean, showed widespread recombination among distantly related members [405]. Therefore, it was proposed that the population homogenization observed in pelagic bacteria due to gene exchanges could be also observed in bacteria from sediment. This conflicts with what was observed for the *Salinispora* genus, where low ratios of inter-species recombination were observed between the *S. arenicola*, *S. tropica* and *S. pacifica* lineages [406]. For instance, the pan-genomic analysis of 119 genomes for the species-level delimitation of the *Salinispora* genus showed that 63% of the core-genome had evidence of recombination events, but with no effect on the inter-species phylogenetic resolution [215]. However, it is important to point out that the inter-species recombination observed in the Southern Ocean strains also included species isolated from non-marine environments, such as plant rhizosphere in the case of *P. broussonetiae* [140] or terrestrial soil in the case of *P. petroleophila* [117]. Hence, in opposition to *Pseudonocardia* strains in symbiosis with fungus-growing ants, for which the ecological differentiation drives the speciation process of lineages [232], *P. abyssalis* and *P. oceani* are part of a taxonomic group with a cosmopolitan distribution.

The effect of homologous recombination over the evolution of *Pseudonocardia* is consistent with previous reports in *Streptomyces*, where evidence of widespread inter-species homologous recombination was demonstrated, but it was not uniform within the genus. In this case, it was hypothesised that the cumulative effect of recombination events was associated with historical biogeography and these dynamics may explain the ancestral formation of mosaic lineages [407, 408]. Overall, the genomic relatedness between Southern Ocean strains and other *Pseudonocardia* spp., as well as the evidence of inter-species homologous recombination, suggests a weak selective pressure over these lineages in particular. A similar hypothesis can be proposed for *Pseudonocardia* sp. KRD291, as it showed strong gene exchange with *P. sediminis*, which was isolated from sediments collected in a geographically distant site from the Southern Ocean [149]. Nevertheless, the performed genomic analysis also identified a close relationship between the *P. abyssalis* and *P. oceani* clade and *Pseudonocardia* sp. CNS-139 [146] and *Pseudonocardia* sp. MH-G8 [159], two strains from marine origin, which may suggests that some members of this *Pseudonocardia* lineage could have a particular selective adaptation to this environment. Although this

situation has been proposed for *Salinispora* and *Streptomyces* [409, 410], a more robust analysis is required to conclude about the evolutions of the Southern Ocean lineages in marine environments in particular. Future efforts should be focused on analysing the impact of intra-species homologous recombination over the Southern Ocean *Pseudonocardia* spp. strains to better understand the evolution of each lineage. Furthermore, as the evidence suggested a non-ecologically restricted distribution for *P. abyssalis* and *P. oceani*, it is predicted that in the future other strains of these species could be isolated from other geographic locations and ecological contexts, which eventually will allow the knowledge of their evolution and ecology to be expanded.

In addition to the evolutionary analysis, the phenotypic characterization of *P. abyssalis* KRD168 and *P. oceani* KRD185 confirmed their identity as two novel species through the polyphasic taxonomy approach. In terms of morphology, the budding substrate and aerial hyphae structure observed for *P. abyssalis* KRD168 and *P. oceani* KRD185 was similar to previously reported for other *Pseudonocardia* spp. [140, 150, 387]. The predominant menaquinone MK-8(H4), and the major fatty acid iso-$C_{16:0}$ were consistent with the genus *Pseudonocardia* [391]. Although it has been suggested that unsaturated or branched fatty acids play a relevant role in the adaptation of deep-sea bacteria to low temperature and high pressure [411], no differences in the ratio of *iso*-branched and unsaturated fatty acids were found between the Southern Ocean strains and *P. broussonetiae* or *P. petroleophila*. These results contrast with observations on *Streptomyces oceani*, for which the presence of 10% of the unsaturated fatty acid anteiso-$C_{17:1}$ w9c helped to distinguish it from their closest related strains [412]. However, previous studies in deep-sea *Alteromonas* and *Vibrio* showed that the ratio of total unsaturated fatty acids increased when the strains were grown at increasing hydrostatic pressures and decreasing temperatures [413–415]. So, further studies in variable pressure conditions are required to conclude on the adaptation of *P. abyssalis* and *P. oceani* to deep-sea conditions. Another important trait analysed for these strains was their halotolerance, which revealed that they are not obligate marine species. In this case, similar results were observed in other *Pseudonocardia* spp. from marine origins such as *P. ammonioxydans* [153], *P. sediminis* [149], *P. antitumoralis* [150], and *P. profundimaris* [151].

In summary, the whole-genome analysis showed that the Southern Ocean strains isolated by Millán-Aguiñaga [261] represent three lineages not previously described. Moreover, the close genomic relatedness of these strains with other *Pseudonocardia* lineages from diverse

environments and geographical origins represent an opportunity to study the evolution of the specialised metabolisms of free-living *Pseudonocardia* spp. Despite the phylogeographic evidence suggesting a sympatric speciation process, there are still reasons to argue that their specialised metabolism could play a relevant role in the adaptation of these lineages to the environment. To address these questions, the specialised metabolisms of the Southern Ocean strains and other *Pseudonocardia* species will be compared in chapter 4, while the role of these metabolomic profiles in the interaction between strains will be explored in chapter 5.

Chapter 4 Multi-omics analysis of the specialised metabolism of *Pseudonocardia* spp. isolated from marine environments.

4.1. Introduction

4.1.1. Genome mining

Genome mining refers to the exploitation of genomic information for the discovery of new processes, targets, and products. Since the first actinomycete genomes were sequenced, genome mining has been considered a potent tool for the discovery and study of specialised metabolites [416]. For instance, the early genome sequence of *Streptomyces coelicolor* revealed an abundance of previously uncharacterized candidate biosynthetic enzymes, showing an unexplored potential hidden in their genome [334]. The sequencing of the *S. coelicolor* A3(2) genome eventually led to the discovery of new specialised metabolites like the tetrapeptide iron chelator, coelichelin [417, 418]. In principle, genome mining for specialised metabolites comprises identifying genes encoding enzymes putatively involved in metabolite biosynthesis, often referred to as BGCs. For example, the analysis of the genome of the marine-derived *Streptomyces* sp. SCSIO 40010 revealed the presence of a putative polycyclic tetrameter macrolactams BGC, which subsequently led to the discovery of six novel macrolactams, five of them with cytotoxic activity against several cancer cell lines [419].

To predict BGCs, the first step in genome mining approaches is commonly to identify conserved biosynthetic genes and then analyse their products to obtain information about the biosynthetic function of the encoded protein (e.g. enzymatic function, regulatory function or transporter function) [420]. This identification process used to be carried out with sequence similarity and homology methods, such as BLAST, and in-house lists of query genes. Besides identifiying genes encoding biosynthetic enzymes, other approaches that include comparative genome mining using sequence similarity networks, phylogeny-based mining methods, and mining of resistance and regulatory genes facilitate the BGCs annotation [185]. Several bioinformatics tools, such as antiSMASH [304] and PRISM [421], have been developed for the automatic detection of BGCs using increasingly sophisticated algorithms. Some examples of specialised metabolites discovered through genome mining include salinilactam A from *Salinispora tropica* [422]; isogermicidin B and germicidin C from *Streptomyces*

coelicolor [423]; and venezuelin from *Streptomyces venezuelae* [424]. These methods identify BGCs using a rule-based system that looks for specific combinations of enzymatic domains, which are detected using profile Hidden Markov Models (pHMMs). Although a possible advantage of this strategy is the low rate of false positives, this model will not detect unknown types of BGCs since it is based on rules used to detect previously characterised biosynthetic genes [169, 420]. For detecting putative new classes of BGCs, algorithms such as ClusterFinder [425] have been proposed. These methods identify genomic regions that are rich in Pfam (Protein families database) domains that occur frequently in BGCs. Since this method is entirely based on Pfam domain frequencies, and nature uses distinct assemblages of the same enzyme superfamilies to construct unrelated natural product classes, ClusterFinder is capable of identifying new BGC classes with relatively little training set bias. The putative function of these annotated genes is sometimes misinterpreted as potential new BGCs. Then, without the proper validation, this could lead to wrong conclusions [52]. More recently, DeepBGC implemented a deep learning approach using Recurrent Neural Networks (RNNs) and vector representations of Pfam for reducing the higher false-positive rate of ClusterFinder [426]. An alternative to identify putative BGC is based on the idea that all secondary metabolic enzymes are distant paralogs of primary metabolic enzymes, with this approach, EvoMining makes a phylogenomic identification of enzymes with expanded substrate specify, and enabled the discovery of novel arseno-organic metabolites from *Streptomyces* sp. [246].

After BGCs are annotated, the evolutionary relationship between them can be mapped by comparing architectural relationships between BGCs in sequence similarity networks and group them into Gene Cluster Families (GCFs), which are groups of BGCs with a similar sequence that encode the production of identical or highly similar molecules. This is the approach that BiG-SCAPE (Biosynthetic Gene Similarity Clustering and Prospecting Engine) applies for the calculation and exploration of BGC similarity networks [306]. Then, CORASON (CORe Analysis of Syntenic Orthologs to prioritize Natural Product-Biosynthetic Gene Cluster) algorithm employs a phylogenomic approach to elucidate evolutionary relationships between BGCs by computing high-resolution, multi-locus phylogenies of BGCs within and across GCFs. CORASON analysis is integrated into the BiG-SCAPE platform [306]. Using BiG-SCAPE/CORASON, Undabarrena *et al*. explored the evolutionary history of BGC of the *Rhodococcus* genus, resulting in the characterisation of corynecin-related BGCs [427]. Similar comparative genomics analysis has also been applied to analyse the specialised metabolisms

of *Nocardia* [178] and *Planomonospora* [428] genera. Recently, BiG-SCAPE was expanded to allow the analysis of larger BGCs datasets with BiG-SLiCE (Biosynthetic Genes Super-Linear Clustering Engine). In short, BiG-SCAPE measures the occurrence and synteny of Pfam domains for each pair of BGCs, along with the sequence similarity of homologous core genes to construct a pairwise-distance network and define GCFs, while BiG-SLiCE vectorises the BGCs into Euclidean space, allowing the use of non-pairwise clustering algorithms to group the GCFs and reducing computing power consumption [188].

Genome sequencing and mining studies in microorganisms have revealed a huge biosynthetic potential encoded in their genomes. For example, a BiG-SLiCE analysis of 170,585 complete and draft bacterial genomes and a dataset of 11,143 metagenome-assembled genomes (MAGs) revealed that the number of chemical classes of microbial natural products documented in NPAtlas corresponds to 2.5% - 3.3% of the predicted potential of the bacterial kingdom according to the number of GCFs. only 3% of genomic potential for specialised metabolites potentially encoded in the annotated BGC has been experimentally characterised [429]. In this context, metabolomics techniques are crucial to better understand the function and expression of BGCs.

4.1.2. Metabolomics

As introduced in chapter 1, mass spectrometry (MS) is an analytical technique to measure metabolites, detected as ions. Moreover, as structurally related metabolites produce similar fragmentation patterns, the alignment of their MS/MS spectra can be used to calculate vector similarities (“cosine scores”) that are used for the generation of molecular networks [257]. For this, the Global Natural Products Social Molecular Networking (GNPS) offers an open-access platform that allows the analysis of MS/MS spectrometry data based on the molecular networking approach, and enables dereplication using the GNPS spectral libraries. Furthermore, GNPS also offers public dataset deposition through the Mass Spectrometry Interactive Virtual Environment (MassIVE) data repository. [323]. In practical terms, the molecular network approach is a versatile dereplication tool to annotate known metabolites from complex mixtures, but also to identify related analogues, which is an advantage over other dereplication strategies [430]. It is therefore possible to cluster metabolites into Molecular Families (MFs), which are defined as related molecules based on their fragmentation pattern that translates to structural similarity [431]. Applications of molecular

networking in the exploration of microbial chemistry include the metabolomics profiling of *Streptomyces* spp. [432] and *Salinispora* spp. [433] strains collections from marine origin.

In addition to the molecular networking and spectral library search ('classical' molecular networking), the GNPS platform also includes other advanced analysis tools as part of its ecosystem. In particular, Feature-Based Molecular Networking (FBMN) expands the approach by including MS1 information like isotope patterns, retention time and relative intensity which allow isomers with similar MS2 spectra to be distinguished, facilitates spectral annotation, and incorporates relative quantitative information [324]. An example of the use of FBMN is illustrated by the work of Zdouc *et al*. which applied it to explore the chemical diversity of the genus *Planomonospora*, allowing the identification of the novel thiopeptide siomycin E [428]. Recently, FBMN incorporated Ion Identity Molecular Networking (IIMN) that integrates chromatographic peak shape correlation analysis using MS1 data into molecular networks to connect different ion species from MS2 data that belong to the same metabolite, reducing complexity and redundancy in the final network. In short, this allows users to connect and collapse different ion species of the same metabolite [326]. This tool has been successfully applied to identify metal-binding compounds using native electrospray mass spectrometry [434].

Other tools integrated into the GNPS ecosystem include Mass Spectrometry Search Tool (MASST) that facilitates the search of query MS/MS data in public metabolomics repositories [435], and Reanalysis of Data User (ReDU) interface, a metadata-filtered repository-scale analysis system that allows the re-analysis of public data in GNPS/MassIVE and the co-analysis of these public data and query MS/MS data [436]. These tools allow mining of stored spectral data to facilitate the analysis of mass spectrometry data. For example, MASST was applied by Wibowo *et al*. to identify Valinomycin analogues from the sea cucumber-associated *Streptomyces* sp. SV 21 [437].

4.1.3. Linking paired 'omics dataset

Genome mining and untargeted metabolomics have become mainstream methodologies for the discovery and analysis of specialised metabolites. As a consequence, computational networking approaches such as GNPS and BiG-SCAPE are widely used tools for the analysis of large-scale metabolomics and BGC data. In this context, many recent efforts have focused on linking these datasets [438]. To cite an example, the recently described NPOmix approach

applies a supervised fingerprint-based strategy to link mass spectrometry fragmentation data to their putative BGCs [439]. In this context, the term “metabologenomics” has been proposed to describe the methodology used to integrated genome sequencing and BGC prediction with MS-based metabolomics [440].

One of the first proposed approaches for the metabologenomic integration was pattern-based genome mining which combines the presence/absence of BGCs profile across strains with molecular networking to identify overlapping distribution patterns [438]. For example, a pattern-based approach applied by Duncan *et al*. to explore the specialised metabolism of 35 *Salinispora* spp. strains allowed the identification of the novel depsipeptide Retimycin A, as well as linking it to the corresponding BGC [441]. These pattern-based linking methods have been applied to the co-occurrence of GCFs and MFs by including correlation scores [438, 442]. For example, Doroghazi *et al*. proposed a scoring method for the pattern-based genome mining analysis of 830 actinomycetes genome. In this method, BGCs from different strains are clustered into GCFs for a subsequent binary correlation with metabolites identified by MS (GCF present, metabolite present = +10; GCF absent, metabolite present = −10; GCF present, metabolite absent = 0; GCF absent, metabolite absent = +1) [190]. In later work, this approach was used to identify the novel non-ribosomal peptides tyrobetaines from *Streptomyces* sp. [443].

Feature-based linking methods predict the structural properties of metabolites from genomics data, which is used to infer possible links with spectral data from metabolomics data sets [438]. NPLinker [444] is a feature-based method based on presence/absence correlations of metabolites and BGCs across strains. Briefly, the NPLinker framework applies the Input-Output Kernel Regression (IOKR) [445] to facilitate analysis of paired genomics and metabolomics datasets by using molecular fingerprints predicted from MS/MS spectra from the application of a function learned from annotated training data [444]. Due to the fact that the scoring method from Doroghazi *et al*. [190] mentioned above is influenced by the size of the GCF and the number of strains that produce the metabolite, NPLinker proposed a standardisation of the strain correlation score that allows comparison of score values across links [444]. This new standardised strain correlation scoring method will be referred to in this thesis as the Metcalf score. This scoring method does not have an upper or lower bound, but a score of zero means that the degree of overlap between any MF and GCF is the same as would be expected if they were chosen at random, while a mean score of 3.672 was observed

in validated links [444]. NPLinker was applied by Soldatou *et al*. to propose putative links between known metabolites and to their producing BGCs in a collection of rare actinomycetes isolated from polar marine environments, showing it to be a useful tool for facilitating the study of the specialised metabolism of bacteria in a large strain collection. Moreover, in this work a new scoring method named Rosetta, based upon a set of collated matches between the GNPS spectral library and the MiBIG database was introduced. As a result, Soldatou *et al*. linked the ectoine metabolite produced by *Rhodococcus* spp. and *Halomonas* spp. with its corresponding BGC [262].

In this chapter, genome mining and metabolomics techniques are applied to explore the specialised metabolism of three species of *Pseudonocardia* spp. isolated from sediments collected in the deep Southern Ocean, their phylogenetic relatives, and other strains collected from distant marine environments. This work aims to explore the chemical potential of the Southern Ocean strains, as well as explore the effect of biogeography and phylogeny in the evolution of their specialised metabolisms. Finally, a feature-based method was applied in an attempt to integrate genomics and metabolomics information.

4.2. Results

The phylogenomic analysis performed in chapter 3 revealed that *P. abyssalis* and *P. oceani* are two different species among a monophyletic group, alongside *P. petrolephila* and *P.* hydrocarbonoxydans, another two species isolated from diverse geographic and ecological origin. Similarly, *Pseudonocardia* sp. KRD291 showed a higher affiliation with *P. sediminis*, also from marine environments. Therefore, the comparative analysis of the biosynthetic and metabolomics profiles of the Southern Ocean strains *P. abyssalis* KRD168, *P. oceani* KRD185, and *Pseudonocardia* sp. KRD291 represent an interesting model to study the evolution of the specialised metabolisms from this genus.

4.2.1. Genomics analysis of the specialised metabolism of the Southern Ocean *Pseudonocardia* spp. strains.

The genome assembly across all strains was improved by linking together non-contiguous contigs into scaffolds to increase the reliability of genome mining. To this end, the graph-based approach MeDuSa was applied using the genome sequences of *P. abyssalis* KRD168

and *P. oceani* KRD185 as reference genomes for scaffolding the assemblies of the strains with only short-read genome sequences. As a result, a hypothesised ordering of the draft genomes was obtained, bridging the gap betwenn contigs (Fig. S5). For instance, the KRD188 genome was reduced from 749 contigs to 66 scaffolds. Similarly, the genome sequence of *Pseudonocardia* sp. KRD291 was scaffolded using the genome of *P. sediminis* DMS 45779 as a reference genome. In this case, the assembly was reduced from 531 contigs to 52 scaffolds, with most of the sequence contained on one scaffold (Fig. S5).

BGC annotation with antiSMASH predicted NRPS (Non-Ribosomal Peptide Synthetase), terpene, PKS (Polyketide Synthase), RiPPs (Ribosomally synthesised and Post-translationally modified Peptides), and ectoine BGCs. Additionally, a betalactone containing protease inhibitor BGC was annotated in the *P. oceani* and *P. abyssalis* genomes. Furthermore, a type I PKS / NRPS-like hybrid BGC was present in the *P. abyssalis* genomes, and an oligosaccharide BGC was present in the *Pseudonocardia* sp. KRD291 genomes (Fig. 4-1). Interestingly, a Type III PKS was annotated for KRD291 and *P. oceani*, while a heterocyst glycolipid synthase-like PKS (hglE-KS) was identified in *P. abyssalis.* In general, between nine and twelve BGCs were annotated in each sequenced genome. In particular, nine BGCs were detected in the two *P. abyssalis* strains and ten in KRD291. In the case of *P. oceani* strains, the number of BGCs was more variable, as ten BGCs were annotated in *P. oceani* KRD185 (the genome with the higher contiguity and completeness), while the rest of strains were reported to contain from nine to twelve BGCs. Furthermore, no RiPPs were identified in KRD182 and KRD184. Due to the high genomic similarity calculated between all *P. oceani* strains, the differences in the number of BGCs are more likely related to the genome fragmentation and completeness in the short-read based genomes (Fig. S6) than inter-strains variability. In general, the number of BGCs annotated in *P. abyssalis* and *P. oceani* is similar to their closest phylogenetic relatives: *P. petroleophila* CGMCC 4.1532 and *P. broussonetiae* Gen 01.

As the antiSMASH results revealed similar BGC classes distributions in *P. abyssalis* and *P. oceani*, an evolutionary analysis was carried out to compare these BGCs. To this end, BiG-SCAPE was used to cluster the BGCs into GCF based on sequence similarity. Besides the Southern Ocean *Pseudonocardia* spp. strains, BGCs from publicly available *Pseudonocardia* spp. genomes and the MIBiG database were used. The network (Fig. 4-2) displayed six GCFs that include BGCs from both *P. abyssalis* and *P. oceani*, which confirmed the similarities between the BGCs annotated across their genomes. Similarly, BGCs annotated in

Pseudonocardia sp. KRD291 and *P. sediminis* genomes shared seven GCFs. Most BGCs from the Southern Ocean strains clustered alongside BGCs from other *Pseudonocardia* spp. genomes, while only one GCF in *P. abyssalis* (*pks3*) and one in *Pseudonocardia* sp. KRD291 (*nrps5*) were not related to any other strain. Furthermore, only one BGC annotated in the Southern Ocean strains (*ect1*) was related to BGCs in the MIBiG database, suggesting biosynthetic novelty (Fig. S7).

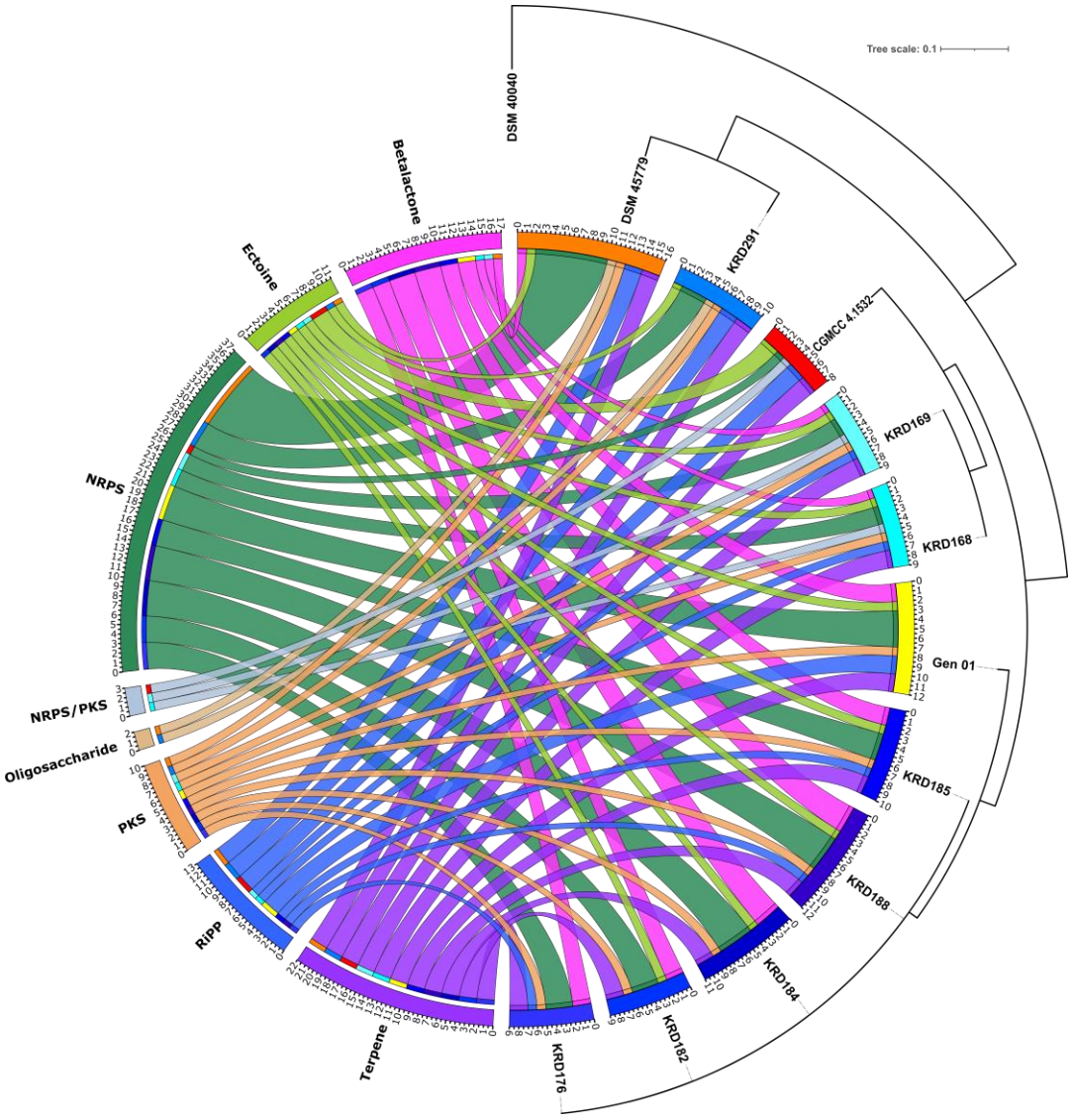


Figure 4-1 Circos diagram of the BGCs diversity of ●*P. abyssalis*, ●*P. oceani*, ●*Pseudonocardia* sp. KRD291, ●*P. petroleophila*, ●*P. broussonetiae*, and ●*P. sediminis*. The left side of the circle shows the BGC type and the right side shows the Southern Ocean *Pseudonocardia* strains. A multi-locus tree based on the autoMLST analysis is shown. The ribbon width represents the number of BGCs annotated using antiSMASH and the colour represents the antiSMASH class: betalactone (▪), ectoine (▪), NRPS-PKS hybrid (▪), NRPS (▪), oligosaccharide (▪), PKS (▪), RiPP (▪), and terpene (▪).

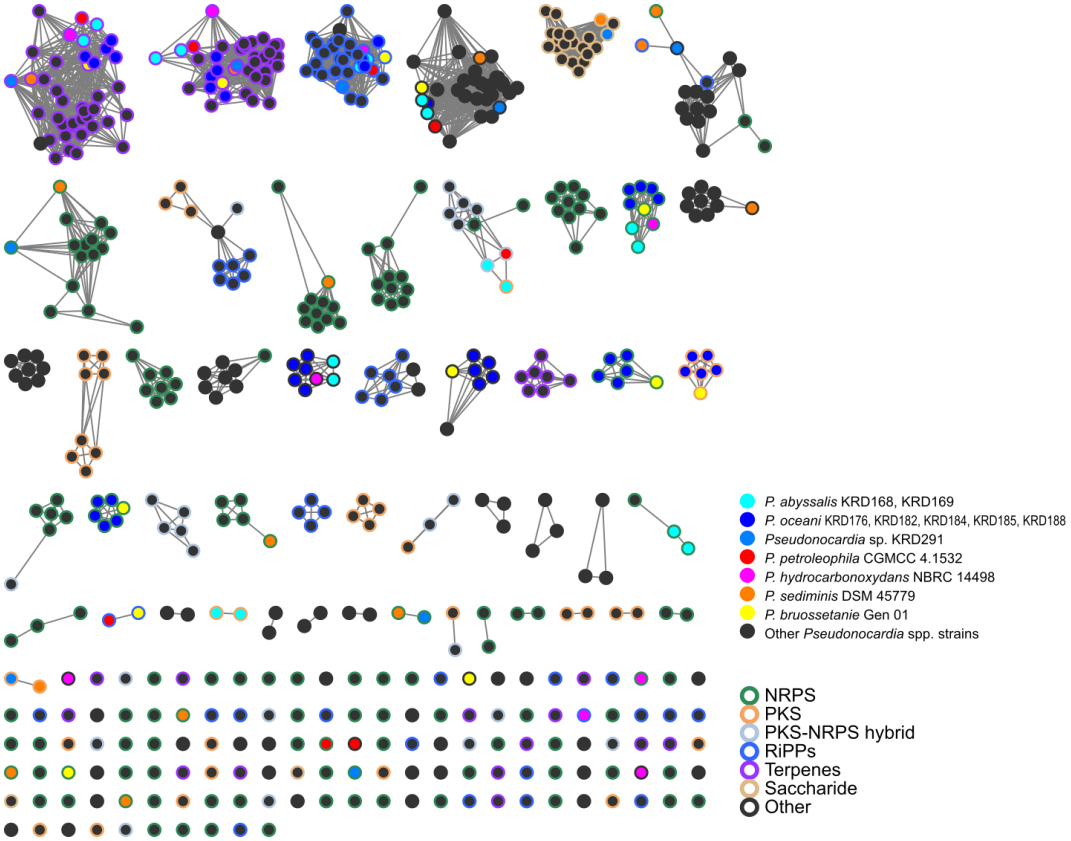


Figure 4-2 Networks of the BGCs grouped into GCFs. Each node represents a BGC and the edges represent a correlation distance of 0.5 or higher. Networks are visualised showing species and assigned natural product classes. The network layout is based on the Jaccard index.

To better understand the evolutionary dynamics behind the BGC clustering into GCF, the GCF absence/presence profile of the analysed strains was collated with their phylogeny reconstructed from a multi-locus analysis (Fig. 4-3). The phylogenetic distribution of BGCs across the genus revealed a close correlation between the specialised metabolisms of *P. abyssalis* KRD168 and *P. oceani* KRD185, which have six GCFs in common. Additionally, both species shared four and five GCFs with *P. petroleophila* CGMCC 4.1532 and *P. hydrocarbonoxydans* NBRC 14498, respectively. Likewise, *Pseudonocardia* sp. KRD291 shared seven GCF with *P. sediminis* DSM 45779. In general, evolutionary distance seems to have a greater influence than geographic or ecological origin of strains on BGC distribution. For example, *P. oceani* KRD185 and *P. broussonetiae* Gen 01, an endophytic strain isolated from plant roots, shared eight of ten GCFs, despite their different ecological origins. Similarly, *Pseudonocardia* sp. KRD291 only shared one GCF with *P. abyssalis* and *P. oceani* (*ripp1*), despite their similar geographic origin.

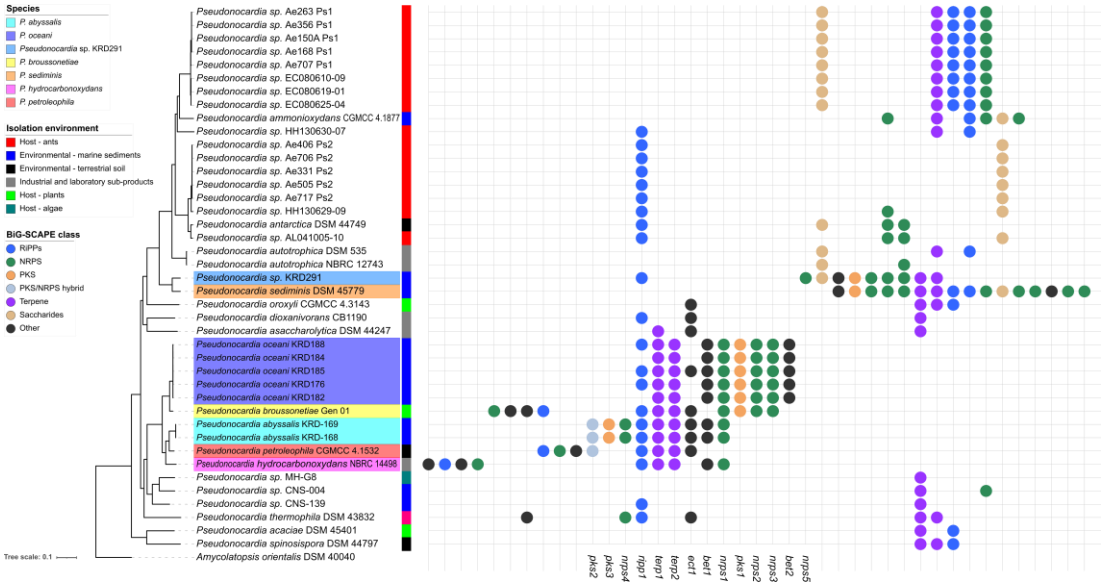


Figure 4-3 BGCs distribution in Pseudonocardia strains grouped by phylogeny using autoMLST. Nodes represent BGC according to natural product class using antiSMASH. Node columns represent BGCs clustered in the same GCF according to the BiG-SCAPE analysis. The "other" BiG-SCAPE class includes betalactone and ectoine. The Southern Ocean *Pseudonocardia* strains and their closest relatives are highlighted according to the clade to which they belong. The isolation environment of each strain is shown as a colour strip. Bar, 0.1 substitutions per nucleotide position.

The evolutionary analysis of the GCF distribution also showed that the RiPP and terpene BGCs are the most conserved across the genus (GCFs have BGCs annotated in several species), with members of the same GCF found across environments. For example, *terp1* and *terp2* were found in all *P. abyssalis* and *P. oceani* strains, as well as in their closest relatives: *P. broussonetiae*, *P. hydrocarbonoxydans*, and *P. petroleophila*. Furthermore, the only GCF present in the three Southern Ocean strains, *ripp1*, was found in strains from terrestrial soil, plants, fungus-growing ants and other marine environments. On the other hand, the species-specific BGCs annotated in the Southern Ocean strains correspond to NRPS, PKS, and betalactone type BGCs, such as *pks3* in *P. abyssalis*, *bet2* in *P. oceani*, and *nrps5* in *Pseudonocardia* sp. KRD291. These findings suggest that these BGCs were potentially acquired later in the evolutionary history, during adaptation to their Southern Ocean environment.

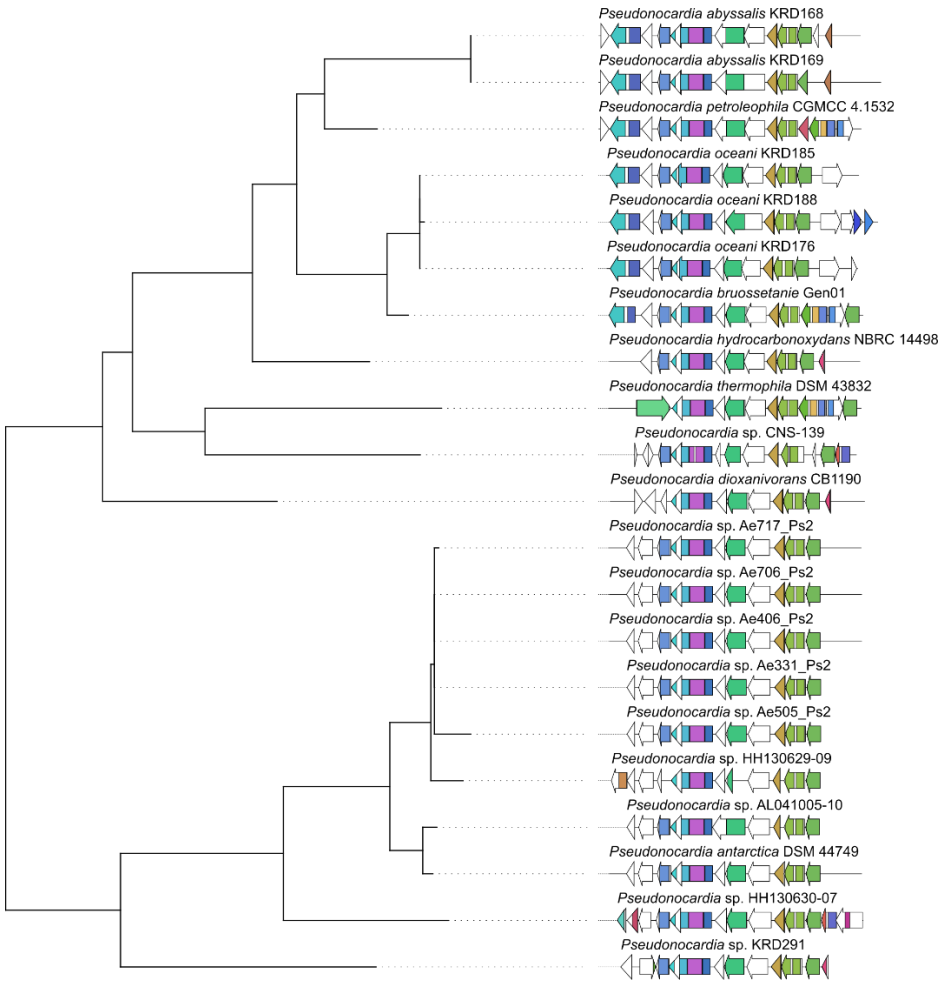


Figure 4-4 Maximum-likelihood multi-locus phylogenetic tree of the GCF *ripp1* reconstructed with CORASON showing the presence of this BGC across 21 *Pseudonocardia* spp. genomes. Genes are represented by arrows facing the direction of transcription. The tree was created using the sequences of the core domains of the GCF.

In addition to the GCFs distribution, the evolution of the *Pseudonocardia* lineages also showed a close correlation to the evolution of individual BGC. By using the well-conserved GCF *ripp1* as a reference (Fig. 4-4), it was observed that the phylogeny corresponded to the multi-locus genome phylogeny for the strains that contained the BGC. For instance, the *ripp1* sequence shows *P. abyssalis* closely related to *P. petroleophila* and *P. oceani* than the more phylogenetically distant strains like *P. antarctica* or *Pseudonocardia* sp. KRD291. Furthermore, the gene structure of conserved GCFs showed that some regions of the BGCs are more conserved than others. For example, *bet1* annotated in *P. abyssalis*, *P. oceani*, and *P. hydrocarbonoxydans*, showed a core group of fifteen genes conserved across the three species, which share a sequence identity above 90% (Fig. 4-5). This core group of genes includes a duplicate copy of a GntR family transcriptional regulator that will be mentioned

later in chapter 5. The accessory genes identified in this GCF were potentially acquired after the speciation process. However, as the metabolite products of these BGCs is still unknown, it is not possible to conclude whether these accessory genes actively participate in their biosynthesis.

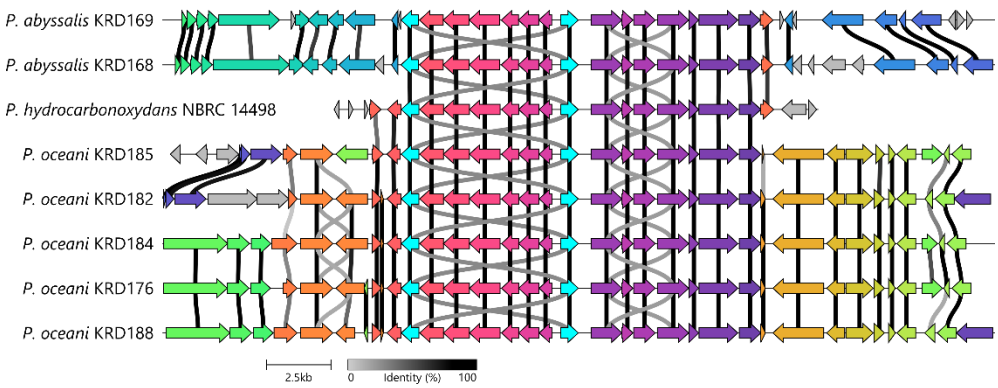


Figure 4-5 Gene alignment of *bet1* annotated in *P. abyssalis*, *P. oceani* and *P. hydrocarbonoxydans* genomes which belong to the same GCF. Genes are represented by arrows facing the direction of transcription, where length is proportional to the gene sequence size. GntR family transcriptional regulator genes are showed in cyan (●).The colour of the lines that are linking the genes represents the level of sequence similarity between them (white = 0% identity, black = 100% identity).

By contrasting the sequence similarity network analysis of BGCs with the multi-locus phylogeny of the analysed strains, a primary role of vertical inheritance in the BGC profiles of *P. abyssalis* and *P. oceani* lineages is inferred. However, in chapter 3 it was discussed that recombination events seems to be common between theses lineages and closely related species. Therefore, the effect of homologous recombination on the evolutionary dynamics of the BGCs was investigated using the ClonalFrameML approach. The sequences of BGCs clustered into the same GCF were aligned to analyse the ratio at which nucleotides are replaced by either recombination or point mutations (r/m) (Table 4-1). By comparing the relative rate of recombination between the core-genome and GCFs of the related species *P. abyssalis*, *P. oceani*, *P. petrolephila*, *P.hydrocabonoxydans*, and *P. broussonetiae*, how the evolutionary process affects BGC diversification can be evaluated. At the genome level, the contribution of recombination relative to mutation in the core genome (chapter 3) was relatively high (r/m > 1.5) among the related species (v=0.04 or 4% genetic divergence among strains), demonstrating that homologous recombination is a cohesive force. However, with the exception of *pks2*, all the analysed GCFs showed even higher values of calculated

homologous recombination (r/m > 1.5) among similar values of genetic divergence (3-6% genetic divergence among strains).

In general, these results show that recombination reduced genetic diversity in BGCs among closely related species. This partially explains the low variability among some BGCs. For example, well-conserved BGCs across the clade such as *ect1*, *ripp1*, *terp1*, and *terp2* showed higher levels of the relative effect of recombination and mutation (r/m) than the core genome. As recombination acts as a cohesive force maintaining sequence similarity, the divergence of this BGC could be limited by a genetic exchange. On the other hand, BGCs like *bet1* and *nrps1* showed lower values of the relative effect of recombination and mutation. This means that the calclated r/m rates for those BGCs do not differ considerably from those in the core genome. In particular, *bet1* showed the lowest value of the relative effect of recombination and mutation, as well as the lowest genetic divergence of imported DNA, as a consequence of a few recombination events that are located in a particular region with a low global effect in the r/m rate (Fig. S8). However, with the exception of *pks2*, it was shown that BGCs experience stronger cohesive forces than the core-genome, suggesting that the diversification of the BGC is less than that of the lineages themselves. This observation agrees with the theory that the specialised metabolites play an essential role for microorganisms' survival. Lastly, the low contribution of recombination relative to mutation observed in *pks2* may be related either to the low genetic divergence among the strains (rate of nucleotide differences in the recombined stretches = 0.25%), the low number on strains analysed, or both. As the major limitation of this analysis is the low number of strains, the inclusion of more *P. abyssalis* and *P. oceani* strains, or related species, from other locations is required to confirm the findings of this work.

Table 4-1 Recombination events detected in shared BGCs annotated in *P. abyssalis* and *P. oceani* strains, as well as *P. petrolephila*, *P.hydrocabonoxydans*, and *P. broussonetiae,* relative to the core-genome.

BGC	No. species	No. strains	Alignment length (bp)	R/θ	δ	ν	r/m
pks2	2	3	69534	0.0115	240.2327	0.0025	0.007
pks1	2	6	58616	0.2696	4942.6162	0.0533	71.0302
nrps2	2	6	83465	0.6889	2794.0063	0.0609	117.1651
nrps3	2	6	51882	0.3055	9181.3032	0.0438	122.7981
ect1	4	5	12448	3.1984	800.6533	0.0470	120.4335
bet1	3	8	64697	0.4600	601.5942	0.0346	9.5893
nrps1	4	9	69712	0.8617	832.2029	0.0435	31.1709
ripp1	5	8	20648	1.6272	1016.6641	0.0454	75.0588
terp1	5	10	30276	3.5482	938.2271	0.0451	150.0822
terp2	5	10	39962	2.3335	1456.3906	0.0418	142.1768
Core genome	5	10	2841693	0.1150	367.3648	0.0439	1.8547

R/θ = relative rate of recombination to mutation; ν = mean import length of recombining DNA; δ = mean divergence of imported recombining DNA; r/m = relative effect of recombination and mutation (r/m = (R/θ) × δ × ν).

4.2.2. Metabolomics analysis of the specialised metabolism of the Southern Ocean *Pseudonocardia* spp. strains.

The genomic analysis revealed that the specialised metabolisms of *P. abyssalis* and *P. oceani* lineages are shaped primarily by phylogeny. Then, it was hypothesised that this phylogenetic relationship might be apparent in metabolomics data. To assess this, a Classical Molecular Networking (CMN) framework was used to investigate the spectral data from metabolomics culture extracts of 16 *Pseudonocardia* strains, including the eight genome-sequenced Southern Ocean *Pseudonocardia* spp. strains. Extracts were prepared in triplicate from biological replicates (one extract from each culture), using the same culture conditions. Furthermore, the MolNetEnhancer workflow was used to characterize the molecular network according to the ClassyFire chemical ontology, revealing that the most populated node clusters correspond to macrolides and analogues (100 nodes), organooxygen compounds (sub-class = ethers) (92 nodes), carboxylic acids and derivatives (75 nodes), glycerolipids (70 nodes), organooxygen compounds (sub-class = alcohols and polyols) (64 nodes), and prenol lipids (45 nodes) (Fig. 4-6). In general, the CMN analysis revealed chemical diversity among the culture collection, where from a total of 4990 parent ions, 1780 were part of a molecular family (cluster). This means that 64% of the parent ions were singletons that did not belong to any molecular family, suggesting chemical novelty within the dataset. Moreover, most singletons did not match any chemical classification. On the other hand,

several populated molecular families were classified as lipids and organic oxygen compounds (e.g. ethers, alcohols and polyols), which is consistent with conserved primary metabolites.

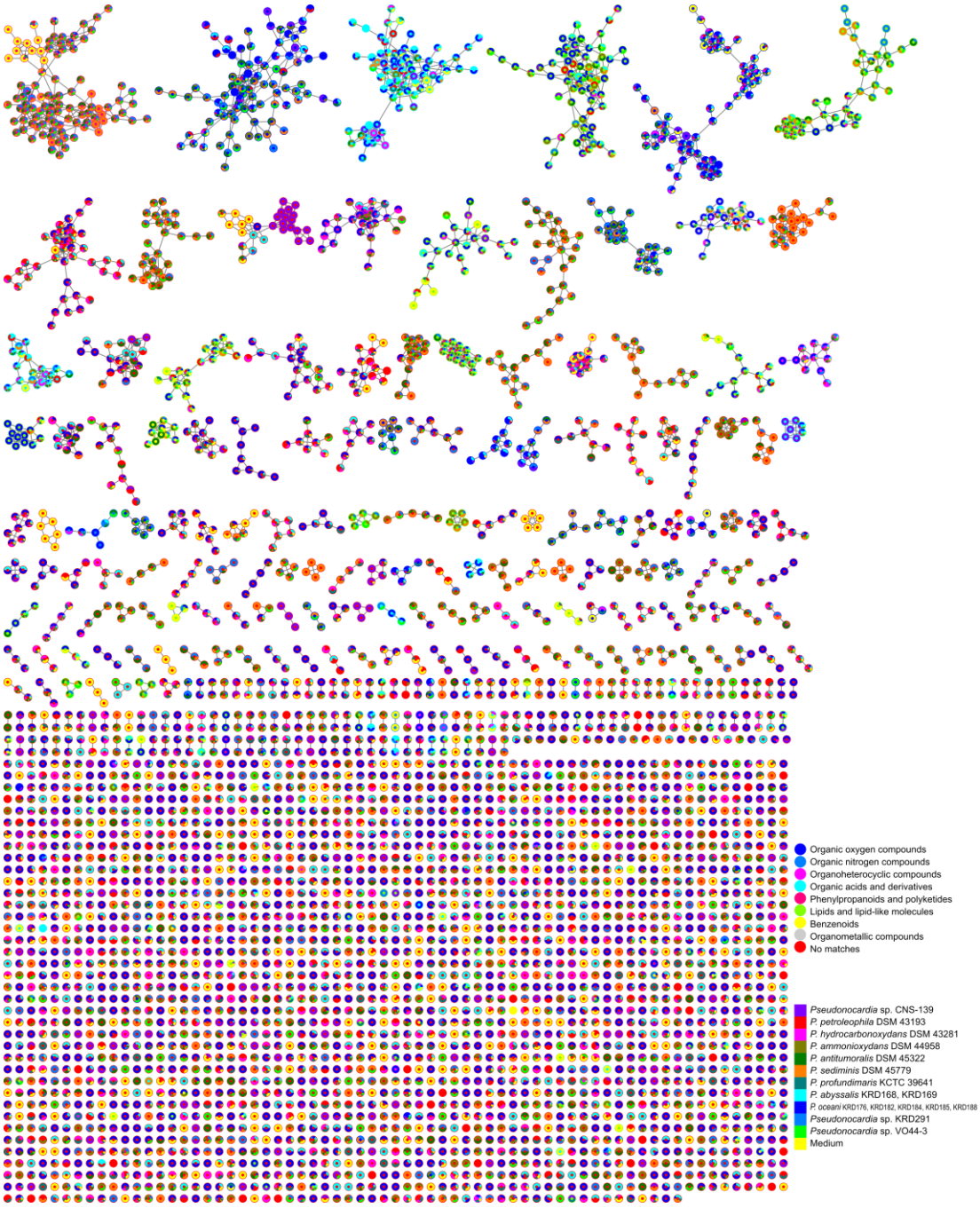


Figure 4-6 Molecular network of 4,990 nodes after applying the MS-Cluster algorithm to 121,088 spectra extracted of the LC-MS data of the eight Southern Ocean *Pseudonocardia* strains and five related *Pseudonocardia* spp. Nodes are coloured according to the ClassyFire chemical taxonomy. Doughnut chart inside each node shows the relative distribution of samples where the particular spectrum was detected. Nodes that included spectra from the solvent blanks were removed from the network before clustering.

The parent ion distribution showed a close relationship with the strain phylogeny. That is, phylogenetically close strains shared a higher number of parent ions than less related strains. For example, *P. abyssalis* KRD168 and *P. oceani* KRD185 shared 637 parent ions. Moreover, *P. abyssalis* KRD168 and *P. oceani* KRD185 shared 689 and 693 parent ions, respectively, with their close relative *P. petroleophila*. On the other hand, *P. abyssalis* KRD168 and *P. oceani* KRD185 share 386 and 362 nodes, respectively, with *P. ammonioxydans*, a more distantly related species (Fig. 4-7). Similarly, *Pseudonocardia* sp. KRD291 shared 904 parent ions with *P. sediminis*, while only 365 and 332 parent ions were shared respectively with *P. abyssalis* KRD168 and *P. oceani* KRD185 (Fig. 4-7). In summary, the metabolomics profiles agreed with the findings retrieved from the genomic analysis. Although these results showcase the value of untargeted metabolomics analysis for the study of the specialised metabolisms, this analysis does not allow the identification of metabolites that are produced by BGCs. As such, the annotation of molecular features and their integration with genomics data remains a great challenge.

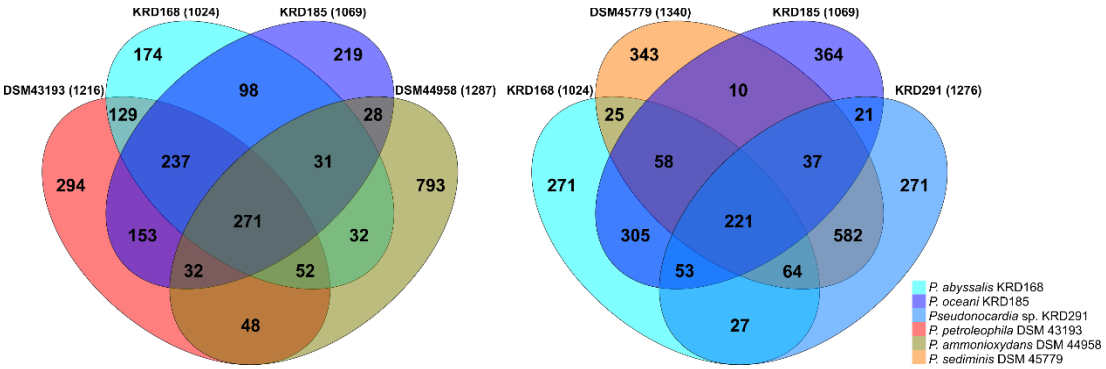


Figure 4-7 Venn diagram showing the node distribution from the GNPS analysis of parent ions detected in selected strains. Parent ions with similar MS/MS spectra (minimum cosine score of 0.7) are clustered in the same node and can be interpreted as metabolites. Shared parent ions potentially mean metabolites in common. The total number of nodes for each strain is shown in brackets.

From the total parent ions, 827 matched with a reference compound in the GNPS spectral library (including analogous ± 100.0 Da in the precursor ion). Among the spectral library hits, a molecular family (Fig. 4-8A) related to the aminoglycoside antibiotic neomycin (Fig. 4-8B) was identified (MS/MS spectral library match, parent ion ±100 Da), for which several spectra from *P. abyssalis* and *P. oceani* strains were part of their nodes. Interestingly, no saccharides class BGCs were identified in either *P. abyssalis* or *P. oceani*. In a deeper spectral search using

DEREPLICATOR+, an *in silico* database search tool that allows the annotation of metabolites in MS/MS data using *in silico* fragmentation graph, a node (Fig. 4-8C) with a precursor ion of *m/z* 639.406 which includes spectra from the three Southern Ocean *Pseudonocardia* spp. and other *Pseudonocardia* strains identified as Filipin (Fig. 4-8D) with a score of 25 (minimum score for significant matches = 12). However, no BGC related to polyene macrolide was identified in the analysed strains. So, more data on the chemistry of these strains is required to conclude that *P. abyssalis* and *P. oceani* are producers of aminoglycosides or polyene macrolides antimicrobials.

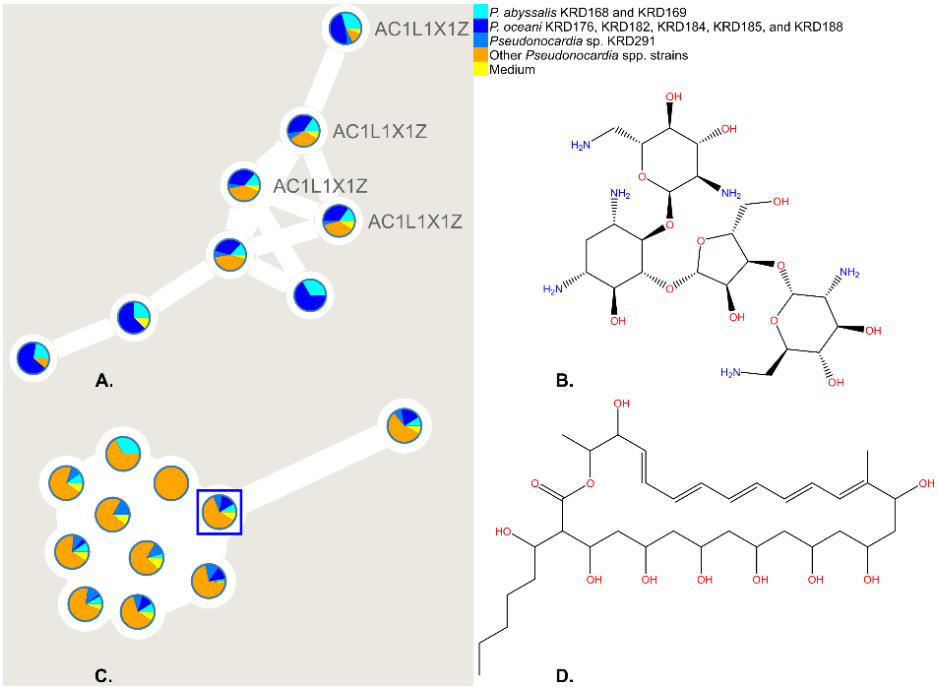


Figure 4-8 Molecular families for which some member was relate to antimicrobial metabolites. Four members of cluster A were annotated as neomycin (AC1L1X1Z) (B) in the spectral library of GNPS. A member of cluster B was annotated as filipin (D). The node annotated with DEREPLICATOR+ is highlighted in blue (C). Pie charts shows to the relative distribution of strains where the particular spectrum was detected.

Although the current mass spectrometry dataset did not allow the identification of metabolites related to an annotated BGC, a pattern-based correlation method can be applied to link the metabolomics and genomics data. To this end, the GNPS and BiG-SCAPE outputs of a subset of strains which genomic data was available (Table S7) were integrated into the NPLinker workflow to analyse the patterns of strain presence and absence between groups of MF and GCF. All links were visually inspected, and those with the highest scores were

selected, as well as those related to the MS/MS spectra in the GNPS library. Among the proposed links by NPLinker (Fig. 4-9A), the GCF *nrps1* which includes BGCs of *P. abyssalis*, *P. oceani*, and *P. hydrocarbonoxydans* (Fig. 4-9B) was putatively related (Metcalf score = 3.4641) to a singleton node with a precursor ion of *m/z* 605.222 produced by the eight strains. This link was one of the higher scored. However, this spectrum, did not show matches in the GNPS spectral library or in the MASST search. The GCF *nrps1* was also linked with the same scoring (Metcalf score = 3.4641) to a spectrum with a precursor ion of *m/z* 415.036, which was annotated as grayanotoxin I (Fig. S9). Sadly, this parent ion was also identified in one solvent blank sample. Additionally, the grayanotoxin I (diterpene) structure is not consistent with an NRPS product. So, this linked spectrum is unlikely to be related to *nrps1*. Furthermore, the third highest scoring for links proposed for BGCs annotated in *P. abyssalis* and *P. oceani*, was associated with a singleton node with a precursor ion of *m/z* 403.202 produced by the seven *P. abyssalis* and *P. oceani* strains that was also putatively linked with *nrps1* (Metcalf score = 2.9580). In this case, the spectrum was annotated in the GNPS library as the dipeptide asperglaucide (Fig. 4-9C). Although there is not enough evidence to accurately identify the parent ion, their spectral relationship with asperglaucide makes its biosynthetic origin consistent with an NRPS product.

For *Pseudonocardia* sp. KRD291 and *P. sediminis*, a t ype III polyketide synthase (T3PKS) BGC (*pks4*) (Fig. 4-9D) was putatively linked (Metcalf score = 3.4641) to a spectrum with a parent ion of *m/z* 249.148 produced by both strains, which was identified in the GNPS library as dihydroauroglaucin (Fig. 4-9E). Interestingly, 42% of the genes of this BGC showed similarity with the BGC reported in MIBiG repository for alkyl-*O*-dihydrogeranyl-methoxyhydroquinones from *Actinoplanes missouriensis* 431 (Fig. S10), whose products show structural similarity to dihydroauroglaucin in the phenolic lipid motif, potentially as a consequence of the T3PKS activity. Similar to *nrps1*, there is insufficient evidence to make conclusions about the identity of the molecular feature linked to *pks4*. However, the structural congruency between the potential product of *pks4* and the GNPS library annotation for the spectrum with parent ion of *m/z* 249.148 gives strong evidence for the putative correlation between them.

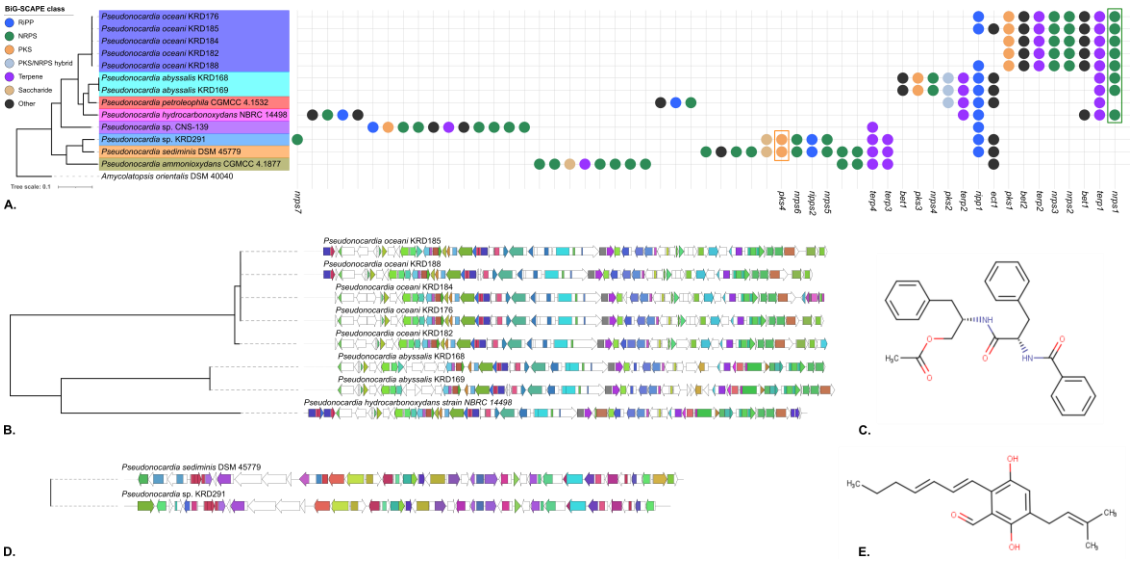


Figure 4-9 Biosynthetic gene cluster (BGC) profiles of *Pseudonocardia* spp. integrated by NPLinker. A. Phylogenetic tree reconstructed based on a multi-locus sequence alignment showing the distribution of BGCs annotated in the Southern Ocean *Pseudonocardia* strains in gene cluster families (GCF). BGCs are represented by nodes, and GCFs by columns. Selected GCFs linked by NPLinker are highlighted (*nrps1* = green box; *pks4* = orange box). CORASON-based multi-locus phylogeneny of *nrps1* (B) and *pks4* (D) are showed. C. Structures of the library hit asperglaucide identified by GNPS for the parent ion m/z 403.202 linked by NPLinker to *nrps1*. E. Structure of the library hit dihydroauroglaucin identified by GNPS for the parent ion m/z 249.148 linked by NPLinker to *pks4*.

4.3. Discussion

In a prior work, Soldatou *et al*. [262] analysed the potential of the Southern Ocean strains by applying an integrative multi-omics approach. However, the *Pseudonocardia* spp. strains were excluded from the genomics analysis to avoid possible gene annotation mistakes due to the fragmentation of open reading frames (ORFs) as a consequence of the low contiguity of the short-reads based assemblies. For instance, in a study with *Streptomyces* draft genomes, it was found that fragmentation is particularly common in protein classes with repeating multi-modular structures such as PKS, NRPS and serine/threonine kinases [397]. In this study, the use of long-reads technology allowed an increase in genome quality of the eight sequenced *Pseudonocardia* strains. The effect of long-reads sequencing technologies over the BGCs fragmentation was reported previously for the *Pseudonocardia* genus. In particular, Goldstein *et al*. (2019) evaluated the fragmentation of BGCs in genomes sequenced by Oxford Nanopore MinION long-reads and Illumina MiSeq short-reads of three *Pseudonocardia* strains isolated from individual *Trachymyrmex septentrionalis* ants, showing that long-reads improved the annotation of insertion sequences and BGCs [398]. Similar to

that observed in Southern Ocean *Pseudonocardia* strains, the differences in the BGC profile in the ant-associated strains were mostly related to NRPS and PKS, while the terpenes were shown to be stable [398]. Our results confirmed that long-reads can disambiguate highly repetitive regions such as NRPS BGCs.

The number of BGCs identified in the genomes of the Southern Ocean *Pseudonocardia* strains was relatively low in comparison to the average number of 21.6 reported for Streptomycetales and 19.8 reported for Pseudonocardiales [190]. In general, *Pseudonocardia* strains associated with ants tend to have a variable number of BGCs, from 11 to 23 [209, 348, 350], while *Pseudonocardia* strains with larger genomes (> 9Mb), such as *P. acacia* and *P. spinosispora* have been reported to have more the 20 BGCs [446]. The genome mining analysis of the phylogenetically close strain *Pseudonocardia* sp. CNS-139, which was also isolated from the marine environment, revealed a total of 14 BGCs including hybrid PKS and many NRPS clusters [194]. However, due to the low contiguity of the CNS-139 assembly, many of the NRPS BGCs were located at contig edges, making it difficult to conclude an accurate number of BGCs for this strain. The recent publication of the complete genome of *P. petroleophila* and *P. broussonetiae* [140], which are the phylogenetically closest relatives of *P. abyssalis* and *P. oceani*, made it possible to integrate these strains in the genomic analysis, showing that the number of BGCs annotated in the Southern Ocean strains is congruent with their closest relatives. However, *P. broussonetiae*, an endophytic bacterium isolated from the roots of *Broussonetia papyrifera* related to *P. oceani* KRD185 with an ANI value of 93.7% (discussed in Chapter 3), had two BGCs more than their marine free-living relative. Although this difference in the number of BGCs may not be significant, it is possible to argue that the ecological context may influence the number and diversity of BGCs, similar to what has been proposed for insect-associated *Streptomyces* [211].

The evolutionary analysis of the BGCs annotated in *P. abyssalis* and *P. oceani* showed that more than 50% of them are part of the same GCFs, which in some cases also included BGCs from *P. hydrocarbonoxydans*, *P. petroleophila* and *P. broussonetiae*. The GCF profile of the analysed *Pseudonocardia* species is more consistent within *Pseudonocardia* than reported for other genera. For example, in the *Salinispora* genus, only five of 124 related BGCs (measured as “Operational Biosynthetic Units”) were present in the common ancestor of the genus after analysing 75 genomes across three species (*S. arenicola, S. pacifica* and *S. tropica*) [179]. However, the number of strains analysed in the *Salinispora* study was higher than the

number of strains in this study, as a consequence, the genetic variability observed in the Southern Ocean *Pseudonocardia* strains might not represent their real biosynthetic richness. In fact, for the *Salinispora* genus, it was shown that the biosynthetic diversity increased with the number of strains for the three species [179]. In this sense, the relatively low number *P. abyssalis* and *P. oceani* strains analysed may influence the estimation of the BGC diversity for these lineages.

The metabolomics analysis also showed a high number of shared molecular features between the metabolomics profiles of *P. abyssalis*, *P. oceani*, *P. hydrocarbonoxydans*, and *P. petroleophila*, confirming that metabolite production in this group is highly influenced by phylogeny, similar to that reported for Myxobacteria [447], *Planomonospora* spp. [428], *Streptomyces* spp. [448], and *Salinispora* spp. [433]. That means that the metabolomics profile observed in *P. abyssalis* and *P. oceani* is closer to related strains like *P. petroleophila* than less related strains like *P. ammonioxydans*. In the same way, *P. abyssalis* KRD168 and *P. oceani* KRD185 shared many GCFs, they also shared more than 50% of the total number of parent ions, making this paired omics dataset attractive for analysing the patterns of strain presence or absence. In fact, this approach allowed the reduction of the number of molecular features to a few candidates by enabling informed decision making, such as selecting molecular features to be analysed in the mass spectrometry imaging (MSI) experiment to be discussed in Chapter 5. The identification of polyketides and non-ribosomal peptides in molecular features linked with PKS and NRPS BGCs supported the results obtained in the genome mining analysis. In general, the analysis of the paired omics dataset revealed a strong correlation between phylogeny and the specialised metabolisms as reported previously for actinomycetes [190]. These results suggested that the BGC in the shared GCFs were present in a common ancestor of *P. abyssalis* and *P. oceani*, as well as of *P. broussonetiae*, *P. hydrocarbonoxydans*, and *P. petroleophila*, while the species-specific BGCs, such as *pks3* and *bet2*, were potentially acquired after the speciation process. A similar interpretation can be made about the GCF distribution between *Pseudonocardia* sp. KRD291 and *P. sediminis*, where *nrps7* were potentially developed after these species diverged. In addition, the congruence observed between the evolutionary relationship in the GCFs and the phylogenetic species tree also supports the correlation between the speciation process and the evolution of the specialised metabolism, in agreement with the observations made in *Salinispora* spp. [179].

So far, the obtained results suggest that clonal expansion played a relevant role in the structure and distribution of BGCs over other evolutionary processes like population isolation or horizontal gene transfer (HGT). However, by following the analysis of homologous recombination on GCFs proposed by Chase *et al*. [449], it was found that the recombination events probably act as a cohesive force in the maintenance of conserved BGCs across related species such as *P. abyssalis* and *P. oceani*. The role of HGT over the evolution of the specialised metabolism in actinomycetes has been widely studied before. For example, in *Salinispora* spp. it was demonstrated that genes related to specialised metabolisms tended to be flanked by genomic islands (GIs) enriched with mobile genetic elements, whose boundaries were highly conserved among all strains, suggesting a horizontal origin for most of the pathways [179, 450], while plasmid-encoded BGCs and GIs in chromosome-integrated BGCs gave evidence of their horizontal origin in ant-associated *Pseudonocardia* spp. strains [348]. Furthermore, the BGCs for rebeccamycin [349], and selvamicin [350], both isolated from *Pseudonocardia* strains related to fungus-growing ants of the genus *Apterostigma* were described as plasmid-encoded BGCs. In particular, the results obtained for *P. abyssalis* and *P. oceani* revealed a high rate of homologous recombination between closely related species, which combined with inter-species gene-flow barriers, conserve cohesion between phylogeny and GCF distribution, in the same way, it was reported for *Salinispora* spp. [449]. Although homologous recombination acts as a cohesive force in conserved BGCs such as *ripp1*, gene gain and loss of BGCs was also observed. This gene gain and loss situation was particularly well-illustrated for *bet1*. These observations are congruent with a model where the sequence evolution (e.g. base-pair substitution and indels) occurs with a delay relative to genome evolution (gain and loss of genes) as a consequence of sequence homogenization by homologous recombination. In other words, the BGC sequence similarity across species could be explained through homologous recombination, but eventually, the gene gain and loss would accelerate the establishment of barriers to recombination, and potentially promote niche differentiation [451]. In the case of *Salinispora*, Chase *et al*. concluded that BCG evolution occurs mainly through gene gain and loss events, while homologous recombination contributes to interspecies gene diversity [449]. This is congruent with the results of the evolutionary analysis performed in the BGCs of the Southern Ocean *Pseudonocardia* spp. strains. However, further analysis is required to assess the origin and evolutionary dynamics of the accessory genes. As one of the main limitations of the present analysis is the low number of strains included, future effort on isolating and sequencing more strains of *P.*

abyssalis and *P. oceani*, as well as of *P. broussonetiae*, *P. hydrocarbonoxydans*, and *P. petroleophila*, will allow a more exhaustive analysis of the evolution and diversification of the BGCs encoded in these species.

The Southern Ocean *Pseudonocardia* strains represent an interesting model to better understand the role of biogeography on the evolution of the specialised metabolism as they are related in both phylogeny and ecology. As vertical inheritance seems to shape the structure of the BGCs in *P. abyssalis* and *P. oceani*, it is possible to hypothesize that their products could play a relevant role in the ecological fitness that is not niche-restricted. However, the acquisition of accessory genes in conserved BGCs, as well as the identification of species-specific BGCs, might be the product of a local adaptation that shaped the differences observed in the specialised metabolism of *P. abyssalis* and *P. oceani*. Similar findings were reported for the study of evolutionary dynamics of BGCs in 24 *Streptomyces* strains distributed in two sister taxa, as phylogenetic distribution agrees with their sample location. In this study, it was proposed that the BGCs that are not present in common ancestors were acquired recently by HGT as a result of differences in the evolutionary pressures during the lineage divergence [231]. Nevertheless, the correlation between geographic location and HGT events has to be analysed carefully, as BGCs appear to have crossed boundaries imposed by geography and ecology [452]. The original hypothesis that biogeography could restrict horizontal acquisition of BGCs in *P. abyssalis* and *P. oceani* from other *Pseudonocardia* spp. is not congruent with the fact that horizontally acquired genes are often shared between closely related species [453]. For instance, the pan-genome analysis of 122 *Streptomyces* genomes showed that most of the GCFs (82.7%) were restricted by phylogeny, even though, 93% of the BGCs exhibited evidence of HGT. So, despite widespread HGT events, the dynamics of these events are highly controlled by phylogeny rather than geographic barriers [454]. In addition, for the *Salinispora* genus, it was shown that the geographic distance between the geographic origins of the strains did not decrease the number of related BGCs, some examples of BGCs that were observed at a single location were reported throughout [215]. Therefore, a plausible model to explain the role of homologous recombination in the evolution of the BGCs annotated in *P. abyssalis* and *P. oceani*, is that, instead of a horizontal transfer event, several recombination events influenced the creation of the sub-clusters exclusively in one of the species that eventually were integrated to the chromosome. This scenario has been proposed before as an evolutionary mechanism for BGCs [170, 248, 455].

Chevrette and Currie (2019) proposed that the evolutionary dynamics of antibiotic biosynthesis is a combination of vertical heritage and HGT, with an influence of changing environmental pressures, and opportunities for genetic exchange bounded by both phylogenetic relatedness and opportunity for gene flow [248]. For instance, the study of the evolution of hybrid isoprenoid clusters in the MAR4 marine *Streptomyces* lineage revealed that a combination of HGT, gene duplication, and gene rearrangement produced a high gene cluster diversity in this particular group, which could be related to the adaptation to the marine environment [456]. This evolutionary dynamic in *P. abyssalis* and *P. oceani* is illustrated by *bet1*, which showed a core group of genes shared across both species, and a group of accessory genes that increased the gene cluster diversity. Even if the evolutionary origin of the differences between the specialised metabolisms of the Southern Ocean *Pseudonocardia* strains have a combination of vertical heritage and HGT events, the role of the environment and geographic distribution over the changes in their specialised metabolism is still unclear.

Even though the Geographic Mosaic Theory of Coevolution (GMC) proposed by Caldera *et al*. (2019) [238] offers a framework for the evolution of the specialised metabolism in fungus-growing ants related *Pseudonocardia* strains, this theory was built for strains with an ecological distribution shaped for their symbiotic relationship [457]. The influence of ecological and geographical variables over the specialised metabolism of free-living *Pseudonocardia* is still understudied. Recent studies have pointed out that specialised metabolites in the natural environment are often produced in non-lethal concentration against other microorganisms. So, it has been suggested that they have more complex ecological functions such as altering phenotypes, fitness, and community composition of microorganisms which are still undetermined [458]. Therefore, even without evidence of coevolution with a host, many other ecological variables may explain or partially explain the effect of biogeography on the evolutionary history of the specialised metabolisms of *P. abyssalis* and *P. oceani*. The free-living condition can be considered an ecological trait that is influenced by the geographic distribution by itself. For instance, in a pan-genomic study of 30 *Streptomyces albidoflavus* strains, three well-defined phylogenomic clades were defined, where the distribution of free-living and entomic (related to insects) strains showed a specific distribution within the clades. Although the geographic origin could not explain the population structure, all the entomic strains were collected from low-latitude areas, while the free-living strains were collected mainly from higher latitudes [459]. In this sense, the

environmental conditions of the Southern Ocean could influence both the gain and loss of genes that eventually drive the evolution of the BGCs identified in *P. abyssalis* and *P. oceani*, as well as in *Pseudonocardia* sp. KRD291. Hence, it is possible to propose that the NRPS, PKS, and betalactone BGCs annotated exclusively in the Southern Ocean *Pseudonocardia* strains could play an important role in the local adaptation of these strains to their environment. A deeper analysis of the potential ecological roles of the specialised metabolism in the inter-species interactions between *P. abyssalis*, *P. oceani*, and their closest relatives will be carried out in chapter 5.

Finally, the GNPS and DEREPLICATOR+ analysis identified the production of metabolites related to known compounds with antimicrobial activity like neomycin and filipin. In particular, filipin is a polyene macrolide with potent antifungal activity isolated originally from *Streptomyces filipinensis* [460]. The *Pseudonocardia* genus is a well-known producer of antifungal polyenes as the nystatin-like *Pseudonocardia* polyenes (NPP) A1 [461] and B1 [462], isolated from *Pseudonocardia autotrophica*, and selvamicin, an antifungal polyene macrolide isolated from *Pseudonocardia* spp. strains related to ants of the genus *Apterostigma* [350]. However, as no BGCs related to polyene macrolides were identified in the genome of the Southern Ocean strains, there is not enough evidence to propose that these strains produce this type of antifungal metabolites. Still, these findings give important inputs for future efforts on isolating and elucidating bioactive compounds from these *Pseudonocardia* strains. Moreover, the novelty observed in the metabolomics and BGC analyses supported the observations of Soldatou *et al*. (2021), who suggested that *P. oceani* KRD185 strains could be the source of antimicrobial metabolites [262]. Further analysis will be performed on the antifungal activity of *P. abyssalis* and *P. oceani* in chapter 5.

Chapter 5 Assessing the effect of culturing conditions on the metabolomic profiles of *P. abyssalis* and *P. oceani*

5.1. Introduction

5.1.1. Antifungal metabolites from *Pseudonocardia* spp.

The *Pseudonocardia* genus is well-known due to the mutualistic relationship developed by some strains with fungus-growing ants. In this mutualistic relationship, it has been shown that antifungal metabolites produced by *Pseudonocardia* strains protect the ants' fungus garden from the fungal parasite *Escovopsis*, as well as other generalist fungal pathogens [141, 201]. As a result, a few antifungal metabolites isolated from members of this genus have been described. For example, the cyclic depsipeptide dentigerumycin isolated from an *Apterostigma* nest collected in Gamboa, Panama, showed inhibitory activity against the wild type and amphotericin-resistant strains of *C. albicans* [463]. Similarly, two *Pseudonocardia* strains from ants of the genus *Apterostigma* collected at La Selva Biological Station, Costa Rica, showed activity against *Candida albicans*, as well as other fungal strains such as *Saccharomyces cerevisiae*, *Trichoderma harzianum*, and *Aspergillus fumigatus*, which was attributed to the polyene macrolide selvamicin [350]. Another example of a polyene macrolide from *Pseudonocardia* is the nystatin-like *Pseudonocardia* polyene (NPP) A1 produced by *P. autotrophica* which showed *in vitro* and *in vivo* antifungal activity against *C. albicans* [462]. In *Pseudonocardia* sp. P1 strains isolated from *Acromyrmex octospinosus* colonies collected in Trinidad and Tobago a not-yet-fully elucidated compound nystatin P1 was identified. This metabolite showed antifungal activity against *C. albicans*. Nystatin P1 and selvamicin have an extra hexose motif in comparison with nystatin A1, according to the MS analysis [464].

In ant-associated *Pseudonocardia*, the antifungal activity appears to be a response to the specific ecological interaction between bacteria, ants, and the parasitic fungus *Escavopsis*. Then, one of the main challenges is to obtain these metabolites under laboratory conditions. For example, selvamicin was isolated using sodium butyrate as a chemical elicitor, which increased its production [350]. Similarly, activity against *C. albicans* by *Pseudonocardia* sp. P1, presumably due to nystatin P1, was increased by the addition of sodium butyrate to the

medium [465]. Therefore, understanding the effect of culturing conditions on the production of bioactive metabolites will help us to exploit the chemical potential of this genus.

5.1.2. Regulation of antibiotic production in actinomycetes

In *Streptomyces*, the life cycle is closely related to the production of antibiotics (Fig. 5-1). Under nutrient depletion, the vegetative mycelium differentiates to form aerial hyphae, and during this moment most antibiotics are produced [100]. For *Streptomyces*, as well as other filamentous microorganisms like *Pseudonocardia*, nutrient depletion leads to an autolytic degradation via programmed cell death (PCD) of the old vegetative mycelia to release new nutrients for the production of aerial hyphae and spores [100, 466]. For example, the production of developmentally-regulated nucleases in *S. antibioticus* degrades DNA and cytoplasmic components, with transitory maintenance of the integrity of the plasma membrane. This occurs twice, first in the substrate mycelium during the emergence of the aerial hyphae, and then in the aerial mycelium during sporulation [467, 468]. During this phase of actinomycete development, the released PCD products like amino acids, amino sugars, and nucleotides offer a nutrient source for other competing microorganisms. Hence, the production of antibiotic specialised metabolites during the maturation and sporulation phase would protect these nutrients from being taken by other microorganisms. To this end, cell-wall peptidoglycan is recycled to release the amino sugars *N*-acetylmuramic acid (MurNAc) and *N*-acetylglucosamine (GlcNAc). Then, after its metabolic transformation to glucosamine-6-phosphate (GlcN-6P) in the cytoplasm, GlcNAc regulates the production of antibiotics through the global GntR-family regulator *DasR*, which represses the activation of BGCs [466, 469, 470]. As such, the use of GlcNAc as a chemical elicitor has been proposed for the activation of cryptic BGCs that are not expressed under standard laboratory conditions in antibiotic discovery programs [265, 471]. Finally, global regulators could additionally be related to the ecological function of the regulated BGC. For instance, as GlcNAc is also a product of fungal cell wall degradation, the BGCs regulated by *DasR* could be related to metabolites with antifungal activity [466].

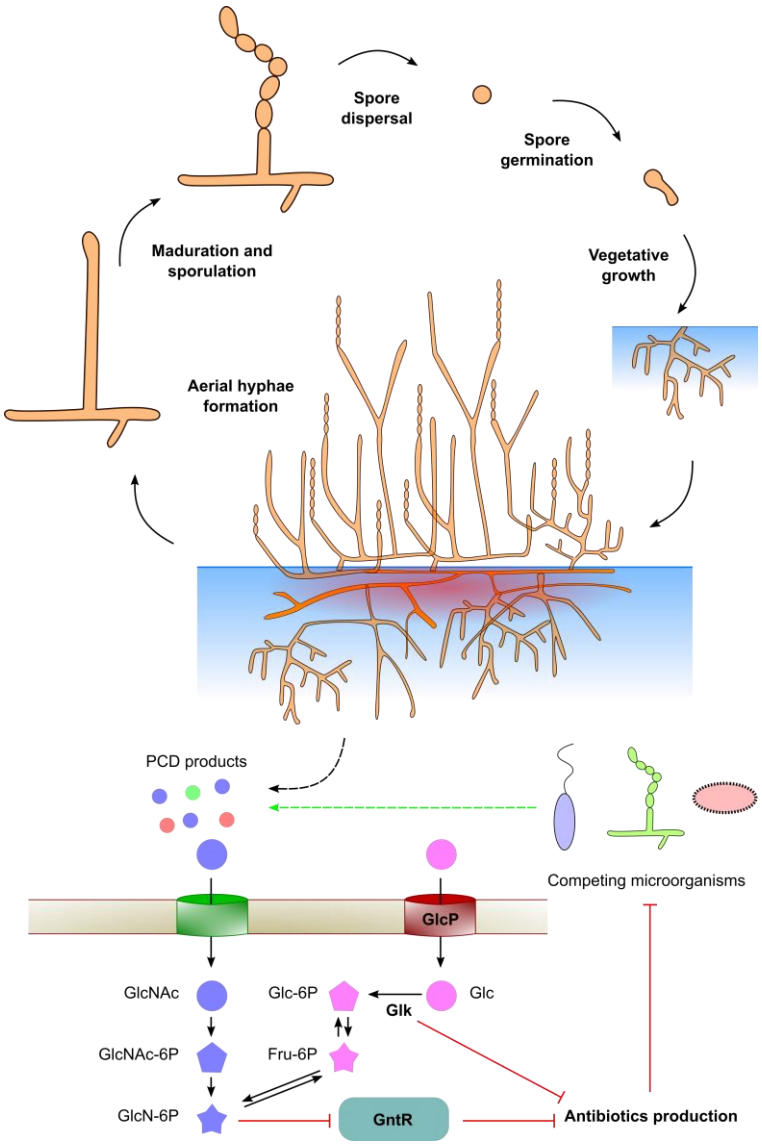


Figure 5-1 Developmental life cycle of actinomycetes, showing the role of nutrient starvation and competing microorganisms in the production of antibiotics.

The nutrient-sensory *DasR* system works as a link between carbon metabolism and development, as well as connecting the control of primary metabolism (carbohydrate metabolism) to the control of specialised metabolisms (BGC activation) [472]. In *Streptomyces*, it was demonstrated that the effect of GlcNAc depends on media conditions. In particular, on nutrient-rich medium it was shown that the addition of GlcNAc blocks morphogenesis and antibiotic production, while on nutrient-depleted media, GlcNAc has the opposite effect. This suggests that under nutrient-rich conditions, the GlcNAc supply is probably derived from chitin, and as such, the peptidoglycan hydrolysis and antibiotics production is repressed as development is undesired. In contrast, under nutrient-depleted

conditions, GlcNAc is probably the product of peptidoglycan hydrolysis, which triggers the production of antibiotics as explained above [470, 472].

Besides its possible relationship with the cellular contractions of GlcNAc, the levels of glucose (Glc) play a central role in the regulation of BGC expression. This is done through carbon catabolite repression (CCR), which ensures that when a primary carbon source, such as Glc, is available the utilization of secondary carbon sources is repressed [472, 473]. The availability of high-energy carbon sources like Glc promote vegetative growth and repress antibiotic production [473]. The negative effect on the production of antibiotics by Glc has been reported for streptomycin production by *S. griseus* [474], actinorhodin by *S. coelicolor* [475], chloramphenicol by *S. venezuelae* [476], and erythromycin by *Saccharopolyspora erythraea* [477]. In *Streptomyces*, CCR is directly regulated by ATP-dependent glucose kinase (*Glk*) activity [472, 473]. For example, the deletion of the *Glk* gene restores the production of antibiotics in the presence of Glc [478, 479]. However, it is important to note that *Glk*-independent pathways of CCR have been also described [480], showcasing the complexity of the role of Glc on specialised metabolism regulation.

5.1.3. Chemical ecology of antibiotic production by actinomycetes

In nature, actinomycetes are part of complex microbial communities, and members of these communities interact with each other through the production of specialised metabolites. In theory, it may therefore be possible to elicit specialised metabolite production in the laboratory, using co-culture experiments to mimic these naturally occurring interactions [466]. For example mycolic acid-containing bacteria have been shown to induce the production of antimicrobial metabolites in *Streptomyces* spp. [481, 482]. Furthermore, interspecies interactions mediated by autoregulators, such as γ-butyrolactones, support the elicitation of metabolites production through actinomycetes-actinomycetes co-cultures [483, 484]. Additionally, for *Streptomyces* it has been shown that the specialised metabolites themselves may regulate the metabolite production of a nearby species [485]. In general, the chemical interaction between actinomycetes in co-cultures involve complex regulatory mechanisms that modify the production of specialised metabolites [259, 483]. For example, using mass spectrometry imaging tools to study the interactions between *S. coelicolor* and another five actinomycetes it was shown that the metabolomic profile of *S. coelicolor* in each interaction was unique. Moreover, the co-culturing elicited the production of metabolites

that were not observed in the *S. coelicolor* monocultures, from among which a new group of siderophores was described [259].

Ecological interactions are a strong force in the regulation of bioactive metabolite production. In this sense, the *Pseudonocardia* genus represents an interesting model to study these. For instance, the *Pseudonocardia* strains in symbiosis with fungus-growing ants show a stronger inhibition against *Escovopsis* spp. than against other fungi. Likewise, the inhibition against *Escovopsis* by the fungus-growing ants' strains is stronger than by free-living *Pseudonocardia* spp. [457]. Furthermore, mass spectrometry imaging analysis showed that *Escovopsis* sp. elicit the production of specific metabolites by *Pseudonocardia* spp. [486]. In the same way, *Pseudonocardia* spp. also elicits the production of antimicrobial metabolites by *Escovopsis* spp. during a fungal infection in response [487].

Besides their antagonistic interactions against fungi, interactions between *Pseudonocardia* strains in symbiosis with fungus-growing ants have been also described. Although there is evidence that fungus-growing ants contain a mixed actinomycete consortium [464, 488], it has been also proven that ants maintain and transmit vertically a single *Pseudonocardia* strain [489, 490]. In this scenario, bioactive compounds against other actinomycetes offer *Pseudonocardia* strains a competitive advantage to protect its ecological niche [200]. In fact, recent reports suggest that vertically transmitted *Pseudonocardia* spp. strains produce antibiotic metabolites that reduce the growth rates of other microorganisms, which selectively maintain an antibiotic-producing bacterial consortium [491]. Based on this ecological hypothesis, competitive assays between *Pseudonocardia* strains led to the discovery of the antimicrobial metabolites 9-methoxyrebeccamycin [349] and the thiopeptide antibiotic GE37468 [492], both presumably related to a niche-defence function.

5.2. Results

Besides their slow growth rate, the low production of metabolites in *P. abyssalis* and *P. oceani* was one of the main challenges of this project. Here, two strategies applied to induce the production of metabolites are presented. First, the effect of the chemical elicitor *N*-acetylglucosamine (GlcNAc) on the clinically relevant antifungal activity of *P. abyssalis* and *P.*

oceani was explored. Then, through a chemical ecology approach, the effect of inter-species interactions on the production of metabolites was analysed.

5.2.1. Effect of *N*-acetyl glucosamine on the production of antifungal metabolites by *P. abyssalis* and *P. oceani*

5.2.1.1. Antifungal activity of *P. abyssalis* and *P. oceani*

Preliminary tests against the environmental fungal strain *Rhizopus oryzae* showed that *P. abyssalis* KRD168 and *P. oceani* KRD185 produce antifungal metabolites and that this activity was induced by GlcNAc (Fig. 5-2). A following test against the clinically relevant human pathogen *Candida albicans* DSM 1386 confirmed the potential of these strains as producers of metabolites with antifungal activity (Table 5-1 and Fig. 5-3). Although the positive effect of GlcNAc over the bioactivity was also observed for *P. oceani*, this was not observed for *P. abyssalis*, *P. hydrocarbonoxydans* DSM 43281, or *P. petroleophila* DSM 43193. On the other hand, removing glucose (Glc) from the medium increased the bioactivity of all the tested strains. This was particularly notable with *P. hydrocarbonoxydans*, as *C. albicans* did not grow at all. Moreover, except for *P. oceani*, in the glucose depleted medium supplemented with GlcNAc the bioactivity of the strains remained similar, while the bioactivity of *P. petroleophila* was higher in the glucose depleted medium without GlcNAc. The strains were also tested against *Aspergillus niger* DSM 1957, but no activity was observed by any strain under the tested conditions.

Table 5-1 Bioactivity (zone of inhibition, mm) of four *Pseudonocardia* spp. against *C. albicans* (n = 2). Strains were cultured in ISP2, ISP2 supplemented with GlcNAc, glucose depleted ISP2, and glucose depleted ISP2 supplemented with GlcNAc.

Strain	ISP2	ISP2 + GlcNAc	ISP2ΔGlc	ISP2ΔGlc + GlcNAc
P. abyssalis	8.0 ± 1.3	7.4 ± 1.2	12.0 ± 4.7	11.6 ± 2.4
P. oceani	4.4 ± 1.1	11.8 ± 0.1	13.1 ± 2.7	20.2 ± 0.1
P. hydrocarbonoxydans	3.4 ± 0.4	3.5 ± 0.1	NG	10.5 ± 7.2
P. petroleophila	6.6 ± 0.8	4.2 ± 0.8	14.6 ± 5.7	9.8 ± 0.1

NG= not grown

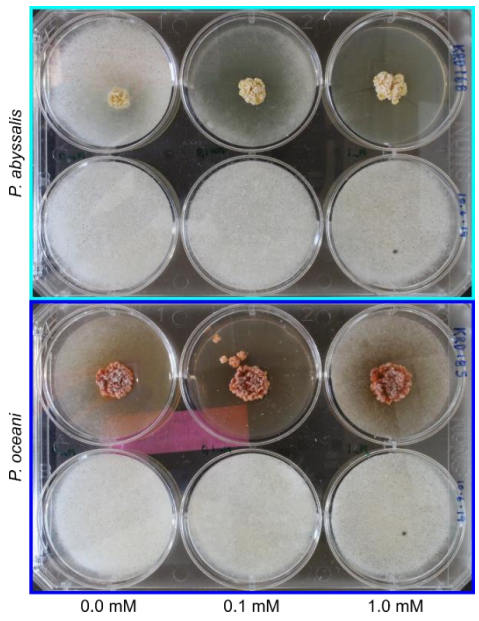


Figure 5-2 Bioactivity of *P. abyssalis* KRD168 and *P. oceani* KRD185 cultured in ISP2 medium, and in ISP2 supplemented with *N*-acetylglucosamine 0.1 mM and 1.0 mM, against *Rhizopus oryzae*. Strains were cultured in 6-well plates (diameter = 35 mm). Controls without Pseudonocardia cultures are in the lower well line of each plate.

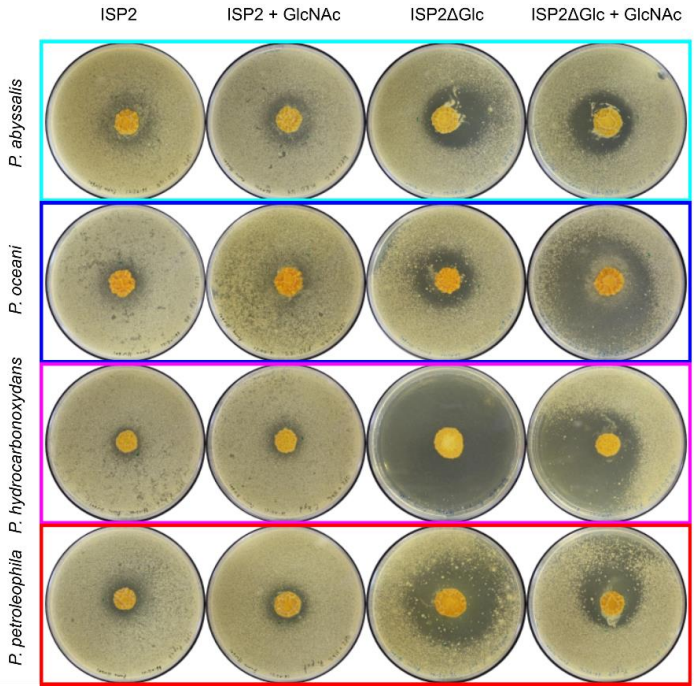


Figure 5-3 Bioactivity of *P. abyssalis* KRD168, *P. oceani* KRD185, *P. hydrocarbonoxydans* DSM43281, and *P. petroleophila* DMS 43193 cultured in ISP2 medium and glucose depleted ISP2 medium (ISP2ΔGlc) supplemented with N-acetylglucosamine 1.0 mM, against *Candida albicans* DSM1386. Strains were cultured in 90 mm Petri dish.

Analysis of genes related to *DasR* which could be related to the effect of GlcNAc was achieved using BLASTP to search protein sequences related to the *dasR* gene described in *Saccharopolyspora erythraea* NRRL 2338 (accession number WP_009951283) from *Pseudonocardia* sp. KRD291, *P. abyssalis* KRD168, and *P. oceani* KRD185. As a result, a HutC/FarR-like transcription factors of the GntR family was identified in the three species (Table S8). To better understand the possible role of this gene in the observed metabolomics profile, the deletion of *dasR* homologous genes may give insight into its role in both bioactivity and metabolite production. Unfortunately, this approach is beyond the scope of this project. Further studies on regulation pathways involving GlcNAc are needed to make conclusions about the effect of GlcNAc over the regulation of specialised metabolites production.

Finally, the antifungal activity of *P. abyssalis* KRD168 and *P. oceani* KRD185 was tested against the clinical isolates *Aspergillus fumigatus* AF293, *Candida albicans* SC5314, and *Candida auris* UACa11 (data not shown). As a result, *P. abyssalis* KRD168 showed inhibitory activity against *C. auris* (Fig. S11), while no growth inhibition was observed for the rest of strains at the tested conditions. Although promising antifungal activity against *Candida* spp. by *P. abyssalis* KRD168 and *P. oceani* KRD185 was observed, the lack of activity against other pathogenic strains could be supported by the fact that the human pathogens tested are not part of the natural environment of the Southern Ocean sediments from which these strains were isolated. As such, a limitation of this work is the minimal bioactivity testing performed. However, the bioactivity observed against the multi-drug resistant clinical isolate *C. auris* demonstrated the potential of *P. abyssalis* and *P. oceani* as source of metabolites with biomedical value.

5.2.1.2. Assessing the effect of *N*-acetyl glucosamine on the metabolomic profiles of *P. abyssalis* and *P. oceani*

To assess the effect of GlcNAc on the metabolomic profiles of *P. abyssalis* and *P. oceani*, as well as related *Pseudonocardia* strains, metabolite culture extracts were analysed using LC-MS/MS. The data generated for 16 *Pseudonocardia* strains, including eight Southern Ocean strains, (Table S7) cultured using both ISP2 agar and ISP2 agar supplemented with GlcNAc was compared using feature-based molecular networking (FBMN). As introduced in chapter 4, this tool introduces information from MS1 data like isotope patterns, retention time and

relative intensity, which allow a relative quantitative analysis to be performed. In the context of LC-MS, a molecular feature corresponds to an ion eluting from the chromatographic separation system and detected by the mass spectrometer. As it will be discussed later, a single metabolite can be related to one or more features that represent different ion species with different retention times. As a result of this analysis, a total number of 2062 features, of which 1291 were part of a network (cluster) and 266 matched with a reference compound in the GNPS spectra library (including analogous ± 100.0 Da in the precursor ion) were obtained (Fig. 5-4). In addition, the MolNetEnhancer workflow was used to characterize the molecular network according to the ClassyFire chemical ontology. This allows categorisation of the molecular features according to their probable chemical classifications. As a whole, this analysis aims to assess what type of metabolite is detected with higher intensity across the two growing conditions (Fig. 5-4). Although most of the molecular families were not classified, some of them related to lipid-like metabolites that showed changes in the relative intensity detected when GlcNAc was added to the ISP2 agar. Other metabolites like benzenoids and organic oxygen compound showed a more equal detection in both growing conditions. However, proper quantitative experiments using external standards are required to confirm these results. Future identification of metabolites produced by the Southern Ocean strains will allow a more detailed analysis.

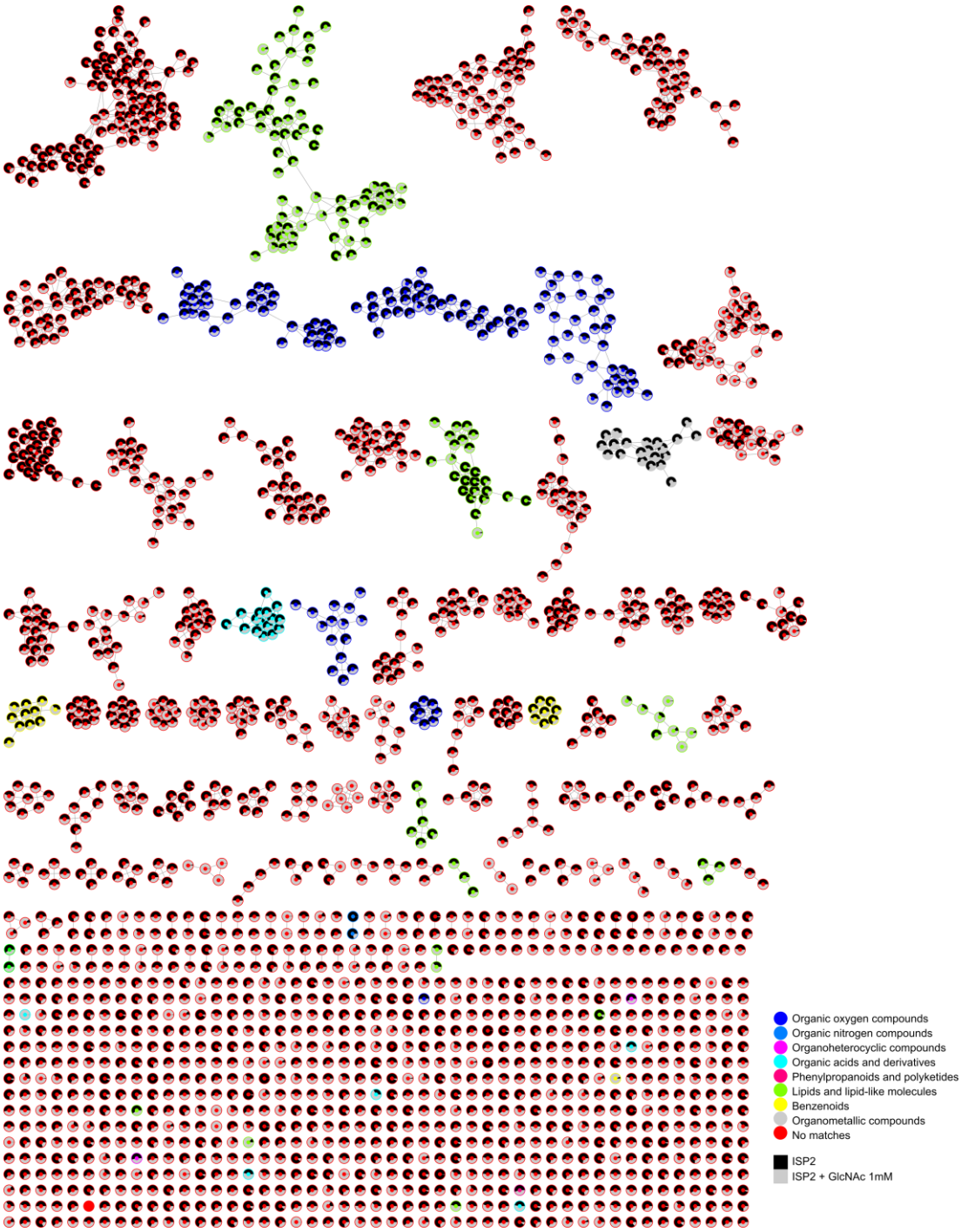


Figure 5-4 Feature-based molecular network of 2,062 features showing the relative abundance of detected ions in ISP2 (black) and ISP2 supplemented with *N*-acetyl glucosamine 1mM (grey). Nodes are coloured according to the ClassyFire chemical taxonomy. Doughnut chart inside each node shows the relative distribution of samples where the particular spectrum was detected, coloured according to the growing conditions.

Through the FBMN analysis, some molecular features were identified to be produced specifically under only one condition. For example, the metabolites of a molecular family of diterpenoids, which includes nodes with a precursor ion of *m/z* 301.215 and *m/z* 314.242

identified in the GNPS database as dehydroabietamide, were detected with higher intensity when *Pseudonocardia* strains were cultured in the presence of GlcNAc. An exception to this was the node with a precursor ion of *m/z* 301.140, identified as (5xi)-abieta-8,11,13-trien-18-oic acid, which was detected with higher intensity in strains cultured using non-supplemented ISP2 (Fig. 5-5). These metabolites are diterpenes that could be sub-products or intermediaries of the terpene biosynthetic pathways, suggesting an effect on this type of specialised metabolites. However, further analysis such as transcriptomics targeting terpene BGCs would be needed to conclude this possible effect.

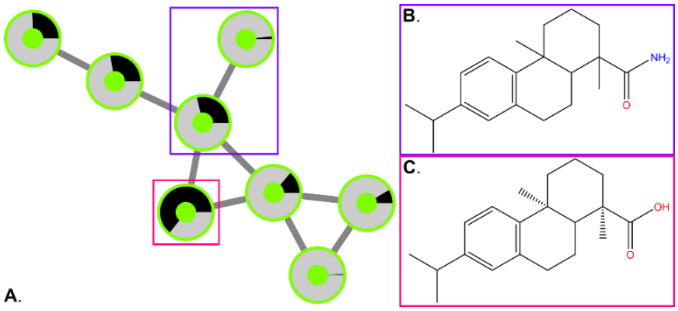


Figure 5-5 Molecular family classified as diterpenoids showing relative abundance of detected ions of each molecular feature detected in ISP2 (black) and ISP2 supplemented with GlcNAc 1 mM (grey) (A). The structures of the library hits dehydroabietamide (B) and (5xi)-abieta-8,11,13-trien-18-oic acid (C) are shown.

To analyse the statistical significance in the relative abundance of each molecular feature, differences in their peak intensities were analysed using volcano plots. This kind of scatterplot shows statistical significance (P value) versus ratio of change (fold change), enabling a visual identification of metabolites with large fold changes that are also statistically significant (p-values < 0.1). The statistical analysis of the feature distribution across the samples showed a species-dependent effect of GlcNAc over the metabolomics profile of the tested *Pseudonocardia* strains. In particular, for *P. abyssalis* KRD168, 1.3% and 7.9% of features showed a fold-change with statistical significance in ISP2 medium and ISP2 supplemented with GlcNAc respectively. While 90.8% of features did not show significant changes (Fig. 5-6A). In the case of *P. oceani* KRD185, 3.1% and 2.5% features showed a fold-change with statistical significance in ISP2 medium and ISP2 supplemented with GlcNAc respectively. In the latter case, 94.4% features did not show significant changes (Fig. 5-6B). As the identity of the metabolites with differential detections in both conditions is unknown, it is not possible to make conclusions about the biological effect of the differences in the

metabolomic profiles. It is possible that only one or few of the up-regulated metabolites observed in the LC-MS experiment were responsible for the differences in bioactivity against *C. albicans* (Fig. 5-3). However, no significant changes were observed in more than 90% of the detected molecular features in both cases, which potentially means that GlcNAc affects fewer metabolic pathways in *P. abyssalis* and *P. oceani* than in the other *Pseudonocardia* strains tested, as discussed below.

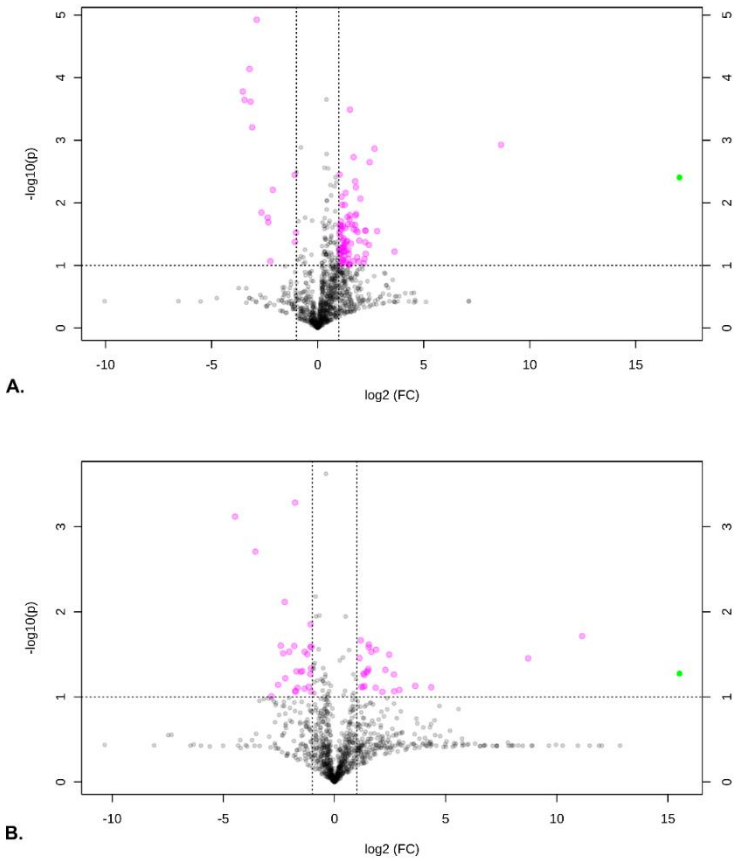


Figure 5-6 Volcano plot showing the differential metabolite identification of *P. abyssalis* KRD168 (A) and *P. oceani* KRD185 (B) cultured in ISP2 and ISP2 supplemented with N-acetyl glucosamine 1mM. Each dot represents a molecular feature detected in the LC-MS/MS analysis. The x-axis represents the ratio of change (fold change), and the y-axis the statistical significance (p-value). Molecular features with p-values < 0.1 (horizontal dotted line) and a fold change < 0.5 or > 2 (vertical dotted lines) are highlighted in magenta. Molecular features considered significantly down-regulated are showed in the left side, and those considered significantly up-regulated are showed in the right side. The molecular feature corresponding to N-acetyl glucosamine is highlighted in green.

In the case of *P. petroleophila*, 15.8% of the molecular features showed a statistically significant fold change in ISP2, while 11.9% in ISP2 supplemented with GlcNAc (Fig. S12A).

Moreover, in the case of *P. hydrocarbonoxydans* 28.0% of the molecular features showed a statistically significant fold change in regular ISP2 and 8.2% did it in ISP2 supplemented with GlcNAc (Fig. S12B). In contrast, in the metabolomics profile of the third Southern Ocean species, *Pseudonocardia* sp. KRD291, 44.8% of the molecular features showed a statistically significant fold change in regular ISP2 and only 5.3% in ISP2 supplemented with GlcNAc that exhibited a fold-change with statistical significance. In this last case, 49.9% of features remain without significant changes (Fig. 5-7A). Similar results were obtained for *P. sediminis* (Fig 5-7B) and *P. ammonioxydans* (Fig. S12C). Once again, the number of differential metabolites can not be directly related to changes in the observed bioactivity. However, it was observed that the response to GlcNAc changes appreciably from one strain to another.

In summary, it was found that the effect of supplementing the medium with GlcNAc was strain-dependent, and affected a greater number of molecular features in *Pseudonocardia* sp. KRD291 than in *P. abyssalis* and *P. oceani*. As mentioned above, these results did not correspond with the observations on the bioactivity against *C. albicans*. So, at least for the antifungal activity against this pathogen, the number of metabolites detected in the different conditions cannot explain the observed bioactivity. Hence, future efforts should focus on structural elucidation of particular metabolites and quantitative analysis of their production in presence of GlcNAc.

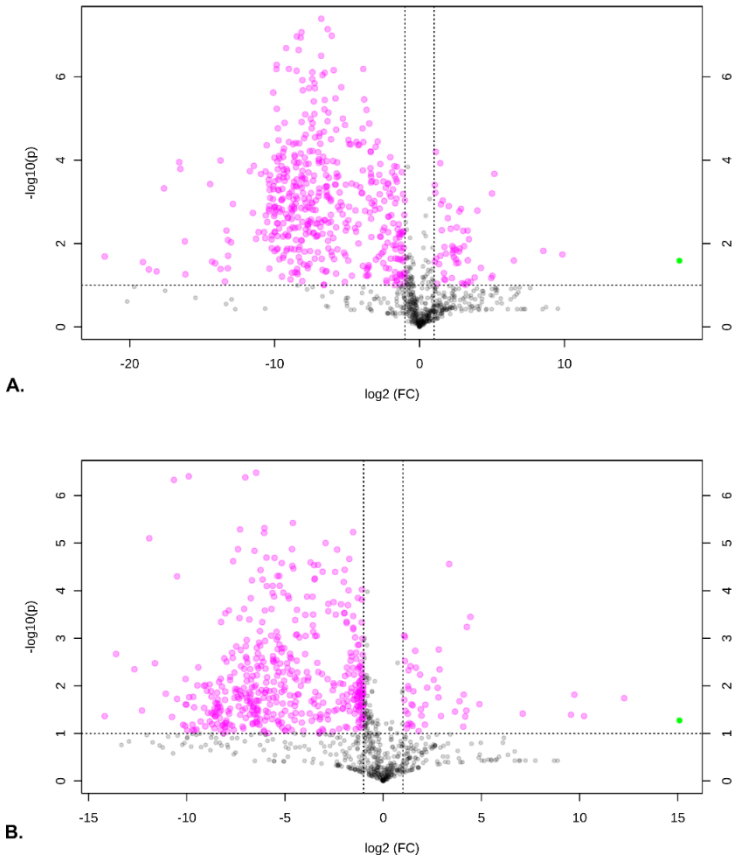


Figure 5-7 Volcano plot showing the differential metabolite identification of *Pseudonocardia* sp. KRD291 (A) and *P. sediminis* DSM 45779 (B) cultured in ISP2 and ISP2 supplemented with N-acetyl glucosamine 1mM. Each dot represents a molecular feature detected in the LC-MS/MS analysis. The x-axis represents the ratio of change (fold change), and the y-axis the statistical significance (p-value). Molecular features with p-values < 0.1 (horizontal dotted line) and a fold change < 0.5 or > 2 (vertical dotted lines) are highlighted in magenta. Molecular features considered significantly down-regulated are showed in the left side, and those considered significantly up-regulated are showed in the right side. The molecular feature corresponding to N-acetyl glucosamine is highlighted in green.

Studying the role of chemical elicitors such as GlcNAc helped to explore both the complex metabolite profiles of *Pseudonocardia* spp. from the Southern Ocean and their antifungal activity against *C. albicans*. However, as far as we know, human pathogenic *Candida* spp. does not share the same ecological niche as *Pseudonocardia*. As such, to have a better comprehension of the culturing variables that influence *Pseudonocardia* metabolomics profiles, a combination of co-culturing strategies was applied to investigate potential metabolite niche-defence roles, as well as having a better understanding of the ecological and phylogenetic variables that affect the specialised metabolisms of the Southern Ocean strains.

5.2.2. Effect of inter-species interactions on the production of bioactive metabolites in *P. abyssalis* and *P. oceani*

5.2.2.1. Using intruder assay to evaluate inter-species interactions

Antagonistic interactions between *P. abyssalis*, *P. oceani*, *P. hydrocarbonoxydans*, and *P. petroleophila*, as well as other *Pseudonocardia* spp. isolated from marine sediments (Table S9) were evaluated. To do this, a challenge assay was performed by co-culturing two strains. First, the resident strain was seeded in the plate centre, and after their sporulation, the challenger strain was inoculated around the resident's colony to measure the inhibition in the growth of the challenger strain. In general, it was found that antagonist interactions occurred more often between phylogenetically distant strains, which means that interactions between species of the same monophyletic group were not observed. For example, although *P. antitumoralis* DSM 45322 was the most aggressive strain interacting against six of the ten challenger strains, including *P. abyssalis* KRD168 and *P. oceani* KRD185, no antagonistic interactions against its phylogenetic neighbours *P. ammonioxydans* DSM 44958 and *Pseudonocardia* sp. VO44-3 nor against the clade of *P. sediminis* DSM 45779 and *Pseudonocardia* sp. KRD291 were observed. Similarly, *P. abyssalis* KRD168, *P. oceani* KRD185, *P. petroleophila* DSM 43193, and *P. hydrocarbonoxydans* DSM 43281, which are members of the same monophyletic group (Fig. 3-6), did not interact with each other (Fig. 5-8). It is important to point out that this experiment was designed to measure growth inhibition. Additional morphological changes or impact as a result of interactions such as sporulation or pigment production were not monitored. Although this approach focussed on the production of antimicrobial metabolites, the production of metabolites with other functions, such as communication, should not be discarded.

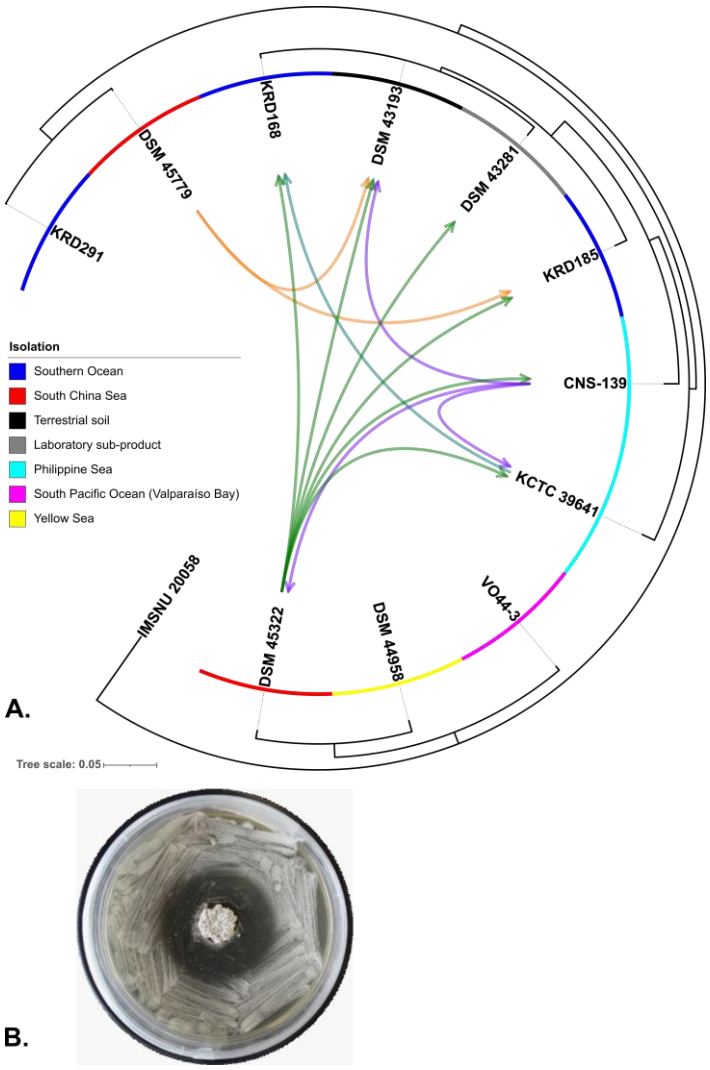


Figure 5-8 Challenge assay between *Pseudonocardia* spp. from different environments. A. Antagonist interactions across phylogeny of *Pseudonocardia* spp. A connecting line represents an antagonistic interaction of the resident strain (tail) against the intruder strain (arrow). Strains are distributed according to the maximum likelihood tree based on the 16S rRNA gene sequence. *Amycolatopsis orientalis* IMSNU 20058 was used as outgroup. The colour around the edge shows the strains' origin. B. Example of an antagonistic interaction between P. sediminis cultured as a single colony (resident strain) and *P. petroleophila* inoculated two weeks later across the whole plate (intruder strains).

Of the strains tested, phylogeny was the main factor to predict antagonist interactions. However, some results suggested that the origin of the strains could also be involved. In particular, *P. sediminis* DSM 45779 (not from the Southern Ocean) interacted against *P. oceani* KRD185 and *P. petroleophila* DSM 43193, while its closest relative *Pseudonocardia* sp. KRD291 (Southern Ocean strain) did not. This may suggest that the Southern Ocean strains may share the same ecological niche despite their phylogenetic distance. A potential explanation for the intruder assay results is the metabolomics profiles of the interacting

strains. From the classical molecular networking analysis discussed in chapter 4, a correlation between the metabolomics profile of studied strains and their phylogenetic neighbours was observed (Fig. 4-7). This means that closely related strains shared a higher number of parent ions, which in principle means that similar metabolites were detected. For example, the lack of interactions between *P. antitumoralis* DSM 45322 and their phylogenetic relatives may be explained through the high number of shared parent ions. For example, *P. antitumoralis* DSM 45322 shared 886 (70.1%) of its parent ions with *P. ammonioxydans* DSM 44958. Similarly, it shared 869 (68.8%) parent ions with *Pseudonocardia* sp. VO44-3, and 872 (69.0%) parent ions with *P. sediminis* DSM 45779 (Fig. 5-9A and Fig. S13A), and 904 (71.5%) nodes with *Pseudonocardia* KRD291 (Fig. 5-9B and Fig. S13B). On the other hand, *P. antitumoralis* shared only 386 (30.5%) and 362 (28.6%) parent ions with *P. abyssalis* KRD168 and *P. petroleophila* DSM 43193, respectively (Fig. 5-9B and Fig. S13B). In these last two cases, antagonistic interactions were observed.

Nonetheless, the metabolite distribution does not explain some of the observed antagonistic interactions. For example, *P. profundimaris* KTCT 39641 shared 684 (57.6%) parent ions with *P. abyssalis* KRD168 (Fig. 5-9C and Fig. S13C), and an antagonistic interaction was observed. Similarly, *Pseudonocardia* sp. CNS-139 shared 503 (50.8%) parent ions with *P. profundimaris* KTCT 39641, and 511 (51.6%) with *P. petroleophila* DSM 43193 (Fig. 5-9C and Fig. S13C). Still *Pseudonocardia* sp. CNS-139 inhibited the growth of both strains. On the other hand, other pairs of strains like and *Pseudonocardia* sp. CNS-139 shared a similar number of parent ions with *P. abyssalis* KRD168, 519 (52.4%), and no antagonistic interactions were observed (Fig. 5-9C and Fig. S13C). These results showed that subtle differences in the metabolomics profiles could have big consequences on bioactivity. As discussed above for the effect of GlcNAc on the bioactivity against *C. albicans*, few metabolites could be responsible for the observed activity. A similar discrepancy was observed in the case of *P. oceani* KRD185, *Pseudonocardia* sp. KRD291, and *P. sediminis* DSM 45779 mentioned above. In particular, *Pseudonocardia* sp. KRD291 shared 332 (26.0%) parent ions with *P. oceani* KRD185 nodes, and no antagonistic interactions was observed, while and *P. sediminis* DSM 45779 shared 326 (24.3%) with *P. oceani* KRD185 and showed an antagonistic interaction (Fig. 5-9D and Fig. S13D). Therefore, the similarities in the metabolomics profiles are related to the phylogeny of the strains that seems to drive antagonistic interactions, but it is not possible to use these similarities to directly predict microbial interactions in co-cultures.

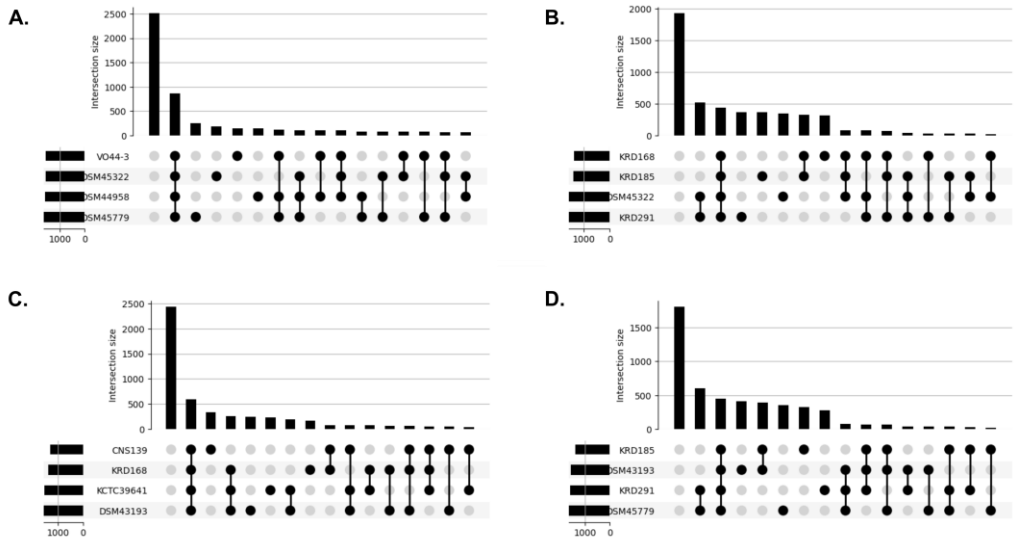


Figure 5-9 UpSet plot showing the node distribution from the GNPS classical molecular networking. Parent ions with similar MS/MS spectra are clustered in the same node and can be interpreted as metabolites. Black dots at the bottom of each vertical bar indicates the intersection, which is made up of strains that share the same nodes. The lined dots indicate that two or more strains shared the same nodes. The vertical bars indicate the number of nodes of the corresponding intersection. The total number of nodes detected in each strain are represented by horizontal bars on the left.

In general, these findings indicate that the metabolomics profiles of monocultures cannot predict interactions between the tested strains. This situation could be expected since the metabolomics profile does not distinguish between basal and specialised metabolism. For instance, it is not possible to state that the metabolites produced in the monocultures and observed in the LC-MS/MS data are also present in the Petri dish during the intruder assay, which represents the major limitation of this analysis. The metabolomic analysis of co-cultures could be interesting to investigate in future projects. Moreover, classical molecular networking does not quantify the production of metabolites, of which the specialised metabolites involved in antagonistic interactions could be produced in higher concentrations during the interaction. To address this, two approaches were applied to assess the metabolite production during the interaction between the strains. First, a mass spectrometry imaging (MSI) technique was used to detect the *in-situ* production of metabolites during the co-culture. Then, FBMN was used to enable the quantitative analysis and resolution of isomers, by including retention time and intensity data.

5.2.2.2. Mass spectrometry imaging

Using matrix-assisted laser desorption ionization (MALDI) as a mass spectrometry imaging technique, the *in-situ* metabolite production of 28 days old co-cultures of *P. abyssalis* KRD168, *P. sediminis* DSM 45779, and *Pseudonocardia* sp. KRD291, were analysed. Using this technique, a list of masses obtained from the NPLinker experiment were scanned over the co-culture plates. Through this approach, a metabolite at *m/z* 756.5568 produced by *P. abyssalis* KRD168 was identified in co-cultures with both *P. sediminis* DSM 45779 and *Pseudonocardia* sp. KRD291 (Fig. 5-10). Other metabolites of interest detected in the MALDI-MSI analysis include *m/z* 650.4431, *m/z* 734.5646, and *m/z* 748.58 (Fig. S14). Although these metabolites were also detected in the *P. abyssalis* monocultures, according to the percentage of intensity visualised, *m/z* 734.5646, *m/z* 748.5798, *m/z* 756.5568, and *m/z* 800.5526 seem to be produced in higher concentration, particularly during the co-cultures with *P. sediminis*. In principle, these findings agree with the intruder assays results that showed *P. sediminis* as a more aggressive strain in comparison to *Pseudonocardia* sp. KRD291, as they suggest that metabolites are produced in higher concentrations, which could explain the elicitation of these metabolites during an interacting co-culture. In particular, *m/z* 800.5526 was observed only when in *P. abyssalis* was co-cultured with *P. sediminis* (Fig. 5-10). Thus, despite the phylogenetic relatedness between *P. sediminis* and *Pseudonocardia* sp. KRD291, their interaction with *P. abyssalis* retrieved different results in terms of metabolites detected. This agrees with the results of the challenge assays. However, the quantitative results of the MSI must be interpreted carefully since the lumpy colony morphology characteristic of *Pseudonocardia* spp. when grown in ISP2 may affect matrix application during sample preparation for the MALDI-MSI experiment. In fact, agar rupture in the co-cultures involving *P. oceani* KRD185 made the addition of the matrix impossible during sample preparation, resulting in the exclusion of this strain from the MSI experiments. This situation could also explain the asymmetric distribution of the detected metabolites between the adjacent colonies in the co-culture, showing a higher ion intensity in the opposite direction to the interacting colony. In this regard, the future optimization of the MSI protocol, in particularly for *Pseudonocardia*, will seek to address this experimental limitation.

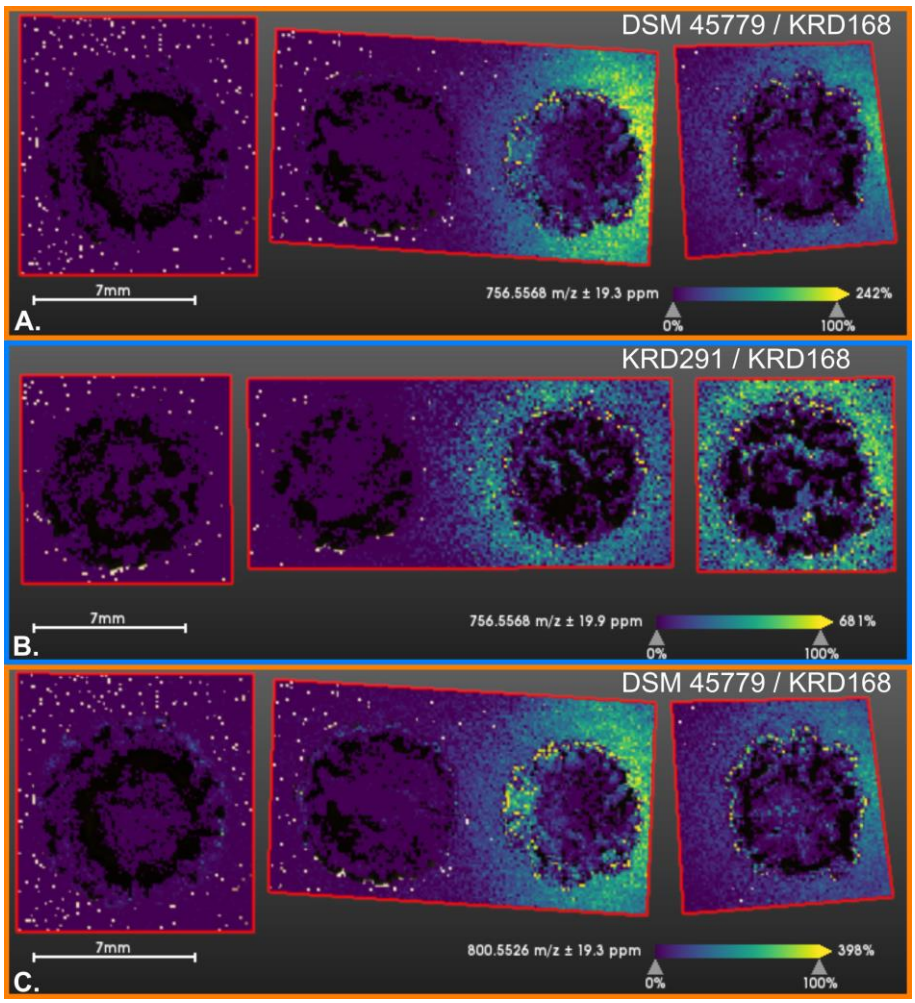


Figure 5-10 Images from MALDI-TOF MSI of *P. abyssalis* KRD168 interacting with *P. sediminis* DSM 45779 (A and C) and Pseudonocardia sp. KRD291 (B), featuring the ion image taken of m/z 756.5568 (A and B) and m/z 800.5526 (C). Monocultures of *Pseudonocardia* sp. KRD291 and *P. sediminis* DSM 45779 are shown in the left-side and *P. abyssalis* KRD168 monocultures in the right side. Each ion signal is normalized by the total ion count, and color scale represents relative intensity for each signal.

In summary, the MSI results suggested a group of metabolites produced by *P. abyssalis* KRD168 that were not produced by *P. sediminis* DSM 45779 or *Pseudonocardia* sp. KRD291. However, the production of these metabolites did not have a direct correlation with the antagonist interactions observed in the challenge experiment as *P. abyssalis* KRD168 did not show antagonistic activity against *P. sediminis* DSM 45779 or *Pseudonocardia* sp. KRD291. Yet, it is possible to propose that these metabolites might be produced as a response to the pressure generated by the aggression of other *Pseudonocardia* sp. strains, as they were observed in a higher intensity when *P. abyssalis* KRD168 was co-cultured with *P. sediminis* DSM 45779, which showed to be an aggressive strain in the challenge assays. Future efforts

should be focussed on the detection of metabolites with masses close to *m/z* 734.5646, *m/z* 748.5798, *m/z* 756.5568, and *m/z* 800.5526 in co-cultures that involve *P. oceani*.

5.2.2.3. Feature-based molecular networking

The differences in the ionization method between the LC-MS/MS and MALDI-MSI experiments make a direct comparison of both results difficult. However, FBMN was used to identify possible molecular features with precursor ions that had masses close to the metabolites observed in the MSI experiment. Later, NPLiker was used to integrate the metabolomics and genomic data, in order to propose possible BGCs involved in the interactions observed between *P. abyssalis* KRD168, *P. sediminis* DSM 45779, and *Pseudonocardia* sp. KRD291. To this end, a subset of LC-MS/MS dataset from sequenced strains used in the challenge assay was analysed by FBMN. This network comprised of 3023 ions, of which 1749 were part of a cluster, and 268 were annotated in the GNPS spectra library (including analogous ± 100.0 Da in the precursor ion) (Fig. 5-11).

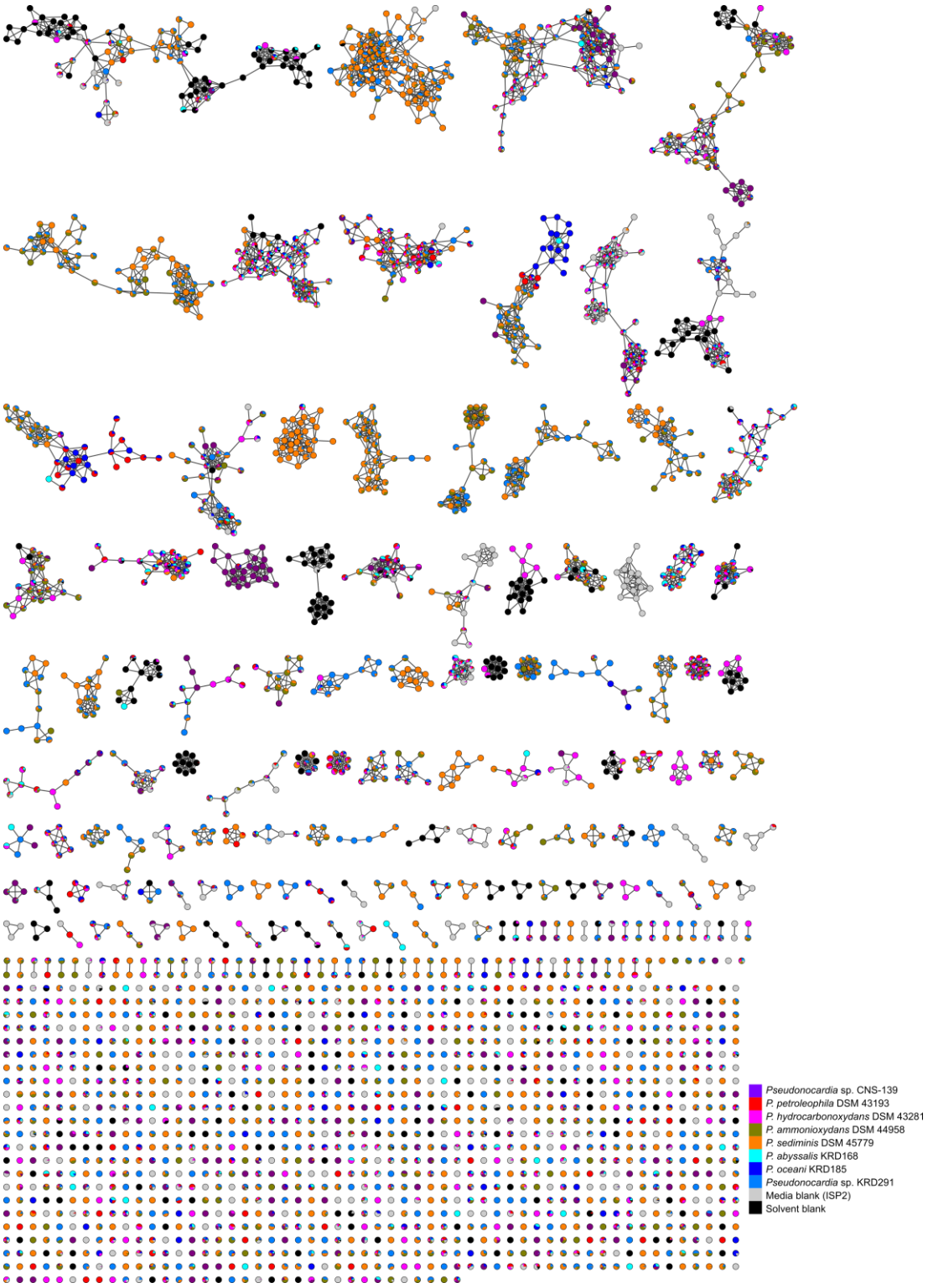


Figure 5-11 Feature-based molecular network of 3,023 features showing the relative abundance of each molecular feature in the samples. Pie chart inside each node shows the relative distribution of samples where the particular spectrum was detected, coloured according to the strain or control sample.

The search for molecular features with masses close to the ones observed in the MSI experiment was successful. A molecular family that includes a node with a mass of *m/z* 750.4981 produced by *P. abyssalis* KRD168, but not observed in *P. sediminis* DSM 45779 or *Pseudonocardia* sp. KRD291 (Fig. 5-12A) was detected. As mentioned above, a direct comparison of the ions is not possible; however, the detection of a molecular feature with a mass close to *m/z* 748.5798 and *m/z* 756.5568 in the same producer strain gives good evidence that these metabolites are likely related. The possibility of different adducts of the same metabolite could explain the mass differences observed, as will be shown later.

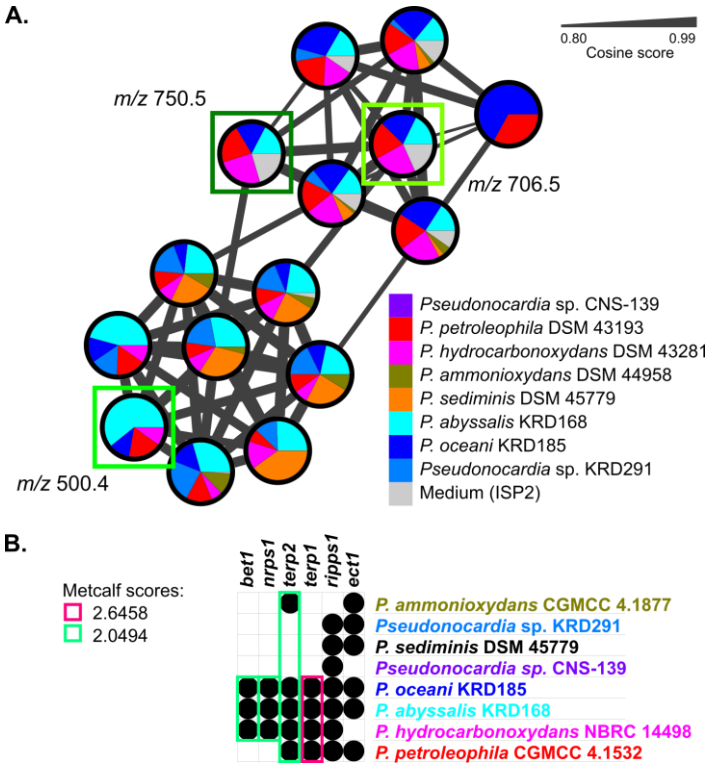


Figure 5-12 FBMN analysis of parent ions linked with NPLinker showing the relative abundance per strain of each molecular feature. A. Molecular family containing three molecular features (m/z 750.4981, m/z 500.3570 and m/z 706.4724) with a mass close to the metabolites observed in the MSI experiment. B. Linked Gene cluster families (GCFs) by NPLinker to the highlighted molecular features. BGCs are represented by nodes, and CGFs by columns.

The strain absence/presence pattern observed for the molecular feature with a mass of *m/z* 750.4981 in the FBMN analysis enabled NPLinker analysis to putatively link it with BGC candidates. Thus, this molecular feature, as well as the other two with parent ion masses of *m/z* 500.3570 and *m/z* 706.4724, were linked to the *terp1* BGC with Metcalf score of 2.6458

and to *terp2*, *nrps1*, and *bet1* with a Metcalf score of 2.0494 (Fig. 5-12B). Although this Metcalf scores were lower than the ones analysed for the classical molecular networking analysis performed in chapter 4, they represent high values for this FBMN analysis. In this way, the proposed BGCs represent good candidates for further analysis to identify the biosynthetic origin of the metabolites observed in the MSI experiment.

As discussed in chapter 4, terpene BGCs are widespread in actinomycetes, while NRPS, PKS and betalactone BGC tend to have a species-specific distribution. As such, *nrps1*, and *bet1* are more likely to play a role in the antagonistic interactions, particularly as no GCFs that include *nrps1* and *bet1* were identified in *P. sediminis* DSM 45779 nor in *Pseudonocardia* sp. KRD291. However, as the products of these BGCs have not been elucidated yet, further experiments are needed to confirm the relationship between them and the metabolites observed in the MSI experiment. Despite the fact that the BGCs proposed by NPLinker are strong candidates for further investigation, these results should be interpreted carefully as no wet-lab experiments have confirmed the proposed links. However, the NPLinker workflow facilitated the analysis of the MSI results and their relationship to the molecular family observed in the LC-MS/MS experiment. For example, the molecular feature with a mass of *m/z* 500.3570 was detected with a higher intensity in *P. abyssalis* KRD168 in the LC-MS/MS experiments (Fig. 5-10A), and it could be related mass to the metabolite with a mass of *m/z* 496.3374 which was also observed in the MSI experiment (Fig. 5-11).

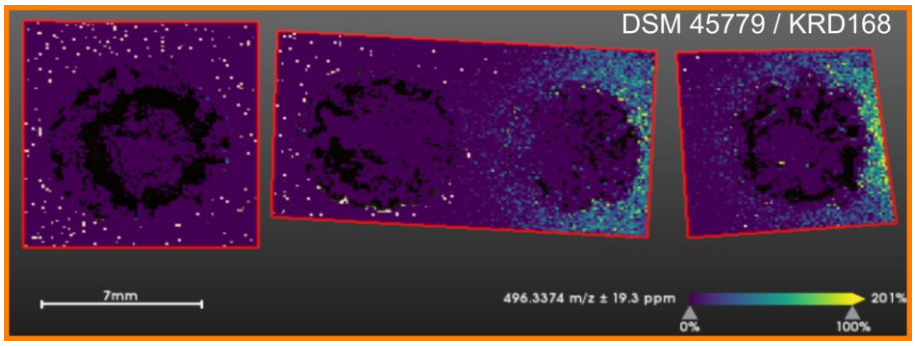


Figure 5-13 Image from MALDI-TOF MSI of *P. abyssalis* KRD168 (right) interacting with and *P. sediminis* DSM 45779 (left), featuring the ion image taken of m/z 496.3374. Each ion signal is normalized by the total ion count, and color scale represents relative intensity for each signal.

To better understand the effect of multiple ion species with different fragmentation patterns, Ion Identity Molecular Networking (IIMN) workflow was applied. This complements the

FBMN analysis by integrating a chromatographic peak shape correlation analysis into molecular networks to connect and collapse different ion species of the same metabolite. This analysis aims to identify parent ions that represent different ion species of the same metabolite. Through this, the number of "edges" (connection inside the network) was increased from 6866 to 11219 (Fig. S15). As a result, it was proposed that the molecular feature with a precursor ion of *m/z* 750.4981 corresponds to the ion $[M\text{-}H_2O+H]^+$ for which a neutral mass is *m/z* 705.5318, while the molecular feature with a precursor ion of *m/z* 706.4724 corresponds to $[M+H]^+$ with a neutral mass of *m/z* 705.4622, suggesting two ion species of the same metabolite (Fig. 5-12) (M^+ represents the molecular ion). A closer view of the mass spectra confirmed both molecular features are related to the same metabolite as the shared peaks in the MS2 spectra suggest the same fragmentation pattern (Fig. S16). Moreover, in these spectra, the peak differences of *m/z* 44 could be related to the loss of an amine due to a α-cleavage (-C2H6N), the loss of an ester via McLafferty rearrangement ($-C_2H_4O$), or the loss of a carbonyl group ($-CO_2$), while the mass difference of *m/z* 42 may be related to the loss of a propyl group ($-C_3H_6$). However, more exhaustive work on the chemistry of this metabolite is needed to interpret its MS2 spectrum. Similarly, the molecular feature with a precursor ion of *m/z* 500.3570 was related to the ion $[M+H]^+$ with a neutral mass of *m/z* 499.3534, suggesting a different metabolite origin for this ion.

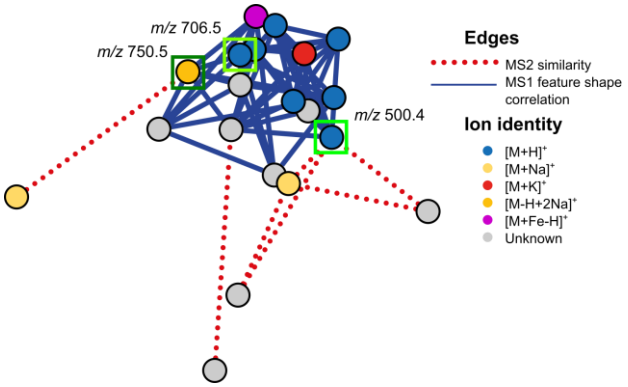


Figure 5-14 Molecular network showing a subset of nodes directly related to the three molecular features (m/z 750.4981, m/z 500.3570 and m/z 706.4724) with a mass close to the metabolites observed in the MSI experiment, as well as their ion identity.

In summary, the combination of FBMN-INN and NPLinker suggested that a metabolite with a mass near to 705.5 Da was produced by the related strains *P. abyssalis* KRD168, *P. oceani*

KRD185, *P. hydrocarbonoxydans* DSM43281, and *P. petroleophila* DMS 43193. Moreover, it is possible to suggest that the observed metabolites in the MSI experiment could be related to this metabolite or other members of its MF. However, this hypothesis must be confirmed due to two main factors. First, with the lack of chemical structural information, it is not possible to confirm that the metabolites observed in the MSI experiment correspond to ion species or adducts of any member of the analysed MF. Then, the molecular features with masses of *m/z* 750.4981 and *m/z* 706.4724 were also detected in the media blanks. This last factor is the main limitation of the analysis, and it is difficult to justify their function as a specialised metabolite. A plausible explanation of this phenomenon is cross-contamination during the sample filtering, but there is no way to test this. Another proposal is that these molecular features could be related to a component of the medium that was not metabolised by these strains but was metabolised by the rest of the strains. In this last scenario, further experiments or replicate are needed to confirm the hypothesis. Despite this major limitation, the hypothesis of a specialised metabolite produced by *P. abyssalis* KRD168 with a mass close to 705.5 Da should not be discarded as metabolites or potential adducts were observed by MSI. Otherwise, future optimised LC-MS/MS experiments could corroborate the presence of these metabolites in the extracts. For example, besides the growing variance between the experiments, it is possible that the observed metabolites in MSI were not extracted with ethyl acetate for the LC-MS/MS experiments. On the other hand, the metabolite with a mass near 499.35 Da was observed in both MSI and LC-MS/MS experiments. Moreover, this metabolite was observed with a higher intensity in the *P. abyssalis* KRD168 samples, and not detected in the media blanks. Overall, the combination of MSI, FBMN and NPLinker facilitated the analysis of large genomics and metabolomics datasets, reducing the number of interesting metabolites from more than a thousand to a small list of candidates for future work.

Finally, co-cultures involving *P. oceani* KRD185 were excluded from the MALDI-MSI measures due to agar breaking, which made the matrix addition impossible. However, future analysis that includes *P. oceani*, as well as *P. petroleophila*, will extend the present results by analysing the production of metabolites close to *m/z* 750 and *m/z* 500 in phylogenetically related strains isolated from other geographic regions.

5.3. Discussion

The observed effects of GlcNAc on *Pseudonocardia* bioactivity varied according to the species and across culture conditions. This is consistent with a previous actinomycetes study that analysed the effect of culture conditions on bioactivity against the ESKAPE pathogens which found that GlcNAc induced the antibiotic activity in some strains, while repressed it in others [493]. Similarly, the present results indicated bioactivity against *C. albicans* increased on addition of GlcNAc for *P. oceani*, while it decreased for *P. hydrocarbonoxydans* and *P. petroleophila*, and remained the same for *P. abyssalis*. Although other chemical elicitors such as sodium butyrate have been shown to increase antifungal activity against *C. albicans* from *Pseudonocardia* [350, 465], no previous reports on GlcNAc were found to compare.

Despite the preliminary results against *R. oryzae* which suggested an increase in the antifungal activity by adding the chemical elicitor GlcNAc, this was not observed against *C. albicans* or *A. niger*. On the other hand, in the absence of glucose, the eliciting effect of GlcNAc on the activity against *C. albicans* was observed only for *P. oceani*. This finding is congruent with the observation made in the *DasR* system in *Streptomyces*, where the eliciting effect of GlcNAc over the bioactivity was observed only in nutrient-depleted conditions [470]. However, in the case of the Southern Ocean *Pseudonocardia* strains, the eliciting effect of GlcNAc was observed in all cases. In fact, the production of metabolites and activity against *C. albicans* were not increased in *P. abyssalis*. Although the *DasR* system has been well described for the *Streptomyces* genus, the available information of this system in the *Pseudonocardia* genus is limited. Nevertheless, the effect of GlcNAc was described for another member of the Pseudonocardiaceae family, *Saccharopolyspora erythraea* [494], for which deletion of the *dasR* gene was shown to result in reduced aerial hyphal and antibiotic production under nutrient-rich conditions [495]. In addition, GlcNAc elicited the production of kigamicins by *Amycolatopsis alba* DSM 44262Δabm9. However, in this latter case, the effect was not directly related to the *DasR* system [496]. Therefore, as genes related to *dasR* were identified in both *P. abyssalis* and *P. oceani*, future molecular biology assays could describe this system in the Southern Ocean *Pseudonocardia* strains.

In terms of the effect of GlcNAc on the production of metabolite profile, changes were observed. However, as with the bioactivity, these changes were species dependent. Moreover, the effect on the metabolomic profiles was not directly related to the observed

effect in the bioactivity against *C. albicans*. This variable effect on the production of metabolites by GlcNAc was previously reported by Dashti *et al.* for *Micromonospora* sp. RV43, *Rhodococcus* sp. RV157, and *Actinokineospora* sp. EG49. This work went a step further, identifying the metabolites that were induced or suppressed by GlcNAc [494]. In this sense, it is possible to propose that the effect on the production of the particular metabolites with antifungal activity against *C. albicans*, did not correlate with the global changes in the metabolomics profile. For instance, the combination of FBMN and MolNetEnhancer evidenced that lipids and lipids-like metabolites, which are not typically related to antifungal activity, were the group the displayed the main changes in relation to the addition of GlcNAc. In fact, it was previously demonstrated for *P. oceani* that the bioactivity against ESKAPE pathogens is not necessarily related to the highest number of produced metabolites [262]. So, here this conclusion is expanded to the antifungal activity, as well as to the use of GlcNAc as a chemical elicitor. Since the effect of GlcNAc was highly variable and strain-dependent, the comparative metabolomics analyses discussed in Chapter 4 and in this chapter were performed in the absence of this chemical elicitor.

Despite all the available information about the chemical ecology of *Pseudonocardia* spp. related to fungus-growing ants [197, 199, 497, 498], still little is known about the interactions between free-living *Pseudonocardia* spp.. Previously, Poulsen *et al*. (2007) proved that the antagonisms between different strains of *Pseudonocardia* spp. related to fungus-growing ants is common [330]. In this works, the experiment of Poulsen was replicated using eleven free-living *Pseudonocardia* strains, of which nine were isolated from marine environments. Poulsen *et al*. [330, 499] reported several antagonistic interactions between *Pseudonocardia* strains, including between the free-living strains *P. saturnae*, *P. halophobica*, and *P. thermophilia*. In contrast, few interactions were observed between the marine-origin *Pseudonocardia* spp. strains, most of them were due to *P. antitumoralis*. However, it was also reported by Poulsen *et al*., that the relative genetic distance between pairs of strains had a significant effect on antagonistic interactions. So, in general, the ecological and biogeographical variables had a small effect on describing antagonistic interactions.

Although phylogeny was more important than geographic origin to explain the observed antagonistic interactions, *P. sediminis* and *Pseudonocardia* sp. KRD291 showed a different bioactivity profile, which implies that the strain origin may play a role in the interaction against other *Pseudonocardia* strains. In fact, none of the three species showed antagonistic

interaction between them, which suggests that for this particular group of actinomycetes, the geographical origin could have a role in the delimitation of the ecological niche. This observation was later confirmed by the MSI experiment, where *P. sediminis* affected the production of certain metabolites in *P. abyssalis* in a different way than *Pseudonocardia* sp. KRD291. These findings contrast with the observations made in soil-derived *Streptomyces* where antagonistic interactions occurred more often between co-occurring microbial populations [214]. However, the lack of information on the bioactivity of the observed metabolites in the MSI experiment is a major limitation of the present work. As closely related species are more likely to have overlapping metabolic niches [500], the lack of antagonistic interactions between the Southern Ocean strains may be related to either a sympatric speciation process by niche differentiation [501], suggesting they would not compete for the same resources. However, this hypothesis should be confirmed by the metabolic characterization of the interacting species. Furthermore, previous reports in *Streptomyces* suggest that the interactions between soil-derived strains tend to be reciprocal, which was particularly enhanced between strains isolated from the same soil sample [502]. This contrasts with the present results, as only the antagonistic interaction between *Pseudonocardia* sp. CNS-139 and *P. antitumoralis* was reciprocal. In fact, only a few strains (*P. antitumoralis*, *P. sediminis*, and *Pseudonocardia* sp. CNS-139) showed two or more antagonistic interactions against other *Pseudonocardia* strains, while most of the combinations did not display any antagonistic interaction. These findings suggest that the free-living *Pseudonocardia* spp. interact with other members of the genus less often than their symbiotic counterparts, suggesting a broader ecological niche as expected.

Using a combination of MSI with different molecular networking experiments, as well as their integration with genomics data through NPLinker [444], it was possible to propose a reduced list of relevant metabolites. In particular, a molecular family that includes the parent ions at *m/z* 750.5, *m/z* 500.4 and *m/z* 706.5, could play a role in the interactions between the Southern Ocean *Pseudonocardia* spp. strains. This strategy, excluding the NPLinker analysis, was successfully applied before to analyse the role of several known metabolites in interactions involving *Streptomyces* spp. [259], and *Salinispora* spp. [503]. However, for the present work, the chemical structure of metabolites isolated from *P. abyssalis* or *P. oceani* have not yet been elucidated, making targeted analysis difficult. Despite this, the metabolite of interest was detected with a higher intensity in interactions between geographically distant strains, suggesting that production is induced during microbial interaction. Similar

results were described in *Pseudonocardia* spp. strains related to fungus-growing ants when interacting with *Escovopsis*, where MSI allows the identification of ecologically relevant metabolites [486].

The reports of *Pseudonocardia* spp. strains in symbiosis with fungus-growing ants showed that the antagonistic interactions observed in Petri dishes do not translate into a reduction of bacteria abundance on the cuticle of sub-colony ant workers [499]. In the same way, it is possible to speculate that the antagonistic interactions observed between the marine-origin *Pseudonocardia* spp. strains will affect the microbial populations in marine sediments. The specilised metabolites detected in the Southern Ocean *Pseudonocardia* strains could have other ecological roles different, like nutrient use modulation. For instance, in *Streptomyces*, antibiotics in sub-inhibitory concentrations modulated nutrient use, minimizing the niche overlap among competitors, which was proposed as an explanation for the low occurrence of antibiotic-resistant phenotypes [504]. Hence, the antagonist interaction among the free-living *Pseudonocardia* spp. alone will not describe entirely the niche shaping in their natural environments. Furthermore, differences in the metabolic rate could affect the dynamic of the free-living *Pseudonocardia* spp. in their natural environment. For example, a similar challenge assay performed in the genus *Salinispora*, showed that the two closely related species *S. arenicola* and *S. tropica* can be differentiated based on their competitive strategies. The higher inhibitory activity occurred in the early stages for *S. arenicola*, while for *S. tropica* it occurred later, suggesting that they adapted different strategies to compete in marine sediments [503].

In summary, as this project focused on changes in metabolomic profile, a broader analysis that includes nutrient utilization is required to properly describe the ecological niche of the Southern Ocean *Pseudonocardia* strains. However, the use of chemical elicitors, as well as the challenge assay, proved to be a successful strategy to modulate the production of metabolites by *Pseudonocardia* spp. from marine origin. In particular, the combination of integrated paired omics datasets, MSI, and challenge assays allowed the identification of a group of promising metabolites whose chemical identity and biological function await to be addressed in future projects.

Chapter 6 General discussion and conclusions

6.1. Discussion

To overcome the continuous rediscovery of known strains and natural products, it has been proposed that the exploration of novel taxa and underexploited environments could lead to the discovery of novel chemical diversity [61, 99]. This work contributes in this direction, by describing three novel actinomycete lineages with antifungal activity, which were isolated from the deep Southern Ocean. Furthermore, combining genomics and metabolomics datasets, the the specialised metabolisms of these and other related strains were explored.

In the Millán-Aguiñaga *et al*. work, a relationship between the Southern Ocean *Pseudonocardia* strains' phylogeny and their geographic origin was noticed. A correlation between geographic distance and BGC distribution has been found in other studies [230, 233]. Therefore, an initial hypothesis of this work proposed that the specialised metabolism of the three Southern Ocean *Pseudonocardia* lineages could be influenced by geographic isolation location, ecological context, or both. However, the analysis performed in chapter 4 demonstrated that the BGC evolution and distribution across strains is mainly shaped by phylogeny over any biogeographical pattern. In this sense, the distribution of BGCs across the Southern Ocean *Pseudonocardia* lineages is explained through the evolutionary history of these lineages. For example, seven of the ten BGCs annotated in *Pseudonocardia* sp. KRD291 were also identified in *P. sediminis* [149] isolated from a distant marine location. Similarly, several BGCs from *P. oceani* were also identified in *P. broussonetiae* [140] which was isolated from a distant location and different ecological context. Despite the broad distribution of the majority of BGCs, few of them like *pks3* and *bet2* were observed exclusively in the Southern Ocean *Pseudonocardia* strains. This suggested they might be relevant for the local adaptation.

Although the hypothesis of geographic isolation was discarded, genetic flow barriers due to ecological differentiation in bacteria can generate patterns of genetic differentiation in populations similar to physical isolation [505]. For example, the population structure of 37 *Streptomyces olivaceus* strains from various geographic origins and ecological context (free-living vs insect-associated), which included marine-derived strains, revealed that the habitat type, rather than the geographic isolation is a strong predictor of genetic distance, suggesting

a significant role of habitat adaptation in the diversification process [506]. However, the effect of geographic isolation on the *Pseudonocardia* genus has been observed before. For example, the analysis of 29 fungus-growing ant-associated *Pseudonocardia* strains from Panama and Costa Rica revealed that populations from a specific location, the Barro Colorado Island, proved to be locally adapted, harbouring diverse and unique BGCs despite low genetic diversity. This diversity on BGCs was not observed in other populations. In particular, from a total of 93 GCFs, 10 unique GCFs were described in the Barro Colorado strains [238]. Although the reduced number of sequenced strains from the Southern Ocean may hide a higher BGC diversity, the results of this work confirmed that the genetic divergence of free-living Pseudonocardia strains from distant geographic locations is lower than what has been observed in insect-associated strains. However, it is possible to hypothesise that the inclusion of more Southern Ocean *Pseudonocardia* strains from different lineages could retrieve other BGCs such *pks3* and *bet2* with a potential role in the local adaptation. Therefore, analysing the potential ecological role of the Southern Ocean *Pseudonocardia* strains is fundamental to understand the evolution of their specialised metabolism.

To assess the potential role of the specialised metabolites in the interactions amongst the Southern Ocean *Pseudonocardia* strains and *Pseudonocardia* strains from different marine environments, a co-culture challenge bioassay was performed. In contrast with the results obtained by Poulsen *et al*. (2007) which demonstrated that antagonistic interactions between fungus-growing ant-associated *Pseudonocardia* strains are common [330], interactions between free-living strains from marine origin were restricted to some strains like *P. antitumoralis*, *P. sediminis*, and *Pseudonocardia* sp. CNS-139. The experiment of Poulsen *et al*. was later extended with sub-colony experiments using *Acromyrmex* leaf-cutting ants to test whether the antagonistic interactions observed in plates affected ant colonies *in vivo*, showing that *in vitro* inter-strain antagonism does not affect the abundance of any of the combined *Pseudonocardia* strains on the cuticle of sub-colony workers [499]. These results illustrate how difficult it can be to extrapolate microbial experiments performed in plates to the real ecological context. However, the results of the present work demonstrated that *P. oceani* and *P. abyssalis* are two phylogenetically close lineages isolated from a similar environment, which did not show antagonistic interactions between them. So, it is possible to propose that members of these lineages might not compete for the same ecological niche, which opens the possibility of a sympatric speciation process. However, to confirm this theory further experiments that involve other variables such as nutrient

requirements, as well as more complex set-ups (e.g. sediments microcosm experiments), are needed.

Finally, the MSI experiment facilitated the interpretation of the interaction experiments, as well as their correlation to the multi-omics analysis performed in chapter 4. For example, a group of parent ions produced by *P. abyssalis* KRD168 were identified as possible metabolites involved in the inter-species interactions. A similar approach was recently applied to identify putative metabolites involved in the antagonistic interactions between fungus-growing ant-related *Pseudonocardia* strains and the entomopathogenic fungus *Metarhizium anisopliae* [143]. This work illustrates that combining genomics and metabolomics techniques to answer evolutionary and ecological questions facilitate the exploration of the chemical diversity of actinomycetes.

6.2. Future work

According to 16S rRNA gene sequence analysis, *P. seranimata* [133] was also identified as part of the clade of which *P. abyssalis* and *P. oceani* belong. However, the lack of a publicly available whole-genome sequence did not allow inclusion of it into the BGC analysis. For future work, to perform a comparative analysis that includes strains such as *P. broussonetiae* and *P. seranimata*, isolated from a plant host, would allow a better understanding of the effect of ecological variables over function and regulation of the specialised metabolism of these lineages. Furthermore, the evolutionary analysis performed in chapters 3 and 4 did not give any reasons to think that *P. abyssalis* and *P. oceani* lineages are geographically or ecologically restricted to the Southern Ocean. So, it is expected that other strains of these species or other related species will be recovered from other locations and environments. Hence, the results of this work are probably just the beginning of the description of many other related lineages that might complete our understanding of the evolution of the *Pseudonocardia* genus in terms of adaption to marine environments. Future isolation efforts that include molecular typing techniques such as BOX-PCR (BOX-A1R-based repetitive extragenic palindromic-PCR) may help to achieve this [507]. This technique was already applied in the recent isolation and description of *P. nigra* from hyper-arid desert soil [118]. Furthermore, the formal taxonomic description of *Pseudonocardia* sp. KRD291 remains a future task.

Although the scope of this work was not focused on microbial ecology, the analysis of the BGCs distribution across *Pseudonocardia* strains from different geographic locations opens the door to many questions on population dynamics. The recombination analysis performed in chapter 3 showed that the genetic barriers among *P. abyssalis*, *P. oceani*, and other phylogenetically related species from different geographic locations are permissive. However, in chapter 4 it was shown that recombination events on BGCs are more common than in the core-genome, suggesting a lower diversification. Then, the application of the fixation index (F_{ST}), a measure of population differentiation due to genetic structure, could help to better understand the role of genetic diversity, particularly from BGCs, in geographically structured populations [508]. This population genetic approach has been applied before to the analysis of the local adaptation of *Pseudonocardia* [238] and *Streptomyces* [506] populations. So, in addition to future isolation efforts of *Pseudonocardia* strains from diverse ecological and geographical origins, a population genetics analysis of these strains is the natural continuation of the present work. To this end, Central America is an ideal scenario to set up this future project as it contains a tropical forest that is the home of the majority of leafcutter ants species [509], as well as two oceans isolated by the Central American Isthmus [510]. This offers the opportunity to isolate free-living and insect-associated *Pseudonocardia* strains from terrestrial and marine environments. Then, the genetic population analysis focussed on the specialised metabolisms will help to understand the ecological processes and evolutionary mechanisms that shape the distribution of BGCs. In this future project, the Southern Ocean strains could play a relevant role as a population from a distant location and different ecological context.

In terms of the bioactive metabolite produced by the Southern Ocean *Pseudonocardia* strains, future isolation efforts should focus on the characterisation of their chemical nature. The integration of fraction libraries and bioassays into the GNPS workflow has been proposed through the concept of 'bioactive molecular networking' to identify candidate bioactive compounds directly from fractionated bioactive extracts [511]. In this sense, extending the NPLinker framework to link bioactive molecular networks to BGCs will potentially accelerate the discovery of antifungal natural products from *P. oceani* and *P. abyssalis*. Likewise, future projects in natural products chemistry should be related to elucidating the chemical nature of the diffusible metabolites observed in the *P. abyssalis* cultures by MSI. Due to experimental difficulties related to the colony morphology of *P. oceani*, it was not possible to identify metabolites produced by this species. So, optimization of the MSI protocol will

allow identified relevant metabolites produced *in situ* in the *P. oceani* cultures in future projects.

The antibacterial activity against ESKAPE pathogens and the metabolomics profiles of the Southern Ocean *Pseudonocardia* strains were previously studied by Soldatou *et al.* (2021), identifying activity against *S. aureus* by *P. oceani* KRD185 [262]. In the present work, the integration of both datasets with NPLinker allowed selecting parent ion and BGCs for future efforts on natural products isolation. Unfortunately, experiments to identify metabolites responsible for the observed antifungal activity were not performed. Although prior work in fungus-growing ant-related strains [350] and *P. autotrophica* [512] correlated polyketide compounds to their antifungal activity, no BGCs related to this type of compounds were detected. So, the metabolite or metabolites responsible for the antifungal activity observed in *P. abyssalis* and *P. oceani* could, in theory, be different from the ones observed in the fungus-growing ant-related strains. Previous work by Cafaro *et al*. (2011) on *Pseudonocardia* showed that the antifungal activity of ant-associated strains against the parasitic fungus *Escovopsis* is higher than free-living strains, while the activity against non-*Escovopsis* fungi was similar in both groups [457]. Then, including the analysis of antifungal activity against a broader fungal collection in the proposed project with Central America *Pseudonocardia* strains, might expand our understanding of antifungal metabolite production in this genus, beyond the well-documented case of the insect-associated strains.

6.3. Conclusions

In summary, the three Southern Ocean *Pseudonocardia* lineages were delimited through comparative genomics with the formal description of two novel species. As a result, the original hypothesis that suggested three monophyletic groups was confirmed. Then, the specialised metabolisms of these three lineages were explored using an integrated multi-omics approach, characterising them as producers of bioactive metabolites. Questions about the chemical identity of these metabolites and their ecological role remain open. However, this work demonstrated the potential of these metabolites as a source of novel antimicrobial compounds. Furthermore, it was shown that phylogeny further explained the distribution of BGCs and metabolomics profiles. Finally, the original proposal of a geographic isolation effect in the evolution of the specialised metabolisms of the Southern Ocean *Pseudonocardia*

strains was discarded. However, the description of genomic and ecological variables that might influence the production of specialised metabolites provides the basis for future projects on their evolutions and ecology.

References

1. **Waksman SA**. Antagonistic relations of microörganisms. *Bacteriol Rev* 1941;5:231–291.

2. **Waksman SA**. What is an Antibiotic or an Antibiotic Substance? *Mycologia* 1947;39:565–569.

3. **Nicolaou KC, Rigol S**. A brief history of antibiotics and select advances in their synthesis. *J Antibiot (Tokyo)* 2018;71:153–184.

4. **Kohanski MA, Dwyer DJ, Collins JJ**. How antibiotics kill bacteria: From targets to networks. *Nat Rev Microbiol* 2010;8:423–435.

5. **Tenson T, Lovmar M, Ehrenberg M**. The mechanism of action of macrolides, lincosamides and streptogramin B reveals the nascent peptide exit path in the ribosome. *J Mol Biol* 2003;330:1005–1014.

6. **Brown D**. Antibiotic resistance breakers: Can repurposed drugs fill the antibiotic discovery void? *Nat Rev Drug Discov* 2015;14:821–832.

7. **Odds FC, Brown AJP, Gow NAR**. Antifungal agents: Mechanisms of action. *Trends Microbiol* 2003;11:272–279.

8. **Ostrosky-Zeichner L, Casadevall A, Galgiani JN, Odds FC, Rex JH**. An insight into the antifungal pipeline: Selected new molecules and beyond. *Nat Rev Drug Discov* 2010;9:719–727.

9. **Bolard J**. How do the polyene macrolide antibiotics affect the cellular membrane properties? *Biochim Biophys Acta* 1986;864:257–304.

10. **Zaffiri L, Gardner J, Toledo-Pereyra LH**. History of Antibiotics. From Salvarsan to Cephalosporins. *J Investig Surg* 2012;25:67–77.

11. **Austrian R, Gold J**. Pneumococcal Bacteremia with Especial Reference to Bacteremic Pneumococcal Pneumonia. *Ann Intern Med* 1964;60:759.

12. **Fleming A**. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzae. *Br J Exp Pathol* 1929;10:226–236.

13. **Laxminarayan R, Sridhar D, Blaser M, Wang M, Woolhouse M**. Achieving global targets for antimicrobial resistance. *Science (80-)* 2016;353:874–875.

14. **Berendonk TU, Manaia CM, Merlin C, Fatta-Kassinos D, Cytryn E, *et al.*** Tackling antibiotic resistance: The environmental framework. *Nat Rev Microbiol* 2015;13:310–317.

15. **Hernando-Amado S, Coque TM, Baquero F, Martínez JL**. Defining and combating antibiotic resistance from One Health and Global Health perspectives. *Nat Microbiol* 2019;4:1432–1442.

16. **Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Robles Aguilar G, *et al.*** Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*

2022;399:629–655.

17. **O'Neill J**. *Tackling drug-resistant infections globally: final report and recommendations*. 2016.

18. **Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, *et al.*** Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect Dis* 2018;18:318–327.

19. **Nnadi NE, Carter DA**. Climate change and the emergence of fungal pathogens. *PLoS Pathog* 2021;17:1–6.

20. **Casadevall A, Kontoyiannis DP, Robert V**. On the Emergence of Candida auris: Climate Change, Azoles, Swamps, and Birds. *MBio* 2019;10:837–850.

21. **Chowdhary A, Sharma C, Meis JF**. Candida auris: A rapidly emerging cause of hospital-acquired multidrug-resistant fungal infections globally. *PLoS Pathog* 2017;13:1–10.

22. **Spivak ES, Hanson KE**. Candida auris: An emerging fungal pathogen. *J Clin Microbiol*;56. Epub ahead of print 2018. DOI: 10.1128/JCM.01588-17.

23. **Aminov RI**. A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future. *Front Microbiol* 2010;1:134.

24. **Ehrlich P, Hata S**. *Die experimentelle Chemotherapie der Spirillosen*. Berlin, Heidelberg: Springer Berlin Heidelberg; 1910. Epub ahead of print 1910. DOI: 10.1007/978-3-642-64926-4.

25. **Bentley R**. Mycophenolic acid: A one hundred year odyssey from antibiotic to immunosuppressant. *Chem Rev* 2000;100:3801–3825.

26. **Gosio B**. Ricerche batteriologiche e chimiche sulle alterazioni del mais. Contributo all'eziologia della pellagra. *Riv d'Igiene e Sanità Pubblica* 1896;7:869–888.

27. **Domagk G**. Ein Beitrag zur Chemotherapie der bakteriellen Infektionen. *DMW - Dtsch Medizinische Wochenschrift* 1935;61:250–253.

28. **Chain E, Florey HW, Adelaide MB, Gardner AD, Oxfd DM, *et al.*** Penicillin As a Chemotherapeutic Agent. *Lancet* 1940;236:226–228.

29. **Abraham EP, Chain E, Fletcher CM, Gardner AD, Heatley NG, *et al.*** Further Observations on Penicillin. *Lancet* 1941;238:177–189.

30. **Lewis K**. Platforms for antibiotic discovery. *Nat Rev Drug Discov* 2013;12:371–387.

31. **Hutchings M, Truman A, Wilkinson B**. Antibiotics: past, present and future. *Curr Opin Microbiol* 2019;51:72–80.

32. **Schatz A, Bugle E, Waksman SA**. Streptomycin, a Substance Exhibiting Antibiotic Activity Against Gram-Positive and Gram-Negative Bacteria. *Exp Biol Med* 1944;55:66–69.

33. **Waksman SA, Lechevalier HA**. Neomycin, a new antibiotic active against streptomycin-resistant bacteria, including tuberculosis organisms. *Science (80-)* 1949;109:305–307.

34. **Waksman SA, Woodruff HB**. Bacteriostatic and Bactericidal Substances Produced by a Soil Actinomyces. *Exp Biol Med* 1940;45:609–614.

35. **Lewis K**. Recover the lost art of drug discovery. *Nature* 2012;485:439–440.

36. **Kolter R, van Wezel GP**. Goodbye to brute force in antibiotic discovery? *Nat Microbiol* 2016;1:15020.

37. **Gould K**. Antibiotics: From prehistory to the present day. *J Antimicrob Chemother* 2016;71:572–575.

38. **Li D, She X, Calderone R**. The antifungal pipeline: the need is established. Are there new compounds? *FEMS Yeast Res* 2020;20:1–8.

39. **Oxford AE, Raistrick H, Simonart P**. Studies in the biochemistry of micro-organisms. *Biochem J* 1939;33:240–248.

40. **Petersen AB, Rønnest MH, Larsen TO, Clausen MH**. The chemistry of griseofulvin. *Chem Rev* 2014;114:12088–12107.

41. **Williams DI, Marten RH, Sarkany I**. Oral Treatment of Ringworm With Griseofulvin. *Lancet* 1958;272:1212–1213.

42. **Hazen EL, Brown R**. Fungicidin, an Antibiotic Produced by a Soil Actinomycete. *Exp Biol Med* 1951;76:93–97.

43. **Dutcher JD**. The Discovery and Development of Amphotericin B. *Dis Chest* 1968;54:296–298.

44. **Brown ED, Wright GD**. Antibacterial drug discovery in the resistance era. *Nature* 2016;529:336–343.

45. **Godefroi EF, Heeres J, van Cutsem J, Janssen PAJ**. The Preparation and Antimycotic Properties of Derivatives of 1-Phenethylimidazole. *J Med Chem* 1969;12:784–791.

46. **Heard SC, Wu G, Winter JM**. Antifungal natural products. *Curr Opin Biotechnol* 2021;69:232–241.

47. **Sorokina M, Steinbeck C**. Review on natural products databases: Where to find data in 2020. *J Cheminform* 2020;12:1–51.

48. **Newman DJ, Cragg GM**. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J Nat Prod* 2020;83:770–803.

49. **Lewis K**. The Science of Antibiotic Discovery. *Cell* 2020;181:29–45.

50. **Hoffman PS**. Antibacterial Discovery: 21st Century Challenges. *Antibiotics* 2020;9:213.

51. **Theuretzbacher U, Bush K, Harbarth S, Paul M, Rex JH, *et al.*** Critical analysis of antibacterial agents in clinical development. *Nat Rev Microbiol* 2020;18:286–298.

52. **Baltz RH**. Natural product drug discovery in the genomic era: realities, conjectures, misconceptions, and opportunities. *J Ind Microbiol Biotechnol* 2019;46:281–299.

53. **Wright GD**. Opportunities for natural products in 21stcentury antibiotic discovery. *Nat Prod Rep* 2017;34:694–701.

54. **Harvey AL, Edrada-Ebel R, Quinn RJ**. The re-emergence of natural products for drug discovery in the genomics era. *Nat Rev Drug Discov* 2015;14:111–129.

55. **Butler MS, Paterson DL**. Antibiotics in the clinical pipeline in October 2019. *J Antibiot (Tokyo)* 2020;73:329–364.

56. **Debono M, Barnhart M, Carrell CB, Hoffmann JA, Occolowitz JL, *et al.*** A21978C, a complex of new acidic peptide antibiotics: Isolation, chemistry, and mass spectral structure elucidation. *J Antibiot (Tokyo)* 1987;40:761–777.

57. **Debono M, Abbott BJ, Molloy RM, Fukuda D, Hunt AH, *et al.*** Enzymatic and chemical modifications of lipopeptide antibiotic A21978C: The synthesis and evaluation of daptomycin (LY146032). *J Antibiot (Tokyo)* 1988;41:1093–1105.

58. **Hochlowski JE, Swanson SJ, Ranfranz LM, Whittern DN, Buko AM, *et al.*** Tiacumicins, a novel complex of 18-membered macrolides. II. Isolation and structure determination. *J Antibiot (Tokyo)* 1987;40:575–88.

59. **Bérdy J**. Thoughts and facts about antibiotics: Where we are now and where we are heading. *J Antibiot (Tokyo)* 2012;65:441–441.

60. **Hug J, Bader C, Remškar M, Cirnski K, Müller R**. Concepts and Methods to Access Novel Antibiotics from Actinomycetes. *Antibiotics* 2018;7:44.

61. **Devine R, Hutchings MI, Holmes NA**. Future directions for the discovery of antibiotics from actinomycete bacteria. *Emerg Top Life Sci* 2017;1:1–12.

62. **Schinke C, Martins T, Queiroz SCN, Melo IS, Reyes FGR**. Antibacterial Compounds from Marine Bacteria, 2010-2015. *J Nat Prod* 2017;80:1215–1228.

63. **Ramirez-Llodra E, Brandt A, Danovaro R, De Mol B, Escobar E, *et al.*** Deep, diverse and definitely different: Unique attributes of the world's largest ecosystem. *Biogeosciences* 2010;7:2851–2899.

64. **Jiménez C**. Marine Natural Products in Medicinal Chemistry. *ACS Med Chem Lett* 2018;9:959–961.

65. Marine Pharmaceuticals: The Clinical Pipeline. https://www.marinepharmacology.org/ (accessed 10 November 2021).

66. **Towle MJ, Salvato KA, Budrow J, Wels BF, Kuznetsov G, *et al.*** In vitro and in vivo anticancer activities of synthetic macrocyclic ketone analogues of halichondrin B. *Cancer Res* 2001;61:1013–1021.

67. **Hirata Y, Uemura D**. Halichondrins - antitumor polyether macrolides from a marine sponge. *Pure Appl Chem* 1986;58:701–710.

68. **Rinehart KL, Holt TG, Fregeau NL, Stroh JG, Keifer PA, *et al.*** Ecteinascidins 729, 743, 745, 759A, 759B, and 770: potent antitumor agents from the Caribbean tunicate Ecteinascidia turbinata. *J Org Chem* 1990;55:4512–4515.

69. **Newman DJ, Cragg GM**. Drugs and drug candidates from marine sources: an assessment of the current 'state of play'. *Planta Med* 2016;82:775–789.

70. **Gerwick WH, Moore BS**. Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. *Chem Biol* 2012;19:85–98.

71. **Tachibana K, Scheuer PJ, Engen D Van, Clardy J**. Okadaic Acid , a Cytotoxic Polyether from Two Marine Sponges of the Genus Halichondria. *J Am Chem Soc* 1981;103:2469–2471.

72. **Yasumoto T, Seino N, Murakami Y, Murata M**. Toxins produced by benthic dinoflagellates. *Biol Bull* 1987;172:128–131.

73. **Schmitz FJ, Vanderah DJ, Hollenbeak KH, Enwall CEL, Gopichand Y, *et al.*** Metabolites from the marine sponge Tedania ignis . A new atisanediol and several known diketopiperazines. *J Org Chem* 1983;48:3941–3945.

74. **Stierle AC, Cardellina JH, Singleton FL**. A marine Micrococcus produces metabolites ascribed to the sponge Tedania ignis. *Experientia* 1988;44:1021.

75. **Takahashi K, Kubo A**. New Antibiotics, Saframycins A, B, C, D and E. *J Antibiot (Tokyo)* 1977;30:1015–1018.

76. **Irschikn H, Gerth K, Reichenbach H, Trowitzsch-Kienast W, Höfle G**. Saframycin mx1, a new natural saframycin isolated from a myxobacterium. *J Antibiot (Tokyo)* 1988;41:993–998.

77. **Ikeda Y, Idemoto H, Hirayama F, Yamamoto K, Iwao K, *et al.*** Safracins, new antitumor antibiotics. I. Producing organism, fermentation and isolation. *J Antibiot (Tokyo)* 1983;36:1279–83.

78. **Rath CM, Janto B, Earl J, Ahmed A, Hu FZ, *et al.*** Meta-omic characterization of the marine invertebrate microbial consortium that produces the chemotherapeutic natural product ET-743. 2011;743:1244–1256.

79. **Schofield MM, Jain S, Porat D, Dick GJ, Sherman DH**. Identification and analysis of the bacterial endosymbiont specialized for production of the chemotherapeutic natural product ET-743. *Environ Microbiol* 2015;17:3964–3975.

80. **Sakai R, Higa T, Kashman Y**. Misakinolide-A, an Antitumor Macrolide from the Marine Sponge Theonella sp. *Chem Lett* 1986;15:1499–1502.

81. **Ueoka R, Uria AR, Reiter S, Mori T, Karbaum P, *et al.*** Metabolic and evolutionary origin of actin-binding polyketides from diverse organisms. *Nat Chem Biol* 2015;11:705–712.

82. **Carroll AR, Copp BR, Davis RA, Keyzers RA, Prinsep MR**. Marine natural products. *Nat Prod Rep* 2021;38:362–413.

83. **Bull AT, Stach JEM**. Marine actinobacteria: new opportunities for natural product search and discovery. *Trends Microbiol* 2007;15:491–499.

84. **Fenical W, Jensen PR**. Developing a new resource for drug discovery: Marine actinomycete bacteria. *Nat Chem Biol* 2006;2:666–673.

85. **Subramani R, Sipkema D**. Marine rare actinomycetes: A promising source of structurally diverse and unique novel natural products. *Mar Drugs* 2019;17:249.

86. **Tortorella E, Tedesco P, Esposito FP, January GG, Fani R, *et al.*** Antibiotics from deep-sea microorganisms: Current discoveries and perspectives. *Mar Drugs* 2018;16:1–16.

87. **Kamjam M, Sivalingam P, Deng Z, Hong K**. Deep Sea Actinomycetes and Their Secondary Metabolites. *Front Microbiol*;8. Epub ahead of print 1 May 2017. DOI: 10.3389/fmicb.2017.00760.

88. **Riedlinger J, Reicke A, Zähner H, Krismer B, Bull AT, *et al.*** Abyssomicins, inhibitors of the para-aminobenzoic acid pathway produced by the marine Verrucosispora strain AB-18-032. *J Antibiot (Tokyo)* 2004;57:271–279.

89. **Bister B, Bischoff D, Ströbele M, Riedlinger J, Reicke A, *et al.*** Abyssomicin C—A Polycyclic Antibiotic from a MarineVerrucosispora Strain as an Inhibitor of thep-Aminobenzoic Acid/Tetrahydrofolate Biosynthesis Pathway. *Angew Chemie Int Ed* 2004;43:2574–2576.

90. **Abdel-Mageed WM, Milne BF, Wagner M, Schumacher M, Sandor P, *et al.*** Dermacozines, a new phenazine family from deep-sea dermacocci isolated from a Mariana Trench sediment. *Org Biomol Chem* 2010;8:2352–2362.

91. **Wagner M, Abdel-Mageed WM, Ebel R, Bull AT, Goodfellow M, *et al.*** Dermacozines H-J isolated from a deep-sea strain of Dermacoccus abyssi from Mariana Trench sediments. *J Nat Prod* 2014;77:416–420.

92. **Zhang F, Zhao M, Braun DR, Ericksen SS, Piotrowski JS, *et al.*** A marine microbiome antifungal targets urgent-threat drug-resistant fungi. *Science (80-)* 2020;370:974–978.

93. **Feling RH, Buchanan GO, Mincer TJ, Kauffman CA, Jensen PR, *et al.*** Salinosporamide A: A Highly Cytotoxic Proteasome Inhibitor from a Novel Microbial Source, a Marine Bacterium of the New Genus Salinospora. *Angew Chemie Int Ed* 2003;42:355–357.

94. **Boccellato C, Kolbe E, Peters N, Juric V, Fullstone G, *et al.*** Marizomib sensitizes primary glioma cells to apoptosis induced by a latest-generation TRAIL receptor agonist. *Cell Death Dis* 2021;12:647.

95. **Braña AF, Sarmiento-Vizcaíno A, Pérez-Victoria I, Otero L, Fernández J, *et al.*** Branimycins B and C, Antibiotics Produced by the Abyssal Actinobacterium Pseudonocardia carboxydivorans M-227. *J Nat Prod* 2017;80:569–573.

96. **Li S, Tian X, Niu S, Zhang W, Chen Y, *et al.*** Pseudonocardians A-C, new diazaanthraquinone derivatives from a deap-sea actinomycete Pseudonocardia sp. SCSIO 01299. *Mar Drugs* 2011;9:1428–1439.

97. **Ye X, Anjum K, Song T, Wang W, Yu S, *et al.*** A new curvularin glycoside and its cytotoxic and antibacterial analogues from marine actinomycete Pseudonocardia sp. HS7. *Nat Prod Res* 2016;30:1156–1161.

98. **Zhang X-M, Zhang D-F, Li W-J, Lu C-H**. Pseudonocardides A - G, New γ - Butyrolactones from Marine-derived Pseudonocardia sp. YIM M13669. *Helv Chim Acta* 2016;99:191–196.

99. **Genilloud O**. Actinomycetes: Still a source of novel antibiotics. *Nat Prod Rep* 2017;34:1203–1232.

100. **Barka EA, Vatsa P, Sanchez L, Gaveau-Vaillant N, Jacquard C, *et al.*** Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiol Mol Biol Rev*

2016;80:1–44.

101. **Goodfellow M**. Actinobacteria phyl. nov. In: *Bergey's Manual of Systematics of Archaea and Bacteria*. Wiley; 2015. pp. 1–2.

102. **Oren A, Garrity GM**. Valid publication of the names of forty-two phyla of prokaryotes. *Int J Syst Evol Microbiol*;71. Epub ahead of print 25 October 2021. DOI: 10.1099/ijsem.0.005056.

103. **Smith I**. Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence. *Clin Microbiol Rev* 2003;16:463–496.

104. **Mcneil MM, Brown JM**. The medically important aerobic actinomycetes: Epidemiology and microbiology. *Clin Microbiol Rev* 1994;7:357–417.

105. **Milani C, Mangifesta M, Mancabelli L, Lugli GA, James K, *et al.*** Unveiling bifidobacterial biogeography across the mammalian branch of the tree of life. *ISME J* 2017;11:2834–2847.

106. **Reddy GSN, Prakash JSS, Srinivas R, Matsumoto GI, Shivaji S**. Leifsonia rubra sp. nov. and Leifsonia aurea sp. nov., psychrophiles from a pond in Antarctica. *Int J Syst Evol Microbiol* 2003;53:977–984.

107. **Davis MJ, Gillaspie AG, Harris RW, Lawson RH**. Ratoon stunting disease of sugarcane: Isolation of the causal bacterium. *Science (80-)* 1980;210:1365–1367.

108. **van Santen JA, Poynton EF, Iskakova D, McMann E, Alsup TA, *et al.*** The Natural Products Atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Res* 2022;50:D1317–D1323.

109. Natural Products Atlas. https://www.npatlas.org/ (2022, accessed 9 January 2022).

110. **Dhakal D, Pokhrel AR, Shrestha B, Sohng JK**. Marine rare actinobacteria: Isolation, characterization, and strategies for harnessing bioactive compounds. *Front Microbiol* 2017;8:1106.

111. **Bérdy J**. Bioactive microbial metabolites. *J Antibiot (Tokyo)* 2005;58:1–26.

112. **Lazzarini A, Cavaletti L, Toppo G, Marinelli F**. Rare genera of actinomycetes as potential producers of new antibiotics. *Antonie Van Leeuwenhoek* 2000;78:399–405.

113. **Tiwari K, Gupta RK**. Diversity and isolation of rare actinomycetes: An overview. *Crit Rev Microbiol* 2013;39:256–294.

114. **Subramani R, Aalbersberg W**. Culturable rare Actinomycetes: Diversity, isolation and marine natural product discovery. *Appl Microbiol Biotechnol* 2013;97:9291–9321.

115. **Pathom-aree W, Stach JEM, Ward AC, Horikoshi K, Bull AT, *et al.*** Diversity of actinomycetes isolated from Challenger Deep sediment (10,898 m) from the Mariana Trench. *Extremophiles* 2006;10:181–189.

116. **Henssen A**. Beiträge zur Morphologie und Systematik der thermophilen Actinomyceten. *Arch Mikrobiol* 1957;26:373–414.

117. **Hirsch P, Engel H**. Über oligocarbophile Actinomyceten. *Ber Dtsch Bot Ges*

1956;69:441–454.

118. **Trujillo ME, Idris H, Riesco R, Nouioui I, Igual JM, *et al.*** Pseudonocardia nigra sp. nov., isolated from atacama desert rock. *Int J Syst Evol Microbiol* 2017;67:2980–2985.

119. **Nolof G, Hirsch P**. Nocardia hydrocarbonoxydans n. spec., ein oligocarbophiler Actinomycete. *Arch Mikrobiol* 1962;44:266–277.

120. **Prabahar V, Dube S, Reddy GSN, Shivaji S**. Pseudonocardia antarctica sp. nov. an Actinomycetes from McMurdo Dry Valleys, Antarctica. *Syst Appl Microbiol* 2004;27:66–71.

121. **Thiemer B, Andreesen JR, Schräder T**. Cloning and characterization of a gene cluster involved in tetrahydrofuran degradation in Pseudonocardia sp. strain K1. *Arch Microbiol* 2003;179:266–277.

122. **Vainberg S, McClay K, Masuda H, Root D, Condee C, *et al.*** Biodegradation of ether pollutants by Pseudonocardia sp. strain ENV478. *Appl Environ Microbiol* 2006;72:5218–5224.

123. **Chen SC, Duan GL, Ding K, Huang FY, Zhu YG**. DNA stable-isotope probing identifies uncultivated members of Pseudonocardia associated with biodegradation of pyrene in agricultural soil. *FEMS Microbiol Ecol* 2018;94:1–10.

124. **Parales RE, Adamus JE, White N, May HD**. Degradation of 1,4-dioxane by an actinomycete in pure culture. *Appl Environ Microbiol* 1994;60:4527–4530.

125. **Mahendra S, Alvarez-Cohen L**. Pseudonocardia dioxanivorans sp. nov., a novel actinomycete that grows on 1,4-dioxane. *Int J Syst Evol Microbiol* 2005;55:593–598.

126. **Sales CM, Mahendra S, Grostern A, Parales RE, Goodwin LA, *et al.*** Genome sequence of the 1,4-dioxane-degrading Pseudonocardia dioxanivorans strain CB1190. *J Bacteriol* 2011;193:4549–4550.

127. **Kämpfer P, Kohlweyer U, Thiemer B, Andreesen JR**. Pseudonocardia tetrahydrofuranoxydans sp. nov. *Int J Syst Evol Microbiol* 2006;56:1535–1538.

128. **Kämpfer P, Kroppenstedt RM**. Pseudonocardia benzenivorans sp. nov. *Int J Syst Evol Microbiol* 2004;54:749–751.

129. **Zhao G-Z, Li J, Zhu W-Y, Li X-P, Tian S-Z, *et al.*** Pseudonocardia bannaensis sp. nov., a novel actinomycete isolated from the surface-sterilized roots of Artemisia annua L. *Antonie Van Leeuwenhoek* 2011;100:35–42.

130. **Zhao G-ZZ, Li J, Huang H-YY, Zhu W-YY, Zhao L-XX, *et al.*** Pseudonocardia artemisiae sp. nov., isolated from surface-sterilized Artemisia annua L. *Int J Syst Evol Microbiol* 2011;61:1061–1065.

131. **Zhao GZ, Li J, Zhu WY, Wei DQ, Zhang JL, *et al.*** Pseudonocardia xishanensis sp. nov., an endophytic actinomycete isolated from the roots of Artemisia annua L. *Int J Syst Evol Microbiol* 2012;62:2395–2399.

132. **Zhao G-Z, Li J, Huang H-Y, Zhu W-Y, Park D-J, *et al.*** Pseudonocardia kunmingensis sp. nov., an actinobacterium isolated from surface-sterilized roots of Artemisia annua L. *Int J Syst Evol Microbiol* 2011;61:2292–2297.

133. **Zhao GZ, Zhu WY, Li J, Xie Q, Xu LH, *et al.*** Pseudonocardia serianimatus sp. nov., a novel actinomycete isolated from the surface-sterilized leaves of Artemisia annua L. *Antonie van Leeuwenhoek, Int J Gen Mol Microbiol* 2011;100:521–528.

134. **Zheng LP, Li XP, Zhou LL, Wang JW**. Endophytes in Artemisia annua L.: new potential regulators for plant growth and artemisinin biosynthesis. *Plant Growth Regul* 2021;95:293–313.

135. **Li J, Zhao G-Z, Varma A, Qin S, Xiong Z, *et al.*** An Endophytic Pseudonocardia Species Induces the Production of Artemisinin in Artemisia annua. *PLoS One* 2012;7:e51410.

136. **Qin S, Zhu WY, Jiang JH, Klenk HP, Li J, *et al.*** Pseudonocardia tropica sp. nov., an endophytic actinomycete isolated from the stem of Maytenus austroyunnanensis. *Int J Syst Evol Microbiol* 2010;60:2524–2528.

137. **Duangmal K, Thamchaipenet A, Matsumoto A, Takahashi Y**. Pseudonocardia acaciae sp. nov., isolated from roots of Acacia auriculiformis A. Cunn. ex Benth. *Int J Syst Evol Microbiol* 2009;59:1487–1491.

138. **Chen HH, Qin S, Li J, Zhang YQ, Xu LH, *et al.*** Pseudonocardia endophytica sp. nov., isolated from the pharmaceutical plant Lobelia clavata. *Int J Syst Evol Microbiol* 2009;59:559–563.

139. **Kaewkla O, Franco CMM**. Pseudonocardia eucalypti sp. nov., an endophytic actinobacterium with a unique knobby spore surface, isolated from roots of a native Australian eucalyptus tree. *Int J Syst Evol Microbiol* 2011;61:742–746.

140. **Mo P, Zhao Y, Liu J, Xu Z, Gao J**. Pseudonocardia broussonetiae sp. nov., an endophytic actinomycete isolated from the roots of Broussonetia papyrifera. *Int J Syst Evol Microbiol* 2021;71:004680.

141. **Currie CR, Summerbell RC, Scott J a., Malloch D**. Fungus-growing ants use antibiotic-producing bacteria to control garden parasites. *Nature* 1999;398:701–704.

142. **Currie CR, Poulsen M, Mendenhall J, Boomsma JJ, Billen J**. Coevolved crypts and exocrine glands support mutualistic bacteria in fungus-growing ants. *Science (80-)* 2006;311:81–83.

143. **Bruner-Montero G, Wood M, Horn HA, Gemperline E, Li L, *et al.*** Symbiont-Mediated Protection of Acromyrmex Leaf-Cutter Ants from the Entomopathogenic Fungus Metarhizium anisopliae. *MBio* 2021;12:e01885-21.

144. **Li H, Sosa-Calvo J, Horn HA, Pupo MT, Clardy J, *et al.*** Convergent evolution of complex structures for ant-bacterial defensive symbiosis in fungus-farming ants. *Proc Natl Acad Sci U S A* 2018;115:10720–10725.

145. **Maldonado LA, Stach JEM, Pathom-aree W, Ward AC, Bull AT, *et al.*** Diversity of cultivable actinobacteria in geographically widespread marine sediments. *Antonie Van Leeuwenhoek* 2005;87:11–18.

146. **Gontang EA, Fenical W, Jensen PR**. Phylogenetic diversity of gram-positive bacteria cultured from marine sediments. *Appl Environ Microbiol* 2007;73:3272–3282.

147. **Zhang G, Cao T, Ying J, Yang Y, Ma L**. Diversity and novelty of actinobacteria in Arctic marine sediments. *Antonie van Leeuwenhoek, Int J Gen Mol Microbiol*

2014;105:743–754.

148. **Claverías FP, Undabarrena A, González M, Seeger M, Cámara B**. Culturable diversity and antimicrobial activity of Actinobacteria from marine sediments in Valparaíso bay, Chile. *Front Microbiol* 2015;6:1–11.

149. **Zhang D-F, Jiang Z, Li L, Liu B-B, Zhang X-M, *et al.*** Pseudonocardia sediminis sp. nov., isolated from marine sediment. *Int J Syst Evol Microbiol* 2014;64:745–750.

150. **Tian X-P, Long L-J, Li S-M, Zhang J, Xu Y, *et al.*** Pseudonocardia antitumoralis sp. nov., a deoxynyboquinone-producing actinomycete isolated from a deep-sea sediment. *Int J Syst Evol Microbiol* 2013;63:893–899.

151. **Zhang G, Wang L, Li J, Zhou Y**. Pseudonocardia profundimaris sp. nov., isolated from marine sediment. *Int J Syst Evol Microbiol* 2017;67:1693–1697.

152. **Chanama S, Janphen S, Suriyachadkun C, Chanama M**. Pseudonocardia mangrovi sp. nov., isolated from soil. *Int J Syst Evol Microbiol* 2018;68:2949–2955.

153. **Liu ZP, Wu JF, Liu ZH, Liu SJ**. Pseudonocardia ammonioxydans sp. nov., isolated from coastal sediment. *Int J Syst Evol Microbiol* 2006;56:555–558.

154. **Xin Y, Kanagasabhapathy M, Janussen D, Xue S, Zhang W**. Phylogenetic diversity of Gram-positive bacteria cultured from Antarctic deep-sea sponges. *Polar Biol* 2011;34:1501–1512.

155. **Sun W, Dai S, Jiang S, Wang G, Liu G, *et al.*** Culture-dependent and culture-independent diversity of Actinobacteria associated with the marine sponge Hymeniacidon perleve from the South China Sea. *Antonie van Leeuwenhoek, Int J Gen Mol Microbiol* 2010;98:65–75.

156. **Lee SD, Kim ES, Min KL, Lee WY, Kang SO, *et al.*** Pseudonocardia kongjuensis sp. nov., isolated from a gold mine cave. *Int J Syst Evol Microbiol* 2001;51:1505–1510.

157. **Baig U, Dahanukar N, Shintre N, Holkar K, Pund A, *et al.*** Phylogenetic diversity and activity screening of cultivable Actinobacteria isolated from marine sponges and associated environments from the western coast of India. *Access Microbiol*;3. Epub ahead of print 21 September 2021. DOI: 10.1099/acmi.0.000242.

158. **Li J, Dong J-D, Yang J, Luo X-M, Zhang S**. Detection of polyketide synthase and nonribosomal peptide synthetase biosynthetic genes from antimicrobial coral-associated actinomycetes. *Antonie Van Leeuwenhoek* 2014;106:623–635.

159. **Alvarado P, Huang Y, Wang J, Garrido I, Leiva S**. Phylogeny and bioactivity of epiphytic Gram-positive bacteria isolated from three co-occurring antarctic macroalgae. *Antonie Van Leeuwenhoek* 2018;111:1543–1555.

160. **Medema MH**. Computational Genomics of Specialized Metabolism: from Natural Product Discovery to Microbiome Ecology. *mSystems* 2018;3:e00182-17.

161. **Medema MH, de Rond T, Moore BS**. Mining genomes to illuminate the specialized chemistry of life. *Nat Rev Genet* 2021;22:553–571.

162. **Davies J**. Specialized microbial metabolites: functions and origins. *J Antibiot (Tokyo)* 2013;66:361–364.

163. **Brien JO, Wright GD**. An ecological perspective of microbial secondary metabolism. *Curr Opin Biotechnol* 2011;22:552–558.

164. **Davies J**. Are antibiotics naturally antibiotics? *J Ind Microbiol Biotechnol* 2006;33:496–499.

165. **Takano E, Chakraburtty R, Nihira T, Yamada Y, Bibb MJ**. A complex role for the γ-butyrolactone SCB1 in regulating antibiotic production in Streptomyces coelicolor A3(2). *Mol Microbiol* 2001;41:1015–1028.

166. **GERTH K, IRSCHIK H, REICHENBACH H, TROWITZSCH W**. The myxovirescins, a family of antibiotics from Myxococcus virescens (Myxobacterales). *J Antibiot (Tokyo)* 1982;35:1454–1459.

167. **Xiao Y, Wei X, Ebright R, Wall D**. Antibiotic production by myxobacteria plays a role in predation. *J Bacteriol* 2011;193:4626–4633.

168. **Kalkreuter E, Pan G, Cepeda AJ, Shen B**. Targeting Bacterial Genomes for Natural Product Discovery. *Trends Pharmacol Sci* 2020;41:13–26.

169. **Medema MH, Fischbach MA**. Computational approaches to natural product discovery. *Nat Chem Biol* 2015;11:639–648.

170. **Chevrette MG, Gutiérrez-García K, Selem-Mojica N, Aguilar-Martínez C, Yañez-Olvera A, *et al.*** Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat Prod Rep* 2020;37:566–599.

171. **Richter TKS, Hughes CC, Moore BS**. Sioxanthin, a novel glycosylated carotenoid, reveals an unusual subclustered biosynthetic pathway. *Environ Microbiol* 2015;17:2158–2171.

172. **Zhang L, Hoshino S, Awakawa T, Wakimoto T, Abe I**. Structural Diversification of Lyngbyatoxin A by Host-Dependent Heterologous Expression of the tleABC Biosynthetic Gene Cluster. *ChemBioChem* 2016;17:1407–1411.

173. **McLean TC, Wilkinson B, Hutchings MI, Devine R**. Dissolution of the Disparate: Co-ordinate Regulation in Antibiotic Biosynthesis. *Antibiotics* 2019;8:83.

174. **Alanjary M, Cano-Prieto C, Gross H, Medema MH**. Computer-aided re-engineering of nonribosomal peptide and polyketide biosynthetic assembly lines. *Nat Prod Rep* 2019;36:1249–1261.

175. **Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, *et al.*** MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res* 2020;48:D454–D458.

176. **Belknap KC, Park CJ, Barth BM, Andam CP**. Genome mining of biosynthetic and chemotherapeutic gene clusters in Streptomyces bacteria. *Sci Rep* 2020;10:2003.

177. **Adamek M, Alanjary M, Sales-Ortells H, Goodfellow M, Bull AT, *et al.*** Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species. *BMC Genomics* 2018;19:1–15.

178. **Männle D, McKinnie SMK, Mantri SS, Steinke K, Lu Z, *et al.*** Comparative Genomics and Metabolomics in the Genus Nocardia. *mSystems* 2020;5:1–19.

179. **Ziemert N, Lechner A, Wietz M, Millan-Aguinaga N, Chavarria KL, *et al.*** Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proc Natl Acad Sci* 2014;111:E1130–E1139.

180. **Kautsar SA, Blin K, Shaw S, Weber T, Medema MH**. BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res* 2021;49:D490–D497.

181. **Kim SY, Ju KS, Metcalf WW, Evans BS, Kuzuyama T, *et al.*** Different biosynthetic pathways to fosfomycin in Pseudomonas syringae and Streptomyces species. *Antimicrob Agents Chemother* 2012;56:4175–4183.

182. **Fischbach MA, Clardy J**. One pathway, many products. *Nat Chem Biol* 2007;3:353–355.

183. **Hoskisson PA, Seipke RF**. Cryptic or silent? The known unknowns, unknown knowns, and unknown unknowns of secondary metabolism. *MBio* 2020;11:1–5.

184. **Baltz RH**. Gifted microbes for genome mining and natural product discovery. *J Ind Microbiol Biotechnol* 2017;44:573–588.

185. **Ziemert N, Alanjary M, Weber T**. The evolution of genome mining in microbes – a review. *Nat Prod Rep* 2016;33:988–1005.

186. **Scherlach K, Hertweck C**. Mining and unearthing hidden biosynthetic potential. *Nat Commun* 2021;12:3864.

187. **Ward AC, Allenby NE**. Genome mining for the search and discovery of bioactive compounds: The Streptomyces paradigm. *FEMS Microbiol Lett* 2018;365:1–20.

188. **Kautsar SA, Van Der Hooft JJJ, De Ridder D, Medema MH**. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *Gigascience* 2021;10:1–17.

189. **Youngblut ND, de la Cuesta-Zuluaga J, Reischer GH, Dauser S, Schuster N, *et al.*** Large-Scale Metagenome Assembly Reveals Novel Animal-Associated Microbial Genomes, Biosynthetic Gene Clusters, and Other Genetic Diversity. *mSystems*;5. Epub ahead of print 2020. DOI: 10.1128/msystems.01045-20.

190. **Doroghazi JR, Albright JC, Goering AW, Ju K-SS, Haines RR, *et al.*** A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol* 2014;10:963–968.

191. **Ju KS, Gao J, Doroghazi JR, Wang KKA, Thibodeaux CJ, *et al.*** Discovery of phosphonic acid natural products by mining the genomes of 10,000 actinomycetes. *Proc Natl Acad Sci U S A* 2015;112:12175–12180.

192. Genomes Online Database. https://gold.jgi.doe.gov/ (2022, accessed 9 January 2022).

193. **Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Sundaramurthi JC, *et al.*** Genomes OnLine Database (GOLD) v.8: Overview and updates. *Nucleic Acids Res* 2021;49:D723–D733.

194. **Schorn MA, Alanjary MM, Aguinaldo K, Korobeynikov A, Podell S, *et al.*** Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. *Microbiology* 2016;162:2075–2086.

195. **Jost L**. Entropy and diversity. *Oikos* 2006;113:363–375.

196. **Schmidt R, Ulanova D, Wick LY, Bode HB, Garbeva P**. Microbe-driven chemical ecology: past, present and future. *ISME J* 2019;13:2656–2663.

197. **Molloy EM, Hertweck C**. Antimicrobial discovery inspired by ecological interactions. *Curr Opin Microbiol* 2017;39:121–127.

198. **Behie SW, Bonet B, Zacharia VM, Mcclung DJ, Traxler MF, *et al.*** Molecules to Ecosystems: Actinomycete Natural Products In situ. *Front Microbiol* 2017;7:2149.

199. **Meij A Van Der, Worsley SF, Hutchings MI, Wezel GP Van, van der Meij A, *et al.*** Chemical ecology of antibiotic production by actinomycetes. *FEMS Microbiol Rev* 2017;41:392–416.

200. **Goldstein SL, Klassen JL**. Pseudonocardia Symbionts of Fungus-Growing Ants and the Evolution of Defensive Secondary Metabolism. *Front Microbiol* 2020;11:1–8.

201. **Currie CR**. A Community of Ants, Fungi, and Bacteria: A Multilateral Approach to Studying Symbiosis. *Annu Rev Microbiol* 2001;55:357–380.

202. **Kaltenpoth M, Göttler W, Herzner G, Strohm E**. Symbiotic Bacteria Protect Wasp Larvae from Fungal Infestation. *Curr Biol* 2005;15:475–479.

203. **Worsley SF, Newitt J, Rassbach J, Batey SFD, Holmes NA, *et al.*** Streptomyces Endophytes Promote Host Health and Enhance Growth across Plant Species. *Appl Environ Microbiol* 2020;86:1–17.

204. **Gerber NN, Lechevalier HA**. Geosmin, an Earthy-Smelling Substance Isolated from Actinomycetes. *Appl Microbiol* 1965;13:935–938.

205. **Gerber NN**. A Volatile Metabolite of Actinomycetes, 2-Methylisoborneol. *J Antibiot (Tokyo)* 1969;22:508–509.

206. **Rabe P, Citron CA, Dickschat JS**. Volatile Terpenes from Actinomycetes: A Biosynthetic Study Correlating Chemical Analyses to Genome Data. *ChemBioChem* 2013;14:2345–2354.

207. **Becher PG, Verschut V, Bibb MJ, Bush MJ, Molnár BP, *et al.*** Developmentally regulated volatiles geosmin and 2-methylisoborneol attract a soil arthropod to Streptomyces bacteria promoting spore dispersal. *Nat Microbiol* 2020;5:821–829.

208. **Smanski MJ, Schlatter DC, Kinkel LL**. Leveraging ecological theory to guide natural product discovery. *J Ind Microbiol Biotechnol* 2016;43:115–128.

209. **Holmes NA, Innocent TM, Heine D, Bassam M Al, Worsley SF, *et al.*** Genome Analysis of Two Pseudonocardia Phylotypes Associated with Acromyrmex Leafcutter Ants Reveals Their Biosynthetic Potential. *Front Microbiol* 2016;7:1–16.

210. **Almeida EL, Carrillo Rincón AF, Jackson SA, Dobson ADW**. Comparative Genomics of Marine Sponge-Derived Streptomyces spp. Isolates SM17 and SM18 With Their Closest Terrestrial Relatives Provides Novel Insights Into Environmental Niche Adaptations and Secondary Metabolite Biosynthesis Potential. *Front Microbiol* 2019;10:1–22.

211. **Chevrette MG, Carlson CM, Ortega HE, Thomas C, Ananiev GE, *et al.*** The

antimicrobial potential of Streptomyces from insect microbiomes. *Nat Commun* 2019;10:516.

212. **Pishchany G**. Applying microbial ecology to antimicrobial discovery. *Curr Opin Microbiol* 2020;57:7–12.

213. **Tracanna V, de Jong A, Medema MH, Kuipers OP**. Mining prokaryotes for antimicrobial compounds: from diversity to function. *FEMS Microbiol Rev* 2017;41:417–429.

214. **Kinkel LL, Schlatter DC, Xiao K, Baines AD**. Sympatric inhibition and niche differentiation suggest alternative coevolutionary trajectories among Streptomycetes. *ISME J* 2014;8:249–256.

215. **Letzel AC, Li J, Amos GCA, Millán-Aguiñaga N, Ginigini J, *et al.*** Genomic insights into specialized metabolism in the marine actinomycete Salinispora. *Environ Microbiol* 2017;19:3660–3673.

216. **Lomolino M V., Riddle BR, Whittaker RJ**. *Biogeography*. 5th ed. Suderland: Sinauer; 2016.

217. **De Wit R, Bouvier T**. 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environ Microbiol* 2006;8:755–758.

218. **O'Malley MA**. The nineteenth century roots of 'everything is everywhere'. *Nat Rev Microbiol* 2007;5:647–651.

219. **Finlay BJ**. Global dispersal of free-living microbial eukaryote species. *Science* 2002;296:1061–3.

220. **Curd EE, Martiny JBH, Li H, Smith TB**. Bacterial diversity is positively correlated with soil heterogeneity. *Ecosphere* 2018;9:e02079.

221. **Logue JB, Findlay SEG, Comte J**. Editorial: Microbial Responses to Environmental Changes. *Front Microbiol* 2015;6:1364.

222. **Logue JB, Lindström ES**. Biogeography of Bacterioplankton in Inland Waters. *Freshw Rev* 2008;1:99–114.

223. **Ramette A, Tiedje JM**. Biogeography: An emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microb Ecol* 2007;53:197–207.

224. **Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH**. Beyond biogeographic patterns: Processes shaping the microbial landscape. *Nat Rev Microbiol* 2012;10:497–506.

225. **Fuhrman JA**. Microbial community structure and its functional implications. *Nature* 2009;459:193–199.

226. **Krause S, Le Roux X, Niklaus PA, Van Bodegom PM, Lennon JT, *et al.*** Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. *Front Microbiol* 2014;5:251.

227. **Green JL, Bohannan BJM, Whitaker RJ**. Microbial biogeography: from taxonomy to traits. *Science* 2008;320:1039–43.

228. **Morlon H, O'Connor TK, Bryant JA, Charkoudian LK, Docherty KM, *et al.*** The Biogeography of Putative Microbial Antibiotic Production. *PLoS One* 2015;10:e0130659.

229. **Lemetre C, Maniko J, Charlop-Powers Z, Sparrow B, Lowe AJ, *et al.*** Bacterial natural product biosynthetic domain composition in soil correlates with changes in latitude on a continent-wide scale. *Proc Natl Acad Sci U S A* 2017;114:11615–11620.

230. **Charlop-Powers Z, Owen JG, Reddy BVB, Ternei M, Guimaraes DO, *et al.*** Global biogeographic sampling of bacterial secondary metabolism. *Elife* 2015;2015:e05048.

231. **Choudoir MJ, Pepe-Ranney C, Buckley DH**. Diversification of Secondary Metabolite Biosynthetic Gene Clusters Coincides with Lineage Divergence in Streptomyces. *Antibiotics* 2018;7:12.

232. **Jensen PR, Mafnas C**. Biogeography of the marine actinomycete Salinispora. *Environ Microbiol* 2006;8:1881–1888.

233. **Edlund A, Loesgen S, Fenical W, Jensen PR**. Geographic distribution of secondary metabolite genes in the marine actinomycete Salinispora arenicola. *Appl Environ Microbiol* 2011;77:5916–5925.

234. **Freel KC, Nam S, Fenical W, Jensen PR**. Evolution of Secondary Metabolite Genes in Three Closely Related Marine Actinomycete Species. *Appl Environ Microbiol* 2011;77:7261–7270.

235. **Goo KS, Tsuda M, Ulanova D**. Salinispora arenicola from temperate marine sediments: New intra-species variations and atypical distribution of secondary metabolic genes. *Antonie van Leeuwenhoek, Int J Gen Mol Microbiol* 2014;105:207–219.

236. **Jensen PR, Williams PG, Oh DC, Zeigler L, Fenical W**. Species-specific secondary metabolite production in marine actinomycetes of the genus Salinispora. *Appl Environ Microbiol* 2007;73:1146–1152.

237. **Mcdonald BR, Chevrette MG, Klassen JL, Horn HA, Caldera EJ, *et al.*** Biogeography and Microscale Diversity Shape the Biosynthetic Potential of Fungus-growing Ant-associated Pseudonocardia. *BioRxiv*. Epub ahead of print 2019. DOI: 10.1101/545640.

238. **Caldera EJ, Chevrette MG, McDonald BR, Currie CR**. Local adaptation of bacterial symbionts within a geographic mosaic of antibiotic coevolution. *Appl Environ Microbiol* 2019;85:1–13.

239. **Caldera EJ, Currie CR**. The population structure of antibiotic-producing bacterial symbionts of apterostigma dentigerum ants: Impacts of coevolution and multipartite symbiosis. *Am Nat* 2012;180:604–617.

240. **Thompson JN**. The Evolution of Species Interactions. *Science (80-)* 1999;284:2116–2118.

241. **Thompson JN**. Specific Hypotheses on the Geographic Mosaic of Coevolution. *Am Nat* 1999;153:S1–S14.

242. **Fukuda TTH, Helfrich EJN, Mevers E, Melo WGP, van Arnam EB, *et al.*** Specialized

metabolites reveal evolutionary history and geographic dispersion of a multilateral symbiosis. *ACS Cent Sci* 2021;7:292–299.

243. **Jensen PR**. Natural Products and the Gene Cluster Revolution. *Trends Microbiol* 2016;24:968–977.

244. **Smith EE, Sims EH, Spencer DH, Kaul R, Olson M V.** Evidence for Diversifying Selection at the Pyoverdine Locus of Pseudomonas aeruginosa. *J Bacteriol* 2005;187:2138–2147.

245. **Cubillos-Ruiz A, Berta-Thompson JW, Becker JW, van der Donk WA, Chisholm SW**. Evolutionary radiation of lanthipeptides in marine cyanobacteria. *Proc Natl Acad Sci* 2017;114:E5424–E5433.

246. **Cruz-Morales P, Kopp JF, Martínez-Guerrero C, Yáñez-Guerra LA, Selem-Mojica N, *et al.*** Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. *Genome Biol Evol* 2016;8:1906–1916.

247. **Sélem-Mojica N, Aguilar C, Gutiérrez-García K, Martínez-Guerrero CE, Barona-Gómez F**. EvoMining reveals the origin and fate of natural product biosynthetic enzymes. *Microb genomics*;5. Epub ahead of print 1 December 2019. DOI: 10.1099/mgen.0.000260.

248. **Chevrette MG, Currie CR**. Emerging evolutionary paradigms in antibiotic discovery. *J Ind Microbiol Biotechnol* 2019;46:257–271.

249. **Thaker MN, Wang W, Spanogiannopoulos P, Waglechner N, King AM, *et al.*** Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nat Biotechnol* 2013;31:922–927.

250. **Culp EJ, Waglechner N, Wang W, Fiebig-Comyn AA, Hsu YP, *et al.*** Evolution-guided discovery of antibiotics that inhibit peptidoglycan remodelling. *Nature* 2020;578:582–587.

251. **Waglechner N, McArthur AG, Wright GD**. Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance. *Nat Microbiol* 2019;4:1862–1871.

252. **Covington BC, McLean JA, Bachmann BO**. Comparative mass spectrometry-based metabolomics strategies for the investigation of microbial secondary metabolites. *Nat Prod Rep* 2017;34:6–24.

253. **Machado H, Tuttle RN, Jensen PR**. Omics-based natural product discovery and the lexicon of genome mining. *Curr Opin Microbiol* 2017;39:136–142.

254. **Bouslimani A, Sanchez LM, Garg N, Dorrestein PC**. Mass spectrometry of natural products: Current, emerging and future technologies. *Nat Prod Rep* 2014;31:718–729.

255. **Floros DJ, Jensen PR, Dorrestein PC, Koyama N**. A metabolomics guided exploration of marine natural product chemical space. *Metabolomics* 2016;12:145.

256. **Hesse M, Meier H, Zeeh B**. *Spectroscopic Methods in Organic Chemistry*. 2nd ed. New York: Thieme; 2008.

257. **Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, *et al.*** Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci U S A* 2012;109:1743–1752.

258. **Fang J, Dorrestein PC**. Emerging mass spectrometry techniques for the direct analysis of microbial colonies. *Curr Opin Microbiol* 2014;19:120–129.

259. **Traxler MF, Watrous JD, Alexandrov T, Dorrestein PC, Kolter R**. Interspecies interactions stimulate diversification of the Streptomyces coelicolor secreted metabolome. *MBio* 2013;4:1–12.

260. **Hoefler BC, Gorzelnik K V., Yang JY, Hendricks N, Dorrestein PC, *et al.*** Enzymatic resistance to the lipopeptide surfactin as identified through imaging mass spectrometry of bacterial competition. *Proc Natl Acad Sci U S A* 2012;109:13082–13087.

261. **Millán-Aguiñaga N, Soldatou S, Brozio S, Munnoch JT, Howe J, *et al.*** Awakening ancient polar Actinobacteria: diversity, evolution and specialized metabolite potential. *Microbiology* 2019;165:1169–1180.

262. **Soldatou S, Eldjárn GH, Ramsay A, van der Hooft JJJ, Hughes AH, *et al.*** Comparative Metabologenomics Analysis of Polar Actinomycetes. *Mar Drugs* 2021;19:103.

263. **Shirling ET, Gottlieb D**. Methods for characterization of Streptomyces species. *Int J Syst Bacteriol* 1966;16:313–340.

264. **Bauer AW, Kirby WMM, Sherris JC, Turck M**. Antibiotic susceptibility testing by a standardized single disk method. *Am J Clin Pathol* 1966;45:493–496.

265. **Moore JM, Bradshaw E, Seipke RF, Hutchings MI, McArthur M**. Use and Discovery of Chemical Elicitors That Stimulate Biosynthetic Gene Clusters in Streptomyces Bacteria. In: *Methods in Enzymology*. Elsevier Inc.; 2012. pp. 367–385.

266. **Pain A, Woodward J, Quail MA, Anderson MJ, Clark R, *et al.*** Insight into the genome of Aspergillus fumigatus: Analysis of a 922 kb region encompassing the nitrate assimilation gene cluster. *Fungal Genet Biol* 2004;41:443–453.

267. **Sharma C, Kumar N, Meis JF, Pandey R, Chowdhary A**. Draft genome sequence of a fluconazole-resistant Candida auris strain from a candidemia patient in India. *Genome Announc* 2015;3:5–6.

268. **Gillum AM, Tsay EYH, Kirsch DR**. Isolation of the Candida albicans gene for orotidine-5'-phosphate decarboxylase by complementation of S. cerevisiae ura3 and E. coli pyrF mutations. *MGG Mol Gen Genet* 1984;198:179–182.

269. **Weisburg WG, Barns SM, Pelletier DA, Lane DJ**. 16S ribosomal DNA amplification for phylogenetic study. *J Bacteriol* 1991;173:697–703.

270. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ**. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.

271. **Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, *et al.*** Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int J Syst Evol Microbiol* 2017;67:1613–1617.

272. **Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, *et al.*** Clustal W

and Clustal X version 2.0. *Bioinformatics* 2007;23:2947–2948.

273. **Saitou N, Nei M**. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;4:406–425.

274. **Tamura K, Nei M**. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 1993;10:512–26.

275. **Kumar S, Stecher G, Li M, Knyaz C, Tamura K**. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 2018;35:1547–1549.

276. **Felsenstein J**. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution (N Y)* 1985;39:783.

277. **Letunic I, Bork P**. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 2016;44:242–245.

278. **Marmur J**. A procedure for the isolation of deoxyribonucleic acid from micro-organisms. *J Mol Biol* 1961;3:208–218.

279. **Salvà Serra F, Salvà-Serra F, Svensson-Stadler L, Busquets A, Jaén-Luchoro D, *et al.*** A protocol for extraction and purification of high-quality and quantity bacterial DNA applicable for genome sequencing: a modified version of the Marmur procedure. *Protoc Exch*. Epub ahead of print 9 August 2018. DOI: 10.1038/protex.2018.084.

280. **Bolger AM, Lohse M, Usadel B**. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.

281. **Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, *et al.*** Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 2013;10:563–569.

282. **Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, *et al.*** Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. *Genome Res* 2017;27:722–736.

283. **Vaser R, Šikić M**. Raven: A de novo genome assembler for long reads. *bioRxiv*. Epub ahead of print 2020. DOI: 10.1101/2020.08.07.242461.

284. **Wick RR, Judd LM, Gorrie CL, Holt KE**. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:1–22.

285. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al.*** SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.

286. **Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, *et al.*** Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One* 2014;9:e112963.

287. **Li H**. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–3100.

288. **García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, *et al.*** Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics*

2012;28:2678–2679.

289. **Gurevich A, Saveliev V, Vyahhi N, Tesler G**. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.

290. **Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM**. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31:3210–3212.

291. **Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, *et al.*** IslandViewer 4: Expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res* 2017;45:W30–W35.

292. **Arndt D, Grant JR, Marcu A, Sajed T, Pon A, *et al.*** PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44:W16–W21.

293. **Kolmogorov M, Yuan J, Lin Y, Pevzner PA**. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;37:540–546.

294. **Langmead B, Salzberg SL**. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.

295. **Bosi E, Donati B, Galardini M, Brunetti S, Sagot MF, *et al.*** MeDuSa: A multi-draft based scaffolder. *Bioinformatics* 2015;31:2443–2451.

296. **Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M**. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 2013;14:60.

297. **Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S**. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;9:1–8.

298. **Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK**. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* 2016;8:12–24.

299. **Alanjary M, Steinke K, Ziemert N**. AutoMLST: an automated web server for generating multi-locus species trees highlighting natural product potential. *Nucleic Acids Res* 2019;47:276–282.

300. **Seemann T**. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.

301. **Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ**. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience* 2019;8:1–9.

302. **Stamatakis A**. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.

303. **Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, *et al.*** Phandango: An interactive viewer for bacterial population genomics. *Bioinformatics* 2018;34:292–293.

304. **Medema MH, Blin K, Cimermancic P, De Jager V, Zakrzewski P, *et al.*** AntiSMASH:

Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 2011;39:339–346.

305. **Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, *et al.*** antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* 2019;47:W81–W87.

306. **Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, *et al.*** A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* 2019;16:1–9.

307. **Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, *et al.*** Circos: An information aesthetic for comparative genomics. *Genome Res* 2009;19:1639–1645.

308. **Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, *et al.*** Cytoscape: A Software Environment for Integrated Models. *Genome Res* 2003;13:2498–2504.

309. **Gilchrist CLM, Chooi Y-H**. Clinker & Clustermap.Js: Automatic Generation of Gene Cluster Comparison Figures. *Bioinformatics* 2021;37:2473–2475.

310. **Didelot X, Wilson DJ**. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Comput Biol* 2015;11:e1004041.

311. **Angiuoli S V., Salzberg SL**. Mugsy: Fast multiple alignment of closely related whole genomes. *Bioinformatics* 2011;27:334–342.

312. **Schwedock J, McCormick JR, Angert ER, Nodwell JR, Losick R**. Assembly of the cell division protein FtsZ into ladder-like structures in the aerial hyphae of Streptomyces coelicolor. *Mol Microbiol* 1997;25:847–858.

313. **Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, *et al.*** Fiji: an open-source platform for biological-image analysis. *Nat Methods* 2012;9:676–682.

314. **Kuykendall LD, Roy MA, O'Neill JJ, Devine TE**. Fatty acids, antibiotic resistance, and deoxyribonucleic acid homology groups of Bradyrhizobium japonicum. *Int J Syst Bacteriol* 1988;38:358–361.

315. **Miller LT**. Single derivatization method for routine analysis of bacterial whole-cell fatty acid methyl esters, including hydroxy acids. *J Clin Microbiol* 1982;16:584–586.

316. **Tindall BJ, Sikorski J, Smibert RA, Krieg NR**. Phenotypic Characterization and the Principles of Comparative Systematics. In: *Methods for General and Molecular Microbiology*. Washington, DC, USA: ASM Press; 2014. pp. 330–393.

317. **Rhuland LE, Work E, Denman RF, Hoare DS**. The Behavior of the Isomers of α,ε-Diaminopimelic Acid on Paper Chromatograms. *J Am Chem Soc* 1955;77:4844–4846.

318. **Staneck JL, Roberts GD**. Simplified Approach to Identification of Aerobic Actinomycetes by Thin-Layer Chromatography. *Appl Microbiol* 1974;28:226–231.

319. **Kessner D, Chambers M, Burke R, Agus D, Mallick P**. ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics* 2008;24:2534–2536.

320. **Pluskal T, Castillo S, Villar-Briones A, Orešič M**. MZmine 2: Modular framework for

processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 2010;11:395.

321. **Myers OD, Sumner SJ, Li S, Barnes S, Du X**. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. *Anal Chem* 2017;89:8696–8703.

322. **Chong J, Soufan O, Li C, Caraus I, Li S, *et al.*** MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res* 2018;46:W486–W494.

323. **Wang M, Carver JJ, Phelan V V., Sanchez LM, Garg N, *et al.*** Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* 2016;34:828–837.

324. **Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, *et al.*** Feature-based molecular networking in the GNPS analysis environment. *Nat Methods* 2020;17:905–908.

325. **Mohimani H, Gurevich A, Shlemov A, Mikheenko A, Korobeynikov A, *et al.*** Dereplication of microbial metabolites through database search of mass spectra. *Nat Commun* 2018;9:4035.

326. **Schmid R, Petras D, Nothias L-F, Wang M, Aron AT, *et al.*** Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nat Commun* 2021;12:3832.

327. **Wandy J, Zhu Y, Van Der Hooft JJJ, Daly R, Barrett MP, *et al.*** Ms2lda.org: Web-based topic modelling for substructure discovery in mass spectrometry. *Bioinformatics* 2018;34:317–318.

328. **Ernst M, Kang K Bin, Caraballo-Rodríguez AM, Nothias L-F, Wandy J, *et al.*** MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. *Metabolites* 2019;9:144.

329. **Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, *et al.*** ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminform* 2016;8:61.

330. **Poulsen M, Erhardt DP, Molinaro DJ, Lin T-L, Currie CR**. Antagonistic Bacterial Interactions Help Shape Host-Symbiont Dynamics within the Fungus-Growing Ant-Microbe Mutualism. *PLoS One* 2007;2:e960.

331. **Yang JY, Phelan V V., Simkovsky R, Watrous JD, Trial RM, *et al.*** Primer on Agar-Based Microbial Imaging Mass Spectrometry. *J Bacteriol* 2012;194:6023–6028.

332. **Levy SE, Myers RM**. Advancements in Next-Generation Sequencing. *Annu Rev Genomics Hum Genet* 2016;17:95–115.

333. **Cole S, R B, J P, T G, C C, *et al.*** Deciphering thebiologyof Mycobacterium tuberculosis from thec omplete genome sequence. *Nature* 1998;393:537–544.

334. **Bentley SD, Chater KF, Cerdeño-Tárraga AM, Challis GL, Thomson NR, *et al.*** Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* 2002;417:141–147.

335. **Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, *et al.*** Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat Biotechnol* 2003;21:526–531.

336. **Gomez-Escribano JP, Alt S, Bibb MJ**. Next generation sequencing of actinobacteria for the discovery of novel natural products. *Mar Drugs* 2016;14:6–8.

337. **Nyrén P, Pettersson B, Uhlén M**. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Analytical Biochemistry* 1993;208:171–175.

338. **Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC**. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLoS One*;8. Epub ahead of print 2013. DOI: 10.1371/journal.pone.0062856.

339. **Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, *et al.*** Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;456:53–59.

340. **Eid J, Fehr A, Gray J, Luong K, Lyle J, *et al.*** Real-time DNA sequencing from single polymerase molecules. *Science (80-)* 2009;323:133–138.

341. **Goodwin S, McPherson JD, McCombie WR**. Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;17:333–351.

342. **Lin H, Liao Y**. Evaluation and Validation of Assembling Corrected PacBio Long Reads for Microbial Genome Completion via Hybrid Approaches. *PLoS One* 2015;10:e0144305.

343. **Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, *et al.*** Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol* 2013;14:R101.

344. **English AC, Richards S, Han Y, Wang M, Vee V, *et al.*** Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS One* 2012;7:e47768.

345. **Lee N, Kim W, Hwang S, Lee Y, Cho S, *et al.*** Thirty complete Streptomyces genome sequences for mining novel secondary metabolite biosynthetic gene clusters. *Sci Data* 2020;7:1–9.

346. NCBI Genome. https://www.ncbi.nlm.nih.gov/genome/ (2021, accessed 8 October 2021).

347. Integrated Microbial Genomes and Microbiomes. https://img.jgi.doe.gov/ (2021, accessed 8 October 2021).

348. **Sit CS, Ruzzini AC, Van Arnam EB, Ramadhar TR, Currie CR, *et al.*** Variable genetic architectures produce virtually identical molecules in bacterial symbionts of fungus-growing ants. *Proc Natl Acad Sci* 2015;112:13150–13154.

349. **Van Arnam EB, Ruzzini AC, Sit CS, Currie CR, Clardy J**. A Rebeccamycin Analog Provides Plasmid-Encoded Niche Defense. *J Am Chem Soc* 2015;137:14272–14274.

350. **Van Arnam EB, Ruzzini AC, Sit CS, Horn H, Pinto-Tomás AA, *et al.*** Selvamicin, an atypical antifungal polyene from two alternative genomic contexts. *Proc Natl Acad*

Sci 2016;113:12940–12945.

351. **Yoshida K, Yasutake Y, Tamura T**. Complete Genome Sequence of an Efficient Vitamin D 3 -Hydroxylating Bacterium, Pseudonocardia autotrophica NBRC 12743. *Microbiol Resour Announc* 2018;7:4–5.

352. **Nouioui I, Carro L, García-López M, Meier-Kolthoff JP, Woyke T, *et al.*** Genome-based taxonomic classification of the phylum actinobacteria. *Front Microbiol* 2018;9:1–119.

353. **Román-Ponce B, Millán-Aguiñaga N, Guillen-Matus D, Chase AB, Ginigini JGM, *et al.*** Six novel species of the obligate marine actinobacterium Salinispora, Salinispora cortesiana sp. nov., Salinispora fenicalii sp. nov., Salinispora goodfellowii sp. nov., Salinispora mooreana sp. nov., Salinispora oceanensis sp. nov. and Salinispora vitien. *Int J Syst Evol Microbiol* 2020;70:4668–4682.

354. **Millán-Aguinãga N, Chavarria KL, Ugalde JA, Letzel AC, Rouse GW, *et al.*** Phylogenomic Insight into Salinispora (Bacteria, Actinobacteria) Species Designations. *Sci Rep* 2017;7:1–11.

355. **Bobay L-M**. The Prokaryotic Species Concept and Challenges. In: Tettelin H, Medini D (editors). *The Pangenome*. Cham: Springer International Publishing; 2020. pp. 21–49.

356. **Stackebrandt E, Frederiksen W, Garrity GM, Grimont PAD, Kämpfer P, *et al.*** Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol* 2002;52:1043–1047.

357. **Rosselló-Mora R, Amann R**. The species concept for prokaryotes. *FEMS Microbiol Rev* 2001;25:39–67.

358. **Nathan MJ, Cracraft J**. The Nature of Species in Evolution. In: Scheiner SM, Mindell DP (editors). *The Theory of Evolution*. London: University of Chicago Press; 2020. pp. 102–122.

359. **Mayr E**. The biological meaning of species. *Biol J Linn Soc* 1969;1:311–320.

360. **Doolittle WF, Zhaxybayeva O**. On the origin of prokaryotic species. *Genome Res* 2009;19:744–756.

361. **Bobay L-M, Ochman H**. Biological Species Are Universal across Life's Domains. *Genome Biol Evol* 2017;9:491–501.

362. **Richter M, Rosselló-Móra R**. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* 2009;106:19126–19131.

363. **Achtman M, Wagner M**. Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol* 2008;6:431–440.

364. **Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, *et al.*** Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *Int J Syst Evol Microbiol* 1987;37:463–464.

365. **Brenner DJ**. Deoxyribonucleic acid reassociation in the taxonomy of enteric bacteria. *Int J Syst Bacteriol* 1973;23:298–307.

366. **Stackebrandt E, Goebel BM**. Taxonomic note: A place for DNA-DNA reassociation

and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* 1994;44:846–849.

367. **Fox GE, Wisotzkey JD, Jurtshuk P**. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol* 1992;42:166–170.

368. **Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, *et al.*** Re-evaluating prokaryotic species. *Nat Rev Microbiol* 2005;3:733–739.

369. **Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC**. Whole-genome prokaryotic phylogeny. *Bioinformatics* 2005;21:2329–2335.

370. **Auch AF, von Jan M, Klenk HP, Göker M**. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand Genomic Sci* 2010;2:117–134.

371. **Yoon S-H, Ha S, Lim J, Kwon S, Chun J**. A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie Van Leeuwenhoek* 2017;110:1281–1286.

372. **Konstantinidis KT, Tiedje JM**. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci* 2005;102:2567–2572.

373. **Cracraft J**. Species Concepts and Speciation Analysis. In: *Current Ornithology*. New York, NY: Springer US; 1983. pp. 159–187.

374. **Cohan FM**. Bacterial Species and Speciation. *Syst Biol* 2001;50:513–524.

375. **Cohan FM, Perry EB**. A Systematics for Discovering the Fundamental Units of Bacterial Diversity. *Curr Biol* 2007;17:373–386.

376. **Shapiro BJ, Polz MF**. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol* 2014;22:235–247.

377. **Cohan FM, Koeppel AF**. The Origins of Ecological Diversity in Prokaryotes. *Curr Biol* 2008;18:1024–1034.

378. **Edwards S V., Hopkins R, Mallet J**. Speciation. In: Scheiner SM, Mindell DP (editors). *The Theory of Evolution*. London: University of Chicago Press; 2020. pp. 296–318.

379. **Vos M**. A species concept for bacteria based on adaptive divergence. *Trends Microbiol* 2011;19:1–7.

380. **Friedman J, Alm EJ, Shapiro BJ**. Sympatric Speciation: When Is It Possible in Bacteria? *PLoS One*;8. Epub ahead of print 2013. DOI: 10.1371/journal.pone.0053539.

381. **Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, *et al.*** Population genomics of early events in the ecological differentiation of bacteria. *Science (80-)* 2012;335:48–51.

382. **Shapiro BJ, Polz MF**. Microbial speciation. *Cold Spring Harb Perspect Biol* 2015;7:1–22.

383. **Shapiro BJ, Leducq JB, Mallet J**. What Is Speciation? *PLoS Genet* 2016;12:1–14.

384. **Cruickshank TE, Hahn MW**. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol* 2014;23:3133–3157.

385. **Vandamme P, Pot B, Gillis M, De Vos P, Kersters K, *et al.*** Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev* 1996;60:407–438.

386. **Gillis M, Vandamme P, Vos P De, Swings J, Kersters K**. Polyphasic Taxonomy. In: *Bergey's Manual of Systematics of Archaea and Bacteria*. Wiley; 2015. pp. 1–10.

387. **Reichert K, Lipski A, Pradella S, Stackebrandt E, Altendorf K**. Pseudonocardia asaccharolytica sp. nov. and Pseudonocardia sulfidoxydans sp. nov., two new dimethyl disulfide-degrading actinomycetes and emended description of the genus Pseudonocardia. *Int J Syst Bacteriol* 1998;48:441–449.

388. **Warwick S, Bowen T, McVeigh H, Embley TM**. A phylogenetic analysis of the family Pseudonocardiaceae and the genera Actinokineospora and Saccharothrix with 16S rRNA sequences and a proposal to combine the genera Amycolata and Pseudonocardia in an emended genus Pseudonocardia. *Int J Syst Bacteriol* 1994;44:293–299.

389. **Huang Y, Wang L, Lu Z, Hong L, Liu Z, *et al.*** Proposal to combine the genera Actinobispora and Pseudonocardia in an emended genus Pseudonocardia, and description of Pseudonocardia zijingensis sp. nov. *Int J Syst Evol Microbiol* 2002;52:977–982.

390. **Park SW, Park ST, Lee JE, Kim YM**. Pseudonocardia carboxydivorans sp. nov., a carbon monoxide-oxidizing actinomycete, and an emended description of the genus Pseudonocardia. *Int J Syst Evol Microbiol* 2008;58:2475–2478.

391. **Huang Y, Goodfellow M**. Pseudonocardia. In: M.E. Trujillo, Dedysh S, DeVos P, Hedlund B, Kämpfer P, et al. (editors). *Bergey's Manual of Systematics of Archaea and Bacteria*. Wiley; 2015. pp. 1–32.

392. **Parte AC, Sardà Carbasse J, Meier-Kolthoff JP, Reimer LC, Göker M**. List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. *Int J Syst Evol Microbiol* 2020;70:5607–5612.

393. Genus Pseudonocardia. https://lpsn.dsmz.de/genus/pseudonocardia (accessed 17 May 2021).

394. **De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, *et al.*** Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb Genomics*;5. Epub ahead of print 1 September 2019. DOI: 10.1099/mgen.0.000294.

395. **Wick RR, Holt KE**. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research* 2019;8:1–22.

396. **Tvedte ES, Gasser M, Sparklin BC, Michalski J, Hjelmen CE, *et al.*** Comparison of long-read sequencing technologies in interrogating bacteria and fly genomes. *G3 Genes|Genomes|Genetics*;11. Epub ahead of print 17 June 2021. DOI: 10.1093/g3journal/jkab083.

397. **Klassen JL, Currie CR**. Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics* 2012;13:14.

398. **Goldstein S, Beka L, Graf J, Klassen JL**. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics* 2019;20:1–17.

399. **Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, *et al.*** DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 2007;57:81–91.

400. **Palmer M, Steenkamp ET, Blom J, Hedlund BP, Venter SN**. All anis are not created equal: Implications for prokaryotic species boundaries and integration of anis into polyphasic taxonomy. *Int J Syst Evol Microbiol* 2020;70:2937–2948.

401. **Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, *et al.*** Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol* 2018;68:461–466.

402. **Konstantinidis KT, Ramette A, Tiedje JM**. Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. *Appl Environ Microbiol* 2006;72:7286–93.

403. **Chevrette MG, Carlos-Shanley C, Louie KB, Bowen BP, Northen TR, *et al.*** Taxonomic and Metabolic Incongruence in the Ancient Genus Streptomyces. *Front Microbiol* 2019;10:2170.

404. **Stott CM, Bobay L-M**. Impact of homologous recombination on core genome phylogenies. *BMC Genomics* 2020;21:829.

405. **López-Pérez M, Haro-Moreno JM, Coutinho FH, Martinez-Garcia M, Rodriguez-Valera F**. The Evolutionary Success of the Marine Bacterium SAR11 Analyzed through a Metagenomic Perspective. *mSystems* 2020;5:1–13.

406. **Freel KC, Millán-Aguiñaga N, Jensen PR**. Multilocus Sequence Typing Reveals Evidence of Homologous Recombination Linked to Antibiotic Resistance in the Genus Salinispora. *Appl Environ Microbiol* 2013;79:5997–6005.

407. **Doroghazi JR, Buckley DH**. Widespread homologous recombination within and between Streptomyces species. *ISME J* 2010;4:1136–1143.

408. **Andam CP, Choudoir MJ, Vinh Nguyen A, Sol Park H, Buckley DH**. Contributions of ancestral inter-species recombination to the genetic diversity of extant Streptomyces lineages. *ISME J* 2016;10:1731–1741.

409. **Jensen PR, Mincer TJ, Williams PG, Fenical W**. Marine actinomycete diversity and natural product discovery. *Antonie van Leeuwenhoek, Int J Gen Mol Microbiol* 2005;87:43–48.

410. **Jensen PR, Gontang E, Mafnas C, Mincer TJ, Fenical W**. Culturable marine actinomycete diversity from tropical Pacific Ocean sediments. *Environ Microbiol* 2005;7:1039–1048.

411. **Russell NJ, Nichols DS**. Polyunsaturated fatty acids in marine bacteria — a dogma rewritten. *Microbiology* 1999;145:767–779.

412. **Tian XP, Xu Y, Zhang J, Li J, Chen Z, *et al.*** Streptomyces oceani sp. nov., a new obligate marine actinomycete isolated from a deep-sea sample of seep authigenic

carbonate nodule in south china sea. *Antonie van Leeuwenhoek, Int J Gen Mol Microbiol* 2012;102:335–343.

413. **Wirsen CO, Jannasch HW, Wakeham SG, Canuel EA**. Membrane lipids of a psychrophilic and barophilic deep-sea bacterium. *Curr Microbiol* 1986;14:319–322.

414. **Kamimura K, Fuse H, Takimura O, Yamaoka Y**. Effects of growth pressure and temperature on fatty acid composition of a barotolerant deep-sea bacterium. *Appl Environ Microbiol* 1993;59:924–926.

415. **DeLong EF, Yayanos AA**. Biochemical function and ecological significance of novel bacterial lipids in deep-sea procaryotes. *Appl Environ Microbiol* 1986;51:730–737.

416. **Challis GL**. Genome mining for novel natural product discovery. *J Med Chem* 2008;51:2618–2628.

417. **Challis GL, Ravel J**. Coelichelin, a new peptide siderophore encoded by the Streptomyces coelicolor genome: Structure prediction from the sequence of its non-ribosomal peptide synthetase. *FEMS Microbiol Lett* 2000;187:111–114.

418. **Lautru S, Deeth RJ, Bailey LM, Challis GL**. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. *Nat Chem Biol* 2005;1:265–269.

419. **Liu W, Zhang W, Jin H, Zhang Q, Chen Y, *et al.*** Genome mining of marine-derived Streptomyces sp. SCSIO 40010 leads to cytotoxic new polycyclic tetramate macrolactams. *Mar Drugs* 2019;17:1–14.

420. **Blin K, Kim HU, Medema MH, Weber T**. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief Bioinform* 2018;20:1103–1113.

421. **Skinnider MA, Johnston CW, Edgar RE, Dejong CA, Merwin NJ, *et al.*** Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proc Natl Acad Sci U S A* 2016;113:E6343–E6351.

422. **Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, *et al.*** Genome sequencing reveals complex secondary metabolome in the marine actinomycete Salinispora tropica. *Proc Natl Acad Sci* 2007;104:10376–10381.

423. **Song L, Barona-Gomez F, Corre C, Xiang L, Udwary DW, *et al.*** Type III polyketide synthase β-ketoacyl-ACP starter unit and ethylmalonyl-CoA extender unit selectivity discovered by Streptomyces coelicolor genome mining. *J Am Chem Soc* 2006;128:14754–14755.

424. **Goto Y, Li B, Claesen J, Shi Y, Bibb MJ, *et al.*** Discovery of unique lanthionine synthetases reveals new mechanistic and evolutionary insights. *PLoS Biol* 2010;8:4–13.

425. **Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, *et al.*** Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* 2014;158:412–421.

426. **Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, *et al.*** A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res* 2019;47:e110.

427. **Undabarrena A, Valencia R, Cumsille A, Zamora-Leiva L, Castro-Nallar E, *et al.*** Rhodococcus comparative genomics reveals a phylogenomic-dependent non-ribosomal peptide synthetase distribution: insights into biosynthetic gene cluster connection to an orphan metabolite. *Microb Genomics*;7. Epub ahead of print 9 July 2021. DOI: 10.1099/mgen.0.000621.

428. **Zdouc MM, Iorio M, Maffioli SI, Crüsemann M, Donadio S, *et al.*** Planomonospora: A Metabolomics Perspective on an Underexplored Actinobacteria Genus. *J Nat Prod* 2021;84:204–219.

429. **Gavriilidou A, Kautsar SA, Zaburannyi N, Krug D, Müller R, *et al.*** Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nat Microbiol* 2022;7:726–735.

430. **Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, *et al.*** Molecular networking as a dereplication strategy. *J Nat Prod* 2013;76:1686–1699.

431. **Nguyen DD, Wu C-H, Moree WJ, Lamsa A, Medema MH, *et al.*** MS/MS networking guided analysis of molecule and gene cluster families. *Proc Natl Acad Sci* 2013;110:E2611–E2620.

432. **Bauermeister A, Pereira F, Grilo IR, Godinho CC, Paulino M, *et al.*** Intra-clade metabolomic profiling of MAR4 Streptomyces from the Macaronesia Atlantic region reveals a source of anti-biofilm metabolites. *Environ Microbiol* 2019;21:1099–1112.

433. **Bauermeister A, Velasco-Alzate K, Dias T, Macedo H, Ferreira EG, *et al.*** Metabolomic Fingerprinting of Salinispora From Atlantic Oceanic Islands. *Front Microbiol* 2018;9:1–13.

434. **Aron AT, Petras D, Schmid R, Gauglitz JM, Büttel I, *et al.*** Native mass spectrometry-based metabolomics identifies metal-binding compounds. *Nat Chem* 2022;14:100–109.

435. **Wang M, Jarmusch AK, Vargas F, Aksenov AA, Gauglitz JM, *et al.*** Mass spectrometry searches using MASST. *Nat Biotechnol* 2020;38:23–26.

436. **Jarmusch AK, Wang M, Aceves CM, Advani RS, Aguirre S, *et al.*** ReDU: a framework to find and reanalyze public mass spectrometry data. *Nat Methods* 2020;17:901–904.

437. **Wibowo JT, Kellermann MY, Köck M, Putra MY, Murniasih T, *et al.*** Anti-Infective and Antiviral Activity of Valinomycin and Its Analogues from a Sea Cucumber-Associated Bacterium, Streptomyces sp. SV 21. *Mar Drugs* 2021;19:81.

438. **Van Der Hooft JJJ, Mohimani H, Bauermeister A, Dorrestein PC, Duncan KR, *et al.*** Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem Soc Rev* 2020;49:3297–3314.

439. **Leao ATF, Wang M, Silva R, Hooft JJJ Van Der, Bauermeister A, *et al.*** A supervised fingerprint-based strategy to connect natural product mass spectrometry fragmentation data to their biosynthetic gene clusters. *bioRxiv* 2021;2021.10.05.463235.

440. **Goering AW, McClure RA, Doroghazi JR, Albright JC, Haverland NA, *et al.*** Metabologenomics: Correlation of Microbial Gene Clusters with Metabolites Drives

Discovery of a Nonribosomal Peptide with an Unusual Amino Acid Monomer. *ACS Cent Sci* 2016;2:99–108.

441. **Duncan KR, Crüsemann M, Lechner A, Sarkar A, Li J, *et al.*** Molecular Networking and Pattern-Based Genome Mining Improves Discovery of Biosynthetic Gene Clusters and their Products from Salinispora Species. *Chem Biol* 2015;22:460–471.

442. **Soldatou S, Eldjarn GH, Huerta-Uribe A, Rogers S, Duncan KR**. Linking biosynthetic and chemical space to accelerate microbial secondary metabolite discovery. *FEMS Microbiol Lett* 2019;366:1–8.

443. **Parkinson EI, Tryon JH, Goering AW, Ju KS, McClure RA, *et al.*** Discovery of the Tyrobetaine Natural Products and Their Biosynthetic Gene Cluster via Metabologenomics. *ACS Chem Biol* 2018;13:1029–1037.

444. **Eldjarn GH, Ramsay A, Van Der Hooft JJJ, Duncan KR, Soldatou S, *et al.*** Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. *PLoS Comput Biol* 2021;17:1–24.

445. **Brouard C, Szafranski M, D'Alché-Buc F**. Input Output Kernel Regression: Supervised and Semi-Supervised Structured Output Prediction with Operator-Valued Kernels. *J Mach Learn Res* 2016;17:1–48.

446. **Mukherjee S, Seshadri R, Varghese NJ, Eloe-Fadrosh EA, Meier-Kolthoff JP, *et al.*** 1,003 Reference Genomes of Bacterial and Archaeal Isolates Expand Coverage of the Tree of Life. *Nat Biotechnol* 2017;35:676–683.

447. **Hoffmann T, Krug D, Bozkurt N, Duddela S, Jansen R, *et al.*** Correlating chemical diversity with taxonomic distance for discovery of natural products in myxobacteria. *Nat Commun* 2018;9:803.

448. **Lee MY, Kim HY, Lee S, Kim JG, Suh JW, *et al.*** Metabolomics-based chemotaxonomic classification of Streptomyces spp. and its correlation with antibacterial activity. *J Microbiol Biotechnol* 2015;25:1265–1274.

449. **Chase AB, Sweeney D, Muskat MN, Guillén-Matus D, Jensen PR**. Vertical inheritance governs biosynthetic gene cluster evolution and chemical diversification. *bioRxiv* 2021;2020.12.19.423547.

450. **Penn K, Jenkins C, Nett M, Udwary DW, Gontang EA, *et al.*** Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. *ISME J* 2009;3:1193–1203.

451. **Iranzo J, Wolf YI, Koonin E V., Sela I**. Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. *Nat Commun* 2019;10:5376.

452. **Fischbach MA, Walsh CT, Clardy J**. The evolution of gene collectives: How natural selection drives chemical innovation. *Proc Natl Acad Sci* 2008;105:4601–4608.

453. **Bolotin E, Hershberg R**. Horizontally Acquired Genes Are Often Shared between Closely Related Bacterial Species. *Front Microbiol* 2017;8:1–10.

454. **McDonald BR, Currie CR**. Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus Streptomyces. *MBio* 2017;8:1–12.

455. **Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA**. A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. *PLoS Comput Biol* 2014;10:e1004016.

456. **Gallagher KA, Jensen PR**. Genomic insights into the evolution of hybrid isoprenoid biosynthetic gene clusters in the MAR4 marine streptomycete clade. *BMC Genomics* 2015;16:960.

457. **Cafaro MJ, Poulsen M, Little AEF, Price SL, Gerardo NM, *et al.*** Specificity in the symbiotic association between fungus-growing ants and protective Pseudonocardia bacteria. *Proc R Soc B Biol Sci* 2011;278:1814–1822.

458. **Pishchany G, Kolter R**. On the possible ecological roles of antimicrobials. *Mol Microbiol* 2020;113:580–587.

459. **Li Y, Pinto-Tomás AA, Rong X, Cheng K, Liu M, *et al.*** Population genomics insights into adaptive evolution and ecological differentiation in streptomycetes. *Appl Environ Microbiol* 2019;85:1–17.

460. **Whitfield GB, Brock TD, Ammann A, Gottlieb D, Carter HE**. Filipin, an Antifungal Antibiotic: Isolation and Properties. *J Am Chem Soc* 1955;77:4799–4801.

461. **Lee MJ, Kong D, Han K, Sherman DH, Bai L, *et al.*** Structural analysis and biosynthetic engineering of a solubility-improved and less-hemolytic nystatin-like polyene in Pseudonocardia autotrophica. *Appl Microbiol Biotechnol* 2012;95:157–168.

462. **Kim HJ, Han CY, Park JS, Oh SH, Kang SH, *et al.*** Nystatin-like Pseudonocardia polyene B1, a novel disaccharide-containing antifungal heptaene antibiotic. *Sci Rep* 2018;8:1–8.

463. **Oh D-C, Poulsen M, Currie CR, Clardy J**. Dentigerumycin: a bacterial mediator of an ant-fungus symbiosis. *Nat Chem Biol* 2009;5:391–393.

464. **Barke J, Seipke RF, Grüschow S, Heavens D, Drou N, *et al.*** A mixed community of actinomycetes produce multiple antibiotics for the fungus farming ant Acromyrmex octospinosus. *BMC Biol* 2010;8:109.

465. **Seipke RF, Grüschow S, Goss RJM, Hutchings MI**. Isolating antifungals from fungus-growing ant symbionts using a genome-guided chemistry approach. *Methods Enzymol* 2012;517:47–70.

466. **van Bergeijk DA, Terlouw BR, Medema MH, van Wezel GP**. Ecology and genomics of Actinobacteria: new concepts for natural product discovery. *Nat Rev Microbiol* 2020;18:546–558.

467. **Miguélez EM, Hardisson C, Manzanal MB**. Hyphal death during colony development in Streptomyces antibioticus: Morphological evidence for the existence of a process of cell deletion in a multicellular prokaryote. *J Cell Biol* 1999;145:515–525.

468. **Fernández M, Sánchez J**. Nuclease activities and cell death processes associated with the development of surface cultures of Streptomyces antibioticus ETH 7451. *Microbiology* 2002;148:405–412.

469. **Rigali S, Nothaft H, Noens EEE, Schlicht M, Colson S, *et al.*** The sugar

phosphotransferase system of Streptomyces coelicolor is regulated by the GntR-family regulator DasR and links N-acetylglucosamine metabolism to the control of development. *Mol Microbiol* 2006;61:1237–1251.

470. **Rigali S, Titgemeyer F, Barends S, Mulder S, Thomae AW, *et al.*** Feast or famine: The global regulator DasR links nutrient stress to antibiotic production by Streptomyces. *EMBO Rep* 2008;9:670–675.

471. **van Wezel GP, McKenzie NL, Nodwell JR**. Chapter 5. Applying the genetics of secondary metabolism in model actinomycetes to the discovery of new antibiotics. *Methods Enzymol* 2009;458:117–41.

472. **Urem M, Świątek-Połatyńska MA, Rigali S, van Wezel GP**. Intertwining nutrient-sensory networks and the control of antibiotic production in Streptomyces. *Mol Microbiol* 2016;102:183–195.

473. **Van Der Heul HU, Bilyk BL, McDowall KJ, Seipke RF, Van Wezel GP**. Regulation of antibiotic production in Actinobacteria: New perspectives from the post-genomic era. *Nat Prod Rep* 2018;35:575–604.

474. **Demain AL, Inamine E**. Biochemistry and regulation of streptomycin and mannosidostreptomycinase (alpha-D-mannosidase) formation. *Bacteriol Rev* 1970;34:1–19.

475. **Kim ES, Hong HJ, Choi CY, Cohen SN**. Modulation of actinorhodin biosynthesis in Streptomyces lividans by glucose repression of afsR2 gene transcription. *J Bacteriol* 2001;183:2198–2203.

476. **Bhatnagar RK, Doull JL, Vining LC**. Role of the carbon source in regulating chloramphenicol production by Streptomyces venezuelae: studies in batch and continuous cultures. *Can J Microbiol* 1988;34:1217–1223.

477. **Bermúdez O, Padilla P, Huitrón C, Flores ME**. Influence of carbon and nitrogen source on synthesis of NADP+- isocitrate dehydrogenase, methylmalonyl-coenzyme a mutase, and methylmalonyl- coenzyme A decarboxylase in Saccharopolyspora erythraea CA340. *FEMS Microbiol Lett* 1998;164:77–82.

478. **Van Wezel GP, McDowall KJ**. The regulation of the secondary metabolism of Streptomyces: New links and experimental advances. *Nat Prod Rep* 2011;28:1311–1333.

479. **Guzmán S, Carmona A, Escalante L, Imriskova I, López R, *et al.*** Pleiotropic effect of the SCO2127 gene on the glucose uptake, glucose kinase activity and carbon catabolite repression in Streptomyces peucetius var. caesius. *Microbiology* 2005;151:1717–1723.

480. **Gubbens J, Janus M, Florea BI, Overkleeft HS, Van Wezel GP**. Identification of glucose kinase-dependent and -independent pathways for carbon control of primary metabolism, development and antibiotic production in Streptomyces coelicolor by quantitative proteomics. *Mol Microbiol* 2012;86:1490–1507.

481. **Onaka H, Mori Y, Igarashi Y, Furumai T**. Mycolic Acid-Containing Bacteria Induce Natural-Product Biosynthesis in Streptomyces Species. *Appl Environ Microbiol* 2011;77:400–406.

482. **Hoshino S, Wakimoto T, Onaka H, Abe I**. Chojalactones A–C, Cytotoxic Butanolides Isolated from Streptomyces sp. Cultivated with Mycolic Acid Containing Bacterium. *Org Lett* 2015;17:1501–1504.

483. **Kim JH, Lee N, Hwang S, Kim W, Lee Y, *et al.*** Discovery of novel secondary metabolites encoded in actinomycete genomes through coculture. *J Ind Microbiol Biotechnol* 2021;48:1–16.

484. **Niu G, Chater KF, Tian Y, Zhang J, Tan H**. Specialised metabolites regulating antibiotic biosynthesis in Streptomyces spp. *FEMS Microbiol Rev* 2016;40:554–573.

485. **Challis GL, Hopwood DA**. Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by Streptomyces species. *Proc Natl Acad Sci U S A* 2003;100:14555–14561.

486. **Gemperline E, Horn HA, Delaney K, Currie CR, Li L**. Imaging with Mass Spectrometry of Bacteria on the Exoskeleton of Fungus-Growing Ants. *ACS Chem Biol* 2017;12:1980–1985.

487. **Heine D, Holmes NA, Worsley SF, Santos ACA, Innocent TM, *et al.*** Chemical warfare between leafcutter ant symbionts and a co-evolved pathogen. *Nat Commun* 2018;9:1–11.

488. **Sen R, Ishak HD, Estrada D, Dowd SE, Hong E, *et al.*** Generalized antifungal activity and 454-screening of Pseudonocardia and Amycolatopsis bacteria in nests of fungus-growing ants. *Proc Natl Acad Sci U S A* 2009;106:17805–17810.

489. **Poulsen M, Cafaro M, Boomsma JJ, Currie CR**. Specificity of the mutualistic association between actinomycete bacteria and two sympatric species of Acromyrmex leaf-cutting ants. *Mol Ecol* 2005;14:3597–3604.

490. **Andersen SB, Hansen LH, Sapountzis P, Sørensen SJ, Boomsma JJ**. Specificity and stability of the Acromyrmex-Pseudonocardia symbiosis. *Mol Ecol* 2013;22:4307–4321.

491. **Worsley SF, Innocent TM, Holmes NA, Al-Bassam MM, Schiøtt M, *et al.*** Competition-based screening helps to secure the evolutionary stability of a defensive microbiome. *BMC Biol* 2021;19:205.

492. **Chang PT, Rao K, Longo LO, Lawton ES, Scherer G, *et al.*** Thiopeptide Defense by an Ant's Bacterial Symbiont. *J Nat Prod* 2020;83:725–729.

493. **Zhu H, Swierstra J, Wu C, Girard G, Choi YH, *et al.*** Eliciting antibiotics active against the ESKAPE pathogens in a collection of actinomycetes isolated from mountain soils. *Microbiology* 2014;160:1714–1725.

494. **Dashti Y, Grkovic T, Abdelmohsen UR, Hentschel U, Quinn RJ**. Actinomycete Metabolome Induction/Suppression with N-Acetylglucosamine. *J Nat Prod* 2017;80:828–836.

495. **Liao CH, Xu Y, Rigali S, Ye BC**. DasR is a pleiotropic regulator required for antibiotic production, pigment biosynthesis, and morphological development in Saccharopolyspora erythraea. *Appl Microbiol Biotechnol* 2015;99:10215–10224.

496. **Li X, Li X, Zhu J, Wang H, Lu C**. Carbamothioic S-acid derivative and kigamicins, the

activated production of silent metabolites in Amycolatopsis alba DSM 44262Δabm9 elicited by N-acetyl-D-glucosamine. *Nat Prod Res* 2020;34:3514–3521.

497. **Batey SFD, Greco C, Hutchings MI, Wilkinson B**. Chemical warfare between fungus-growing ants and their pathogens. *Curr Opin Chem Biol* 2020;59:172–181.

498. **Menegatti C, Fukuda TTH, Pupo MT**. Chemical Ecology in Insect-microbe Interactions in the Neotropics. *Planta Med* 2021;87:38–48.

499. **Poulsen M, Currie CR**. Symbiont Interactions in a Tripartite Mutualism: Exploring the Presence and Impact of Antagonism between Two Fungus-Growing Ant Mutualists. *PLoS One* 2010;5:e8748.

500. **Zelezniak A, Andrejev S, Ponomarova O, Mende DR, Bork P, *et al.*** Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc Natl Acad Sci* 2015;112:6449–6454.

501. **Oakley BB, Carbonero F, van der Gast CJ, Hawkins RJ, Purdy KJ**. Evolutionary divergence and biogeography of sympatric niche-differentiated bacterial populations. *ISME J* 2010;4:488–497.

502. **Vetsigian K, Jajoo R, Kishony R**. Structure and Evolution of Streptomyces Interaction Networks in Soil and In Silico. *PLoS Biol* 2011;9:e1001184.

503. **Patin N V., Duncan KR, Dorrestein PC, Jensen PR**. Competitive strategies differentiate closely related species of marine actinobacteria. *ISME J* 2016;10:478–490.

504. **Vaz Jauri P, Bakker MG, Salomon CE, Kinkel LL**. Subinhibitory antibiotic concentrations mediate nutrient use and competition among soil streptomyces. *PLoS One* 2013;8:8–13.

505. **Krause DJ, Whitaker RJ**. Inferring speciation processes from patterns of natural variation in microbial genomes. *Syst Biol* 2015;64:926–935.

506. **Wang J, Li Y, Pinto-Tomás AA, Cheng K, Huang Y**. Habitat Adaptation Drives Speciation of a Streptomyces Species with Distinct Habitats and Disparate Geographic Origins. *MBio*. Epub ahead of print 11 January 2022. DOI: 10.1128/mbio.02781-21.

507. **Brusetti L, Malkhazova I, Gtari M, Tamagnini I, Borin S, *et al.*** Fluorescent-BOX-PCR for resolving bacterial genetic diversity, endemism and biogeography. *BMC Microbiol* 2008;8:220.

508. **Holsinger KE, Weir BS**. Genetics in geographically structured populations: Defining, estimating and interpreting FST. *Nat Rev Genet* 2009;10:639–650.

509. **Mueller UG, Mikheyev AS, Hong E, Sen R, Warren DL, *et al.*** Evolution of cold-tolerant fungal symbionts permits winter fungiculture by leafcutter ants at the northern frontier of a tropical ant-fungus symbiosis. *Proc Natl Acad Sci* 2011;108:4053–4056.

510. **Lessios HA**. The Great American Schism: Divergence of Marine Organisms After the Rise of the Central American Isthmus. *Annu Rev Ecol Evol Syst* 2008;39:63–91.

511. **Nothias LF, Nothias-Esposito M, Da Silva R, Wang M, Protsyuk I, *et al.*** Bioactivity-

Based Molecular Networking for the Discovery of Drug Leads in Natural Product Bioassay-Guided Fractionation. *J Nat Prod* 2018;81:758–767.

512. **Kim BG, Lee MJ, Seo J, Hwang Y Bin, Lee MY, *et al.*** Identification of functionally clustered nystatin-like biosynthetic genes in a rare actinomycetes, Pseudonocardia autotrophica. *J Ind Microbiol Biotechnol* 2009;36:1425–1434.

Appendices

Table S1. Parameters used in the bioinformatics analysis.

Analysis	Program	Parameters
16S rRNA gene evolutionary analysis using Neighbor-Joining method	Mega X	Analysis Statistical Method = Neighbor-joining Phylogeny Test Test of Phylogeny = Bootstrap method No. of Bootstrap Replications = 1000 Substitution Model Substitutions Type = Nucleotide Model/Method = Kimura 2-parameter model Substitutions to Include = d: Transitions + Transversions Rates and Patterns Rates among Sites = Uniform Rates Pattern among Lineages = Same (Homogeneous) Data Subset to Use Gaps/Missing Data Treatment = Pairwise deletion System Resource Usage Number of Threads = 7 No. of Sites:1633 No Of Bootstrap Reps = 1000
16S rRNA gene evolutionary analysis using Maximum Likelihood method	Mega X	Analysis Statistical Method = Maximum Likelihood Phylogeny Test Test of Phylogeny = Bootstrap method No. of Bootstrap Replications = 1000 Substitution Model Substitutions Type = Nucleotide Model/Method = Kimura 2-parameter model Rates and Patterns Rates among Sites = Uniform Rates Data Subset to Use Gaps/Missing Data Treatment = Use all sites Tree Inference Options ML Heuristic Method = Nearest-Neighbor-Interchange (NNI) Initial Tree for ML = Make initial tree automatically (Default - NJ/BioNJ) Branch Swap Filter = None System Resource Usage Number of Threads = 7 No. of Seqs:37 No. of Sites:1633 No Of Bootstrap Reps = 1000
16S rRNA gene evolutionary analysis using Maximum Parsimony method	Mega X	Analysis Statistical Method = Maximum Parsimony Phylogeny Test Test of Phylogeny = Bootstrap method No. of Bootstrap Replications = 1000 Substitution Model Substitutions Type = Nucleotide Data Subset to Use Gaps/Missing Data Treatment = Use all sites Tree Inference Options MP Search Method = Subtree-Pruning-Regrafting (SPR) No. of Initial Trees (random addition) = 10 MP Search level = 1 Max No. of Trees to Retain = 100 System Resource Usage Number of Threads = 7 No. of Sites:1633 No Of Bootstrap Reps = 1000 CI = 0.516923 RI = 0.618932 RCI = 0.319940(for all sites) iCI = 0.376984(for parsimony informative sites) iRI = 0.618932(for parsimony informative sites) iRCI = 0.233328(for parsimony informative sites)

Illumina short-reads genome assembly using Unicycler	Unicycler	bin/unicycler -1 reads_1.fastq -2 reads_2.fastq -o PATH --spades_path PATH/SPAdes-3.14.1/bin/spades.py --pilon_path PATH/pilon-1.23.jar
Illumina short-reads genome assembly using SPAdes	SPAdes	spades.py -1 reads_1.fastq -2 reads_2.fastq -o PATH --bio Assembly parameters: k: [21, 33, 55] Repeat resolution is enabled Mismatch careful mode is turned OFF MismatchCorrector will be SKIPPED Coverage cutoff is turned OFF Other parameters: Dir for temp files: /home/hsb18158/data/KRD168_SPAdes/tmp Threads: 16 Memory limit (in Gb): 250
Pac-Bio long-reads genome assembly using Flye assembler	Flye	bin/flye --pacbio-raw reads.fastq --trestle --genome-size SIZE --iterations 10 --out-dir PATH
PacBio long-reads genome assembly using Raven assembler	Raven	raven ./reads.fastq 1> ./assembly.fa 2> ./assembly.log --graphical-fragment-assembly ./graph.gfa
PacBio long-reads genome assembly using Canu assembler	Canu	bin/canu -p LABEL -d ./PATH genomeSize=SIZE minReadLength=1000 minOverlapLength=500 corOutCoverage=40 contigFilter='2 5000 1.0 0.5 5' -pacbio ./reads.fastq
Consensus genome assembly using Flye assembler	Flye	bin/flye --subassemblies assembly1.fasta assembly2.fasta assembly3.fasta assembly4.fasta --genome-size SIZE --out-dir PATH --iterations 0 --plasmid --trestle
Illumina short-reads and PacBio long-reads genome hybrid assembly using Unicycler	Unicycler	bin/unicycler -1 reads_1.fastq -2 reads_2.fastq -l reads.fastq --existing_long_read_assembly assembly_graph.gfa -o PATH --spades_path PATH/SPAdes-3.14.1/bin/spades.py --pilon_path PATH/pilon-1.23.jar --mode normal --min_fastalength 1000 --keep 0
Reads alignment against genome assembly using Minimap2	Minimap2	minimap2 -ax sr assembly.fasta reads_1.fastq reads_2.fastq > PATH/aln.bam
Reads alignment against genome assembly using Bowtie2	Bowtie2	bowtie2-build assembly.fasta PATH bowtie2 -x PATH -1 reads_1.fastq -2 reads_2.fastq | \ samtools view -bS -o PATH/.bam
Genome shining using Pilon	Pilon	java -Xmx16G -jar PATH/pilon-1.23.jar --genome PATH/assembly.fasta --bam PATH/.bam --outdir PATH_pilon --output NAME_pilon --changes --verbose
Pairwise whole-genome Average Nucleotide Identity	fastANI	./fastANI --ql QUERY_LIST --rl REFERENCE_LIST -o PATH
Alignment coverage and percentage identity (ANIm)	PYANI	average_nucleotide_identity.py -v -i PATH/Pseudonocardia_ncbi -o PATH/Pseudonocardia_ANIm_output -g --gformat png,pdf,eps --labels PATH/Pseudonocardia_ncbi/labels.txt --classes PATH/Pseudonocardia_ncbi/classes.txt
Orthologous genes shared across genomes	PIRATE	bin/PIRATE -i Prokka -o PATH -a -r -t 4 Threshold(s): 50 60 70 80 90 95 98 MCL inflation value: 1.5 Homology test cutoff: 1E-6 Loci file contains 131624 loci from 19 genomes
Maximum Likelihood tree based on the core-genome	RAxML	bin/raxmlHPC -s core_alignment.fasta -p 1234567890 -m GTRGAMMA
Sequence similarity networks of BGCs	BiG-SCAPE	bigscape.py -i PATH/antiSMASH -o PATH/BiG-SCAPE --label Pseudonocardia --cutoffs 0.5 --clan_cutoff 0.5 0.8 --mix --include_singletons --mibig --verbose

Recombination events inferred using ClonalFrameML	ClonalFrameML	./ClonalFrameML tree.nhx alingment.fasta PATH –embranch true –emsim 100 > PATH.log.txt Rscript cfml_results.R PATH
BGC sequences alignment	Mugsy	mugsy --directory PATH --prefix NAME sequence1.fa sequence2.fa sequence2.fa … sequencen.fa

Table S2. Overview of the whole-genome assembly sequences of *Pseudonocardia* spp. used in the comparative genomic analysis. Sequences used in the pan-genome analysis are underlined.

Organism Name	Strain	Assembly	Level	Size (Mb)	GC%
Pseudonocardia abyssalis	KRD168	GCA_019263705.1	Complete	6.31	73.4
Pseudonocardia abyssalis	KRD169	GCA_019263715.1	Contig	5.98	73.3
Pseudonocardia oceani	KRD185	GCA_019263585.1	Scaffold	6.82	74.0
Pseudonocardia oceani	KRD176	GCA_019263645.1	Contig	6.62	74.0
Pseudonocardia oceani	KRD182	GCA_019263685.1	Contig	6.51	73.9
Pseudonocardia oceani	KRD184	GCA_019263655.1	Contig	6.43	73.9
Pseudonocardia oceani	KRD188	GCA_019263625.1	Contig	6.34	73.9
Pseudonocardia sp.	KRD291	GCA_019263595.1	Contig	6.57	73.0
Pseudonocardia acaciae	DSM 45401	GCA_000620785.1	Scaffold	9.93	72.3
Pseudonocardia ammonioxydans	CGMCC 4.1877	GCA_900115005.1	Scaffold	7.36	73.5
Pseudonocardia antarctica	DSM 44749	GCA_013408715.1	Contig	6.24	74.1
Pseudonocardia asaccharolytica	DSM 44247	GCA_000423625.1	Scaffold	5.06	71.8
Pseudonocardia autotrophica	DSM 535	GCA_004361965.1	Contig	7.57	72.8
Pseudonocardia autotrophica	NBRC 12743	GCA_003945385.1	Complete	7.54	72.8
Pseudonocardia broussonetiae	Gen 01	GCA_013155125.1	Complete	7.29	74.4
Pseudonocardia dioxanivorans	CB1190	GCA_000196675.1	Complete	7.30	73.2
Pseudonocardia hydrocarbonoxydans	NBRC 14498	GCA_006539565.1	Contig	5.29	74.5
Pseudonocardia oroxyli	CGMCC 4.3143	GCA_900102195.1	Scaffold	6.11	73.0
Pseudonocardia petroleophila	CGMCC 4.1532	GCA_014235185.1	Complete	6.49	74.0
Pseudonocardia sediminis	DSM 45779	GCA_004217185.1	Contig	6.61	72.9
Pseudonocardia spinosispora	DSM 44797	GCA_000429025.1	Scaffold	9.54	69.4
Pseudonocardia thermophila	DSM 43832	GCA_900142365.1	Contig	6.10	72.9
Pseudonocardia sp.	CNS-139	GCA_001942415.1	Scaffold	7.12	74.2
Pseudonocardia sp.	CNS-004	GCA_001942185.1	Contig	9.20	72.6
Pseudonocardia sp.	MH-G8	GCA_002262885.1	Scaffold	10.18	72.6
Pseudonocardia sp.	AL041005-10	GCA_001294605.1	Complete	6.14	74.4
Pseudonocardia sp.	EC080625-04	GCA_001294425.1	Complete	6.55	73.5
Pseudonocardia sp.	HH130629-09	GCA_001294645.1	Complete	6.30	73.5
Pseudonocardia sp.	Ae150A_Ps1	GCA_001932315.1	Scaffold	6.39	73.7
Pseudonocardia sp.	Ae505_Ps2	GCA_001932425.1	Scaffold	6.39	73.6
Amycolatopsis orientalis	KCTC 9412	GCA_000478275.1	Contig	9.06	69.0

Table S3. List of orthologous genes with confirmed conserved functions that were used to construct the multi-locus tree using autoMLST.

Accession number	Gene	Name
TIGR00133	gatB	aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase, B subunit
TIGR01798	cit_synth_I	citrate (Si)-synthase
TIGR00135	gatC	aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase, C subunit
TIGR00138	rsmG_gidB	16S rRNA (guanine(527)-N(7))-methyltransferase RsmG
TIGR00033	aroC	chorismate synthase
TIGR00031	UDP-GALP_mutase	UDP-galactopyranose mutase
TIGR00036	dapB	4-hydroxy-tetrahydrodipicolinate reductase
TIGR00234	tyrS	tyrosine--tRNA ligase
TIGR01203	HGPRTase	hypoxanthine phosphoribosyltransferase
TIGR00331	hrcA	heat-inducible transcription repressor HrcA
TIGR01044	rplV_bact	ribosomal protein uL22
TIGR00652	DapF	diaminopimelate epimerase
TIGR00338	serB	phosphoserine phosphatase SerB
TIGR01327	PGDH	phosphoglycerate dehydrogenase
TIGR01146	ATPsyn_F1gamma	ATP synthase F1, gamma subunit
TIGR01280	xseB	exodeoxyribonuclease VII, small subunit
TIGR01966	RNasePH	ribonuclease PH
TIGR00521	coaBC_dfp	phosphopantothenoylcysteine decarboxylase / phosphopantothenate--cysteine ligase
TIGR03594	GTPase_EngA	ribosome-associated GTPase EngA
TIGR00088	trmD	tRNA (guanine(37)-N(1))-methyltransferase
TIGR00855	L12	ribosomal protein bL12
TIGR00083	ribF	riboflavin biosynthesis protein RibF
TIGR00244	TIGR00244	transcriptional regulator NrdR
TIGR01039	atpD	ATP synthase F1, beta subunit
TIGR01134	purF	amidophosphoribosyltransferase
TIGR00086	smpB	SsrA-binding protein
TIGR01030	rpmH_bact	ribosomal protein bL34
TIGR00690	rpoZ	DNA-directed RNA polymerase, omega subunit
TIGR01032	rplT_bact	ribosomal protein bL20
TIGR00019	prfA	peptide chain release factor 1
TIGR00396	leuS_bact	leucine--tRNA ligase
TIGR00012	L29	ribosomal protein uL29
TIGR01978	sufC	FeS assembly ATPase SufC
TIGR00150	T6A_YjeE	tRNA threonylcarbamoyl adenosine modification protein YjeE
TIGR00152	TIGR00152	dephospho-CoA kinase
TIGR00154	ispE	4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase
TIGR00313	cobQ	cobyric acid synthase CobQ
TIGR00981	rpsL_bact	ribosomal protein uS12
TIGR00090	rsfS_iojap_ybeB	ribosome silencing factor
TIGR00959	ffh	signal recognition particle protein
TIGR01308	rpmD_bact	ribosomal protein uL30
TIGR01302	IMP_dehydrog	inosine-5'-monophosphate dehydrogenase
TIGR01029	rpsG_bact	ribosomal protein uS7
TIGR01021	rpsE_bact	ribosomal protein uS5
TIGR01024	rplS_bact	ribosomal protein bL19
TIGR00382	clpX	ATP-dependent Clp protease, ATP-binding subunit ClpX
TIGR00061	L21	ribosomal protein bL21
TIGR00639	PurN	phosphoribosylglycinamide formyltransferase
TIGR01410	tatB	twin arginine-targeting protein translocase TatB
TIGR00635	ruvB	Holliday junction DNA helicase RuvB
TIGR00581	moaC	molybdenum cofactor biosynthesis protein C

TIGR01011	rpsB_bact	ribosomal protein uS2
TIGR01850	argC	N-acetyl-gamma-glutamyl-phosphate reductase
TIGR00577	fpg	DNA-formamidopyrimidine glycosylase
TIGR01169	rplA_bact	ribosomal protein uL1
TIGR00928	purB	adenylosuccinate lyase
TIGR00482	TIGR00482	nicotinate (nicotinamide) nucleotide adenylyltransferase
TIGR00962	atpA	ATP synthase F1, alpha subunit
TIGR00922	nusG	transcription termination/antitermination factor NusG
TIGR01164	rplP_bact	ribosomal protein uL16
TIGR01009	rpsC_bact	ribosomal protein uS3
TIGR00168	infC	translation initiation factor IF-3
TIGR02012	tigrfam_recA	protein RecA
TIGR00281	TIGR00281	segregation and condensation protein B
TIGR00166	S6	ribosomal protein bS6
TIGR00414	serS	serine--tRNA ligase
TIGR02422	protocat_beta	protocatechuate 3,4-dioxygenase, beta subunit
TIGR00560	pgsA	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase
TIGR00419	tim	triose-phosphate isomerase
TIGR01171	rplB_bact	ribosomal protein uL2
TIGR00810	secG	preprotein translocase, SecG subunit
TIGR00952	S15_bact	ribosomal protein uS15
TIGR00119	acolac_sm	acetolactate synthase, small subunit
TIGR00114	lumazine-synth	6,7-dimethyl-8-ribityllumazine synthase
TIGR00116	tsf	translation elongation factor Ts
TIGR00059	L17	ribosomal protein bL17
TIGR00708	cobA	cob(I)yrinic acid a,c-diamide adenosyltransferase
TIGR01464	hemE	uroporphyrinogen decarboxylase
TIGR00468	pheS	phenylalanine--tRNA ligase, alpha subunit
TIGR01737	FGAM_synth_I	phosphoribosylformylglycinamidine synthase I
TIGR01066	rplM_bact	ribosomal protein uL13
TIGR02727	MTHFS_bact	5-formyltetrahydrofolate cyclo-ligase
TIGR02729	Obg_CgtA	Obg family GTPase CgtA
TIGR00020	prfB	peptide chain release factor 2
TIGR00228	ruvC	crossover junction endodeoxyribonuclease RuvC
TIGR00670	asp_carb_tr	aspartate carbamoyltransferase
TIGR00029	S20	ribosomal protein bS20
TIGR03263	guanyl_kin	guanylate kinase
TIGR03635	uS17_bact	ribosomal protein uS17
TIGR00184	purA	adenylosuccinate synthase
TIGR03632	uS11_bact	ribosomal protein uS11
TIGR01051	topA_bact	DNA topoisomerase I
TIGR01358	DAHP_synth_II	3-deoxy-7-phosphoheptulonate synthase

Table S4. Buffers composition used in the modified media for the pH tolerance assay.

pH	Component	Molarity (mol/L)
4	Sodium citrate dihydrate	0.0335
	Citric acid	0.0665
5	Sodium citrate dihydrate	0.0577
	Citric acid	0.0423
6	K_2HPO_4	0.0138
	KH_2PO_4	0.0862
7	K_2HPO_4	0.0536
	KH_2PO_4	0.0464
8	K_2HPO_4	0.0935
	KH_2PO_4	0.0065
9.2	Sodium bicarbonate	0.0910
	Sodium carbonate	0.0090
10	Sodium bicarbonate	0.0461
	Sodium carbonate	0.0539

Table S5. Contiguity and quality evaluation of consensus long-reads assembly produced by the subassembly mode of Flye for KRD185.

Contigs	Largest contig (bp)	Length (bp)	N50 (bp)	Mismatches per 100 kbp*	Indels per 100 kbp*	BUSCO (% complete single-copy genes)
3	6661213	6807015	6661213	16.07	2.22	99.7%

* Mismatched statistics calculated by aligning assemblies with the short-reads assembly.

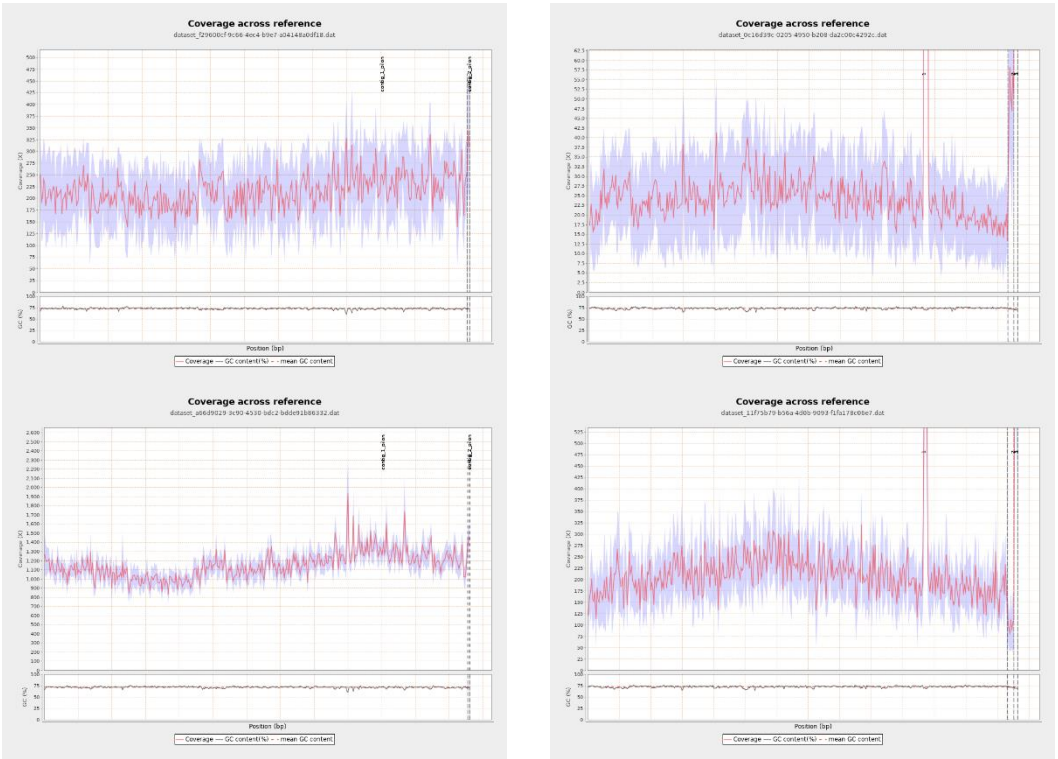


Figure S1. Coverage of Illumina short-reads (upper) and PacBio long-reads (lower) across the final assembly of *P. abyssalis* KRD-168 (left) and *P. oceani* RKD-185 (right).

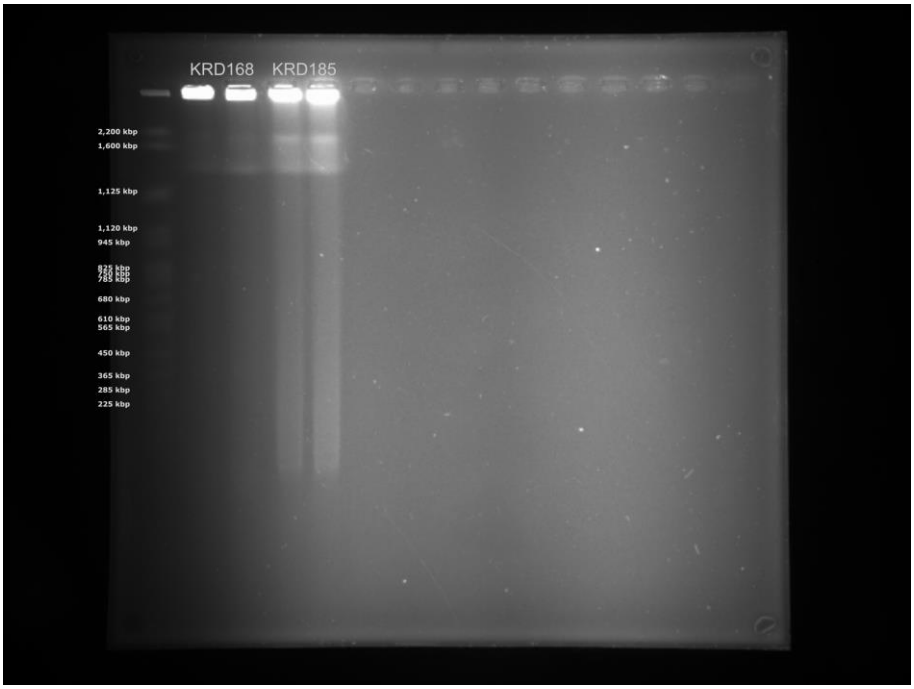


Figure S2. Pulsed field gel electrophoresis profile of *P.abyssalis* KRD168 (lanes 2 and 3, duplicate) and *P. oceani* KRD185 (lanes 4 and 5, duplicate). CHEF DNA Size Marker, 0.2–2.2 Mb, *S. cerevisiae* Ladder (Bio-Rad) is shown in lane 1.

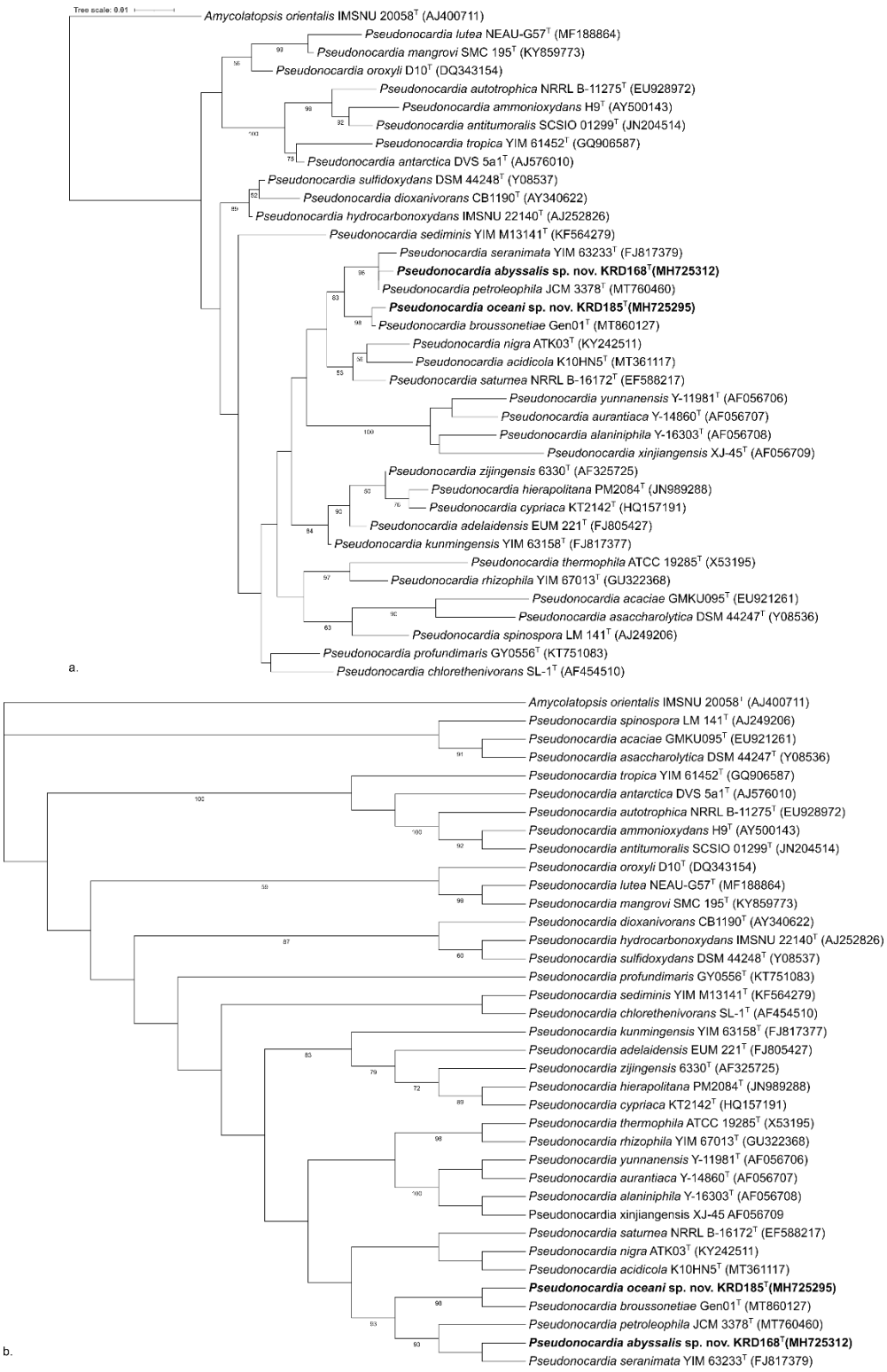


Figure S3. Bootstrap consensus tree inferred from 1000 replicates. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The evolutionary history was inferred using the Maximum Likelihood (a), and Maximum Parsimony (b).

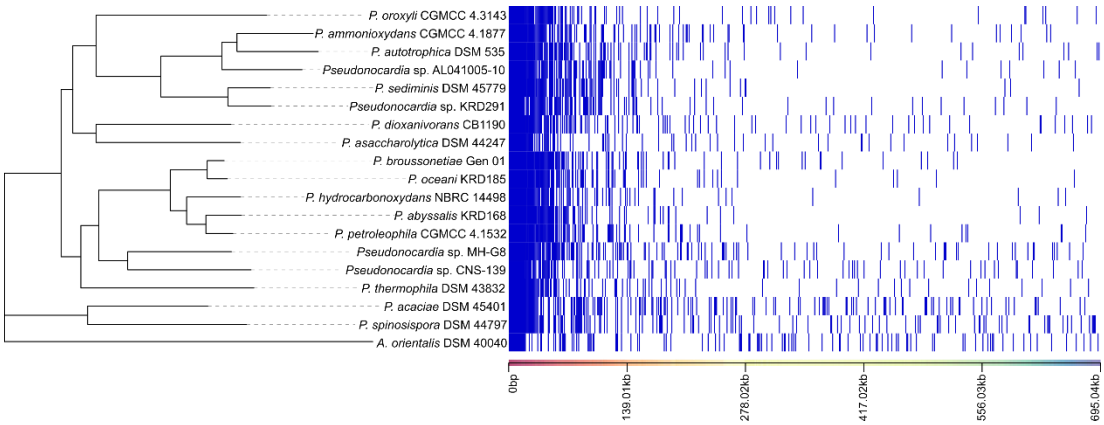


Figure S4. Shared gene presence per genome ordered alongside a phylogenetic tree generated from the core gene alignment. Gene family presence is indicated by a blue block per column.

Table S6. Recombination events per branch in the maximum-likelihood reconstructed phylogeny from the core-genome.

Branch	R/θ	ν	δ	r/m
Mean	1.12426	0.032267	98.31778	3.56657
A. orientalis DSM 40040	0.019742	0.092411	0.67787	0.001237
P. abyssalis KRD168	0.885051	1.56866	1.886661	2.619334
P. acacia DSM 45401	0.049955	0.134595	0.732113	0.004922
P. ammonioxydans CGMCC 4.1877	0.080107	0.270664	0.824606	0.017879
P. asaccharolytica DSM 44247	0.026116	0.077032	0.891742	0.001794
P. autotrophica DSM 535	0.060578	0.208001	0.884588	0.011146
P. broussonetiae Gen01	0.952231	0.956078	1.385792	1.261635
P. dioxanivorans CB1190	0.053316	0.224755	0.66565	0.007977
P. hydrocarbonoxydans NBRC 14498	1.51084	2.20046	1.456109	4.840895
P. oceani KRD185	1.10156	0.988462	1.774651	1.932329
P. oroxyli CGMCC 4.3143	0.038391	0.145905	0.623076	0.00349
P. petroleophila CGMCC 4.1532	0.790546	1.42476	1.327826	1.495581
P. sediminis DSM 45779	1.30581	1.59262	2.291943	4.76646
Pseudonocardia sp. AL041005-10	0.046078	0.092069	1.246665	0.005289
Pseudonocardia sp. CNS-139	3.79723	3.10502	4.068746	47.97244
Pseudonocardia sp. KRD291	1.35645	1.59535	2.316477	5.012885
Pseudonocardia sp. MH-G8	0.038759	0.109268	0.885112	0.003749
P. spinosispora DSM 44797	0.031164	0.12765	0.819383	0.00326
P. thermophila DSM 43832	0.038762	0.131119	0.608702	0.003094
P. oceani and *P. abyssalis* clade	0.168896	0.201924	0.546293	0.018631
P. sediminis clade	0.164837	0.291064	0.641976	0.030801

R/θ = relative rate of recombination to mutation; ν = mean import length of recombining DNA; δ = mean divergence of imported recombining DNA; r/m = relative effect of recombination and mutation (r/m = (R/θ) × δ × ν).

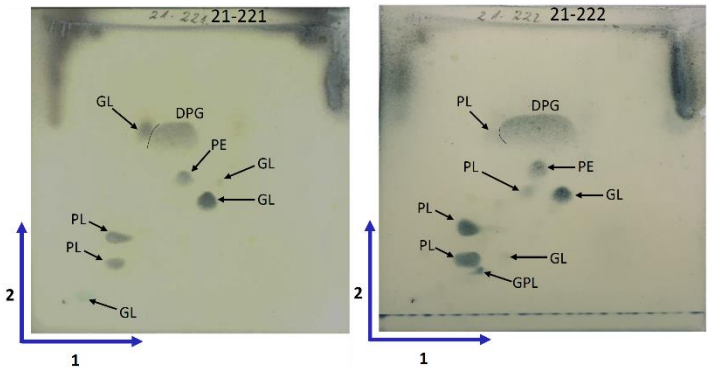


Figure S5. Polar lipid profile of 1, *Pseudonocardia abyssalis* sp. nov. KRD168^{T}; and 2, *Pseudonocardia oceani* sp. nov. KRD185^{T}; after two-dimensional thin-layer chromatography (TLC). PE, phosphatidylethanolamine; DPG, diphosphatidylglycerol; PL, unidentified phospholipid; GL, unidentified glycolipid; GPL, unidentified glycophospholipid.

Table S7. Strain collection used in the metabolomics analysis. Accession numbers for sequenced strains are shown in Table S2.

Organism Name	Strain	WGS availability
Pseudonocardia abyssalis	KRD168	Yes
Pseudonocardia abyssalis	KRD169	Yes
Pseudonocardia oceani	KRD176	Yes
Pseudonocardia oceani	KRD182	Yes
Pseudonocardia oceani	KRD184	Yes
Pseudonocardia oceani	KRD185	Yes
Pseudonocardia oceani	KRD188	Yes
Pseudonocardia sp.	KRD291	Yes
Pseudonocardia hydrocarbonoxydans	DSM 43281	Yes
Pseudonocardia petroleophila	DSM 43193	Yes
Pseudonocardia sediminis	DSM 45779	Yes
Pseudonocardia ammonioxydans	DSM 44958	Yes
Pseudonocardia antitumoralis	DSM 45322	No
Pseudonocardia profundimaris	KCTC 39641	No
Pseudonocardia sp.	CNS-139	Yes
Pseudonocardia sp.	VO44-3	No

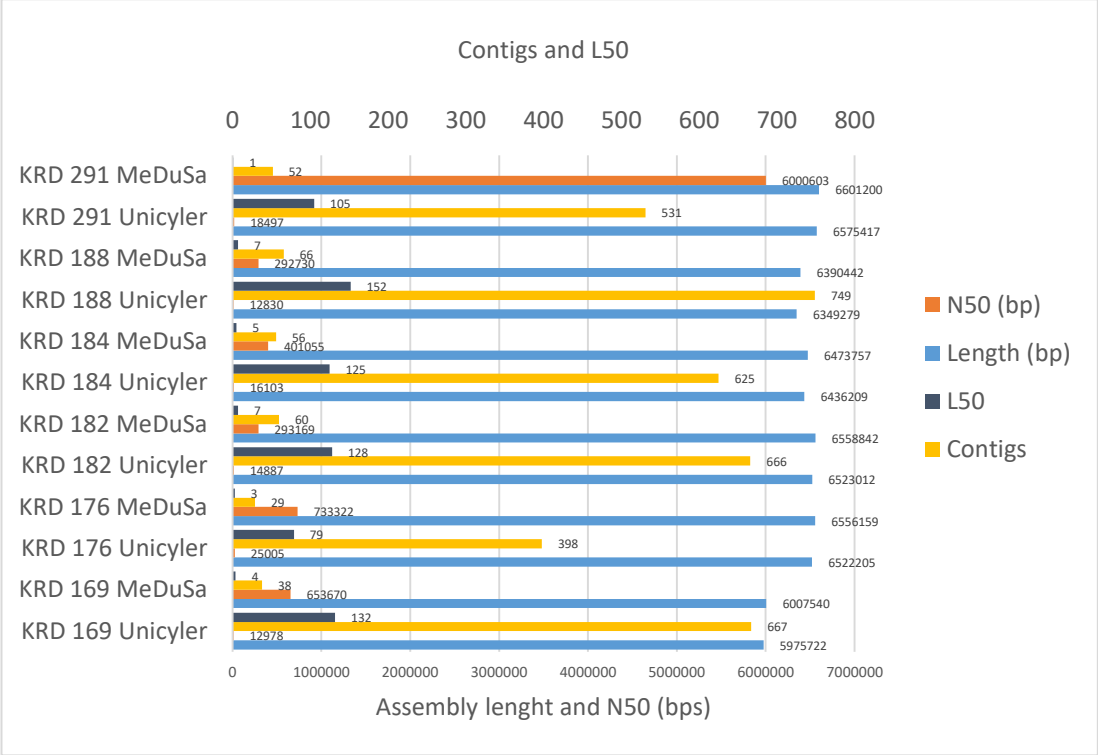


Figure S6. Genome quality evaluation for the Southern Ocean strains assemblies before and after scaffolding by MeDuSa.

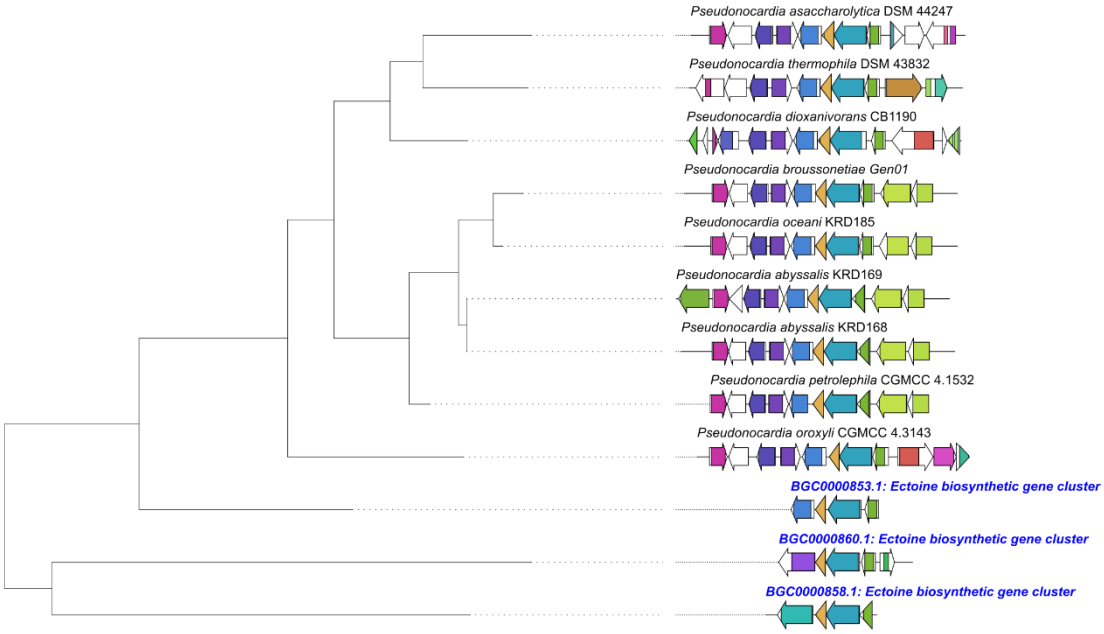


Figure S7. Maximum-likelihood multi-locus phylogenetic tree of the GCF *ect1* reconstructed with CORASON. Genes are represented by arrows facing the direction of transcription. The tree was created using the sequences of the core domains in the GCF.

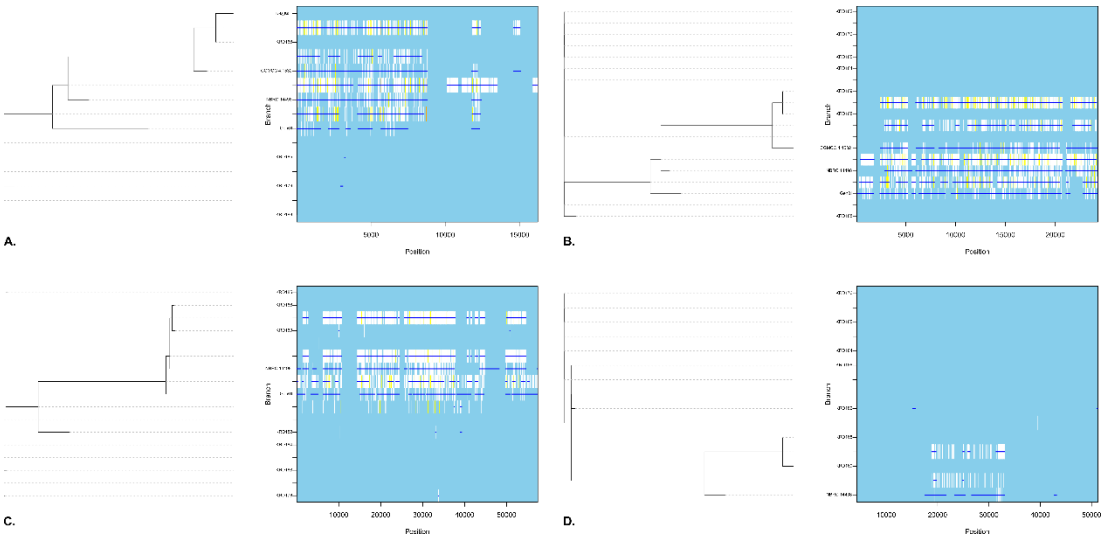


Figure S8. Clonal phylogeny showing the inferred recombination events as dark blue horizontal bars in *ripp1* (**A**), *terp1* (**B**), *nrps1* (**C**), and *bet1* (**D**).

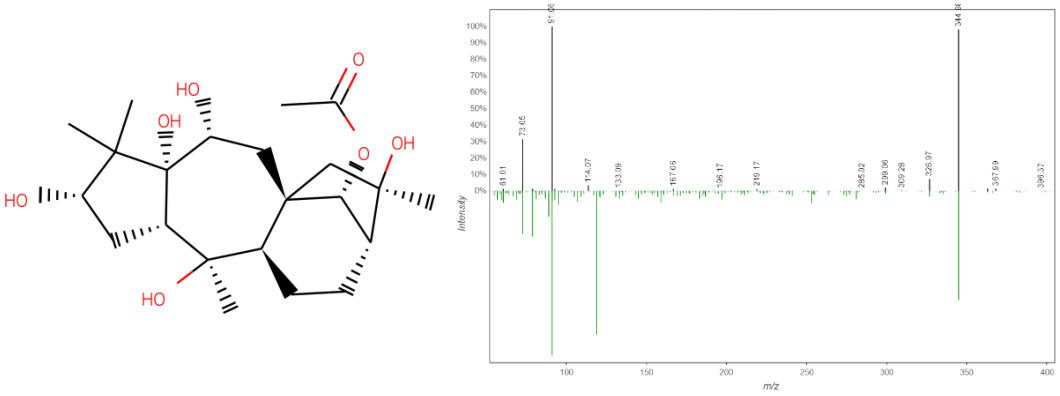


Figure S9. Structure of grayanotoxin I and the mirror plot visualising the matching spectra (library spectrum in green).

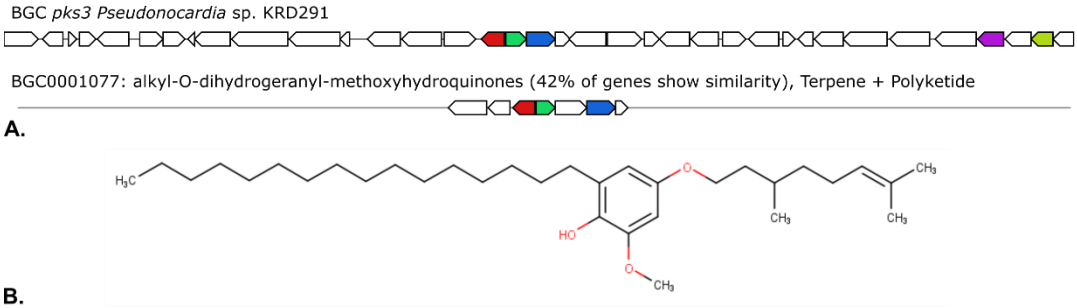


Figure S10. Gene similarity between the BGC *pks4* from *Pseudonocardia* sp. KRD291 and BGC0001077 from *Actinoplanes missouriensis* 431 (**A**), alongside a prenylated phenolic lipid product of the latter.

Figure S11. Bioactivity of *P. abyssalis* KRD168 cultivate in ISP2 against *Candida auris* UACa11. Strain was cultured in a 90 mm Petri dish.

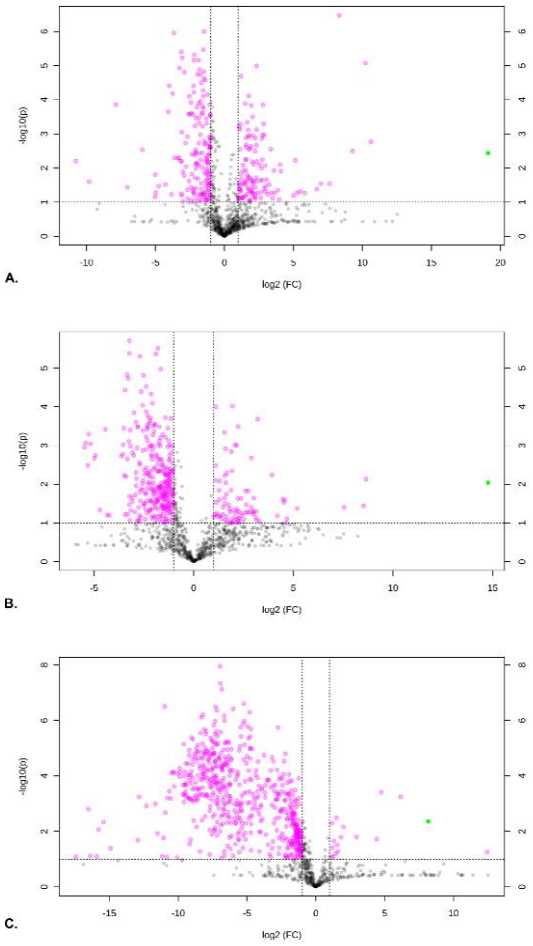


Figure S12. Volcano plot showing the differential metabolite identification of *P. hydrocarbonoxydans* DSM 43281, *P. petroleophila* DSM 43193 (**B**), and *P. ammonioxydans* DSM 44958 (**C**) cultured in ISP2 and ISP2 supplemented with *N*-acetyl glucosamine 1mM. Each dot represents a molecular feature detected in the LC-MS/MS analysis. The x-axis represents the ratio of change (fold change), and the y-axis the statistical significance (p-value). Molecular features with p-values < 0.1 (horizontal dotted line) and a fold change < 0.5 or > 2 (vertical dotted lines) are highlighted in magenta. Molecular features considered significantly down-regulated are showed in the left side, and those considered significantly up-regulated are showed in the right side. The molecular feature corresponding to *N*-acetyl glucosamine is highlighted in green.

Table S8. Homologous proteins to *dasR* (WP_009951283) identified in the Southern Ocean strains genomes by standard protein BLAST

Total score	Query cover (%)	Sequence
93.2	89	>WP_218611789.1 UTRA domain-containing protein [Pseudonocardia sp. KRD291] MGERTLVRGNGQPLWQQLLDDLRARLDGGDFDAALPSEFALAAEYSVSRHTVREA LRRMRDDGAVVATRGRVARRAEPTPIEQPLSEPYSLYRAVEAAGRTQRSVVRQFDV RADGVIAARLGLEESTPLVHLERLRLADSEPLAVDRVWLPEELAAPLLDVDFSHTGLY DEYRRRCGIRVTGGRENIRAVVPGEGERQLLGIGDGVAALVIERCGMAEGRAVEW RHTVVRGDRFTVTAQFSDRTGYRVDLAE
72.0	80	>WP_218605969.1 UTRA domain-containing protein [Pseudonocardia abyssalis] MAAPLHRTLAAELRRRIRSGAIGIGESLPSEAQLCREFAASRGPVRQALAALRDEGLI GGGQGKRPVVLDAVPAQPFESFQSFTRRAELTGHVPGQRLQEIALRRPEPAVAAAL QLDDDTPAVQLLRLRLLDGRPAMLERMTYVESVGRALLDADLDAGSIYALLTSHGV DLHAARHTFDAVAADATDSALLEVPVGSPLLRERRLTSDSSGRPLEWSDDRYRSDV ATVTVTNTRSSRAPVTRL
67.0	89	>WP_218591783.1 UTRA domain-containing protein [Pseudonocardia oceani] MPEIQKVLPKYLQIAGHIRDQIVRGDLSPGDEVPSERELAARWSVARPTAARALESL RVQGLVESRQGSGTYVRASQSAPRARERYERGRDLGTMYGATESVEFLATDIVAGV DHVVEALQLPPGSDVIRRIRLLRGEDGKPIELSTSWFAPHLAEQAPRLLHPERLLGGT GKYIAEVTGRESEYARDQVAARLATADERRLLELPRPAAVLVYRLTAYDSSDLPIQFD DATYPSDRWAFRQEYPLAR
63.9	81	>WP_218601160.1 UTRA domain-containing protein [Pseudonocardia abyssalis] MTSSEGGDRRPASERLADELRVSIESGELAAGAKLPSERELAEFHGIARNTARQAIRL LAESGLVIAEHGRGVFVRPAISVIRLGNDRYSPRYRDTGLSPFLLECAKQGKAGRFEV LSVERVTPPADVAARLKISPDTPSALRRENVFWADADPVQRVTTWLPWALADSTAL LRDEVGHPFGIHGILEERGHLMARIHDEISARMPTPDEREHLQLSPGVPVLDVLHTSI DTEGEPYELTRFVMRADLSGLRYDAPVE

Table S9. Origin of *Pseudonocardia* spp. strains used in the chemical ecological experiment showing environment, location, depth and GPS coordinates of their origin.

Species	Strain	Environment	Location	Depth	Coordinates
P. abyssalis	KRD168	Marine sediments	Southern Ocean	4539 m	62° 57.56’ S 27° 53.23’ W
P. oceani	KRD185	Marine sediments	Southern Ocean	4060 m	65° 19.88’ S 48° 5.58’ W
Pseudonocardia sp.	KRD291	Marine sediments	Southern Ocean	4060 m	65° 19.88’ S 48° 5.58’ W
P. sediminis	DSM 45779	Marine sediments	South China Sea	652 m	21° 23.20’ N 117° 37.20’ E
P. petroleophila	DSM 43193	Soil	-	-	-
P. hydrocarbonoxydans	DSM 43281	Laboratory/Industrial sub-product	-	-	-
Pseudonocardia sp.	CNS-139	Marine sediments	Palau	8 m	7° 30’ N, 134° 30’ E
Pseudonocardia sp.	VO44-3	Marine sediments	Valparaíso Bay	29.4 m	33° 1.18' S 71° 38.72' W
P. profundimaris	KCTC 39641	Marine sediments	Western Pacific Ocean	7118 m	141° 59.7’ E 11° 59.7’ N
P. antitumoralis	DSM 45322	Marine sediments	Northern South China Sea	3258 m	120° 0.98' E 19° 0.66' N
P. ammonioxydans	DSM 44958	Coastal sediments	Jiao-Dong peninsula	0 m	35° 58.8’ N 120° 14.2’ E

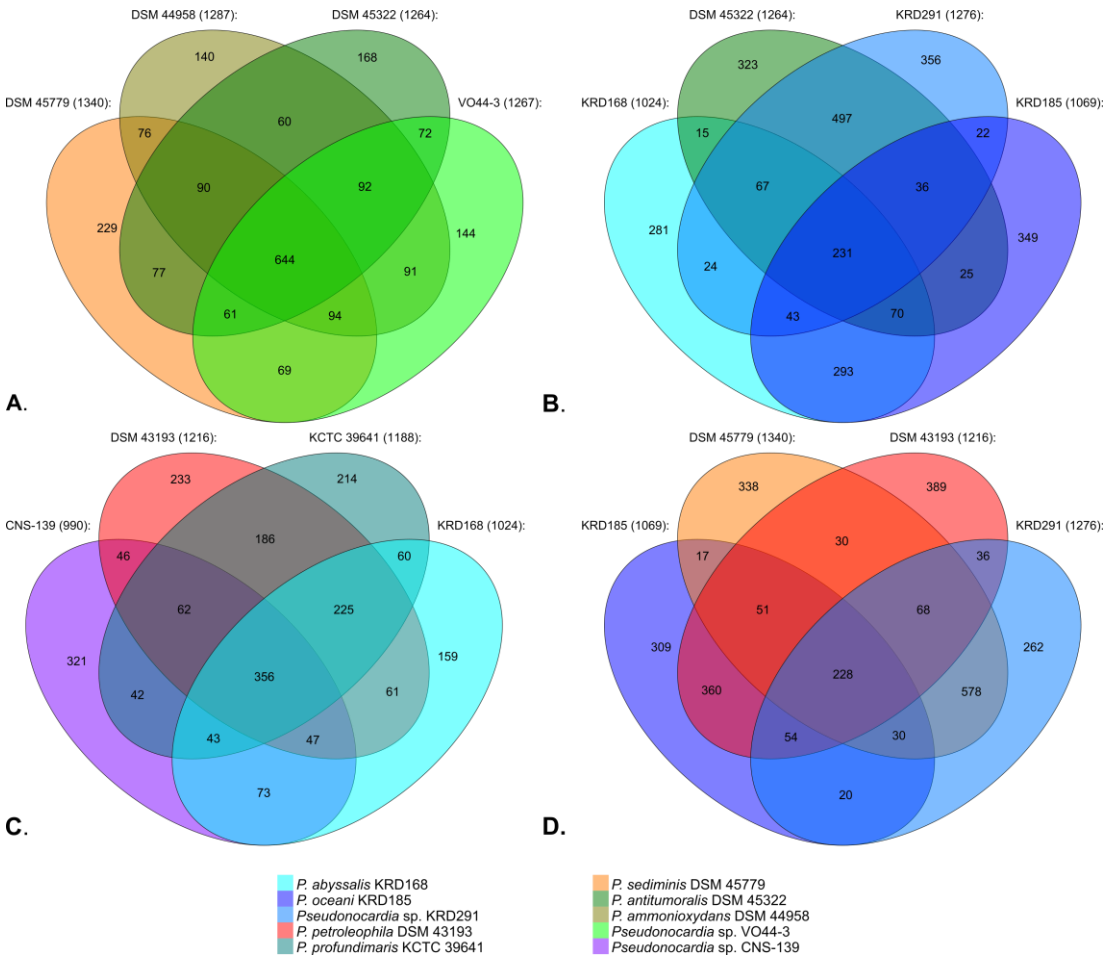


Figure S13. Venn diagram showing the node distribution from the GNPS classical molecular networking showing parent ions detected. Parent ions with similar MS/MS spectra are clustered in the same node and can be interpreted as metabolites. Shared parent ions potentially mean metabolites in common. The total number of parent ions for each strain are shown in brackets.

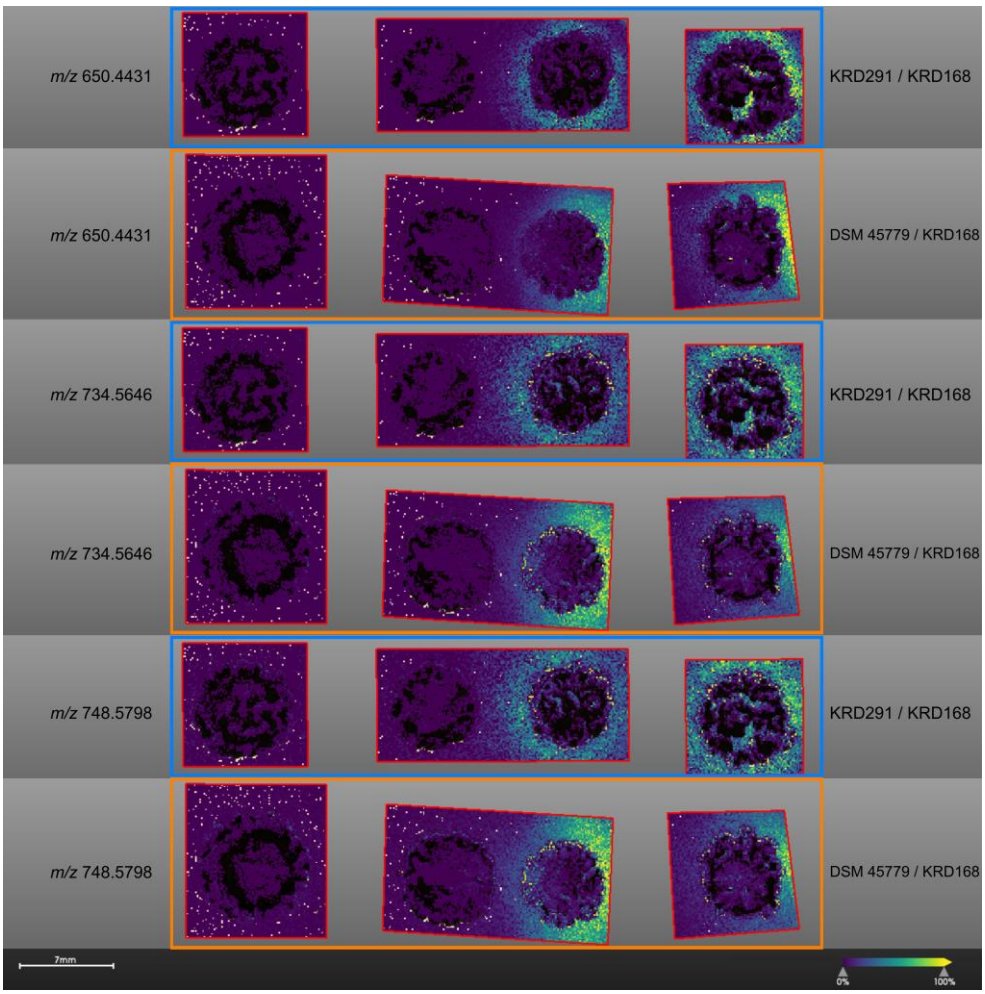


Figure S14. Image from MALDI-TOF MSI of *P. abyssalis* KRD168 (right) interacting with *Pseudonocardia* sp. KRD291 and *P. sediminis* DSM 45779 (left), featuring the ion image taken of *m/z* 650.4431, *m/z* 734.5646, and *m/z* 748.5798.

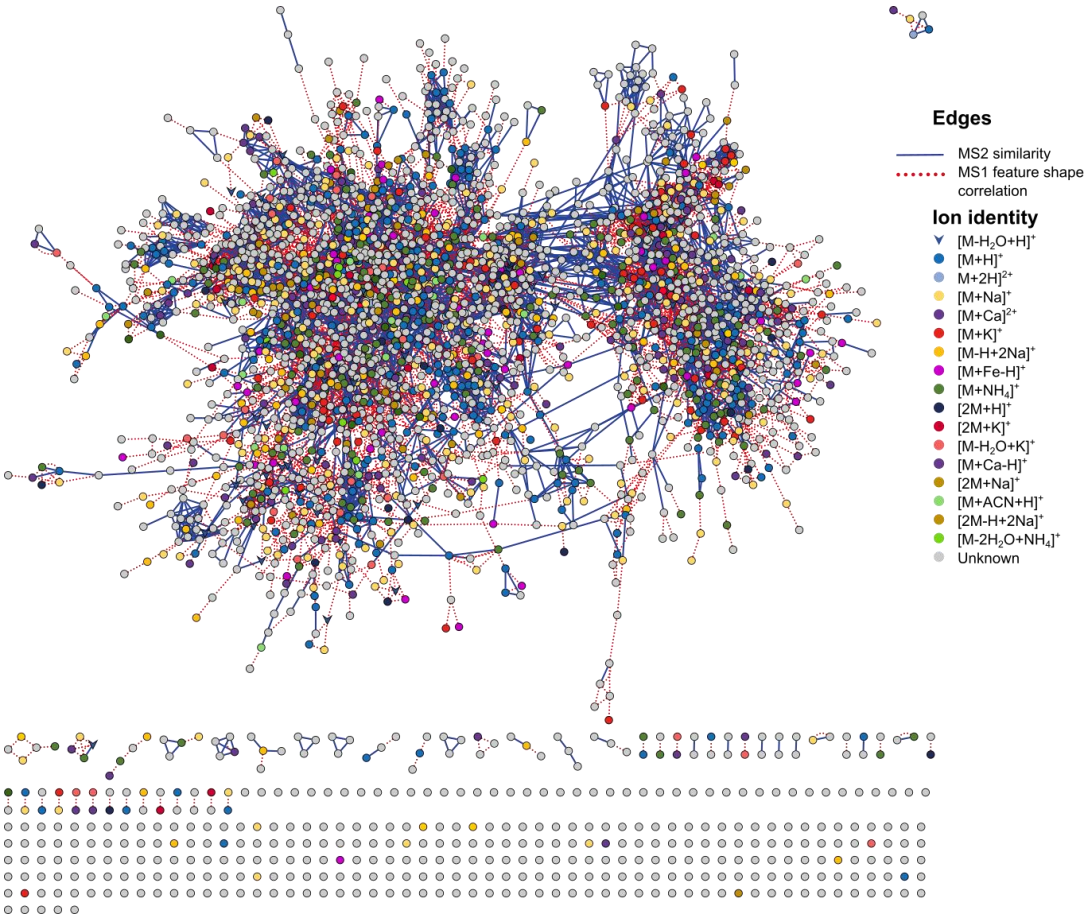


Figure S15. Ion identity molecular network of 3023 features showing the predicted ion identity for each node. Molecular features are coloured according to their ion species. The network contains MS2 similarity edges from molecular networking as solid blue lines combined with MS1 ion identity networking edges as dotted red line.

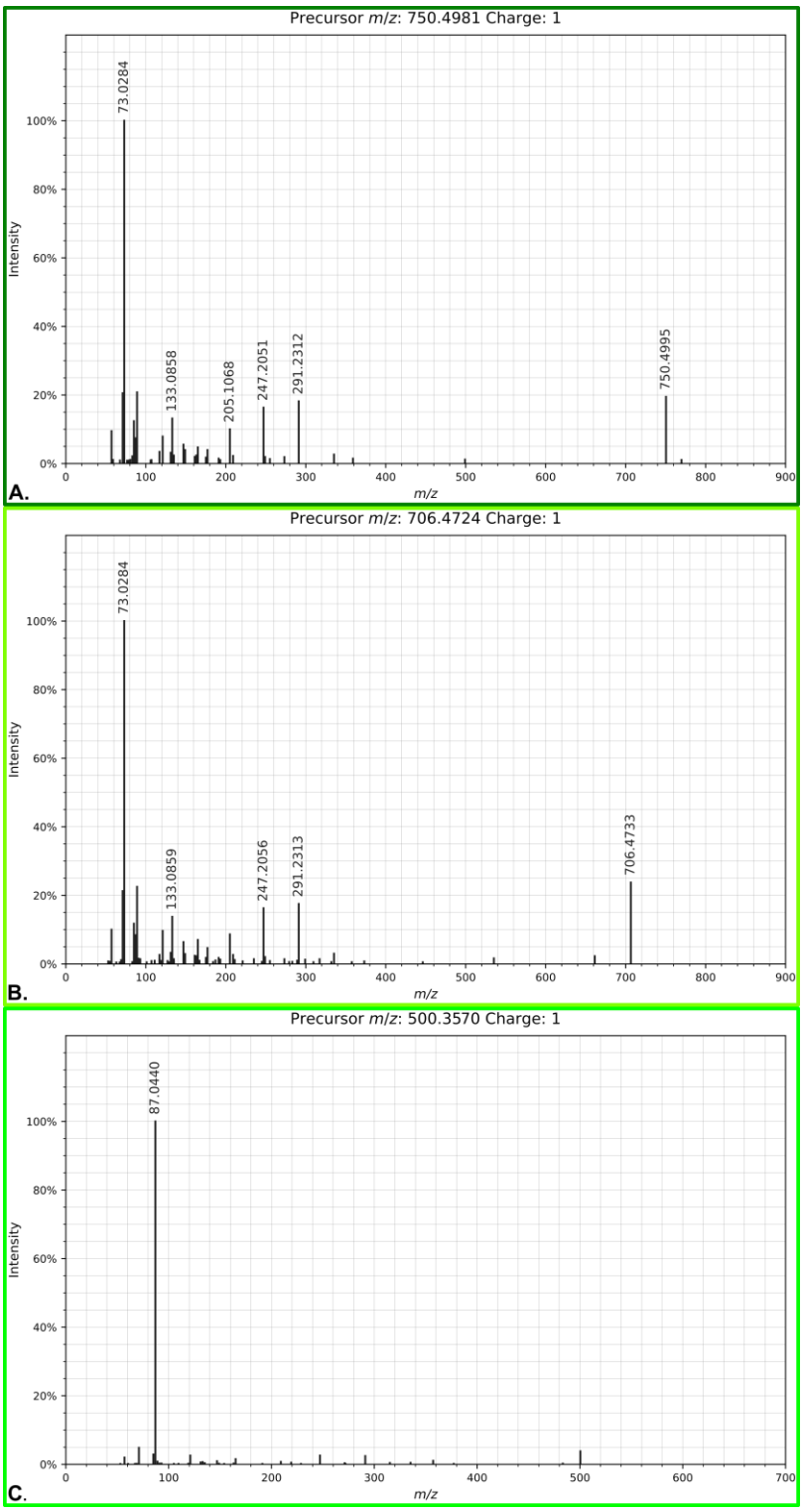


Figure S16. MS2 spectra of the molecular nodes with precursor ions of *m/z* 750.4981 (**A**), *m/z* 706.4724 (**B**), and *m/z* 500.3570 (**C**) from the IIMN analysis. The y-axis represents the percentage of intensity for each signal, and the x-axis the mass-to-charge ratio (*m/z*).