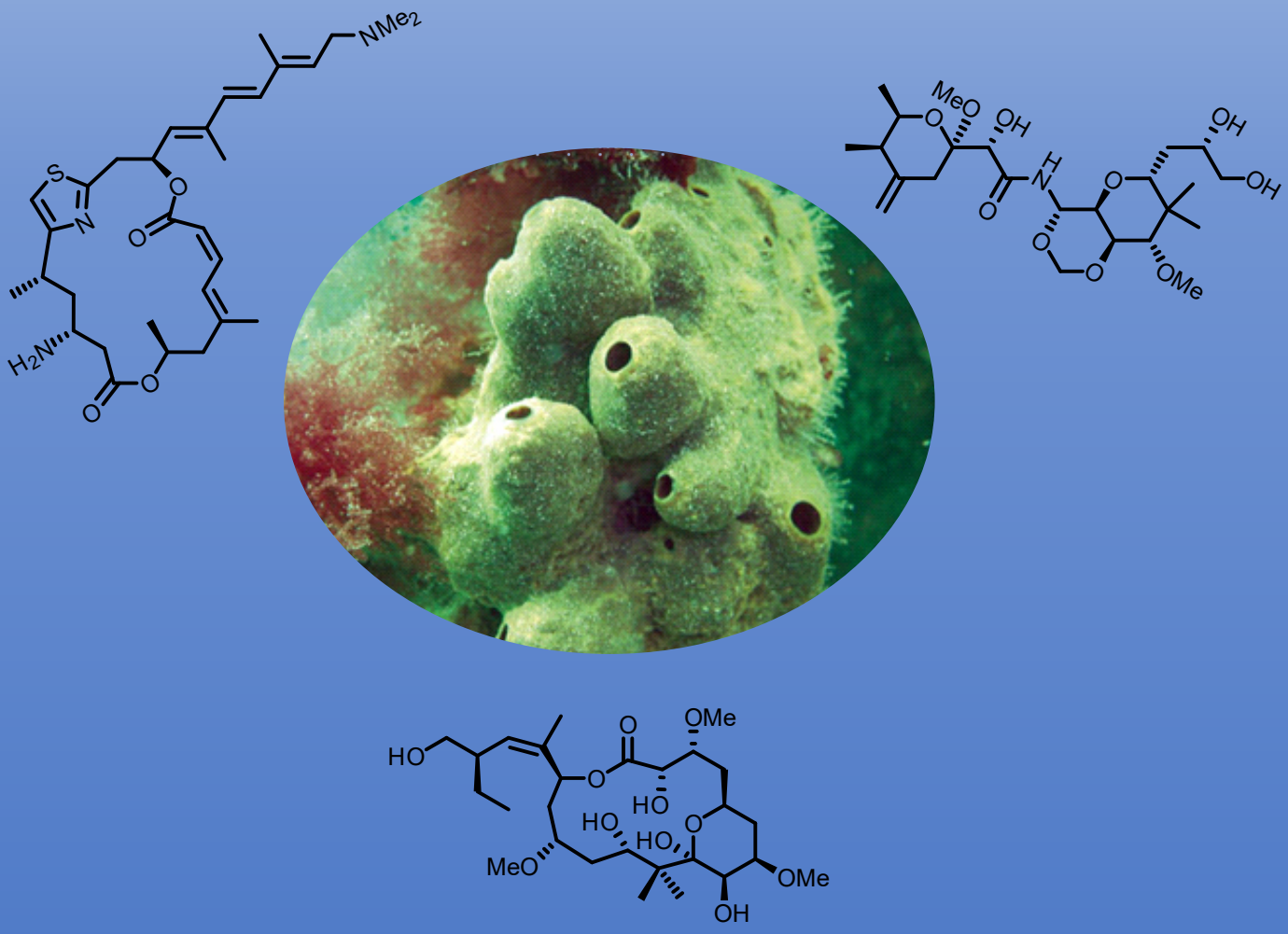
DISS ETH NO. 27437

IDENTIFICATION AND CHARACTERIZATION OF NATURAL PRODUCT BIOSYNTHETIC PATHWAYS ENCODED IN THE MICROBIOME OF THE SPONGE *MYCALE HENTSCHELI*

MICHAEL RUST



2021

DISS ETH NO. 27437

IDENTIFICATION AND CHARACTERIZATION OF NATURAL PRODUCT BIOSYNTHETIC PATHWAYS ENCODED IN THE MICROBIOME OF THE SPONGE *MYCALE HENTSCHELI*

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

MICHAEL RUST

MSc Interdisciplinary Sciences, ETH Zurich

born on 11.11.1991

citizen of Walchwil ZG, Switzerland

accepted on the recommendation of

Prof. Dr. Jörn Piel

Prof. Dr. Shinichi Sunagawa

Prof. Dr. Peter Kast

2021

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor **Prof. Jörn Piel** for his continuous support and motivation throughout my time in his lab. Ever since attending his inaugural lecture in 2013, I was fascinated by natural products and joined the lab as a bachelor student soon after, came back for my master thesis, and continued to pursue a PhD. I am thankful for his guidance on all aspects of my projects and feel honored to have worked with a person having an immense knowledge and being so humble at the same time. I thank my co-referees **Prof. Shinichi Sunagawa** and **Prof. Peter Kast** for valuable input and advice. I am deeply grateful to **Prof. Shinichi Sunagawa** for his continuous support and new ideas related to bioinformatics and the fruitful collaborations with his lab.

My biggest thanks go to ***current and former members of the Piel group*** for creating an exceptional working atmosphere. I could not have envisioned a better place for my PhD. First, I would like to sincerely thank **Dr. Eric Helfrich** who was my supervisor during my student projects and remained a mentor and good friend throughout my PhD. You combine all qualities to become a successful group leader. A special thank goes to **Dr. Michael Freeman** for his teaching and guidance during my time as a student. I want to thank **Silke Probst** ☕ for making our time inside and outside the lab truly special. Having started in the group at the same time, you have become an amazing friend and significantly contributed to my research, sportiness, and wellbeing. I am very thankful to **Dr. Alexander Brachmann** for insightful conversations and all his help with mass spectrometry-related questions. I would like to thank **Dr. Amy Fraley** and **Dr. Franziska Hemmerling** for inspiring discussions about polyketides and the great collaborative work. Additionally, I am very thankful to all polyketide-people who provided valuable ideas for numerous projects: **Dr. Eric Helfrich**, **Stefan Leopold-Messer**, **Dr. Roy Meoded**, **Hannah Minas**, **Silke Probst**, and **Dr. Silke Reiter**. I had the great pleasure of working with **Dr. Pakjira Nanudorn**, thank you for the great work. Additionally, I would like to thank my students **Nemo Milos**, **Tomas Kündig**, and **Mariella Greutmann** for their contributions to many projects. It was fun working with you and I learned a lot about teaching.
I was lucky to share an office with wonderful people who always found time to discuss science or other things over a cup of coffee – **Dr. Alexander Brachmann**, **Dr. Ana Flávia Canovas Martinez**, **Dr. Franziska Hemmerling**, **Dr. Florian Hubrich**, **Sarolt Magyari**, **Dr. Silke Reiter** and **Dr. Serina Robinson**.
I am very thankful to my sporty colleagues for keeping me fit and healthy: **Edgars**, **Florian**, **Niklas**, **Roy**, **Sarolt**, **Silke**, and **Stefan** for TRX and muscle pump sessions; **Alessandro**, **Christine** and **Eric** for fun on the tennis court; **Chandrashekhar** and **Pakjira** for being awesome badminton partners; **Agneya** and **Silke** for intense squash games; **Alessandro**, **Silke**, **Stefan**, and **Tom** for being great hiking and boating buddies.

A special thank goes to **Dr. Jeremy Owen** for hosting me in his lab in New Zealand and organizing our collection trip. It was very special to see the native environment of the sponge I had been working on the past years. Thanks to **Dr. Robert Keyzers** and **Dr. Mike Page** for joining the trip and bringing the samples to the surface.
A huge thank goes to **Dr. Christopher Field** for all his support with analyzing sequencing data. Many thanks to **Dr. Charlotte Carlström** and **Salome Hegi** for their help with establishing the Samplix workflow.

I am deeply grateful to ***the whole staff team at the Institute of Microbiology*** for dealing with countless things that made my daily work much easier.

My sincere gratitude goes to **Dr. Alexander Brachmann**, **Dr. Amy Fraley**, **Dr. Serina Robinson**, and **Dr. Thomas Scott** for highly appreciated advice for the improvement of this thesis.

I want to thank my flat mates **Rea** and **Stefan** for joining me on this journey and for vivid discussions on science and many other things.

Most importantly, I would like to thank **my family**, **Franzi**, and **friends** for their loving support and continuous motivation. **My parents** deserve the biggest thanks for the infinite energy they invested in my education and wellbeing throughout my life. This thesis is dedicated to them.

Table of Contents

Abstract

Specialized metabolites from natural sources (animals, plants, microbes) have been pillars of traditional and modern medicine alike. The structural diversity of natural products gives rise to various bioactivities, which are exploited to combat infectious diseases, cancer, and numerous other illnesses. Traditional bioactivity-guided approaches led to the identification of numerous blockbuster molecules, but rapidly became limited by the frequent re-isolation of known metabolites. Technological advances have enabled the exploration of new habitats and the mining of genome data for the discovery of novel natural products. Within the marine environment, sponges have been a particularly rich source of specialized metabolites. Sponges, which represent the oldest extant animals, co-exist with astonishing numbers of microorganisms. For some species, it has been shown that microbial symbionts produce bioactive molecules known from the sponge host.
This work centers around the biosynthetic potential encoded in the microbiome of the chemically rich sponge *Mycale hentscheli*. We pursued diverse strategies to (i) identify biosynthetic genes, (ii) link them to complete biosynthetic gene clusters (BGCs) and predict the structure of encoded products, (iii) assign the BGCs to producing organisms, and (iv) provide functional data by expressing individual enzymes in heterologous hosts.

In **Chapter II**, we demonstrated that multiple symbionts jointly generate the rich chemistry in *M. hentscheli*, a contrasting scenario to other sponges, in which single "superproducer" symbionts within more complex microbiomes synthesize most of the bioactive compounds known from their host. In addition to bacterial pathways for the three cytotoxic polyketides (mycalamide, pateamine, and peloruside), we identified numerous orphan BGCs distributed across a broad phylogenetic range of bacteria. Characterization of enzymes from an orphan polytheonamide-type pathway suggests a novel member of this rare family as its product. The microbiome is particularly rich in pathways of a distinct family of polyketide synthases (PKSs), termed *trans*-acyltransferase (*trans*-AT) PKSs, that are often found in uncultivated symbiotic bacteria, reinforcing uncultured microbes as promising sources of chemical novelty.

Uncultivated sponge symbionts often harbor orphan BGCs, raising intriguing questions about their origin and function. Strategies to study them are usually limited to cultivation approaches or heterologous expression studies. In **Chapter III**, we pursued an alternative strategy by searching for similar clusters in sequenced genomes of cultured bacteria. We analyzed a small hybrid cluster consisting of non-ribosomal peptide synthetase and PKS parts in the metagenome of *M. hentscheli* and identified orthologous clusters in bacteria of the *Rhodobacteraceae* family. Establishing a transformation protocol for one of the cultured strains enabled a reporter assay to identify conditions under which the cluster is active. The results provide the basis for targeted product isolation and will facilitate the discovery of similar compounds in sponge extracts.

The identification of the BGC for the pharmaceutically relevant pelorusides paved the way toward biotechnological production. However, the lack of an assigned producer genome restricts targeted isolation and heterologous expression experiments. In **Chapter IV**, we applied two strategies to obtain insights into bacterial symbionts in *M. hentscheli*, a novel indirect sequence capture method to enrich DNA fragments that contain parts of a target BGC and long-read sequencing. We putatively linked the peloruside pathway to the mycalamide producer affiliated with a marine taxon previously not known as source of natural products. Extending both ends of a large orphan polyketide cluster to a member of the Verrucomicrobiota phylum further demonstrated the potential of long-read sequencing to assign BGCs to producing organisms in complex datasets.

Trans-AT PKSs are giant multimodular enzymes that are responsible for the biosynthesis of polyketides exhibiting remarkably diverse structures and bioactivities. These systems often originate in uncultivated and symbiotic bacteria, rendering their analysis particularly challenging. In **Chapter V**, we established a cloning strategy to assemble complete *trans*-AT PKS clusters into plasmids compatible with expression in different hosts. We initially captured clusters with known products from cultivated sources that can be used as model systems to study PKS function. Cloning orphan clusters from a sponge metagenome expanded the set of expression plasmids and opened avenues for the characterization of unknown pathways. Finally, we implemented an inducible expression system in two non-canonical host strains that provides a basis for the development of a heterologous production system.

This thesis sheds light on the biosynthetic potential encoded in the microbiome of the sponge *M. hentscheli* and presents strategies to characterize natural product pathways from uncultivated sources. A better understanding of the biosynthetic machineries will spur the development of synthetic biology tools to expand the natural product chemical space.

Zusammenfassung

Spezialisierte Metaboliten aus natürlichen Quellen (Tiere, Pflanzen, Mikroben) stellen Grundpfeiler der traditionellen und modernen Medizin gleichermaßen dar. Die strukturelle Vielfalt der Naturstoffe widerspiegelt sich in verschiedenen Bioaktivitäten, die zur Bekämpfung von Infektionen, Krebs und zahlreichen anderen Krankheiten genutzt werden. Klassische bioaktivitätsgeleitete Ansätze zur Isolierung von Naturstoffen führten zur Identifizierung zahlreicher Blockbuster-Moleküle, wurden jedoch rasch durch die häufige Re-Isolierung bereits bekannter Metaboliten eingeschränkt. Inzwischen haben technologische Fortschritte die Erkundung neuer Lebensräume und die Auswertung von Genomdaten die Entdeckung neuer Naturstoffe ermöglicht. Innerhalb der marinen Umwelt sind Schwämme eine besonders reiche Quelle an spezialisierten Metaboliten. Schwämme repräsentieren die ältesten lebenden Tiere und koexistieren mit einer erstaunlichen Vielzahl von Mikroorganismen. Für einige Schwammarten wurde gezeigt, dass mikrobielle Symbionten die bioaktiven Naturstoffe produzieren, die von diesem Wirt bekannt sind.
Diese Arbeit konzentriert sich auf das biosynthetische Potential, das im Mikrobiom des chemisch reichen Schwammes *Mycale hentscheli* kodiert ist. Wir verfolgten verschiedene Strategien, um (i) biosynthetische Gene zu identifizieren, (ii) sie mit kompletten biosynthetischen Genclustern (BGC) zu verknüpfen und die Struktur der kodierten Produkte vorherzusagen, (iii) die BGC den produzierenden Organismen zuzuordnen und (iv) funktionelle Daten durch Expression einzelner Enzyme in heterologen Wirten zu bekommen.

In **Kapitel II** zeigten wir, dass mehrere Symbionten gemeinsam die Fülle an Naturstoffen in *M. hentscheli* erzeugen. Dies steht im Gegensatz zu anderen Schwämmen, in denen einzelne "Superproduzenten" als Teil komplexerer Mikrobiome den Grossteil der bioaktiven Substanzen ihres Wirtes synthetisieren. Zusätzlich zu den bakteriellen Biosynthesewegen für die drei zytotoxischen Polyketide (Mycalamid, Pateamin und Pelorusid) identifizierten wir zahlreiche unbekannte BGC, die über einen grossen phylogenetischen Bereich von Bakterien verteilt sind. Die Charakterisierung von Enzymen eines unbekannten Polytheonamid-ähnlichen Biosyntheseweges deutet auf ein neues Produkt dieser seltenen Naturstofffamilie hin. Das Mikrobiom von *M. hentscheli* ist besonders reich an Stoffwechselwegen einer bestimmten Familie von Polyketid-Synthasen (PKS). Diese werden als *trans*-Acyltransferase (trans-AT) PKS bezeichnet und treten häufig in unkultivierten symbiotischen Bakterien auf, was unkultivierte Mikroben zu einer vielversprechenden Quelle für neuartige Moleküle macht.

Das Auftreten neuartiger BGC in unkultivierten Schwammsymbionten wirft Fragen über deren Ursprung und Funktion auf. Strategien zur Untersuchung dieser BGC beschränken sich normalerweise auf Kultivierungsansätze oder heterologe Expressionsstudien. In **Kapitel III** verfolgten wir eine alternative Strategie, in der wir sequenzierte Genome von kultivierten Bakterien nach ähnlichen Clustern durchsuchten. Wir analysierten einen kleinen Hybridcluster bestehend aus nicht-ribosomalen Peptid-Synthetase und PKS-Elementen aus dem Metagenom von *M. hentscheli* und identifizierten orthologe Cluster in Bakterien der Familie *Rhodobacteraceae*. Die Etablierung eines Transformationsprotokolls für einen der kultivierten Stämme ermöglichte einen Reporter-Assay zur Identifizierung von Bedingungen, unter denen der Cluster aktiv ist. Die Ergebnisse bilden die Grundlage für eine gezielte Produktisolierung und erleichtern zukünftig die Entdeckung ähnlicher Verbindungen in Schwammextrakten.

Die Identifizierung des BGC für die pharmazeutisch relevanten Peloruside ebnete den Weg zu deren biotechnologischen Produktion. Da jedoch bisher dem BGC kein Produzentengenom zugeordnet ist, sind die gezielte Isolierung und heterologe Expressionsexperimente nur eingeschränkt machbar. In **Kapitel IV** verfolgten wir zwei Strategien, um Einblicke in bakterielle

Symbionten in *M. hentscheli* zu erhalten: eine neuartige indirekte Methode zur Anreicherung von DNA-Fragmenten die Teile eines Ziel-BGC enthalten und Long-Read-Sequenzierung. Dadurch haben wir den Pelorusid-Biosyntheseweg mit dem Mycalamid-Produzenten in Verbindung gebracht. Dieser gehört zu einem marinen Taxon, das bisher nicht als Quelle von Naturstoffen bekannt war. Ausserdem konnten wir beide Enden eines großen unbekannten Polyketid-Clusters einem Mitglied des Verrucomicrobiota-Phylums zuordnen. Unsere Daten demonstrieren das Potenzial der Long-Read-Sequenzierung, um BGC in komplexen Datensätzen produzierenden Organismen zuzuordnen.

Trans-AT PKS sind riesige multimodulare Enzyme, die für die Biosynthese von Polyketiden mit erstaunlicher struktureller Vielfalt und verschiedener Bioaktivitäten verantwortlich sind. Diese Systeme stammen oft aus unkultivierten und symbiotischen Bakterien, was ihre Analyse besonders herausfordernd macht. In **Kapitel V** etablierten wir eine Klonierungsstrategie, um komplette *trans*-AT PKS-Gencluster in Plasmiden zu assemblieren, die mit der Expression in unterschiedlichen Wirten kompatibel sind. Wir starteten mit Clustern aus kultivierten Bakterien deren Produkte bekannt sind, da diese als Modellsysteme zur Untersuchung von PKS-Funktionen verwendet werden können. Das Klonieren von unbekannten Genclustern aus einem Schwamm-Metagenom erweiterte den Satz an Expressionsplasmiden und eröffnete Wege zur Charakterisierung neuer Biosynthesewege. Schliesslich implementierten wir ein induzierbares Expressionssystem in zwei nicht-typischen Wirtsstämmen, welches eine Grundlage für die Entwicklung eines heterologen Produktionssystems darstellt.

Diese Arbeit beleuchtet das biosynthetische Potenzial, das im Mikrobiom des Schwamms *M. hentscheli* kodiert ist und stellt Strategien zur Charakterisierung von Naturstoffbiosynthese-Wegen aus unkultivierten Quellen vor. Ein besseres Verständnis der biosynthetischen Mechanismen wird die Entwicklung von Werkzeugen der synthetischen Biologie vorantreiben, um die chemische Vielfalt der Naturstoffe zu erweitern.

Chapter I

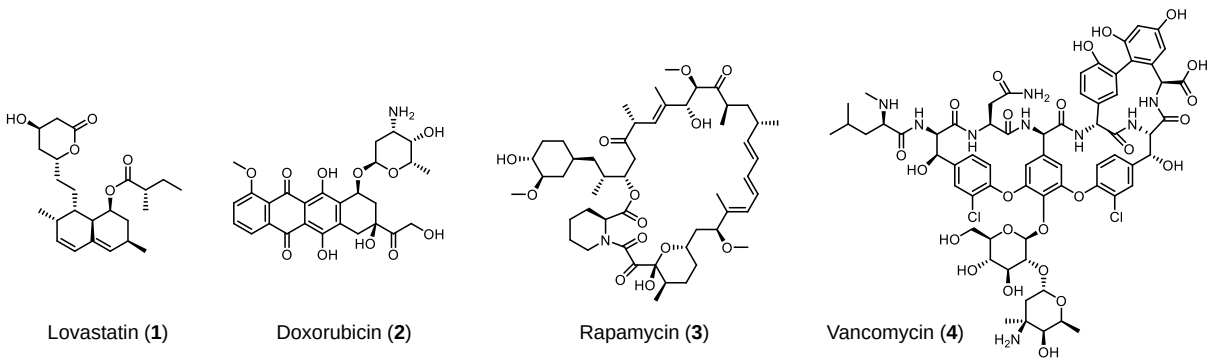
Introduction

1. Introduction

1.1 Natural Products

The biosynthesis and breakdown of nucleic acids, proteins, lipids, and carbohydrates is known as primary metabolism and defines a hallmark of living organisms. In addition to these universal biomolecules, there are substances called specialized metabolites that are not strictly required for growth, development and reproduction of the producing organism.[1] In contrast to primary metabolites, production of specialized metabolites is often restricted to a single species or a narrow taxonomic group.[2] The term secondary metabolites has widely been used to differentiate these molecules from primary metabolites, but implies a minor importance for the producing organism. Therefore, the terms natural products and specialized metabolites are used synonymously throughout this work. Although the native functions of many natural products remain elusive,[3] several studies have shown that these chemicals equip producers with distinctive advantages in their interaction with the environment.[4, 5] Natural products play a critical role in a broad range of biological functions such as social signaling, defense against competitors or predators, transportation of nutrients, and adaptation to changing environmental conditions.[4, 6]

Apart from the numerous functions in their native environments, natural products have played important roles in treating human diseases. Herbal extracts formed the basis of most traditional medicines, and compounds isolated from plants are still widely used as treatments against various diseases.[7] Striking examples are the anticancer drug paclitaxel derived from the yew (*Taxus*) tree[8] and the antimalarial compound artemisinin from the plant *Artemisia annua*.[9] Another prominent source of natural products are microorganisms such as bacteria and fungi. The serendipitous discovery of penicillin[10] kickstarted the golden era of antibiotics, during which mainly soil-dwelling bacteria and fungi were intensively studied, and shifted the focus from plants to microorganisms.[11] For example, the cholesterol-lowering agent lovastatin (**1**) was isolated from the fungus *Aspergillus terreus* in 1980,[12] while Actinobacteria are the source of numerous blockbuster drugs including the antitumor agent doxorubicin (**2**), the immunosuppressant rapamycin (**3**), and the antibiotic vancomycin (**4**) (**Fig. 1**).[13]



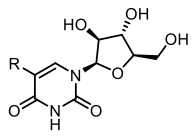
Lovastatin (**1**) Doxorubicin (**2**) Rapamycin (**3**) Vancomycin (**4**)

Figure 1. Selection of medically relevant natural products isolated from bacteria and fungi.

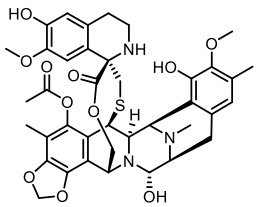
A considerable number of drugs approved over the last four decades consist of natural products or their (semi-)synthetic derivatives,[14] of which the numbers are particularly striking (>50%) for antibacterial and anticancer drugs.[15] Natural product-based therapeutics usually display greater chemical diversity and stereochemical content than compounds of purely synthetic origin.[16] The evolutionary pressure that optimizes binding of a specialized metabolite to a macromolecular target translates into the improved bioactivity of many natural compounds compared to molecules from synthetic libraries.[17] However, the complex molecular structures of natural products and the rediscovery of already known compounds have led to a decline of natural product-based screening programs in the pharmaceutical industry.[18] More recently, technological advances and the exploration of new habitats have considerably expanded the natural product chemical space.[19] Among others, the marine environment has proven a particularly promising resource for novel chemical functionalities and thus novel biosynthetic enzymology.

1.2 Marine Natural Products

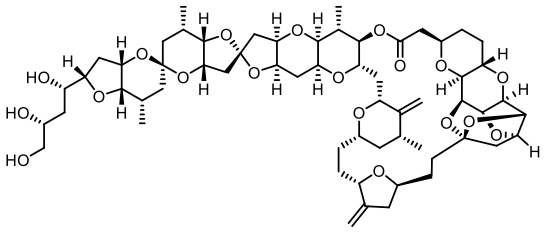
The marine environment, unprecedented in terms of both scale and biodiversity, is an extensive but understudied source of natural products.[20] A majority of the world's oceans remain inaccessible or unexplored, limiting the chemical space to that produced by a slim portion of marine organisms. The unusual nucleosides spongothymidine (**5**) and spongouridine (**6**) (**Fig. 2**) isolated from the sponge *Tethya crypta* in the 1950s[21] were the first biologically active marine compounds, leading to the synthesis of the anticancer agent Ara-C and the antiviral drug Ara-A.[22] Sponges are a particularly rich source of specialized metabolites.[23] These sessile, filter-feeding organisms are among the oldest known forms of animal life, populating a broad range of aquatic environments,[24] and providing a nutrient-rich habitat for various microorganisms.[25] The prokaryotic diversity in sponges is remarkably high, with 40% of the biomass consisting of microbial symbionts in some species.[26] Many sponge-associated natural products display structural similarities to microbial specialized metabolites and exhibit biosynthetic gene architectures similar to bacteria. Thus, it was hypothesized that the true producers of these compounds are symbiotic microbes.[27] It is thought that the production of these molecules could serve a protective role in the host, as a chemical defense against predators or pathogenic bacteria.[28]



Spongothymidine (**5**) (R=CH_3)
Spongouridine (**6**) (R=H)



Trabectedin (**7**)



Halichondrin B (**8**)

Figure 2. Representative marine natural products used as therapeutics.

Despite the presumably minor fraction of marine natural products known today, they have proven superior to their terrestrial counterparts in terms of chemical novelty.[29] Marine natural products exhibit a wide range of biological activities, such as cytotoxic,[30] neurotoxic,[31] anti-inflammatory,[32] and anti-infective,[33] that make them exceptional candidates for pharmaceutical applications. Trabectedin (**7**) is an alkaloid from the sea squirt *Ecteinascidia turbinata* that is used as an anticancer agent against soft-tissue sarcoma and ovarian cancer,[34] while halichondrin B (**8**) is a large polyether macrolide with potent anticancer activity originally isolated from the rare sponge *Halichondria okadai* (**Fig. 2**).[35] As for many other marine natural products, difficulties in supplying sufficient material considerably hampered the development of halichondrin B into a marketable product.[36] The total chemical synthesis, twelve years after the compound was first isolated, finally opened avenues for clinical trials.[37] However, chemical synthesis of promising natural products is often not feasible or economical.[38] A sustainable supply of these molecules therefore requires alternative strategies such as the development of biological production systems. A prerequisite for such an endeavor is the profound understanding of the enzymatic machineries that nature uses to generate molecular complexity.

1.3 Assembly Line Polyketide Synthases

Polyketides are a family of highly diverse natural products with broad biological activities.[39] They are synthesized by multi-domain enzymes or enzyme complexes called polyketide synthases (PKSs).[40] Based on the architecture and mode of action, PKSs are classified into various types. Type I PKSs are multifunctional enzymes consisting of covalently fused catalytic domains.[41] These systems are further grouped into iterative and modular type I PKSs. Type II PKSs are dissociable multi-enzyme complexes consisting of monofunctional proteins,[42] while type III PKSs form simple homodimers with a single catalytic center.[43] Numerous polyketides in microorganisms are generated by modular type I PKSs.[44] These large, multifunctional enzymes consist of various catalytic domains that act in a concerted fashion on protein-bound intermediates to orchestrate the assembly of complex molecules from simple acyl building blocks.[41] The set of catalytic domains responsible for the incorporation of one building block into the growing polyketide chain is termed a module. One round of chain elongation requires the action of a minimum of three distinct domains, an acyl carrier protein (ACP), an acyltransferase (AT), and a ketosynthase (KS). The activated ACP contains a conserved serine residue that is posttranslationally equipped with the 4'-phosphopantetheinyl arm (Ppant) of coenzyme A by an enzyme called 4'-phosphopantetheinyl transferase (PPTase).[45] The AT domain selects an acyl-CoA building block and loads it onto the free thiol of the ACP-bound Ppant moiety via a thioester linkage. The KS domain receives the polyketide chain from the upstream module and catalyzes a decarboxylative Claisen-like condensation with the ACP-bound extender unit, resulting in an elongated ACP-bound β-keto thioester intermediate (**Fig. 3a**).[46] In addition, PKS modules can harbor β-keto processing domains catalyzing the following reactions: a ketoreductase (KR) reduces the β-keto function to a hydroxyl group, a dehydratase (DH) generates an olefinic residue by the elimination of water, and an enoylreductase (ER) reduces a double bond to a fully saturated intermediate (**Fig. 3b**).[41] Once the polyketide chain has reached its final length, a terminal thioesterase (TE) catalyzes the release from the assembly line by various mechanisms including hydrolysis, macrocyclization, transesterification, or dimerization.[47] Tailoring enzymes can further

decorate the polyketide product by reactions such as alkylations, hydroxylations, halogenations, and glycosylations.[48]

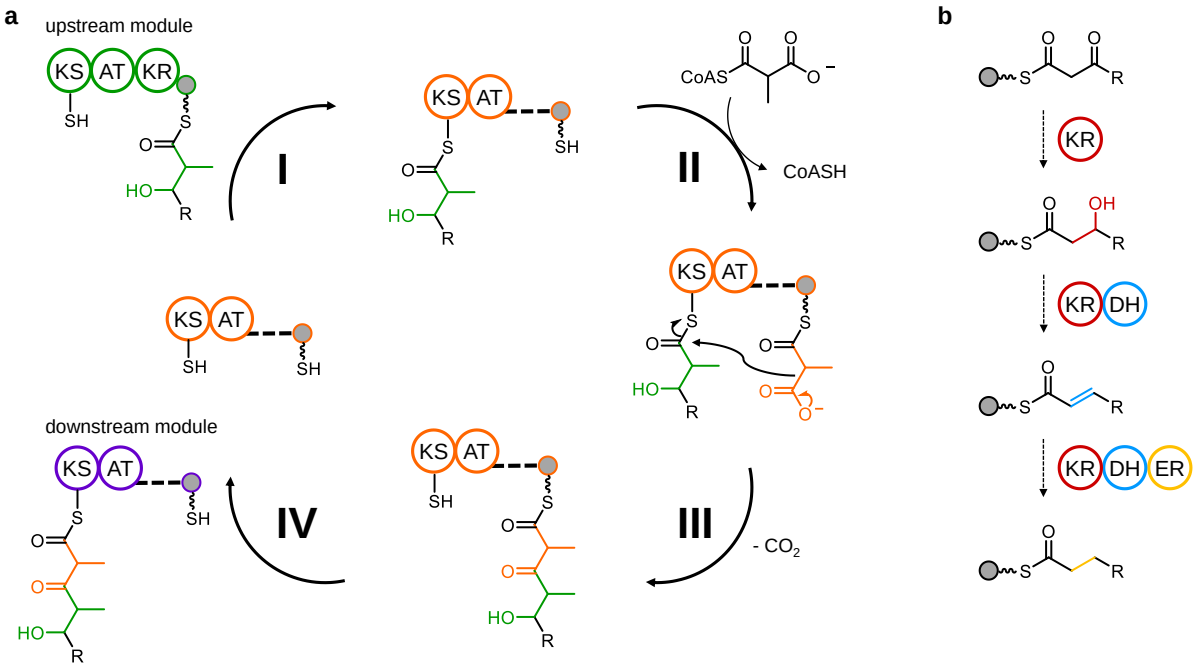


Figure 3. Catalytic cycle for a representative PKS module and optional reductive steps. a) I: The ketosynthase (KS) accepts the polyketide intermediate from the acyl carrier protein (ACP) of the upstream module (green). II: The acyltransferase (AT) loads the ACP with an extender unit from a CoA thioester. III: The KS catalyzes a decarboxylative Claisen-like condensation reaction between its substrate and the ACP-bound extender unit. IV: Translocation of the elongated chain to the KS of the downstream module (purple). b) Optional reductive processing steps of the elongated product depending on the tailoring domains present in the module. DH, dehydratase; ER, enoylreductase; KR, ketoreductase (adapted from Lowry *et al.*, *ACS Cent. Sci.* 2016[46]).

The PKS responsible for the biosynthesis of the antibiotic erythromycin (**9**) is by far the best studied system and has become the prototypical assembly line PKS (**Fig. 4**). The pathway consists of 28 active sites distributed across a loading module, six extension modules and a release module on three large proteins.[49] The released product undergoes post-PKS tailoring reactions including glycosylation and hydroxylation to yield the final product erythromycin.[50] Early studies on PKS systems focused on erythromycin and other pathways from filamentous actinomycetes, leading to the development of the textbook model of type I PKSs. However, analysis of bacteria from different habitats and taxonomic groups resulted in the discovery of a second large family of multimodular PKSs. These systems were termed *trans*-AT PKSs in order to distinguish them from the textbook *cis*-AT PKSs containing online AT domains.[51]

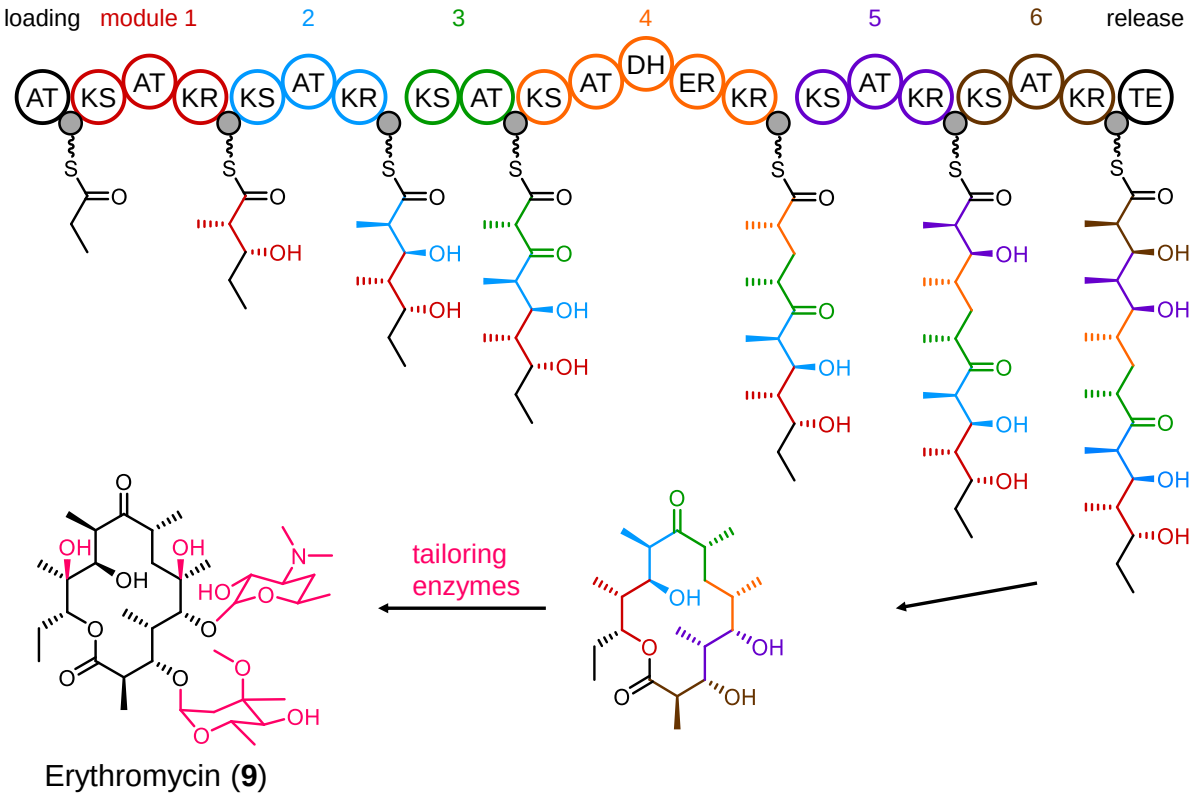


Figure 4. Erythromycin biosynthesis. The colors of the individual modules correlate with the colors of the processed moieties. The growing polyketide chain is attached to acyl carrier proteins (ACPs) and the wavy lines symbolize the 4'-phosphopantetheinyl arms (Ppant). AT, acyltransferase; DH, dehydratase; ER, enoylreductase; KR, ketoreductase, KS, ketosynthase; TE, thioesterase (adapted from Menzella *et al.*, *Nat. Biotechnol.* 2005[52]).

1.3.1 *Trans*-Acyltransferase Polyketide Synthases

Initially overlooked due to their presence in less studied microbes, *trans*-AT PKSs are widely distributed and make up almost 38% of all bacterial modular PKSs.[53] Their polyketide products exhibit an astonishing structural diversity and are involved in a broad range of biological functions including symbiosis, pathogenicity, and regulation.[54-57] The eponymous difference to *cis*-AT PKSs is the lack of the AT domain within each module.[58] Instead, a freestanding AT usually loads the ACP of each module, leading to the incorporation of the same type of building block, typically malonyl-CoA. In some pathways, non-malonyl-specific ATs load ACPs of individual PKS modules.[51] β-branching is another type of polyketide modification catalyzed by *trans*-acting enzymes. It regularly occurs in *trans*-AT PKS assembly lines, while being rare in *cis*-AT PKS systems. The conversion of the β-carbonyl function into a carbon branch is catalyzed by a set of enzymes including a freestanding ACP, a 3-hydroxy-3-methylglutaryl-CoA synthase-like protein (HMGS), a KS, and one to three enoyl-CoA hydratase-like enzymes (ECHs).[59] In contrast to *cis*-AT PKS systems, *trans*-AT PKSs harbor an unprecedented diversity of module variants including modules with unusual domain types or domain repetitions, non-elongating modules, and modules that are split between two proteins.[60] In addition, *trans*-AT PKSs are often present as hybrid systems with modular nonribosomal peptide synthetases (NRPSs), further adding structural diversity to their products (**Fig. 5**).[60]

The current evolutionary model supports the differentiation between *cis*- and *trans*-AT PKSs stating that the two systems evolved independently, with gene duplication being the main driver of *cis*-AT PKS evolution,[61] while *trans*-AT PKSs commonly undergo recombination and horizontal gene transfer.[62] However, a more recent model identifies a number of features, such as AT domain remnants in *trans*-AT PKSs or the existence of hybrid systems, which hint towards a more unified evolutionary path than previously expected.[63] The presence of these systems in a broad microbial range, *cis*-AT PKSs in well-studied bacteria and *trans*-AT PKSs in uncultivated sources, leads to new questions regarding the (co-)evolution of megasynthases in diverse producers. However, the widespread distribution of *trans*-AT PKSs in symbiotic and uncultivated strains makes them particularly challenging to study. Nevertheless, the presence in less studied bacteria also highlights the potential for finding chemical novelty in these systems. This effort has been facilitated by the identification of biosynthetic rules based on protein sequences and architectures, leading to the *de novo* prediction of polyketide structures from genomic data.[64]

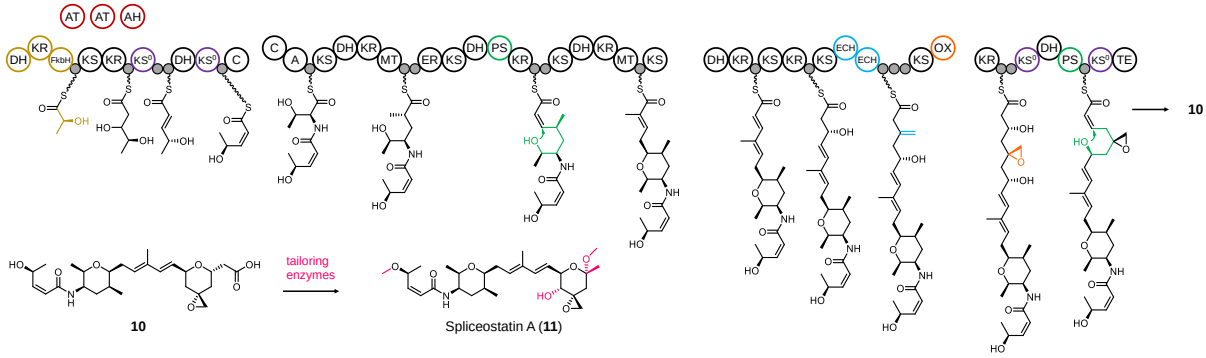


Figure 5. Spliceostatin A biosynthesis. The domains/modules that are characteristic for *trans*-AT PKSs are colored. The unusual starter module in yellow, *trans*-acting domains in red, non-elongating KSs (KS^0) in purple, pyran synthases (PS) in green, ECHs (enoyl-CoA hydratase) in light blue, and the oxidoreductase (OX) in orange. A, adenylation domain; AH, acyl hydrolase; C, condensation domain; FkbH, hydroxylase; MT, methyltransferase (adapted from Eustáquio *et al.*, *PNAS* 2014[65]).

1.3.2 Prediction of Polyketide Synthase Products

The advent of genome sequencing revealed that the majority of biosynthetic genes responsible for the production of specialized metabolites including polyketides cluster in microbial genomes.[66] These biosynthetic gene clusters (BGCs) typically encode enzymes that synthesize the core structure, tailoring enzymes that modify this core structure, and regulatory components such as transporters and resistance proteins.[67] The clustering of biosynthetic genes enabled the development of knowledge-based algorithms, such as antiSMASH, for the automated detection of BGCs in genome data.[68] Due to the modular architecture of *cis*-AT PKS pathways, the succession of chemical moieties in the polyketide backbone correlates with the order and composition of modules in the PKS. This correlation is known as colinearity principle and allows the prediction of the polyketide structure from the PKS architecture and *vice versa*.[40] In contrast, the plethora of unusual features in *trans*-AT PKSs renders the colinearity rule ineffective for the prediction of the chemical structure. However, analysis of *trans*-AT PKS domains revealed that phylogenetically similar KSs process polyketide intermediates harboring the same chemical moieties vicinal to the thioester.[62] In other words, there is a close correlation between the KS sequence and the polyketide modifications introduced by the upstream PKS module (**Fig. 6**). This correlation facilitates the prediction of

polyketide intermediates from individual KS sequences as well as the discovery of *trans*-AT PKSs by structure-based genome mining.[69] Using this method, the prediction and successful isolation of numerous polyketides, including thailandamide,[62] nosperin,[70] and lacunalide[71] was achieved. However, the process of integrating individual KS sequences into reference trees and manually assigning substrates is laborious and requires profound biosynthetic knowledge. The TransATor prediction pipeline automates this workflow by generating KS specificity predictions and structural proposals based on input PKS sequences.[64] The web-based application also facilitates the assignment of partial clusters to known natural products and guides the completion of polyketide BGCs.[72] Automated workflows for the prediction of polyketide structures from genomic data become increasingly important for the analysis of complex data sets such as metagenomes. Since *trans*-AT PKSs often originate in uncultivated and symbiotic bacteria, mining metagenomes for these pathways has become a particularly promising strategy to discover novel enzymology.

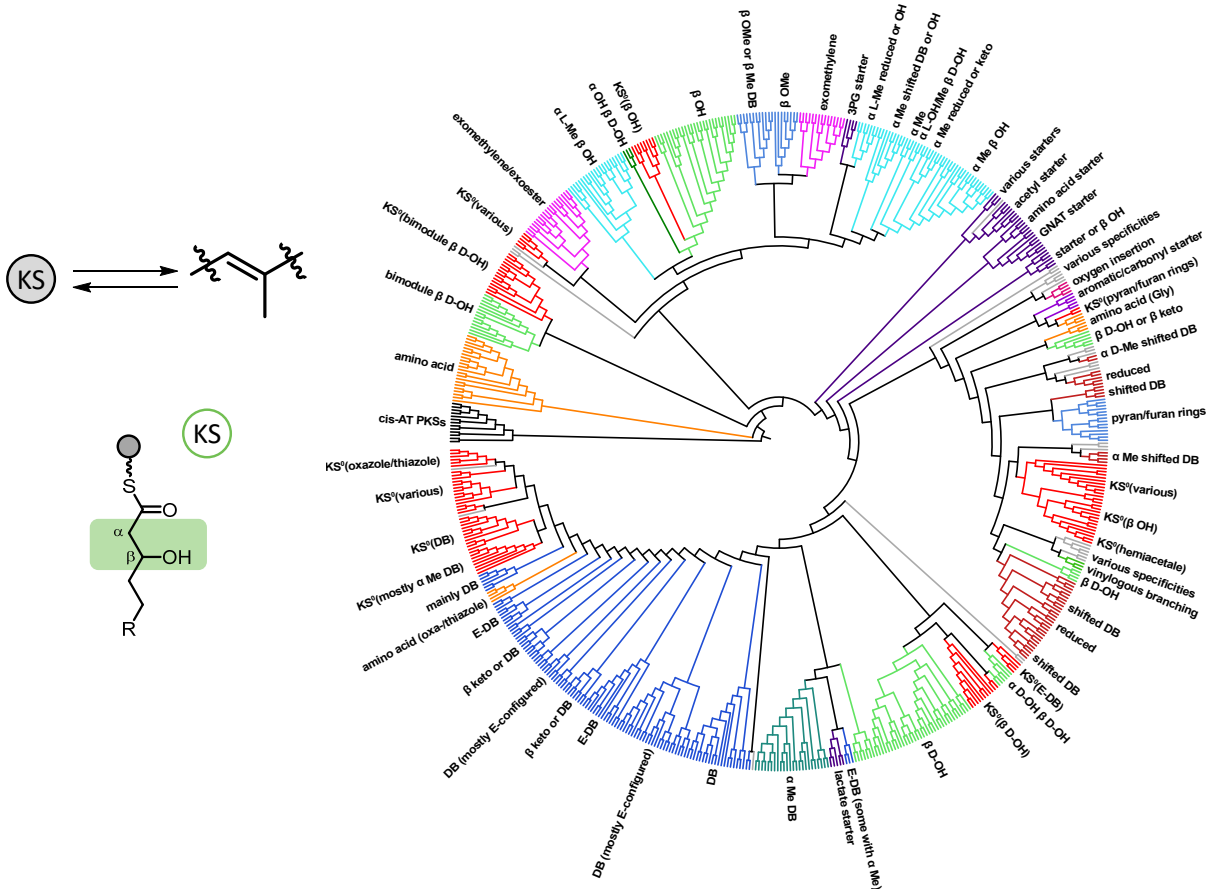


Figure 6. Correlation between KS sequence and incoming intermediate in *trans*-AT PKSs. Phylogenetically similar KSs process intermediates with the same chemical moieties in the α- to γ-region around the thioester. The cladogram represents KS domains from *trans*-AT PKSs with known products. Distinct ketide clades are labeled with the predicted chemical moiety of the incoming substrate (adapted from Helfrich *et al. Nat. Chem. Biol.* 2019[64]).

1.4 Metagenome Mining

This chapter was published in:
Comprehensive Natural Products III, H.-W. Liu, T. P. Begley, Eds. (Elsevier, Oxford, 2020)
DOI: 10.1016/B978-0-12-409547-2.14722-5[73]

Microorganisms are universally present and indispensable for life on Earth. They occupy almost every habitat on this planet, from the deep sea to deserts to the digestive systems of almost all multicellular organisms. A gram of soil can harbor up to 10^{10} bacterial cells comprising between 4,000 and 50,000 distinct species.[74, 75] A milliliter of the Sargasso sea harbors around one million bacterial cells with an average genome size of 2 megabases.[76] However, in spite of their enormous numbers and diversity, and their importance for biogeochemical cycles, biotechnology, and drug discovery, virtually nothing is known about functional properties of the large majority of microbes. Only a small percentage (below 1% in many habitats)[77] of bacteria have been cultivated in the laboratory, and the majority of the more than 90 currently known bacterial (candidate) phyla lack cultured representatives.[78-80] Such "microbial dark matter" encompasses most portions of the bacterial tree of life, comprising numerous deep-branching taxa.[80, 81]

For uncultivated bacteria, the term "candidate taxon" is used (e.g. candidate species, candidate genus, etc.), since traditional taxonomic classification also requires cultivation-based assessment of functional traits.[82, 83] Spectacularly, such taxa lacking cultivated representatives include not only more than 50 (ca. 60%) of the known phylum-like divisions, but even a recently proposed new higher-order superphylum (Candidate Phyla Radiation, CPR) that covers a vast region of bacterial diversity.[84] Based on the wealth of natural products derived from a minute fraction of the cultivated diversity, it has been hypothesized that uncultivated prokaryotes represent a tremendous resource of chemical novelty. This hypothesis is supported by a growing number of studies that revealed natural product types previously unknown from cultured microbes,[85-87] as well as chemically gifted producer taxa with a rich specialized metabolism.[88] Exploring this uncharted natural product space holds considerable promise to discover new chemical entities and biosynthetic enzymes for pharmaceutical and biotechnological applications, and to obtain clues about metabolic and ecological functions of microbial dark matter. Not only soil as the initially proposed metagenomic ecosystem[89] is a proven resource of new natural products, but also host-microbe interactions from very different environments, such as fungi, plants, insects, marine invertebrates, and mammals are influenced or even governed by a multitude of specialized metabolites. For an increasing number of cases, it has been shown that the true producers of many host-derived bioactive metabolites are associated bacteria, and the frequency at which such cases are encountered suggest that natural products-driven symbiosis is widespread in nature.

The study of microbial dark matter is challenging and still in its infancy despite remarkable technical advances that have been made over the past two decades. The first part of this chapter will discuss techniques that have been developed to assess diversity and to obtain sequence and functional information with a focus on natural product discovery. In the second part, we will provide an overview of the current state of knowledge on the chemistry of microbial dark matter. Of central importance for this chapter is the concept of metagenomics, i.e., the study of the combined genomic information present in a given habitat. Metagenomic studies are based on the analysis of DNA derived from environmental species communities (**Fig. 7**), thus circumventing the need to isolate individual microbes. Initial studies focused on

cloning of environmental DNA (eDNA) and analysis of those genome fragments that were identified as positives in function- or sequence-based screens. In addition, advanced technologies now provide high-throughput, deep sequencing data from even high-diversity biomes. This methodology as well as complementary techniques to sequence single bacterial cells are currently transforming the field.

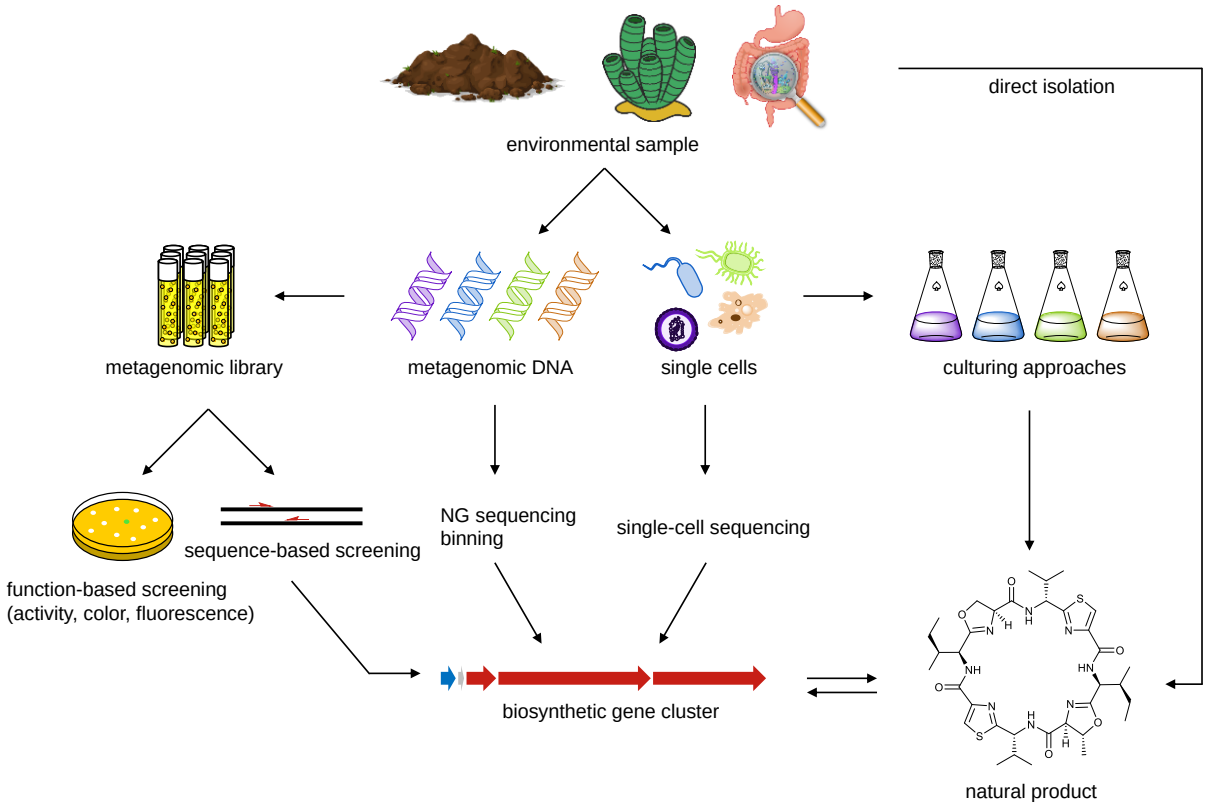


Figure 7. Overview of metagenome mining workflows. Shown are different steps involved in the mining of environmental samples for biosynthetic gene clusters in order to get access to known or novel natural products. NG sequencing, next-generation sequencing.

1.4.1 Diversity Analyses

To determine the microbial diversity in an environmental sample, sequences of the ribosomal RNA (rRNA) are commonly analyzed.[90] Due to the essential and conserved function in every living cell, rRNA sequences are seldom transferred horizontally[91] and therefore well-suited as molecular marker to determine evolutionary relationships between organisms.[92] For prokaryotes, the 16S rRNA is the preferred biomolecule because its sequence contains regions of high conservation that permit PCR-based amplification of near-full-length rRNA genes from eDNA using degenerate primers.[93] In addition to these highly conserved parts, the sequence also harbors hypervariable, species-specific regions that allow bacterial classification.[94] Sequences that are obtained from an environmental sample can then be used to generate alignments and phylogenetic trees with sequences of characterized organisms.[95] These inform about the next relatives and allow one to taxonomically assign organisms present in an environmental sample. Despite the wide use of amplicon sequencing techniques, these analyses are biased.[96] For example, it was shown that genomic GC content correlates negatively with the observed relative abundances, likely due to PCR bias against GC-rich

species.[97] In addition, organisms from novel lineages can be overlooked in such experiments due to sequence divergence relative to the primers used for 16S gene amplification.[84] The strongest bias during sample processing is introduced by the DNA extraction method and PCR amplification.[98] More recently, primer-free 16S rRNA extraction, reverse transcription, and cDNA sequencing have been used to circumvent such biases.[99] Diversity analyses based on PCR amplicons can be applied to many studies beyond taxonomy, for example to assess the diversity of biosynthetic genes, such as NRPSs and PKSs.[100] These and many other enzymes contain conserved sequences that permit the detection of gene homologs in eDNA.[101] Since the phylogeny of biosynthetic enzymes often correlates with structural features in the natural product, large phylogenetic distances to characterized homologs provide clues about chemical novelty even for short amplified gene fragments.

High-diversity samples, such as many soils, often contain a large number of organisms at low abundance. A method that provides insights into the total diversity of such challenging communities at moderate sequencing depth is rarefaction analysis.[102] This method estimates the proportion of a sample within the total species richness. Rarefaction graphs are created by plotting the number of species found as a function of the number of samples. The resulting rarefaction curve usually shows a steep slope at the beginning and flattens out as fewer species per sample are detected at increasing sample numbers. By calculation of the asymptote, the total number of genes is extrapolated. This method can be applied to 16S rRNA as well as natural product genes.

1.4.2 Metagenomic Libraries

Before the advent of cost-effective next-generation sequencing technologies, the construction of metagenomic DNA libraries has been the main method to study natural products of uncultivated bacteria. To generate such libraries, fragments of metagenomic DNA extracted from an environmental sample are cloned into a vector and introduced into a bacterial host. These libraries are then subjected to sequence- or function-based screenings to identify clones that contain biosynthetic genes or exhibit a specific, phenotypic trait.

1.4.2.1 DNA Extraction

An efficient protocol for the extraction of high-quality, high-molecular-weight (HMW) metagenomic DNA from a limited amount of environmental sample is crucial for all downstream applications. The physical and chemical composition of an environmental sample influences the size, quantity, and quality of metagenomic DNA that can be extracted. Challenges during the DNA extraction process are manifold and highly dependent on the sample type. The analysis of aquatic communities often requires the collection and filtration of large amounts of water to obtain sufficient DNA.[103] Extractions from soil samples are complicated by the presence of contaminants, such as humic acids, which interfere with downstream processes.[104] DNA extraction from host-associated target communities is often preceded by physical fractionation or selective cell lysis to minimize the amount of host DNA.[105]

For cell lysis, a combination of physical, chemical and mechanical methods is usually used to efficiently lyse the different bacterial cell types.[106] Often, there is a tradeoff between recovery of DNA from diverse microorganisms and high-quality DNA, because harsh lysis conditions affect DNA quality.[107] The community composition also influences the diversity of

recovered metagenomic DNA. Microbes harboring different types of cell walls and membranes behave differently under specific cell lysis conditions, and the method can thus bias the composition of the extracted DNA. Assessment and optimization of cell lysis methods is therefore crucial to obtain DNA from all community members. Despite the development of guidelines for DNA extractions from human fecal samples[108] and soil,[109] there is no general standard for metagenomic DNA extraction.[106] DNA extraction is usually followed by gel electrophoresis to separate HMW fragments from smaller fragments and contaminants.

1.4.2.2 Vector Systems and Cloning

The size of many natural product BGCs requires vectors that are capable of stably replicating large DNA fragments. Depending on the size of the fragments to be cloned, cosmids/fosmids (both up to 45 kb) or bacterial artificial chromosomes (BACs, up to 300 kb) are most commonly used. Cosmids are hybrid vectors containing *cos* genes of the λ bacteriophage that allow them to be packaged into viral particles.[110] Fosmids are based on the *Escherichia coli* F-factor replicon and therefore maintained as single copy, which is suggested to offer improved stability compared to cosmids (not observed in the authors' laboratory).[111] Fosmids with a second inducible replicon have been developed to facilitate vector amplification and insert sequencing.[112] BACs are also based on the F-factor replicon, but are able to maintain inserts larger than 300 kb.[113] In contrast to cosmids and fosmids, BACs are introduced by electroporation, therefore overcoming the limiting insert size but also showing much lower cloning efficiency.[114]

Libraries containing small DNA inserts are generally used to identify individual enzymatic activities or natural product types encoded by small gene clusters, such as terpenes, type II PKS products, ribosomally synthesized and posttranslationally modified peptides (RiPPs), or some alkaloids. Large-insert libraries are useful to capture more complex biosynthetic pathways, including NRPSs and modular type I PKS pathways. Generally, large inserts are preferred because fewer clones are needed to sufficiently cover a metagenome, thus increasing the probability of hit detection. However, many eDNA samples are so complex that only a fraction of the total metagenomic DNA is present in the library. This effect becomes particularly dominant in host-microbe samples in which host DNA is overrepresented or if target genes stem from low-abundance organisms. Therefore, microbiome DNA enrichment strategies have been developed for numerous eDNA samples, such as plants[115] and sponges[116], to increase hit detection during library screens.

1.4.2.3 Library Hosts

E. coli is the most commonly used host for creating metagenomic libraries. Engineered *E. coli* strains have useful properties for metagenomic applications, including mutations that inhibit recombination (*recA*) and DNA degradation (*endA*). Furthermore, efficient methods exist to generate the large clone numbers required to capture complex eDNA samples. Sequence-based analyses are expression-independent and therefore almost exclusively performed in *E. coli*. Functional screening, i.e., screening depending on the production of functional proteins, has also resulted in various natural product discoveries with *E. coli* hosts. A common challenge, however, is the expression of genes from remotely related organisms. Potential incompatibility with the transcription machinery, different codon usage, and absence of biosynthetic precursors must be considered. To overcome these limitations, various non-

E. coli hosts have been used, such as *Streptomyces albus* and a range of Gram-negative bacteria. Phenotypic screenings often show significant differences depending on the host regarding the expressed BGC sets, showcasing the importance of testing different host systems for functional screenings.[117] A drawback of these alternative hosts is however to achieve the large clone numbers obtained with *E. coli*.

1.4.2.4 Function-Based Library Screening

Functional metagenomics is based on the identification of clones that exhibit a certain phenotype based on the bioactivity encoded by eDNA. Such functional studies do not require homology to genes with known function and are thus well-suited to discover metabolic novelty. In the context of natural product studies, screening has been conducted based on various types of readouts. These include direct detection of metabolites by high performance liquid chromatography (HPLC) and mass spectrometry (MS), inhibition zones caused by antibiotic activity, or pigmentation. To simplify activity-based screens of large libraries with rare positives, assays with more easily identifiable phenotypic traits, such as antibiotic production or pigmentation, are commonly used. More sophisticated complementation and reporter systems have been developed to detect or enrich for PKSs and NRPSs,[118-120] siderophores,[118] or quorum-sensing factors.[121] Similar to complementation assays, genes conferring resistance towards a given compound can be identified by exposing the library to inhibitory concentrations of a particular toxin.

Despite the significant advantages of functional metagenomics over homology/sequence-based screening, there are a number of challenges in the context of natural product discovery. It is commonly observed that the frequency of active clones is much lower than what would be expected from BGC frequencies in metagenomic datasets.[122] Among the many reasons for this observation is the challenge to heterologously express complete pathways to access the full metabolic potential of an environmental sample.[123] The need for complete gene clusters harbored in one clone is complicated by the large size of many biosynthetic pathways. In contrast to functional studies that usually screen for a single enzymatic activity, many natural product biosynthetic pathways require concerted expression of multiple genes for the successful production of a given compound.[124] Some gene clusters that encode bioactive compounds are known to remain silent under laboratory conditions, likely because environmental signals that trigger activation are missing.[125]

1.4.2.5 Sequence-Based Library Screening

An alternative approach to identify natural product gene clusters in metagenomic DNA libraries is homology/sequence-based screening. This method involves the design of degenerate primers that target conserved regions of biosynthetic genes, such as KSs of modular PKSs and adenylation domains of NRPSs.[118] The primers are used to PCR-amplify homologs from eDNA, which then serve as templates to generate specific probes for library screening. To efficiently screen large libraries, pooling methods have been developed in which thousands of clones are screened in a single PCR, followed by serial dilutions until the target clone is identified.[109, 126, 127] Despite the limited capability to detect enzymatic novelty beyond homologs, a significant advantage of sequence-based screens is that BGCs can be identified without being present in a single clone and without functional expression. A major challenge of DNA-based screening methods is the identification of a suitable host system to functionally

analyze the captured cluster. In many cases, neither the compound that is encoded by the cluster nor the producing organism is known. However, if there is prior information about the target natural products, phylogeny-based screening strategies can be used to target biosynthetic genes of a particular pathway (family), thereby significantly reducing the screening effort. This strategy led to the identification of BGCs for specific polyketide, peptide, and alkaloid scaffolds of pharmaceutical interest from soil as well as for previously known sponge and ascidian compounds. Because most vectors used for metagenomic library preparation are not suitable for heterologous expression studies and many natural product BGCs are too large to be captured in a single clone, methods have been developed to change vector backbones and to reassemble genes that were separated during library preparation. The *E. coli*-based Red/ET cloning technology[128] has been used for the reconstitution of numerous BGCs and expression in heterologous host strains.[129, 130] Transformation-associated recombination (TAR) in yeast is another important technique to reassemble complete gene clusters from multiple overlapping fragments and therefore functionally study large biosynthetic pathways.[131] More recent methods, such as RecET direct cloning combined with Redαβ-mediated recombination, will further facilitate the functional analysis of larger clusters from uncultured organisms once these methods are reliably applicable to metagenomic DNA samples.[132]

1.4.3 Direct Metagenomic DNA Sequencing

Advances in high-throughput sequencing and assembly technologies have made the direct sequencing of enriched or metagenomic DNA an attractive alternative for homology-based gene discovery. Analysis of a microbial community by metagenome shotgun sequencing provides a much broader and less biased functional profile compared to 16S rRNA gene studies. The major challenge of *de novo* metagenome sequencing is the assembly of multiple unknown genomes from short DNA sequences. A number of specialized tools have been developed to assemble such complex metagenomes with varying performance dependent on sample type and research goal.[133] Metagenome assemblies can be improved by combining short sequence reads with long-read sequencing data, such as Oxford Nanopore or Pacific Biosciences reads.[134, 135] The large size and repetitive nature of many natural product gene clusters render the assembly from metagenomic sequencing reads particularly challenging. Bioinformatic tools, such as antiSMASH[136] that scan assembled fragments and annotate biosynthetic genes are used to predict the structures of encoded natural products and therefore facilitate the assignment of gene clusters to known compounds or guide the structure elucidation process of novel scaffolds.

To broaden the genomic context of strains harboring interesting biosynthetic genes, assembled fragments can be sorted in a process called binning.[137] The goal of this procedure is to group DNA sequences that likely belong to the same species. Binning is particularly useful for complex DNA samples harboring multiple biosynthetic pathways from different producers. DNA fragments encoding partial biosynthetic pathways within the same bin can then be connected by gap closing PCR to complete entire gene clusters. Additionally, binning can provide valuable information about the phylogeny and the metabolic requirements of the producing organism, thereby guiding heterologous expression and cultivation studies. Improvements in long-read sequencing technologies as well as assembly and binning algorithms will not only accelerate the identification of natural product gene clusters from environmental samples, but also guide their assignment to chemical structures and individual organisms.

1.4.4 Single-Cell Sequencing

Although metagenomic sequencing provides valuable information about the composition of a microbial community and putative functions of some members, challenges in assembly, binning and functional annotation rarely allow a complete link between the functional genes of a microorganism and its phylogeny.[138] A more recent approach for studying microbial communities is single-cell metagenomics. The technique relies on the amplification and sequencing of DNA from a single bacterial cell obtained from an environmental sample.[139] The major advantage of this method is that it can link metabolic functions to a specific species. In addition, high-quality genomes of low-abundance species that might be lost in metagenomic studies can be recovered. The procedure starts with the separation and isolation of a single cell by techniques such as serial dilution, microfluidic encapsulation, micromanipulation or fluorescence-activated cell sorting (FACS).[140] After lysis of individual cells, the low amount of genomic DNA is amplified by a high-fidelity polymerase in a process called multiple displacement amplification (MDA) in order to generate enough material for subsequent sequencing.[141] MDA generates long, overlapping amplicons in a simple reaction setup that minimizes the risk of handling errors and permits automation. However, the method also exhibits limitations including uneven genome amplification, bias against GC-rich templates, and formation of chimeric reads.[142] Optimized protocols using thermostable mutant polymerases have been developed to overcome some of these limitations and improve genome recovery from single cells.[143] Due to the minimal amount of input DNA, MDA is highly susceptible to contamination from extracellular DNA.[140] As for metagenomic assembly, several softwares have been developed for single-cell genome assemblies that deal with the limitations mentioned above. Furthermore, it has been shown that the combined assembly of metagenome and single-cell genome data can greatly improve assembly continuity and completeness.[138]

Single-cell sequencing combined with metagenomics resulted in the successful isolation of biosynthetic pathways and a broader characterization of the producing organism. For example, a combination of single-cell sequencing and metagenomic library screening has led to the identification of the apratoxin pathway in filamentous marine cyanobacteria associated with heterotrophic bacteria.[144] Furthermore, metagenomic sequencing combined with single-cell genomics revealed members of the candidate phylum ‘Tectomicrobia’ as chemically rich producers in sponges.[88]

Single-cell sequencing is becoming an important tool with implications for microbial ecology, evolutionary biology and biotechnology. It is highly complementary to other approaches, such as culture-based methods or metagenomics, and has expanded our view on microbial and functional diversity.[138] Regarding natural product research, single-cell genomics will facilitate the assessment of the biosynthetic potential of uncultivated bacteria from various habitats and accelerate the link between gene clusters and individual species. In addition, targeted single-cell resolution screens could help prioritizing metabolically talented bacteria for sequencing and cultivation studies.

1.4.5 Cultivation

Cultivation of a novel microorganism is often time consuming and tedious. Despite the fast-growing field of culture-independent techniques, cultivation approaches remain essential for the characterization of physiological properties. Reasons why certain bacteria are not growing under laboratory conditions are manifold (nutrients, cofactors, temperature, pH, osmotic conditions, quorum sensing, shared metabolism with another organism, host signals, etc.) and summarized by the failure to replicate essential aspects of their natural environment. One

important finding was that some slow-growing species thrive in nutrient-poor environments and might be inhibited by substrate-rich media commonly used. The use of media containing low concentrations of organic carbon have been successfully used for the cultivation of previously unculturable bacteria from aquatic and terrestrial environments.[145] Two other related strategies that have led to significant progress in culturing microbes are co-cultures of bacteria from the same habitat, simulated environments,[146] and *in-situ* cultivation using devices such as the ichip.[147]

Cultivation of native producers remains important for natural product discovery. As long as there are no general heterologous expression hosts, cultivation and characterization of native strains can lead to a sustainable supply and furthermore guide the identification of essential components for the development of heterologous production systems. With the cultured bacteria representing only a minority of the global microbial diversity, advances in the cultivation of novel species will likely lead to the identification of new natural products and improved biological production systems.

1.5 Heterologous Expression of Biosynthetic Gene Clusters

The wealth of sequencing data and bioinformatic tools for BGC mining sets the stage for the identification of a myriad of natural products with novel bioactivities. However, there is a large discrepancy between the number of identified BGCs and the number of characterized pathways with known products.[53] A reason for this observation is the challenging task of identifying a natural product in a native host and unambiguously linking it to a BGC. Cultivation of respective organisms under laboratory conditions is often not feasible. Even in case of successful cultivation, the metabolic cost of producing specialized metabolites often limits high titers or the activation of cryptic BGCs requires substantial effort, for example by genetically modifying the host strain or screening various cultivation conditions.[148,149] A more applicable method is the development of robust heterologous hosts with good growth characteristics and established genetic engineering tools for the production of natural products. Such systems not only provide access to novel metabolites but also serve as a platform to access valuable chemicals in an economical and sustainable fashion.

E. coli is by far the most commonly used host for heterologous production across different natural product families.[150] The fast growth kinetics, scalable culture conditions, the available genetic tools, and the lack of endogenous natural product pathways that could interfere with the heterologous system have primed *E. coli* as universal production chassis.[150] Nevertheless, the successful expression of the best studied assembly-line PKS generating the precursor of the antibiotic erythromycin A required a highly engineered *E. coli* strain and substantial optimization procedures.[151] The production of the final product erythromycin A required the expression of 17 additional genes.[152]

For the same reasons as for *E. coli*, the yeast *Saccharomyces cerevisiae* has widely been used as heterologous expression platform, mainly for the production of fungal natural products.[153] *Streptomyces* hosts are commonly used for expression of polyketide BGCs from other Actinobacteria. Due to native polyketide pathways, they harbor the enzymatic machinery for precursor supply and sophisticated post-PKS processing steps. The development of gene editing systems as well as metabolic engineering strategies have further primed *Streptomyces* as host organisms.[154]

A long-term goal of natural product research is the generation of altered natural products by combinatorial biosynthesis.[155] The stable production of a natural compound in a heterologous host is the ideal starting point for pathway engineering and product diversification. Modular systems such as NRPSs and PKSs are predestined for combinatorial

approaches, potentially leading to large compound libraries from a few simple building blocks.[156]

1.6 Aims of this Thesis

The overarching aim of this work was to identify and characterize biosynthetic pathways encoded in the microbiome of the sponge *Mycale hentscheli* and apply strategies to exploit the biosynthetic potential encoded in uncultivated bacteria. The chemically rich sponge *M. hentscheli* is the known source of at least three bioactive polyketides. The identification of these polyketide BGCs and the assignment to microbial producers provided insights into the role of bacterial symbionts in the production of therapeutic small molecules. Additionally, metagenome mining revealed numerous BGCs with unassigned natural products. Structural predictions based on the architecture of the biosynthetic machinery combined with characterization of individual enzymes have provided a basis for the targeted isolation of new natural products (Chapter II). An alternative approach involved the analysis of homologous pathways in cultivated strains with the aim to characterize these products and screen sponge extracts for similar metabolites (Chapter III). The assignment of unbinned BGCs to producing organisms aids the identification of missing enzymes and provides an overview of the biosynthetic potential in environmental strains. We employed a targeted sequence capture method combined with long-read sequencing to extend biosynthetic sequences and link them to bacterial producers (Chapter IV). Since *M. hentscheli* is particularly rich in polyketides generated by a distinct class of PKSs often associated with uncultivated and symbiotic bacteria, we aimed to identify a cloning strategy and suitable hosts for the heterologous expression of such pathways. We started with two small model clusters with known products from cultivated bacteria before expanding the strategy to orphan pathways from the sponge metagenome. A successful implementation could pave the way towards a broadly applicable biotechnological production system (Chapter V).

References

1. F. Marinelli, G. L. Marcone, "3.26 - Microbial Secondary Metabolites" in Comprehensive Biotechnology (Third Edition), M. Moo-Young, Ed. (Pergamon, Oxford, 2011), vol. 3, pp. 312-323.
2. A. L. Demain, J. L. Adrio, Contributions of microorganisms to industrial biology. *Mol. Biotechnol.* **38**, 41-55 (2008).
3. J. Davies, G. B. Spiegelman, G. Yim, The world of subinhibitory antibiotic concentrations. *Curr. Opin. Microbiol.* **9**, 445-453 (2006).
4. A. L. Demain, A. Fang, "The Natural Functions of Secondary Metabolites" in History of Modern Biotechnology I. Advances in Biochemical Engineering/Biotechnology, A. Fiechter, Ed. (Springer, Berlin, Heidelberg, 2000), vol. 69, pp. 1-39.
5. A. Fajardo, J. L. Martínez, Antibiotics as signals that trigger specific bacterial responses. *Curr. Opin. Microbiol.* **11**, 161-167 (2008).
6. J. Davies, Specialized microbial metabolites: functions and origins. *J. Antibiot.* **66**, 361-364 (2013).
7. A. G. Atanasov *et al.*, Discovery and resupply of pharmacologically active plant-derived natural products: a review. *Biotechnol. Adv.* **33**, 1582-1614 (2015).
8. M. C. Wani, H. L. Taylor, M. E. Wall, P. Coggon, A. T. McPhail, Plant antitumor agents. VI. Isolation and structure of taxol, a novel antileukemic and antitumor agent from *Taxus brevifolia*. *J. Am. Chem. Soc.* **93**, 2325-2327 (1971).
9. D. L. Klayman *et al.*, Isolation of artemisinin (Qinghaosu) from *Artemisia annua* growing in the United States. *J. Nat. Prod.* **47**, 715-717 (1984).

10. A. Fleming, On the antibacterial action of cultures of a Penicillium, with special reference to their use in the isolation of *B. influenzæ*. *Br. J. Exp. Pathol.* **10**, 226-236 (1929).
11. M. I. Hutchings, A. W. Truman, B. Wilkinson, Antibiotics: past, present and future. *Curr. Opin. Microbiol.* **51**, 72-80 (2019).
12. A. W. Alberts *et al.*, Mevinolin: a highly potent competitive inhibitor of hydroxymethylglutaryl-coenzyme A reductase and a cholesterol-lowering agent. *Proc. Natl. Acad. Sci. U.S.A.* **77**, 3957-3961 (1980).
13. A. L. Demain, S. Sanchez, Microbial drug discovery: 80 years of progress. *J. Antibiot.* **62**, 5-16 (2009).
14. D. J. Newman, G. M. Cragg, Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* **79**, 629-661 (2016).
15. D. J. Newman, G. M. Cragg, Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770-803 (2020).
16. C. F. Stratton, D. J. Newman, D. S. Tan, Cheminformatic comparison of approved drugs from natural product versus synthetic origins. *Bioorg. Med. Chem. Lett.* **25**, 4802-4807 (2015).
17. P. Hunter, Harnessing nature's wisdom: turning to nature for inspiration and avoiding her follies. *EMBO Rep.* **9**, 838-840 (2008).
18. G. T. Carter, Natural products and Pharma 2011: strategic changes spur new opportunities. *Nat. Prod. Rep.* **28**, 1783-1789 (2011).
19. Y. Chen, M. Garcia de Lomana, N.-O. Friedrich, J. Kirchmair, Characterization of the chemical space of known and readily obtainable natural products. *J. Chem. Inf. Model.* **58**, 1518-1532 (2018).
20. A. R. Carroll, B. R. Copp, R. A. Davis, R. A. Keyzers, M. R. Prinsep, Marine natural products. *Nat. Prod. Rep.* **37**, 175-223 (2020).
21. W. Bergmann, D. C. Burke, Contributions to the study of marine products. XXXIX. The nucleosides of sponges. III.[1] Spongothymidine and Spongouridine[2]. *J. Org. Chem.* **20**, 1501-1507 (1955).
22. S. Sagar, M. Kaur, K. P. Minneman, Antiviral lead compounds from marine sponges. *Mar. Drugs* **8**, 2619-2638 (2010).
23. J. W. Blunt, B. R. Copp, R. A. Keyzers, M. H. G. Munro, M. R. Prinsep, Marine natural products. *Nat. Prod. Rep.* **32**, 116-211 (2015).
24. J. A. Zumberge *et al.*, Demosponge steroid biomarker 26-methylstigmastane provides evidence for Neoproterozoic animals. *Nat. Ecol. Evol.* **2**, 1709-1714 (2018).
25. M. W. Taylor, R. Radax, D. Steger, M. Wagner, Sponge-associated microorganisms: evolution, ecology, and biotechnological potential. *Microbiol. Mol. Biol. Rev.* **71**, 295-347 (2007).
26. N. S. Webster, M. W. Taylor, Marine sponges and their microbial symbionts: love and other relationships. *Environ. Microbiol.* **14**, 335-346 (2012).
27. J. Piel *et al.*, Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge *Theonella swinhoei*. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16222-16227 (2004).
28. V. J. Paul, M. P. Puglisi, R. Ritson-Williams, Marine chemical ecology. *Nat. Prod. Rep.* **23**, 153-180 (2006).
29. J. Shang *et al.*, Cheminformatic insight into the differences between terrestrial and marine originated natural products. *J. Chem. Inf. Model.* **58**, 1182-1193 (2018).
30. S. A. M. Khalifa *et al.*, Marine natural products: a source of novel anticancer drugs. *Mar. Drugs* **17**, 491 (2019).
31. R. Sakai, G. T. Swanson, Recent progress in neuroactive marine natural products. *Nat. Prod. Rep.* **31**, 273-309 (2014).
32. R. C. F. Cheung, T. B. Ng, J. H. Wong, Y. Chen, W. Y. Chan, Marine natural products with anti-inflammatory activity. *Appl. Microbiol. Biotechnol.* **100**, 1645-1666 (2016).
33. M. Donia, M. T. Hamann, Marine natural products and their potential applications as anti-infective agents. *Lancet Infect. Dis.* **3**, 338-348 (2003).
34. B. J. Monk, H. Dalton, I. Benjamin, A. Tanovic, Trabectedin as a new chemotherapy option in the treatment of relapsed platinum sensitive ovarian cancer. *Curr. Pharm. Des.* **18**, 3754-3769 (2012).
35. Y. Hirata, D. Uemura, Halichondrins—antitumor polyether macrolides from a marine sponge. *Pure Appl. Chem.* **58**, 701-710 (1986).
36. U. Swami, I. Chaudhary, M. H. Ghalib, S. Goel, Eribulin—a review of preclinical and clinical studies. *Crit. Rev. Oncol. Hematol.* **81**, 163-184 (2012).

37. T. D. Aicher *et al.*, Total synthesis of halichondrin B and norhalichondrin B. *J. Am. Chem. Soc.* **114**, 3162-3164 (1992).
38. J. Li, A. Amatuni, H. Renata, Recent advances in the chemoenzymatic synthesis of bioactive natural products. *Curr. Opin. Chem. Biol.* **55**, 111-118 (2020).
39. K. J. Weissman, P. F. Leadlay, Combinatorial biosynthesis of reduced polyketides. *Nat. Rev. Microbiol.* **3**, 925-936 (2005).
40. J. Staunton, K. J. Weissman, Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.* **18**, 380-416 (2001).
41. C. Hertweck, The biosynthetic logic of polyketide diversity. *Angew. Chem. Int. Ed. Engl.* **48**, 4688-4716 (2009).
42. J. Wang *et al.*, Biosynthesis of aromatic polyketides in microorganisms using type II polyketide synthases. *Microb. Cell Fact.* **19**, 110 (2020).
43. I. Abe, H. Morita, Structure and function of the chalcone synthase superfamily of plant type III polyketide synthases. *Nat. Prod. Rep.* **27**, 809-838 (2010).
44. A. T. Keatinge-Clay, The structures of type I polyketide synthases. *Nat. Prod. Rep.* **29**, 1050-1073 (2012).
45. J. Beld, E. C. Sonnenschein, C. R. Vickery, J. P. Noel, M. D. Burkart, The phosphopantetheinyl transferases: catalysis of a post-translational modification crucial for life. *Nat. Prod. Rep.* **31**, 61-108 (2014).
46. B. Lowry, X. Li, T. Robbins, D. E. Cane, C. Khosla, A turnstile mechanism for the controlled growth of biosynthetic intermediates on assembly line polyketide synthases. *ACS Cent. Sci.* **2**, 14-20 (2016).
47. M. E. Horsman, T. P. A. Hari, C. N. Boddy, Polyketide synthase and non-ribosomal peptide synthetase thioesterase selectivity: logic gate or a victim of fate? *Nat. Prod. Rep.* **33**, 183-202 (2016).
48. M. A. Fischbach, C. T. Walsh, Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.* **106**, 3468-3496 (2006).
49. C. Khosla, Y. Tang, A. Y. Chen, N. A. Schnarr, D. E. Cane, Structure and mechanism of the 6-deoxyerythronolide B synthase. *Annu. Rev. Biochem.* **76**, 195-221 (2007).
50. L. Fang *et al.*, Broadened glycosylation patterning of heterologously produced erythromycin. *Biotechnol. Bioeng.* **115**, 2771-2777 (2018).
51. J. Piel, Biosynthesis of polyketides by *trans*-AT polyketide synthases. *Nat. Prod. Rep.* **27**, 996-1047 (2010).
52. H. G. Menzella *et al.*, Combinatorial polyketide biosynthesis by *de novo* design and rearrangement of modular polyketide synthase genes. *Nat. Biotechnol.* **23**, 1171-1176 (2005).
53. R. V. O'Brien, R. W. Davis, C. Khosla, M. E. Hillenmeyer, Computational identification and analysis of orphan assembly-line polyketide synthases. *J. Antibiot.* **67**, 89-97 (2014).
54. R. Ueoka, M. Bortfeld-Miller, B. I. Morinaka, J. A. Vorholt, J. Piel, Toblerols: cyclopropanol-containing polyketide modulators of antibiosis in *Methylobacteria*. *Angew. Chem. Int. Ed. Engl.* **57**, 977-981 (2018).
55. N. Moebius *et al.*, Biosynthesis of the respiratory toxin bongkrekic acid in the pathogenic bacterium *Burkholderia gladioli*. *Chem. Biol.* **19**, 1164-1174 (2012).
56. L. P. Partida-Martinez, C. Hertweck, A gene cluster encoding rhizoxin biosynthesis in "*Burkholderia rhizoxina*", the bacterial endosymbiont of the fungus *Rhizopus microsporus*. *ChemBioChem* **8**, 41-45 (2007).
57. S. P. Niehs *et al.*, Insect-associated bacteria assemble the antifungal butenolide gladiofungin by non-canonical polyketide chain termination. *Angew. Chem. Int. Ed. Engl.* **59**, 23122-23126 (2020).
58. J. Piel, A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of *Paederus* beetles. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 14002-14007 (2002).
59. P. D. Walker, A. N. M. Weir, C. L. Willis, M. P. Crump, Polyketide β-branching: diversity, mechanism and selectivity. *Nat. Prod. Rep.*, Advance Article (2021).
60. E. J. N. Helfrich, J. Piel, Biosynthesis of polyketides by *trans*-AT polyketide synthases. *Nat. Prod. Rep.* **33**, 231-316 (2016).
61. H. Jenke-Kodama, T. Börner, E. Dittmann, Natural biocombinatorics in the polyketide synthase genes of the actinobacterium *Streptomyces avermitilis*. *PLoS Comput. Biol.* **2**, e132 (2006).
62. T. Nguyen *et al.*, Exploiting the mosaic structure of *trans*-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat. Biotechnol.* **26**, 225-233 (2008).

63. A. Nivina, K. P. Yuet, J. Hsu, C. Khosla, Evolution and diversity of assembly-line polyketide synthases. *Chem. Rev.* **119**, 12524-12547 (2019).
64. E. J. N. Helfrich *et al.*, Automated structure prediction of *trans*-acyltransferase polyketide synthase products. *Nat. Chem. Biol.* **15**, 813-821 (2019).
65. A. S. Eustáquio, J. E. Janso, A. S. Ratnayake, C. J. O'Donnell, F. E. Koehn, Spliceostatin hemiketal biosynthesis in *Burkholderia* spp. is catalyzed by an iron/α-ketoglutarate–dependent dioxygenase. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E3376-E3385 (2014).
66. J. J. J. van der Hooft *et al.*, Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.* **49**, 3297-3314 (2020).
67. A. Crits-Christoph, N. Bhattacharya, M. R. Olm, Y. S. Song, J. F. Banfield, Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity. *Genome Res.* **31**, 239-250 (2020).
68. K. Blin *et al.*, antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81-W87 (2019).
69. K. M. Fisch *et al.*, Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting. *Nat. Chem. Biol.* **5**, 494-501 (2009).
70. A. Kampa *et al.*, Metagenomic natural product discovery in lichen provides evidence for a family of biosynthetic pathways in diverse symbioses. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E3129-E3137 (2013).
71. R. Ueoka *et al.*, Genome-based identification of a plant-associated marine bacterium as a rich natural product source. *Angew. Chem. Int. Ed. Engl.* **57**, 14519-14523 (2018).
72. M. Rust *et al.*, A multiproducer microbiome generates chemical diversity in the marine sponge *Mycale hentscheli*. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9508-9518 (2020).
73. J. Piel, M. Rust, "6.05 - Metagenome Mining" in Comprehensive Natural Products III, H.-W. Liu, T. P. Begley, Eds. (Elsevier, Oxford, 2020), vol. 6, pp. 50-89.
74. V. Torsvik, J. Goksoyr, F. L. Daae, High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.* **56**, 782-787 (1990).
75. L. F. Roesch *et al.*, Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J.* **1**, 283-290 (2007).
76. K. B. Heidelberg, J. A. Gilbert, I. Joint, Marine genomics: at the interface of marine microbial ecology and biodiscovery. *Microb. Biotechnol.* **3**, 531-543 (2010).
77. J. Overmann, "Principles of Enrichment, Isolation, Cultivation, and Preservation of Prokaryotes" in The Prokaryotes, E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, F. Thompson, Eds. (Springer, Berlin, Heidelberg, 2013), pp. 149-207.
78. M. Achtman, M. Wagner, Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.* **6**, 431-440 (2008).
79. P. Yarza *et al.*, Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635-645 (2014).
80. L. A. Hug *et al.*, A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
81. D. H. Parks *et al.*, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533-1542 (2017).
82. P. Hugenholtz, B. M. Goebel, N. R. Pace, Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180**, 4765-4774 (1998).
83. K. T. Konstantinidis, R. Rosselló-Móra, Classifying the uncultivated microbial majority: a place for metagenomic data in the *Candidatus* proposal. *Syst. Appl. Microbiol.* **38**, 223-230 (2015).
84. C. T. Brown *et al.*, Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208-211 (2015).
85. M. F. Freeman *et al.*, Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. *Science* **338**, 387-390 (2012).
86. R. W. King, J. D. Bauer, S. F. Brady, An environmental DNA-derived type II polyketide biosynthetic pathway encodes the biosynthesis of the pentacyclic polyketide erdacin. *Angew. Chem. Int. Ed. Engl.* **48**, 6257-6261 (2009).
87. Z. Feng, J. H. Kim, S. F. Brady, Fluostatins produced by the heterologous expression of a TAR reassembled environmental DNA derived type II PKS gene cluster. *J. Am. Chem. Soc.* **132**, 11902-11903 (2010).
88. M. C. Wilson *et al.*, An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58-62 (2014).

89. J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, R. M. Goodman, Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, R245-R249 (1998).
90. G. J. Olsen, D. J. Lane, S. J. Giovannoni, N. R. Pace, D. A. Stahl, Microbial ecology and evolution: a ribosomal RNA approach. *Annu. Rev. Microbiol.* **40**, 337-365 (1986).
91. R. Jain, M. C. Rivera, J. A. Lake, Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3801-3806 (1999).
92. C. R. Woese, Bacterial evolution. *Microbiol. Rev.* **51**, 221-271 (1987).
93. K. Boye, E. Høgdall, M. Borre, Identification of bacteria using two degenerate 16S rDNA sequencing primers. *Microbiol. Res.* **154**, 23-26 (1999).
94. G. C. Baker, J. J. Smith, D. A. Cowan, Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods* **55**, 541-555 (2003).
95. S. G. Tringe, P. Hugenholtz, A renaissance for the pioneering 16S rRNA gene. *Curr. Opin. Microbiol.* **11**, 442-446 (2008).
96. A. Klindworth *et al.*, Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, e1 (2013).
97. M. F. Laursen, M. D. Dalgaard, M. I. Bahl, Genomic GC-content affects the accuracy of 16S rRNA gene sequencing based microbial profiling due to PCR bias. *Front. Microbiol.* **8**, 1934 (2017).
98. J. P. Brooks *et al.*, The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol.* **15**, 66 (2015).
99. S. M. Karst *et al.*, Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nat. Biotechnol.* **36**, 190-195 (2018).
100. Z. Charlop-Powers *et al.*, Urban park soil microbiomes are a rich reservoir of natural product biosynthetic diversity. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 14811-14816 (2016).
101. Z. Charlop-Powers *et al.*, Global biogeographic sampling of bacterial secondary metabolism. *eLife* **4**, e05048 (2015).
102. N. J. Gotelli, R. K. Colwell, Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* **4**, 379-391 (2001).
103. D. A. Cowart, K. R. Murphy, C.-H. C. Cheng, Metagenomic sequencing of environmental DNA reveals marine faunal assemblages from the West Antarctic Peninsula. *Mar. Genomics* **37**, 148-160 (2018).
104. C. C. Tebbe, W. Vahjen, Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. *Appl. Environ. Microbiol.* **59**, 2657-2665 (1993).
105. T. Thomas, J. Gilbert, F. Meyer, "Metagenomics: A guide from sampling to data analysis" in The Role of Bioinformatics in Agriculture, S. Kumar, Ed. (Apple Academic Press, New York, 2014), pp. 357-383.
106. S. Bag *et al.*, An Improved Method for High Quality Metagenomics DNA Extraction from Human and Environmental Samples. *Sci. Rep.* **6**, 26775 (2016).
107. S. G. Tringe, E. M. Rubin, Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* **6**, 805-814 (2005).
108. A. Santiago *et al.*, Processing faecal samples: a step forward for standards in microbial community analysis. *BMC Microbiol.* **14**, 112 (2014).
109. S. F. Brady, Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. *Nat. Protoc.* **2**, 1297-1305 (2007).
110. J. Collins, B. Hohn, Cosmids: a type of plasmid gene-cloning vector that is packageable *in vitro* in bacteriophage lambda heads. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 4242-4246 (1978).
111. U. J. Kim, H. Shizuya, P. J. de Jong, B. Birren, M. I. Simon, Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res.* **20**, 1083-1085 (1992).
112. J. Wild, Z. Hradecna, W. Szybalski, Conditionally amplifiable BACs: switching from single-copy to high-copy vectors and genomic clones. *Genome Res.* **12**, 1434-1444 (2002).
113. H. Shizuya *et al.*, Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8794-8797 (1992).
114. K. S. Kakirde, L. C. Parsley, M. R. Liles, Size does matter: application-driven approaches for soil metagenomics. *Soil Biol. Biochem.* **42**, 1911-1923 (2010).
115. H. X. Wang, Z. L. Geng, Y. Zeng, Y. M. Shen, Enriching plant microbiota for a metagenomic library construction. *Environ. Microbiol.* **10**, 2684-2691 (2008).

116. Y. Ouyang *et al.*, Isolation of high molecular weight DNA from marine sponge bacteria for BAC library construction. *Mar. Biotechnol.* **12**, 318-325 (2010).
117. J. J. Banik, J. W. Craig, P. Y. Calle, S. F. Brady, Tailoring enzyme-rich environmental DNA clones: a source of enzymes for generating libraries of unnatural natural products. *J. Am. Chem. Soc.* **132**, 15661-15670 (2010).
118. Z. Charlop-Powers, J. J. Banik, J. G. Owen, J. W. Craig, S. F. Brady, Selective enrichment of environmental DNA libraries for genes encoding nonribosomal peptides and polyketides by phosphopantetheine transferase-dependent complementation of siderophore biosynthesis. *ACS Chem. Biol.* **8**, 138-143 (2013).
119. J. K. Bitok, C. Lemetre, M. A. Ternei, S. F. Brady, Identification of biosynthetic gene clusters from metagenomic libraries using PPTase complementation in a *Streptomyces* host. *FEMS Microbiol. Lett.* **364**, fnx155 (2017).
120. J. G. Owen, K. J. Robins, N. S. Parachin, D. F. Ackerley, A functional screen for recovery of 4'-phosphopantetheinyl transferase and associated natural product biosynthesis genes from metagenome libraries. *Environ. Microbiol.* **14**, 1198-1209 (2012).
121. C. H. Guan *et al.*, Signal mimics derived from a metagenomic analysis of the gypsy moth gut microbiota. *Appl. Environ. Microbiol.* **73**, 3669-3676 (2007).
122. D. M. Ekkers, M. S. Cretoiu, A. M. Kielak, J. D. van Elsas, The great screen anomaly—a new frontier in product discovery through functional metagenomics. *Appl. Microbiol. Biotechnol.* **93**, 1005-1020 (2012).
123. M. Ferrer, A. Beloqui, K. N. Timmis, P. N. Golyshin, Metagenomics for mining new genetic resources of microbial communities. *J. Mol. Microbiol. Biotechnol.* **16**, 109-123 (2009).
124. M. Trindade, L. J. van Zyl, J. Navarro-Fernández, A. A. Elrazak, Targeted metagenomics as a tool to tap into marine natural product diversity for the discovery and production of drug candidates. *Front. Microbiol.* **6**, 890 (2015).
125. R. Montaser, H. Luesch, Marine natural products: a new wave of drugs? *Future Med. Chem.* **3**, 1475-1489 (2011).
126. S. Hrvatin, J. Piel, Rapid isolation of rare clones from highly complex DNA libraries by PCR analysis of liquid gel pools. *J. Microbiol. Methods* **68**, 434-436 (2007).
127. J. G. Owen *et al.*, Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11797-11802 (2013).
128. Y. Zhang, J. P. P. Muyrers, G. Testa, A. F. Stewart, DNA cloning by homologous recombination in *Escherichia coli*. *Nat. Biotechnol.* **18**, 1314-1317 (2000).
129. O. Perlova *et al.*, Reconstitution of the myxothiazol biosynthetic gene cluster by Red/ET recombination and heterologous expression in *Myxococcus xanthus*. *Appl. Environ. Microbiol.* **72**, 7485-7494 (2006).
130. S. C. Wenzel *et al.*, Heterologous expression of a myxobacterial natural products assembly line in Pseudomonads via Red/ET recombineering. *Chem. Biol.* **12**, 349-356 (2005).
131. J. H. Kim *et al.*, Cloning large natural product gene clusters from the environment: piecing environmental DNA gene clusters back together with TAR. *Biopolymers* **93**, 833-844 (2010).
132. H. Wang *et al.*, RecET direct cloning and Redαβ recombineering of biosynthetic gene clusters, large operons or single genes for heterologous expression. *Nat. Protoc.* **11**, 1175-1190 (2016).
133. J. Vollmers, S. Wiegand, A. K. Kaster, Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! *PLoS One* **12**, e0169662 (2017).
134. J. Risse *et al.*, A single chromosome assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore sequencing data. *GigaScience* **4**, s13742–13015–10101–13746 (2015).
135. R. D. Stewart *et al.*, Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, 870 (2018).
136. K. Blin *et al.*, antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* **45**, W36-W41 (2017).
137. I. Saeed, S.-L. Tang, S. K. Halgamuge, Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res.* **40**, e34 (2012).
138. Y. Xu, F. Zhao, Single-cell metagenomics: challenges and applications. *Protein Cell* **9**, 501-510 (2018).
139. R. Stepanauskas, Single cell genomics: an individual look at microbes. *Curr. Opin. Microbiol.* **15**, 613-620 (2012).

140. P. C. Blainey, The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol. Rev.* **37**, 407-427 (2013).
141. R. S. Lasken, Single-cell genomic sequencing using Multiple Displacement Amplification. *Curr. Opin. Microbiol.* **10**, 510-516 (2007).
142. J. Sabina, J. H. Leamon, "Bias in Whole Genome Amplification: Causes and Considerations" in Whole Genome Amplification. Methods in Molecular Biology, T. Kroneis, Ed. (Humana Press, New York, 2015), vol. 1347, pp. 15-41.
143. R. Stepanauskas *et al.*, Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat. Commun.* **8**, 84 (2017).
144. R. V. Grindberg *et al.*, Single cell genome amplification accelerates identification of the apratoxin biosynthetic pathway from a complex microbial assemblage. *PLoS One* **6**, e18565 (2011).
145. S. R. Vartoukian, R. M. Palmer, W. G. Wade, Strategies for culture of 'unculturable' bacteria. *FEMS Microbiol. Lett.* **309**, 1-7 (2010).
146. E. J. Stewart, Growing unculturable bacteria. *J. Bacteriol.* **194**, 4151-4160 (2012).
147. B. Berdy, A. L. Spoering, L. L. Ling, S. S. Epstein, *In situ* cultivation of previously uncultivable microorganisms using the ichip. *Nat. Protoc.* **12**, 2232-2242 (2017).
148. A. G. Atanasov *et al.*, Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discovery* **20**, 200-216 (2021).
149. H. B. Bode, B. Bethe, R. Höfs, A. Zeeck, Big effects from small changes: possible ways to explore nature's chemical diversity. *ChemBioChem* **3**, 619-627 (2002).
150. D. Park, G. Swayambhu, B. A. Pfeifer, Heterologous biosynthesis as a platform for producing new generation natural products. *Curr. Opin. Biotechnol.* **66**, 123-130 (2020).
151. B. A. Pfeifer, S. J. Admiraal, H. Gramajo, D. E. Cane, C. Khosla, Biosynthesis of complex polyketides in a metabolically engineered strain of *E. coli*. *Science* **291**, 1790-1792 (2001).
152. H. Zhang, Y. Wang, J. Wu, K. Skalina, B. A. Pfeifer, Complete biosynthesis of erythromycin A and designed analogs using *E. coli* as a heterologous host. *Chem. Biol.* **17**, 1232-1240 (2010).
153. J. M. Billingsley, A. B. DeNicola, Y. Tang, Technology development for natural product biosynthesis in *Saccharomyces cerevisiae*. *Curr. Opin. Biotechnol.* **42**, 74-83 (2016).
154. R. Liu, Z. Deng, T. Liu, *Streptomyces* species: ideal chassis for natural product discovery and overproduction. *Metab. Eng.* **50**, 74-84 (2018).
155. H. G. Floss, Combinatorial biosynthesis—potential and problems. *J. Biotechnol.* **124**, 242-257 (2006).
156. F. T. Wong, C. Khosla, Combinatorial biosynthesis of polyketides—a perspective. *Curr. Opin. Chem. Biol.* **16**, 117-123 (2012).

Chapter II

This chapter was published in Proceedings of the National Academy of Sciences of the United States of America (DOI: 10.1073/pnas.1919245117)

A Multiproducer Microbiome Generates Chemical Diversity in the Marine Sponge *Mycale hentscheli*

Michael Rust[1], Eric J. N. Helfrich[1], Michael F. Freeman[1,2+], Pakjira Nanudorn[1+], Christopher M. Field[1], Christian Rückert[3], Tomas Kündig[1], Michael J. Page[4], Victoria L. Webb[5], Jörn Kalinowski[3], Shinichi Sunagawa[1], Jörn Piel[1]

[1] Institute of Microbiology, ETH Zurich, 8093 Zurich, Switzerland.
[2] Department of Biochemistry, Molecular Biology, and Biophysics, and BioTechnology Institute, University of Minnesota–Twin Cities, St. Paul, MN 55108, USA.
[3] Institute for Genome Research and Systems Biology, Center for Biotechnology, Universität Bielefeld, 33594 Bielefeld, Germany.
[4] National Institute of Water and Atmospheric Research Ltd. (NIWA), Nelson 7010, New Zealand.
[5] National Institute of Water and Atmospheric Research Ltd. (NIWA), Wellington 6021, New Zealand.

[+] These authors contributed equally.

Author Contributions

MR, EJNH, MFF and JP designed research. MJP and VLW collected sponge samples. MR, EJNH and MFF constructed and screened metagenomic libraries. CMF and SS performed assembly and binning. MR, EJNH and JP identified and assigned biosynthetic gene clusters and predicted natural products. MR and CMF performed phylogenetic analysis and constructed phylogenetic trees. CR and JK performed nanopore sequencing and assembly. MR, PN and TK performed PCRs to complete and clone gene clusters. MR and PN performed heterologous expression experiments and HPLC-MS analyses. MR, EJNH and JP wrote the manuscript with contributions from all authors.

Significance

Sponges, one of the oldest extant animal phyla, stand out among marine organisms as sources of structurally diverse bioactive natural products. Previous work on chemically rich sponges identified single “superproducer” symbionts in their microbiomes that generate the majority of the bioactive compounds known from their host. Here, we present a contrasting scenario for the New Zealand sponge *Mycale hentscheli* in which a multiproducer consortium is the basis of chemical diversity. Other than the known cocktail of cytotoxins, metagenomic and functional data support further chemical diversity originating from various uncultivated bacterial lineages. The results provide a rationale for distinct patterns of chemical variation observed within sponge species and reinforce uncultured microbes as promising source of compounds with therapeutic potential.

Abstract

Bacterial specialized metabolites are increasingly recognized as important factors in animal–microbiome interactions: for example, by providing the host with chemical defenses. Even in chemically rich animals, such compounds have been found to originate from individual members of more diverse microbiomes. Here, we identified a remarkable case of a moderately complex microbiome in the sponge host *Mycale hentscheli* in which multiple symbionts jointly generate chemical diversity. In addition to bacterial pathways for three distinct polyketide families comprising microtubule-inhibiting peloruside drug candidates, mycalamide-type contact poisons, and the eukaryotic translation-inhibiting pateamines, we identified extensive biosynthetic potential distributed among a broad phylogenetic range of bacteria. Biochemical data on one of the orphan pathways suggest a previously unknown member of the rare polytheonamide-type cytotoxin family as its product. Other than supporting a scenario of cooperative symbiosis based on bacterial metabolites, the data provide a rationale for the chemical variability of *M. hentscheli* and could pave the way toward biotechnological peloruside production. Most bacterial lineages in the compositionally unusual sponge microbiome were not known to synthesize bioactive metabolites, supporting the concept that "microbial dark matter" harbors diverse producer taxa with as yet unrecognized drug discovery potential.

There is strong evidence that microbiome-derived specialized metabolites play key roles in health, disease, reproductive success, evolutive diversification, and the survival of macroorganisms.[1, 2] An important example is defensive symbiosis, in which hosts benefit from protective substances synthesized by a microbial partner.[3] Since few symbiotic producers have been successfully cultivated, such interactions have to date been uncovered in a relatively small number of cases, but their identification in taxonomically diverse hosts and symbionts suggests that defensive symbiosis is rather prevalent in nature.

As the oldest extant metazoans[4] and prolific sources of bioactive natural products,[5] marine sponges offer particularly intriguing opportunities to study symbiotic interactions. Although featuring a simple body plan that lacks specialized tissues, many sponges are complex multispecies organisms containing hundreds to thousands of bacterial phylotypes at remarkable collective cell numbers.[6] Little is known about which of the prokaryotes detected by 16S rRNA gene surveys establish stable associations with sponges rather than being accumulated by filter-feeding or derived from non-specific colonization,[7, 8] and few experimentally validated functions in sponge-bacterial symbiosis have been uncovered.[9] For the lithistid sponge *Theonella swinhoei*, a source of an unusually wide array of natural products, we and collaborators recently identified symbiotic "Entotheonella" bacteria of the candidate phylum "Tectomicrobia" as the key producers of bioactive metabolites.[10] *T. swinhoei* comprises several sponge variants with diverse and mostly non-overlapping sets of bioactive metabolites.[11] In each of the Japanese chemotypes *T. swinhoei* Y and W, a single symbiont, "*Candidatus* Entotheonella factor"[10, 12, 13] or "*Ca.* Entotheonella serta",[14] respectively, produces all or almost all of the polyketide and peptide natural products known from the sponges. Each symbiont harbors diverse and almost orthogonal sets of biosynthetic gene clusters (BGCs), providing a rationale for the distinct chemistry of *T. swinhoei* variants. "Entotheonella" symbionts were also assigned to bioactive metabolites in the lithistid sponges *Discodermia calyx*[15, 16] and a Palauan chemotype of *T. swinhoei*.[17] In addition to "Entotheonella", which seems to be a widespread producer taxon,[10] the cyanobacterium *Oscillatoria spongeliae* has recently been identified as a source of halogenated natural products in dysideid sponges.[18-20] Many sponge natural products play suspected or proven roles in chemical defense[6] and have attracted much attention as sources for new therapeutics.[5] Commonly, these metabolites are exclusively known from sponges and exhibit pharmacological profiles that impart high drug potential.[21] Their low natural abundance, however, represents a major obstacle to drug development that might be overcome by developing bacterial production systems.[6]

Similar to *Theonella*, the sponge genus *Mycale* (order Poecilosclerida) is known as a rich and varied source of bioactive substances.[22] From New Zealand specimens of *Mycale* (*Carmia*) *hentscheli*,[23] three groups of cytotoxic polyketides with distinct modes of action have been reported, represented by the ribosome-inhibiting[24] contact poison mycalamide A (**1**),[25] the translation initiation inhibitor[26] pateamine A (**2**),[27] and the microtubule inhibitor[28] peloruside A (**3**) (**Fig. 1**).[29] Peloruside A has attracted attention as a promising anticancer agent, since it binds to a microtubule site distinct from inhibitors in clinical use.[30] However, attempts to establish a sponge mariculture system for peloruside production were abandoned after invasion of a destructive nudibranch grazer.[31] Further impeding drug development, the chemistry of *M. hentscheli* is highly variable, with individual specimens containing all, two, or one of the three polyketide groups.[32] Considering the

insights gained from the *T. swinhoei* model, understanding microbiome functions in *M. hentscheli* will be crucial for the sustainable use of this resource.

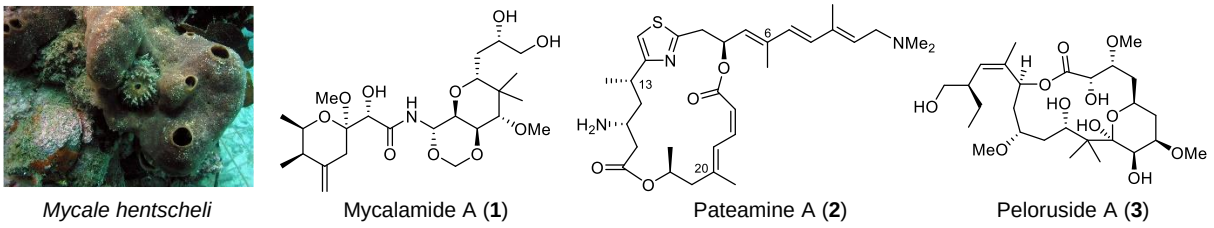


Figure 1. Representative natural products from the marine sponge *Mycale hentscheli*. Structures of selected congeners for the three polyketide families are shown. The methyl groups at numbered positions of pateamine A are thought to be introduced by a β-branching mechanism common for *trans*-AT PKSs.

Here, we provide evidence that the known natural products of *M. hentscheli* originate from its microbiota, thus further supporting the hypothesis that chemical functions contributed by bacteria play widespread and fundamental roles in this animal phylum. However, metagenomic data revealed a bacterial consortium in *M. hentscheli* that is, to our knowledge, phylogenetically distinct from that of any previously analyzed sponge. In an opposite scenario to that of *T. swinhoei*, biosynthetic genes are distributed among almost all microbiome members, with a multi-producer consortium being the collective source of the host chemistry. In addition to genes for the three known polyketide types, we identified BGC candidates for various further bioactive compounds previously unknown from this sponge, attributed to at least 19 distinct bacteria. The data show that sponges use multiple strategies to acquire and utilize bacterial chemicals. Moreover, they reinforce "microbial dark matter" as a rich discovery resource that harbors a wide range of previously unrecognized lineages with distinct and biomedically relevant chemistry.

Results

Identification of Gene Candidates for the *M. hentscheli* Polyketides

We initiated our study on *M. hentscheli* with consideration to the possible enzymatic origin of its known natural products, the mycalamides, pateamines, and pelorusides. Mycalamides belong to the pederin family, a group of defensive polyketides produced in remarkably diverse host-symbiont systems, comprising beetles,[33] psyllids,[34] lichens,[35] and sponges,[36, 37] as well as free-living bacteria[38, 39] with producers from at least four different phyla. We previously showed that the compounds from non-*Mycale* holobionts are enzymatically generated by a family of polyketide synthases (PKSs) termed *trans*-acyltransferase (*trans*-AT) PKSs.[33, 36, 40] For as-yet unknown reasons, *trans*-AT PKSs are the predominant enzyme family for complex polyketide biosynthesis in uncultivated symbionts studied to date.[41] The structure of pateamine contains three methyl groups (at C6, C13, and C20) (Fig. 1) at positions that suggest attachment by a β-branching mechanism.[42] This feature is ubiquitous for *trans*-AT PKSs, but

rare in *cis*-AT PKS pathways, the second large enzymatic source of complex polyketides. Peloruside lacks structural moieties that indicate its PKS type. Hypothesizing that *trans*-AT PKSs generate at least two of the three polyketide series, we performed metagenomic sequencing.

A Medium-Sized Microbiome Containing Diverse Natural Product Biosynthesis Genes

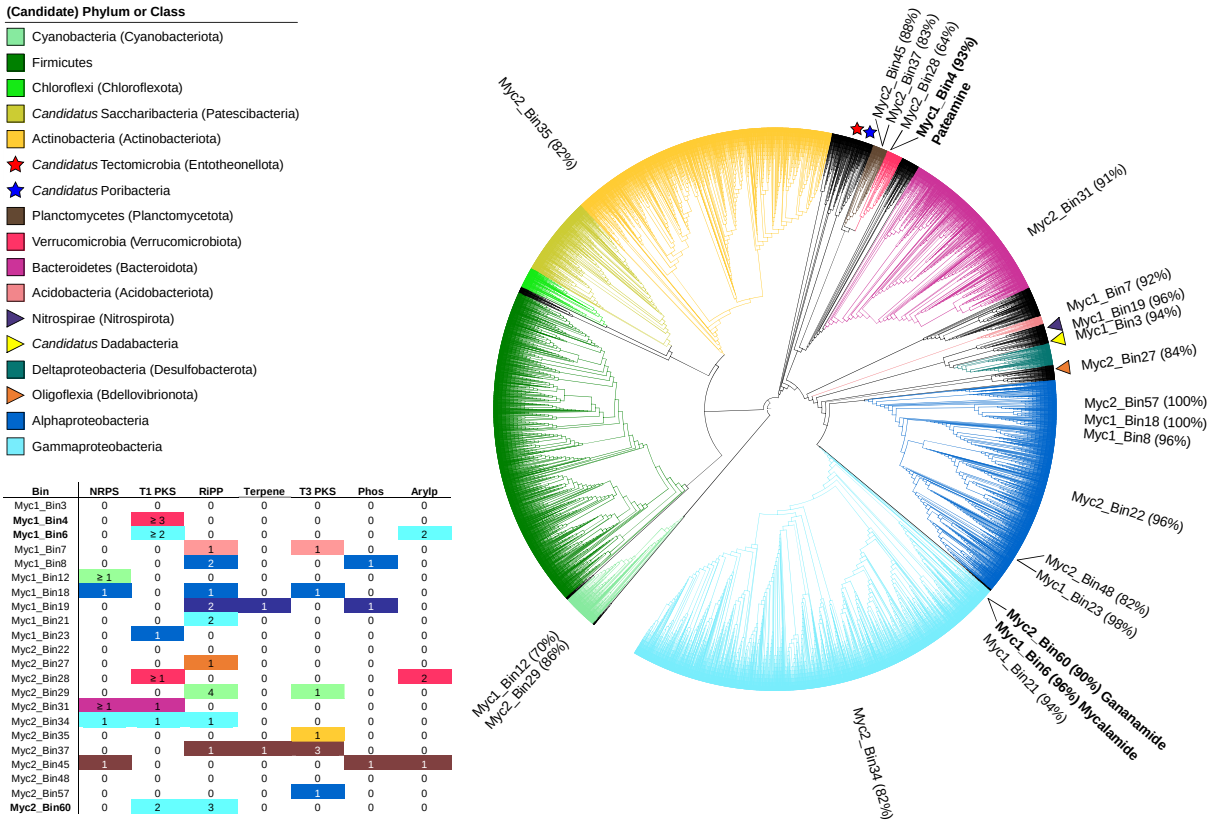
DNA was extracted from different *M. hentscheli* specimens that were positive for at least two of the three polyketides. Fast degradation of some DNA samples during purification posed a major challenge for downstream processing. Ultimately, we obtained high-molecular-weight DNA from two sponge specimens, designated as Myc1 (sponge ID: 1MJP40-24) and Myc2 (1MJP3-79.12) containing all three polyketides (**Table S1**), which was subjected to Illumina sequencing. Reads were assembled separately for the two sponge samples using metaSPAdes[43] (**Table S2**). We obtained initial insights into the general biosynthetic potential of the *M. hentscheli* microbiome by analyzing contigs >1,500 bp with the automated BGC detection tool antiSMASH[44] in combination with extensive manual analyses to identify non-canonical BGCs not detected by the software. The data revealed a plethora of BGC-containing contigs from diverse biosynthetic families (**Tables S3 and S4**). Since there is a close correlation between the module architecture of a PKS and the polyketide structure,[45] assumptions about the core BGC are possible if the structure is known and *vice versa*. Gratifyingly, BGC regions on some contigs correlated well with the structures of mycalamide and pateamine (see next section). In addition, many other contigs contained PKS or NRPS genes that could not be assigned to known compounds, suggesting that the *M. hentscheli* microbiome harbors a greater biosynthetic potential than previously expected. Since repetitive modular PKS regions prevented the assembly of some BGC regions, we also attempted sequencing one of the sponge specimens with a nanopore-based sequencing platform (MinION). However, this step did not improve the overall contig length and quality of the Illumina dataset.

Assembled contigs were binned to gain insights into the diversity of microorganisms and their BGC content. The relatively limited number of bins (14 for Myc1 and 20 for Myc2 with >60% estimated genome completeness, **Table S5**) revealed a microbiome of much lower diversity in *M. hentscheli* than previous 16S rRNA studies had suggested for other *Mycale* species[46] or for example, *T. swinhoei*.[47] Of 14 genomes in the Myc1 dataset, 12 were at least partially present in the Myc2 dataset. The 22 binned genomes were taxonomically classified (https://github.com/Ecogenomics/GTDBTk) and affiliated with eleven bacterial phyla (**Fig. 2**).

14 out of 22 closest bacterial neighbors identified by whole-genome phylogeny originated from cultivation-independent sequencing studies (**Fig. S1**). Based on the chemical richness of *M. hentscheli*, we had initially hypothesized that it might contain an "Entotheonella" relative as a talented producer. However, no "Entotheonella"-related genes were detected in our data and none of the *Mycale* bins were assigned to the candidate phylum "Tectomicrobia" to which "Entotheonella" belongs.[10] Further analysis showed that in contrast to the "Entotheonella" superproducers in *T. swinhoei*, BGCs are distributed across a large phylogenetic range of bacteria in *M. hentscheli* with most microbiome members harboring only few biosynthetic pathways (Fig. 2). The bins with a >90% estimated genome completeness were analyzed for the presence of central metabolism and amino acid biosynthesis pathways (**Table S6**).

Although some of these pathways are partially or completely absent in some bins, we did not observe any cases of extreme genome reduction as found for various intracellular symbionts.[48, 49] Furthermore, we analyzed these bins for the presence of clusters of

orthologous genes associated with symbiosis (**Table S7**). All bins harbor several of these 'symbiosis factors' with eukaryotic-like proteins, such as ankyrin repeats (COG0666), tetratricopeptide repeats (COG0457, COG0790), and WD40 proteins (COG2319, COG1520), being the most abundant.



Bin	NRPS	T1 PKS	RiPP	Terpene	T3 PKS	Phos	Arylp
Myc1_Bin3	0	0	0	0	0	0	0
Myc1_Bin4	0	≥ 3	0	0	0	0	0
Myc1_Bin6	0	≥ 2	0	0	0	0	2
Myc1_Bin7	0	0	1	0	1	0	0
Myc1_Bin8	0	0	2	0	0	1	0
Myc1_Bin12	≥ 1	0	0	0	0	0	0
Myc1_Bin18	1	0	1	0	1	0	0
Myc1_Bin19	0	0	2	1	0	1	0
Myc1_Bin21	0	0	2	0	0	0	0
Myc1_Bin23	0	1	0	0	0	0	0
Myc2_Bin22	0	0	0	0	0	0	0
Myc2_Bin27	0	0	1	0	0	0	0
Myc2_Bin28	0	≥ 1	0	0	0	0	2
Myc2_Bin29	0	0	4	0	1	0	0
Myc2_Bin31	≥ 1	1	0	0	0	0	0
Myc2_Bin34	1	1	1	0	0	0	0
Myc2_Bin35	0	0	0	0	1	0	0
Myc2_Bin37	0	0	1	1	3	0	0
Myc2_Bin45	1	0	0	0	0	1	1
Myc2_Bin48	0	0	0	0	0	0	0
Myc2_Bin57	0	0	0	0	1	0	0
Myc2_Bin60	0	2	3	0	0	0	0

Figure 2. Taxonomic classification of 22 *Mycale* bins and distribution of identified BGCs. Maximum-likelihood placement of the 22 bins with >60% estimated genome completeness (values given in parentheses behind each bin number) in the GTDB-Tk reference tree [50] consisting of 17,435 leaves. Cyanobacteria were used as outgroup. Colored stars denote the placement of the two candidate phyla "Poribacteria" and "Tectomicrobia" known from previous sponge microbiome studies. Colored triangles denote positions of *Mycale* bins affiliated with phyla with few known genomes (Nitrospirae, *Candidatus* Dadabacteria, Oligoflexia). The three bins harboring the mycalamide, pateamine, and gananamide BGCs are shown in bold. The number of BGCs per bin is listed in the table, color-coded according to the bin's assigned phylum. The ≥ sign was used if multiple biosynthetic contigs were found that might belong to the same pathway. Contigs encoding NRPS-PKS hybrid systems were listed as PKSs. Arylp, arylpolyene; Phos, phosphonate; T1, type 1; T3, type 3.

Insights into Mycalamide, Pateamine, and Peloruside Biosynthesis

The structural similarity of pederin-type compounds (diaphorin, nosperin, onnamide A, pederin, psymberin) is reflected in their shared PKS architectures.[41] This correlation, in combination with the binning data, facilitated the assignment of partial BGCs in the metagenomic dataset to the mycalamide pathway. In an initial assembly, we identified four contigs in the same bin with protein architectures that match the predicted mycalamide biosynthesis. These contigs were subsequently connected by PCR to yield the complete *myc* BGC (**Fig. 3a**, **Table S8**). Additionally, the BGC architecture was confirmed during an improved assembly, in which the complete locus was present on a single contig. The architecture of the hybrid *trans*-AT PKS–NRPS assembly line closely resembles those of the pederin[33] and onnamide[36] PKSs. An interesting feature shared with the pederin system is the large PKS gene *mycH* that does not correspond to any part of the polyketide structure. MycG encoded directly upstream of *mycH* belongs to a group of PKS-associated monooxygenases, for which biochemical data support a function as oxygen-inserting Baeyer–Villigerase acting on growing polyketide chains.[51] In the case of pederin and mycalamide, hydrolytic cleavage of the resulting ester moiety would be in agreement with the oxygenated polyketide terminus and missing polyketide portion. Based on the 16S rRNA gene sequence in the *myc* BGC-containing bin, the producer is affiliated with the marine group UBA10353 (94% sequence identity to its nearest neighbor[52]), an uncultivated gammaproteobacterial taxon that was previously not known as source of pederin-type compounds or other natural products. In agreement, the nearest neighbor in the genome-based phylogeny is an uncultivated gammaproteobacterium from the UBA10353 order associated with a glass sponge (Fig. S1).[53] The name "*Candidatus* Entomycale ignis" is proposed for the mycalamide producer.

Pateamine contains a rare *N*-methylated glycine starter. As a match for this diagnostically useful moiety, we identified in another bin the NRPS-PKS gene *pamA* encoding a predicted glycine-specific NRPS module with an *N*-methyltransferase (NMT) domain. To further interrogate a role in pateamine biosynthesis, the downstream PKS modules were analyzed by TransATor, a recently developed web application that allows functional assignment of *trans*-AT PKS pathways.[54] The software employs the correlation between phylogenetically similar ketosynthase (KS) domains and incoming intermediates carrying similar chemical moieties in the α- to γ-region around the thioester.[45] Predicted intermediates for the PKS modules in PamA showed good agreement with the C1 to C10 moiety of pateamine (**Table S9**). Two additional NRPS cyclization (Cy) domains at the C-terminus of PamB suggested the presence of an oxazol(in)e or thiazol(in)e moiety,[55] consistent with the pateamine thiazole unit. NRPS cyclization modules for aromatic azoles normally also contain serine- or cysteine-specific adenylation (A) domains and an oxidoreductase domain. A single, incomplete PKS gene, *pamC*, encoding such domains was identified in the same bin on another contig. We were unable to connect the two contigs by PCR or reassembly, and since *pamC* was preceded by a series of genes unrelated to natural product biosynthesis, the contigs might be located in different genome regions. Candidates for further *pam* PKS genes were identified on three additional contigs in that bin, which encoded enzymes consistent with the predicted missing biosynthetic steps. Successful connection of these contigs by PCR generated a continuous 50 kb fragment ending with a thioesterase (TE) region (**Fig. 3b**, **Table S10**). PamC contains various rare features; however, these are consistent with pateamine biosynthesis and include (i) an enoylreductase (ER) domain associated with the β-branching module for KS7 and needed to reduce an initially generated sp^2 center at C13, (ii) an aminotransferase (AMT) domain in

the downstream module that matches the amino group at C15, and (iii), as candidates for the C17 ester function that interrupts the pateamine chain, an NRPS condensation (C) domain in the terminal PamC module and an N-terminal KS on PamD predicted to accept acetyl starters. A similar C domain is also proposed to introduce an ester moiety in the malleilactone (= burkholderic acid) pathway.[56, 57] These features suggest that pateamine is biosynthesized by esterification of two separate polyketide chains rather than a mycalamide-type Baeyer–Villiger oxidation. Pateamine contains three β-branches at C6, C13, and C20, for which accessory biosynthetic genes, termed β-branching cassette, are required. These were identified on a third contig that could not be connected to the other fragments by PCR (*pamEFGHIM*). The 16S rRNA gene of the pateamine producer, with the proposed name "*Candidatus* Patea custodiens" (from the Latin word custos for guardian), was verified by PCR reamplification and sequencing. SILVA[52] analysis of the 16S rRNA gene revealed low sequence identity (90.3%) to a member of the Kiritimatiellaeota, a recently proposed phylum previously assigned to Verrucomicrobia.[58] This affiliation is consistent with whole genome-based phylogeny that identified members of the Kiritimatiellaeota as closest neighbors (Fig. S1).[50] Like for the mycalamide-assigned UBA10353 taxon, Kiritimatiellaeota were previously not known as a natural product source.

In the unbinned metagenomic fraction, a 55 kb contig encoding a seemingly complete *trans*-AT PKS assembly line (*pel* PKS, **Fig. 3c**, **Table S11**) attracted our attention. For the large PKS portion between KS4 and the terminal TE, the TransATor core structure prediction suggested a polyketide structure that is almost identical to the peloruside portion covering the macrocycle (**Fig. 3d**, **Table S12**). In contrast, the first PKS modules contained numerous non-canonical features that made a functional assignment challenging. More detailed analysis revealed that the unusual domain series starting with KS2 (KS^0-acyl carrier protein [ACP]-TE_B-KS^0-ACP-C) occurs in an almost identical sequence (KS^0-ACP-ACP-TE_B-KS^0-ACP-C) in the spliceostatin and thailanstatin PKSs,[59, 60] where it was assigned to a *Z* double bond, a feature also found at the matching peloruside moiety (C4 to C5). Interestingly, all three PKSs contain members of a phylogenetically distinct group of internal TE-like domains, previously misannotated as a dehydratase (DH) in the spliceostatin PKS, for which we proposed the name TE_B (for branching TE). Work to be published elsewhere on a TE_B homolog from the oocydin PKS suggests that the domain attaches an acetyl side chain to a hydroxyl function. In the biosynthesis of peloruside and the statins, this modification might facilitate elimination as a DH-independent mechanism to introduce double bonds. Moving further upstream along the *pel* PKS, the module containing KS1 is predicted to introduce an α-methyl-ß-hydroxyl moiety based on its domain architecture and the phylogeny of KS2, which again agrees with the peloruside moiety. This assignment suggests two alternatives for the origin of the remaining C_5 unit composed of C1 to C4 and C21 of peloruside: incorporation of a 2-methylbutyryl starter or iterative action of the KS1 module to elongate an acetyl starter twice.

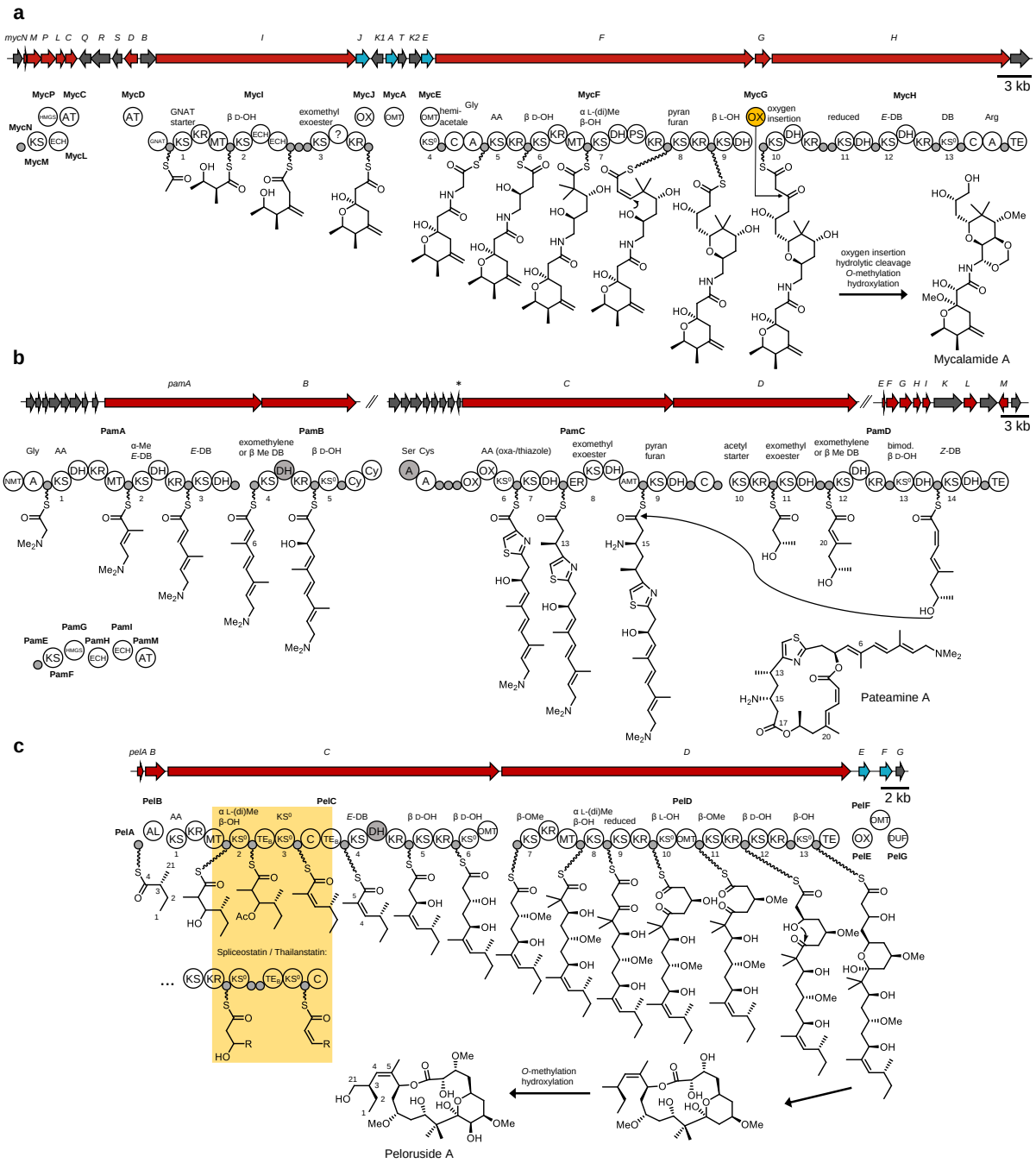

a
mycN M P L C Q R S D B
I
J K1 A T K2 E
F
G
H
3 kb
MycP MycC
MycN
MycM
MycL
MycD
MycI
MycJ MycA
MycE
MycF
MycG
MycH
GNAT starter
β D-OH
exomethyl exoester
hemi-acetale
Gly
AA
α L-(di)Me β-OH
pyran furan
β L-OH
oxygen insertion
reduced
E-DB
DB
Arg
oxygen insertion
hydrolytic cleavage
O-methylation
hydroxylation
Mycalamide A
b
pamA
B
*
C
D
E F G H I K L M
3 kb
PamA
PamB
PamC
PamD
Gly
AA
α-Me E-DB
E-DB
exomethylene or β Me DB
β D-OH
Ser Cys
AA (oxa-/thiazole)
exomethyl exoester
pyran furan
acetyl starter
exomethyl exoester
exomethylene or β Me DB
bimod. β D-OH
Z-DB
PamE
PamG
PamH
PamI
PamM
PamF
Pateamine A
c
pelA B
C
D
E F G
2 kb
PelA
PelB
PelC
PelD
PelE
PelF
PelG
AA
α L-(di)Me β-OH
E-DB
β D-OH
β-OMe
α L-(di)Me β-OH reduced
β L-OH
Spliceostatin / Thailanstatin:
O-methylation
hydroxylation
Peloruside A
d
TransATor prediction
linearized peloruside

Figure 3. Gene clusters and biosynthetic models for mycalamide, pateamine, and peloruside. Core PKS/NRPS genes are shown in red, tailoring biosynthetic genes in blue, and genes with unknown function in grey. Gaps between domains denote protein boundaries. Biosynthetic intermediates are shown tethered to the acyl/peptidyl carrier protein (ACP/PCP) domains (small grey circles). Predicted substrates are shown above the KS and A domains (exomethyl/exoester refers to a clade of KSs that mainly accept intermediates containing β-branched methyl groups or a β-branched ester in case of bryostatin). Domains depicted in grey are predicted to be nonfunctional. a) The *myc* gene cluster and biosynthetic model for mycalamide. The oxygenase thought to introduce oxygen into the growing polyketide chain is highlighted in orange. b) The *pam* gene cluster and biosynthetic model for pateamine. Double slashes denote separate contigs. c) The *pel* gene cluster and biosynthetic model for peloruside. d) TransATor-based structure prediction for the putative *pel* BGC. The linearized peloruside structure is shown for comparison. AA, amino acid; AL, acyl-CoA/ACP ligase; DB, double bond; DUF, domain of unknown function 955; ECH, enoyl-CoA hydratase; GNAT, GCN5-related *N*-acetyl transferase superfamily; HMGS, 3-hydroxy-3-methylglutaryl-CoA synthase homolog; KR, ketoreductase; KS^0, non-elongating KS; MT, *C*-methyltransferase; OMT, *O*-methyltransferase; OX, oxidoreductase; PS, pyran synthase; *, transposase gene.

To obtain initial biochemical insights into the *pel* pathway, we focused on the monodomain protein PelB. Its resemblance to acyl-CoA/ACP ligases initially suggested that it might load the starter unit onto the free-standing ACP PelA. We tested its substrate preference by producing PelB in *E. coli* and performing *in vitro* assays with a range of carboxylic acids (**4** to **17**) and coenzyme A (**Fig. S2a**). Monitoring reactions by high-performance liquid chromatography–mass spectrometry (HPLC-MS) revealed that PelB efficiently converted acetic acid (**4**) to acetyl-CoA but did not accept acids **5** to **17** (**Fig. S2b**). Unexpectedly, however, co-incubations with PelA failed to generate acylated ACP species for acetate or any other test substrate. One explanation for this result might be that PelB provides its acetyl substrate not to the ACP but to another acceptor, such as the TE_B. PelA could be loaded with an isobutyryl unit by an as-yet unknown enzyme, a scenario that is also supported by the predicted specificity of KS1 for an amino acid-type substrate rather than acetyl. Experiments to clarify this issue are underway but out of the scope of the current study, considering the challenge to characterize an assembly line with multiple non-canonical features from an uncultivated bacterium. Since the peloruside sequence was not binned, it might be located on a plasmid with a different tetranucleotide frequency than the rest of the producer genome.

High Natural Product Potential Collectively Encoded in Diverse Microbiome Members

Deeper analyses of the metagenome revealed many additional BGCs for PKSs, NRPSs, ribosomally synthesized and posttranslationally modified peptides (RiPPs), and other compounds that could not be assigned to reported *M. hentscheli* metabolites (**Fig. 4**). In the bin of "Patea custodiens", the Kiritimatiellaeota pateamine producer, we identified a small *trans*-AT PKS cluster (Fig. 4, cluster 3) that appears complete based on the presence of loading and termination modules. The terminal PKS modules are architecturally almost identical to those of the psymberin PKS,[37] suggesting an isocoumarin-type compound as the product. The same bin harbors four additional *trans*-AT PKS contigs (Fig. 4, clusters 1, 2, 4, and 5), including a large 34 kb fragment ending with a TE. "P. custodiens" therefore likely generates at least two additional polyketides that are currently unknown. The bin of the mycalamide producer "Entomycale ignis" harbors two small *cis*-AT PKS contigs (Fig. 4, clusters 7 and 8), suggesting the production of an additional polyketide. Another bin (assigned to "Caria hoplita",

see next section) contains a small, seemingly complete NRPS/PKS hybrid cluster (Fig. 4, cluster 9). In addition, the analyses revealed multiple *trans*-AT PKS contigs in the unbinned fraction (Fig. 4, clusters 14 to 18), including one encoding a large *trans*-AT PKS (Fig. 4, cluster 14) and a glycosyltransferase, suggesting a glycosylated polyketide. Another small and apparently complete *trans*-AT PKS BGC (Fig. 4, cluster 16) encodes an architecturally unusual PKS containing a domain of unknown function (DUF) and pyridoxal phosphate-dependent enzyme (PLP) that are similar to those assigned to sulfur heterocycle biosynthesis in leinamycin,[61] as well as a PedG/MycG-type Baeyer–Villigerase homolog. In summary, the *M. hentscheli* microbiome is particularly rich in *trans*-AT PKSs, pathways that are often found in uncultivated symbiotic bacteria, but also harbors various BGCs from additional natural product classes (Tables S3 and S4).

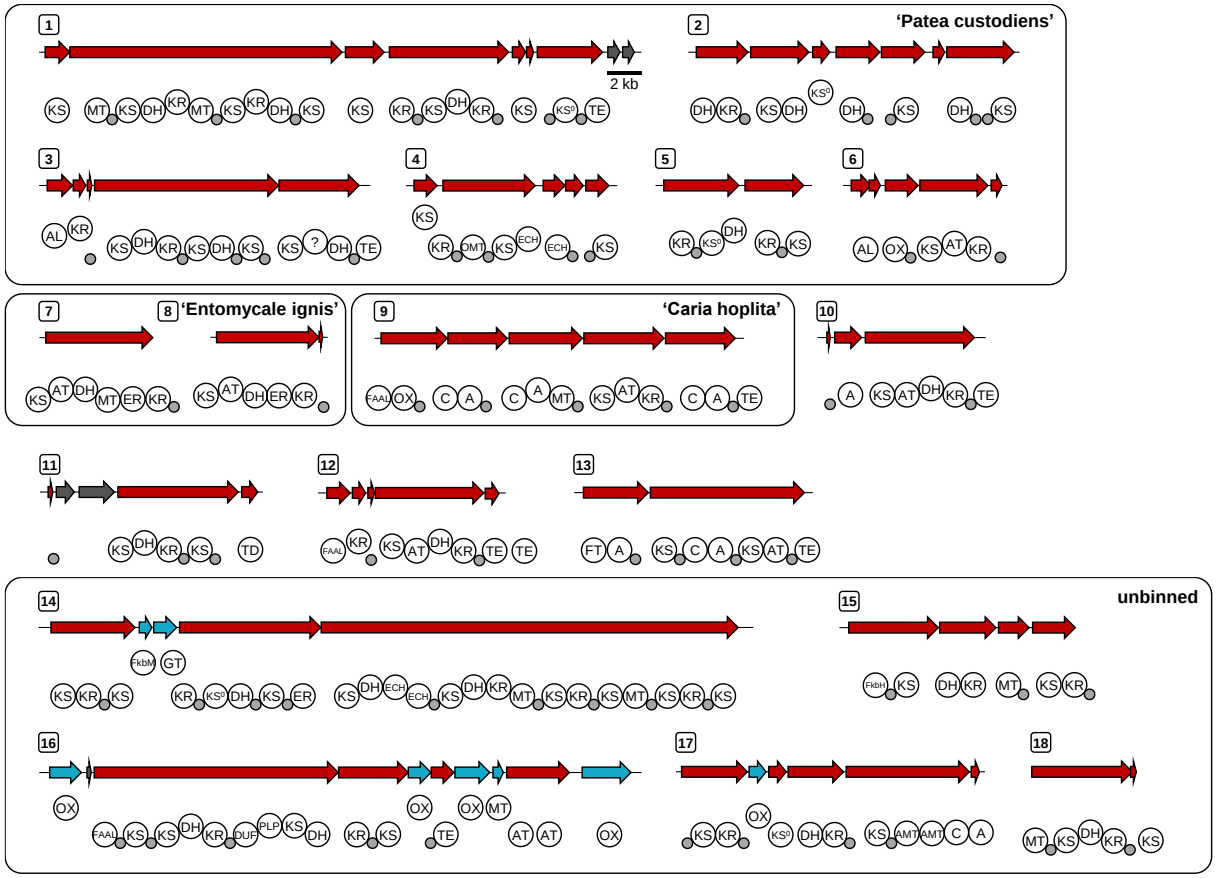


Figure 4. Orphan PKS gene clusters. Core PKS/NRPS genes are shown in red, tailoring biosynthetic genes in blue, and genes with unknown function in grey. Gaps between domains denote protein boundaries. Small grey circles represent ACP/PCP domains. Clusters 1 to 6 were identified in the pateamine producer "Patea custodiens", clusters 7 and 8 in the mycalamide producer "Entomycale ignis", cluster 9 in the gananamide producer "Caria hoplita", and clusters 14 to 18 in the unbinned fraction. DUF, domain of unknown function 2156; FAAL, fatty acyl-AMP ligase; FkbH, hydroxylase; FkbM, methyltransferase; FT, formyltransferase; GT, glycosyltransferase; PLP, pyridoxal phosphate-dependent enzyme; TD, terminal domain.

Functional Insights into Polytheonamide-Type Pathway in the *Nitrosococcaceae* Symbiont "Caria hoplita"

An unexpected orphan locus identified in the *M. hentscheli* metagenome (**Fig. 5a, Table S13**) closely resembles the polytheonamide (*poy*) cluster of "E. factor", the multiproducer of the sponge *T. swinhoei* Y.[62] Polytheonamides[63] are extraordinarily complex and rare cytotoxic peptides generated by a RiPP pathway involving 49 posttranslational modifications.[62, 64] The unmodified precursor consists of an N-terminal leader region and a C-terminal core that is processed by maturases during biosynthesis and ultimately cleaved off by proteolysis. For five of six polytheonamide maturation enzymes, homologs were identified in the *M. hentscheli* BGC (*gan* cluster). These comprise a radical *S*-adenosyl methionine (rSAM) epimerase (generating 18 D-amino acids in polytheonamides), a Ser/Thr dehydratase (*poy*: Thr1 dehydration), two rSAM *C*-methyltransferases (*poy*: 17 *C*-methylations), and an *N*-methyltransferase (*poy*: 8 Asn side-chain *N*-methylations). The core sequence of the precursor GanA likewise resembles the *poy* core in length and composition, but contains a characteristic GANANA repeat. We therefore provisionally named the orphan natural product gananamide.

To test whether the *gan* cluster is functional, we co-expressed in *E. coli* genes for the dehydratase homolog GanF and the precursor GanA, carrying an added N-terminal His_6 tag. As in the polytheonamide pathway, GanF efficiently dehydrated the threonine residue at core position 1 (**Fig. 5b and 5c**). Further co-expression trials with the epimerase gene *ganD*, however, failed to generate active enzyme. Since we previously showed that the core sequence largely dictates the D-amino acid pattern introduced by rSAM epimerases,[65] we performed individual co-expression experiments with PoyD and AerD to obtain clues about D-residues in gananamide. AerD is a recently identified epimerase with improved processivity in *E. coli* from the polytheonamide-type aeronamide pathway in *Microvirgula aerodenitrificans*.[66] For both epimerase homologs, additional peaks with different retention times appeared in the HPLC-MS data (**Fig. S3**). Subsequent application of a previously developed method that permits quantification and localization of D-residues through deuteration (ODIS, orthogonal D_2O-based induction system)[67] revealed six epimerized amino acids for PoyD at the N-terminal core portion of GanA and ten epimerizations for AerD (**Fig. 5d and 5e** and **Fig. S4-S6**). Experiments to identify the mature natural product in *M. hentscheli* are underway.

The partial 16S rRNA gene of the proposed gananamide producer, for which we suggest the name "*Candidatus* Caria hoplita" (based on hoplites, armed foot-soldiers that played a role in defending the ancient region Caria during the battle of Mycale), was completed by PCR amplification and sequencing. It was affiliated with the *Nitrosococcaceae* and has the highest sequence identity of 94.7% to an uncultured sediment bacterium belonging to this family. The closest neighbor based on available whole genomes is an unclassified gammaproteobacterium from a water purification plant metagenome.[68]

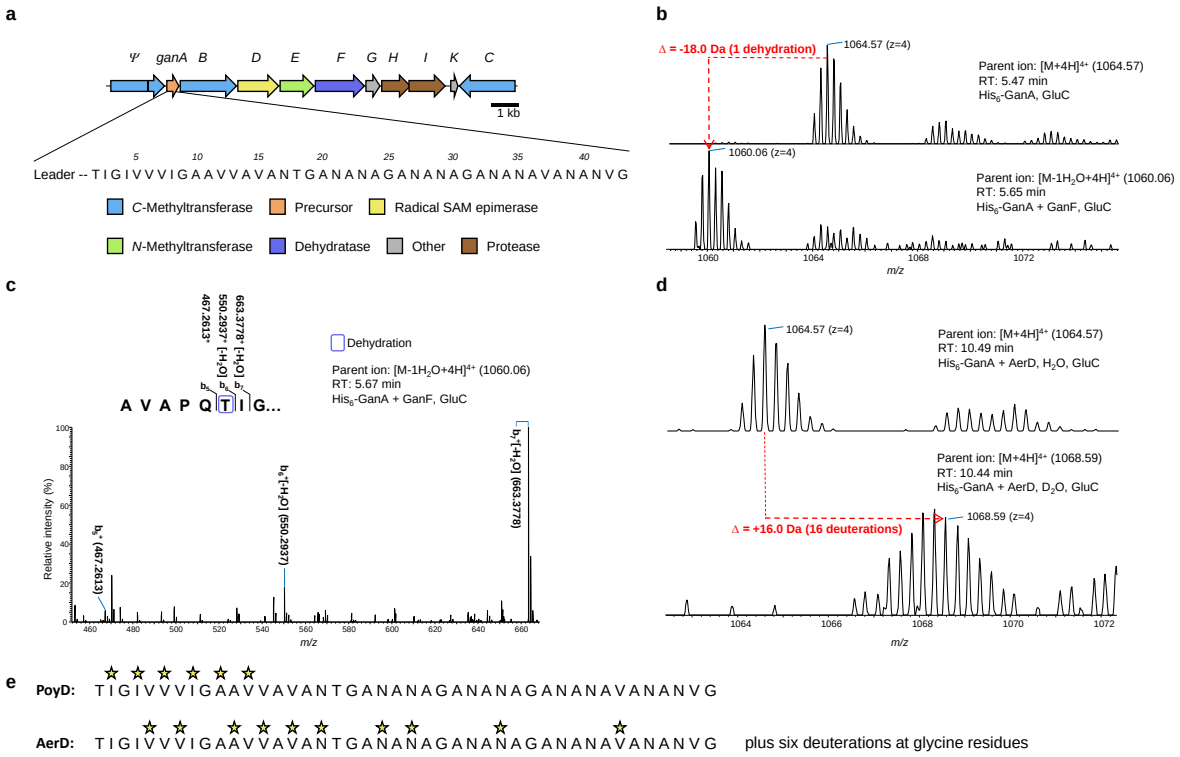


Figure 5. The *gan* BGC and functional characterization of modifying enzymes. a) The *gan* cluster with genes color-coded according to known enzymatic functions of the polytheonamide homologs. The core sequence of the precursor GanA (43 amino acids) containing the repetitive GANANA motif is shown. Ψ: rSAM *C*-methyltransferase pseudogene. b) MS spectra of the GluC-digested precursor GanA produced in *E. coli* without (top) and with (bottom) the dehydratase GanF. c) MS^2 spectrum of the GluC-digested precursor GanA co-produced with GanF, localizing the dehydration to Thr1 at the N-terminus of the core. d) MS spectra of GluC-digested precursor GanA co-produced with the epimerase AerD in H_2O and D_2O. A mass shift corresponding to a total of 16 deuterations was detected. e) Localization of epimerized amino acids in the core sequence of GanA introduced by the epimerases PoyD and AerD. RT, retention time.

Discussion

With more than 8,000 species being globally distributed, sponges comprise a successful and evolutionarily ancient animal phylum.[69] Many sponges harbor remarkably diverse microbial communities, for which various contributions to the host physiology and ecology have been proposed or, in fewer cases, experimentally demonstrated.[9] Among the latter is the provision of bioactive natural products that can be present as rich arrays in some holobionts.[70] Such studies have identified single members within more diverse microbiomes that generate some or all of the specialized metabolites known from the host.[10, 14, 20, 48] Here, we reveal a contrasting scenario, in which the source of three distinct cytotoxin families, the mycalamides,[25] pateamines,[27] and pelorusides,[29] is a complex chemistry-based symbiosis in the poecilosclerid sponge *M. hentscheli*. Rather than production being localized in individual bacteria within BGC-depleted microbiomes,[10, 14, 48] the data support a producer consortium, in which multiple phylogenetically diverse members contribute to the overall rich chemistry of the holobiont. Besides BGCs assigned to the three known polyketide classes, this microbiome

was found to contain numerous additional loci from distinct natural product families. Most of these BGCs appear to be intact, and functional studies on the cryptic polytheonamide-type gananamides provide additional evidence that richer chemistry is to be expected. Current work in our and collaborating groups aims at revealing the identity of these unknown metabolites.

M. hentscheli has attracted much attention based on the therapeutic potential of its compounds, with pelorusides being particularly promising anticancer drug candidates.[71] One of several obstacles in the pharmaceutical development[31] of this resource has been the high chemical variability among sponge specimens regarding the presence of individual compounds.[32] This complex variation contrasts with the association with distinct "superproducers" as in the case of *T. swinhoei* with a complete switch of metabolic profiles,[10, 14] or variations of *Prochloron* phylotypes associated with tropical ascidians.[72, 73] In *M. hentscheli*, a multiproducer consortium of variable composition is the likely reason for the diverse chemotypes. With knowledge on producers and their BGCs available, further development of *M. hentscheli* compounds could be expedited by permitting rapid PCR profiling of specimens, targeted isolation of producers from the host or alternative sources,[66, 74] or the establishment of heterologous expression systems.

Among all chemically assigned BGCs, the *pel* cluster assigned to pelorusides is the only one that could not be linked to a specific producer by binning, perhaps because it is located on a plasmid. To address this issue, collection of fresh specimens will be required to localize *pel* genes by *in situ* hybridization or single-cell genomics. Biosynthetic assignments of the architecturally aberrant *pel* BGC, as well as the fragmented pateamine (*pam*) PKS loci, were possible by retrobiosynthetic dissection and analysis of the KS domains,[45, 54] showcasing the value of *in silico* biosynthetic predictions using *trans*-AT PKS correlations. These correlations provide reasonably sound biosynthetic models that can be used as basis for further functional analyses and heterologous expression studies. The functions of the acetyl transferase PelB within the peloruside assembly line and the unusual domains at the start of PelC are currently under investigation in our laboratory. The relatively small size and apparent plasmid-based localization of the six-gene *pel* cluster and the pharmacological relevance of peloruside render this pathway a promising system for heterologous expression studies.

Metagenomic and single-cell datasets provide a valuable data resource to study how natural products evolve and disperse in symbiotic systems, and these are beginning to reveal intriguing patterns.[75] The identification of the mycalamide producer "Entomycale ignis" adds a further bacterial lineage to an already astonishing diversity of organisms producing pederin-type compounds, reinforcing intriguing questions about the evolution and dispersal of these natural products. Previously identified producers belong to Alpha-,[38] Beta-[34] and Gammaproteobacteria[33] as well as two Cyanobacteria[35, 39] and "Tectomicrobia",[10] and are mostly symbionts of unrelated hosts comprising sponges,[10] beetles,[33] psyllids,[34] and a lichen fungus.[35] The data support extensive horizontal gene transfer and retention in extraordinarily diverse host–symbiont systems, including intracellular organelle-like bacteria with minimalistic genomes[34] as well as multicellular prokaryotes featuring genomes of around 10 Mb.[10, 14] Pederin-type metabolites might thus represent ancient, "symbiotically privileged" natural products that have driven the evolution of multiple symbioses through host protection.[76, 77] However, more recently, pederin-type compounds were also reported from two free-living bacteria.[38, 39]

"Microbial dark matter", which comprises numerous deep-branching clades lacking cultivated members, has been proposed as a rich and largely untapped resource of novel chemistry.[78, 79] Studies on microbiomes of marine invertebrates provide direct experimental support for this hypothesis: in addition to the talented producer taxon "Entotheonella" within an uncultured candidate phylum,[10, 14] they have uncovered lineages such as "Endobugula",[80] "Endohaliclona",[48] "Endolissoclinum",[49] "Didemnitutus",[81] and "Endobryopsis"[82] as new natural product sources from uncultivated life. The current study further expands the range of producers by members of Kiritimatiellaeota,[58] the UBA10353 taxon, *Nitrosococcaceae*, Verrucomicrobia, and other groups. This growing prokaryotic diversity collectively generates a wide range of bioactive compounds with chemical scaffolds that were largely unknown from conventionally screened bacterial taxa such as actinomycetes. The data suggest that, with cultivation-independent studies becoming more routine, continued functional exploration of "microbial dark matter" will substantially change our understanding of bacterial specialized metabolism.

Materials and Methods

Instrumentation

Ultra-performance liquid chromatography-heated electrospray ionization mass spectrometry was performed on a Thermo Scientific Q Exactive mass spectrometer coupled to a Dionex Ultimate 3000 UPLC system.

Sample Collection

Mycale hentscheli specimens were collected in Pelorus Sound at the northern end of the South Island of New Zealand and stored in *RNA*later™ as described previously.[83]

Isolation of Metagenomic DNA

Metagenomic DNA was isolated from one gram of sponge sample as described previously[84] with slight modifications. Briefly, the sponge material was ground to a fine powder under liquid nitrogen, transferred to a 50 mL falcon tube containing 10 mL sponge lysis buffer and incubated at 60 °C for 20 min. The sample was extracted two times with phenol-/chloroform/isoamyl alcohol (25:24:1, v:v:v) and the aqueous phase was extracted with chloroform. The DNA was precipitated by adding one-tenth volume of 3 M sodium acetate (pH 7) and 1 volume isopropanol. The sample was gently mixed by inverting the tube several times and then incubated at room temperature for 30 min. The tube was centrifuged at 10,000× g and 4 °C for 30 min. The supernatant was removed and the DNA pellet was washed twice with 70% ice-cold ethanol. The sample was centrifuged at 10,000× g and 4 °C for 20 min after each washing step. The pellet was air-dried for 5 min and then resuspended in 500 µL pre-warmed Tris-HCl buffer (5 mM, pH 8.5). The reasons for the fast degradation of some DNA samples is currently unknown. A fast downstream processing after the sponge lysis step was crucial for obtaining high-molecular-weight DNA.

Metagenome Sequencing, Assembly and Binning

Sequencing libraries (2× 250 bp, paired-end reads) were prepared from purified metagenomic DNA and sequenced using an Illumina HiSeq2500 platform. BBDuk (v37.55, Joint Genome Institute) was first used in right-trimming mode with a kmer length of 23 down to 11 and a Hamming distance of 1 to filter out sequencing adapters. A second pass with a kmer length of 31 and a Hamming distance of 1 was used to filter out PhiX sequences. A third and final pass performed quality trimming on both read ends with a Phred score cutoff of 15 and an average quality score cutoff of 20, with reads under 30 bp or containing Ns subsequently rejected.

For the initial assembly (fast and memory-efficient), all paired-end reads from both sponge specimens (Myc1 and Myc2) were combined and assembled using MEGAHIT[85] (v1.1.1) with "--k-list 39,59,79,99" and otherwise, default parameters. Assembly with metaSPAdes[43] (usually better assembly statistics) required more RAM than the 3 TB we had available, so the paired-end reads of each readset were first normalized to reduce the complexity of the assembly graph. BBNorm (v37.55, Joint Genome Institute) was used with a default kmer length of 31, a minimum depth of 2 and a target depth of 80 to downsample sequences of high depth and filter out unique kmers.

The normalized paired-end and unnormalized singleton reads of each read set were assembled using metaSPAdes (v3.11.0) without the error correction module, but otherwise default parameters. Scaffolds smaller than 1,500 bp were then filtered out. The quality-controlled paired-end reads were aligned to the assembled metaSPAdes scaffolds using BWA[86] (v0.7.15-r1140) and the alignments sorted by SAMtools[87] (v1.3.1). Coverage depth across the scaffolds was calculated using the MetaBAT2[88] (v2.12.1) jgi_summarize_bam_contig_depths script and this information was then used by MetaBAT2 to bin the scaffolds with default parameters. The quality of the bins was assessed using the CheckM[89] (v1.0.11) lineage workflow, which included taxonomic assignment, with plots to visualize these results also produced by CheckM. To phylogenetically locate the bins, GTDBTk's classify workflow (v0.16; *https://github.com/Ecogenomics/GTDBTk*) was run with default parameters. The resulting tree was then manipulated in R (v3.5.1) with the ape package (v5.1) to, for instance, find the nearest neighbours of each bin. Trees were visualised with the Interactive Tree of Life tool[90] (v4.4.2).

Genome Annotations

For Table S6, bins of interest were annotated by RASTtk workflow[91] and subsystems were manually analyzed in the SEED viewer. For Table S7, bins of interest were annotated with eggNOG-mapper[92] (v1.0.3) and the number of genes matching OGs of interest were counted. HMMER (v3.2.1; *http://hmmer.org/*) program hmmscan was used to identify and count matches for TIGRFAMs of interest, using the trusted cutoff (--cut_tc) for potential hits.

Secondary Metabolite Cluster Prediction

Bioinformatic analysis of natural product genes was conducted as described previously.[10] Briefly, all assembled contigs >1500 bp were subjected to the antiSMASH standalone toolkit (v4.0.2) combined with manual BLAST analysis and conserved domain searches of uncertain regions. All manual annotation and routine bioinformatic analysis were performed using Geneious (v7.1.9) created by Biomatters (available from *http://www.geneious.com*). *Trans*-AT PKS gene clusters were further analyzed by TransATor (*https://transator.ethz.ch/*).

PCR-Based Verification of Clusters and 16S rRNA Sequences

PCRs for gap closing and gene amplification were performed with Q5 High-Fidelity DNA Polymerase (New England BioLabs). A typical PCR (25 µL) contained 1× Q5 reaction buffer, 200 µM dNTPs, 0.5 µM of each primer (**Table S14** no. 45 to 80 for the pateamine cluster, no. 81 to 89 for 16S rRNA genes, no. 90 to 105 for orphan cluster no. 16, no. 106 to 115 for orphan cluster no. 9, no. 116 to 123 for orphan cluster no. 3), 20 to 50 ng template DNA, and 0.5 U Q5 High-Fidelity DNA Polymerase. The reaction was heated to 98 °C for 30 s, followed by 30 cycles of: 98 °C for 10 s, 62 °C for 20 s, 72 °C for 30 s/kb DNA target sequence. At the end, a final incubation at 72 °C for 2 min was performed. DNA fragments were either directly sequenced or subcloned into the pCR-Blunt II-TOPO vector using the Zero Blunt TOPO PCR Cloning Kit (Invitrogen). Plasmids were transformed into *E. coli* DH5α and sequenced using the M13 forward and reverse primers.

Gene Expression and Protein Production and Purification

Expression constructs were constructed using either typical PCR and restriction-endonuclease-mediated cloning techniques, fusion PCR or Gibson assembly. A typical PCR (25 µL) contained 1× Q5 reaction buffer, 200 µM dNTPs, 0.5 µM of each primer (Table S14 no. 1 to 30 for the gananamide cluster, no. 31 to 44 for the peloruside cluster), 20-50 ng template DNA, and 0.5 U Q5 High-Fidelity DNA Polymerase. The reaction was heated to 98 °C for 30 s, followed by 30 cycles of: 98 °C for 10 s, 62 °C for 20 s, 72 °C for 30 s/kb DNA target sequence. At the end, a final incubation at 72 °C for 2 min was performed. For fusion PCR, 0.5 µL of the initial PCRs were used as template and only the outermost flanking primers were used. Gibson assembly was performed as per manufacturer's instructions (Gibson Assembly Master Mix; New England BioLabs). *E. coli* BL21 (DE3) were transformed with plasmids and expression cultures were inoculated from overnight cultures in a 1:100 (v:v) dilution in TB medium. Cultures were grown at 37 °C, 200 rpm to an OD_{600} of 1.2 to 1.5, cooled on ice for 30 min, gene expression was induced by adding 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and 0.2% (w/v) L-arabinose, and the cultures were incubated at 16 °C, 200 rpm overnight. After harvesting the cells by centrifugation at 3,220× g and 4 °C for 10 min, the pellet was resuspended in lysis buffer (50 mM NaH_2PO_4 pH 8, 300 mM NaCl, 10 mM imidazole, 10% glycerol), lysed by sonication and centrifuged at 12,000× g and 4 °C for 30 min. Proteins were purified using the Protino Ni-NTA resin (Macherey-Nagel) according to the manufacturer's protocol.

PelA and PelB *in Vitro* Assays and HPLC-MS Analysis

Nhis-pelA in pCDFDuet-1 was expressed in *E. coli* BAP1 [(93)] in order to obtain activated (phosphopantetheinylated) ACPs. *Nhis-pelB* in pET28b was expressed in *E. coli* BL21 (DE3). Cultures were grown in 50 mL TB medium at 30 °C, 200 rpm to an OD_{600} of 1.2 in 250 mL flasks. Gene expression was induced by adding 0.5 mM IPTG. Eluted proteins were desalted using a PD MiniTrap G-25 column (GE Healthcare Biosciences) in 50 mM Tris buffer, pH 7.8, containing 100 mM NaCl, 50 mM KCl, and 5% glycerol. Incubations were performed in 50 mM Tris pH 7.8, containing 50 mM $MgCl_2$, 1 mM carboxylic acid substrate, 1 mM ATP, 0.5 mM CoASH, and 1 µM PelA/PelB in a total reaction volume of 50 µL. After incubation at 25 °C for 3 h, 50 µL of acetonitrile were added to quench the reaction. The reaction mixture was centrifuged at 20,000× g for 10 min and the supernatant was analyzed by HPLC-MS on a Phenomenex Kinetex 2.6 µm XB-C18 100 Å (150 × 2.1 mm) column for PelB assays and a Phenomenex Aeris WIDEPORE 3.6 µm C4 (50 × 2.1 mm) for assays including PelA. The column was heated to

50 °C and the solvents used were water with 0.1% (v/v) formic acid (solvent A) and acetonitrile with 0.1% (v/v) formic acid (solvent B). For the PelB assays, a flow rate of 0.5 mL/min with solvent B at 5% from 0 to 2 min, 5% to 98% from 2 to 12 min, 98% from 12 to 15 min, 98% to 5% from 15 to 17 min, and 5% from 17 to 19 min was used. ESI-MS was performed in positive ion mode, with a spray voltage of 3,500 V, a capillary temperature of 280 °C, probe heater temperature of 475 °C and an S-Lens RF level of 50. Full MS was performed at a resolution of 140,000 (AGC target 1e6, maximum IT 150 ms, range 100 to 1,000 *m/z*). For the assays including PelA, 0.1% trifluoroacetic acid was added to the solvents instead of formic acid. A flow rate of 0.2 mL/min with solvent B at 10% from 0 to 2 min, 10% to 50% from 2 to 5 min, 50% to 98% from 5 to 12 min, 98% from 12 to 18 min, 98% to 10% from 18 to 19 min and 10% from 19 to 20 min was used. ESI-MS was performed in positive ion mode, with a spray voltage of 3,500 V, a capillary temperature of 280 °C, probe heater temperature of 475 °C and an S-Lens RF level of 100. Full MS was performed at a resolution of 140,000 (AGC target 1e6, maximum IT 150 ms, range 870 to 2,000 *m/z*).

Expression Conditions for Gananamide Proteins

Nhis-ganA in pCDFDuet-1 with or without *ganF* in pACYCDuet-1 or *ganD* in pBAD-Myc-HisA were expressed in *E. coli* BL21 (DE3). Cultures were grown at 37 °C, 250 rpm to an OD_{600} of 1.6 to 2.0, cooled on ice for 30 min, gene expression was induced with 0.1 mM IPTG and 0.2% L-arabinose for epimerase induction. The cultures were then incubated at 16 °C, 200 rpm overnight. The cells were harvested by centrifugation at 3,220× g and 4 °C for 10 min. The cell pellets were resuspended, lysed and the Nhis-GanA precursor purified as described above. The elution fractions were concentrated using Amicon Ultra 0.5 mL centrifugal filters with a 3 kDa MWCO (Merck).

ODIS Experiments

E. coli Tuner (DE3) was co-transformed with *Nhis-ganA* precursor in pACYCDuet-1, and *poyD* or *aerD* in pCDFBAD-Myc-HisA,[66] which is derived from pBAD/Myc-His A with the native origin of replication replaced by that of pCDFDuet, and plated on LB agar containing chloramphenicol (25 µg/mL) and ampicillin (100 µg/mL). Two separate 50 mL Falcon tubes containing TB medium (15 mL), chloramphenicol (25 µg/mL) and ampicillin (100 µg/mL) were inoculated with overnight culture in a 1:100 (v:v) dilution, and incubated at 250 rpm, 37 °C to an OD_{600} of 1.5 to 2.0. The cultures were then cooled on ice for 30 min, induced with 0.1 mM IPTG, and incubated at 250 rpm, 16 °C for 16 h. The cultures were centrifuged at 3,220× g and 4 °C for 10 min, and the supernatant removed. The cell pellets were then washed with TB medium (2× 15 mL) to remove any residual IPTG. In the second wash, the cells were incubated at 200 rpm, 16 °C for 1 h to further metabolize intracellular IPTG. The washed cell pellet was reconstituted in TB medium in D_2O (15 mL) containing 0.2% (w/v) L-arabinose in D_2O and incubated at 250 rpm, 16 °C for 24 h. The cells were harvested by centrifugation at 3,220× g and 4 °C for 10 min, and subjected to protein purification. The elution fractions were concentrated as described above.

Proteolytic Digests and HPLC-MS/MS Analysis

Endoproteinase GluC digests (40 µL) were typically conducted using 19 µL of concentrated elution fraction, 20 µL of 2× GluC reaction buffer, and 1 µL endoproteinase GluC (0.2 mg/mL). The reaction mixtures were incubated at 37 °C overnight. After incubation, 40 µl of acetonitrile were added to quench the reaction. The reaction mixtures were centrifuged at 20,000× g for

10 min and 15 µL of the mixtures were analyzed by HPLC-MS on a Phenomenex Aeris WIDEPORE 3.6 µm C4 (50 × 2.1 mm) column. The column was heated to 50 °C and the solvents used were water with 0.1% (v/v) formic acid (solvent A) and acetonitrile with 0.1% (v/v) formic acid (solvent B). A flow rate of 0.8 mL/min with solvent B at 5% from 0 to 2 min, 5% to 20% from 2 to 5 min, 20% to 65% from 5 to 15 min, 65% to 98% from 15 to 17 min, 98% from 17 to 18 min, and 98% to 5% from 18 to 20 min was used. ESI-MS was performed in positive ion mode, with a spray voltage of 3,500 V, a capillary temperature of 280 °C, probe heater temperature of 475 °C and an S-Lens RF level of 100. Full MS was performed at a resolution of 35,000 (AGC target 1e6, maximum IT 250 ms, range 300 to 2,000 *m/z*). Parallel reaction monitoring (PRM) or data-dependent MSMS was performed at a resolution of 17,500 (AGC target between 2e5 and 1e6, maximum IT between 100 ms and 500 ms, isolation windows in the range of 1.1 to 2.2 *m/z*) using a stepped NCE of 18, 20 and 22. Scan ranges, inclusion lists, charge exclusions, and dynamic exclusions were adjusted as needed.

For proteinase K digests (40 µL), 16 µL of concentrated elution fraction were added to 20 µL of 2× proteinase K buffer (50 mM Tris, pH 8.0, 2 mM $CaCl_2$, final concentration), and 4 µL of proteinase K (0.2 mg/mL). The reaction mixtures were incubated at 50 °C for 12 h, quenched by addition of 40 µL acetonitrile, and 15 µL of the mixture analyzed by HPLC-MS on a Phenomenex Kinetex 2.6 µm XB-C18 100 Å (150 × 2.1 mm) column. The column was heated to 50 °C and the solvents used were water with 0.1% (v/v) formic acid (solvent A) and acetonitrile with 0.1% (v/v) formic acid (solvent B). A flow rate of 0.5 mL/min with solvent B at 5% from 0 to 2 min, 5% to 50% from 2 to 5 min, 50% to 65% from 5 to 15 min, 65% to 98% from 15 to 17 min, 98% from 17 to 18 min, and 98% to 5% from 18 to 20 min was used. ESI-MS was performed in positive ion mode, with a spray voltage of 3,500 V, a capillary temperature of 280 °C, probe heater temperature of 475 °C and an S-Lens RF level of 100. Full MS was performed at a resolution of 35,000 (AGC target 1e6, maximum IT 250 ms, range 300 to 2,000 *m/z*). Parallel reaction monitoring (PRM) or data-dependent MSMS was performed at a resolution of 17,500 (AGC target between 2e5 and 1e6, maximum IT between 100 ms and 500 ms, isolation windows in the range of 1.1 to 2.2 *m/z*) using a stepped NCE of 18, 20 and 22. Scan ranges, inclusion lists, charge exclusions, and dynamic exclusions were adjusted as needed.

MS/MS Analysis - MaxQuant

For MS-based identification of proteinase K-treated peptide fragments, the program MaxQuant[94] (v1.5.2.8) was used. The following parameters were changed from default settings in a typical run: variable modifications, as needed with customized masses made in the Andromeda configuration tab; digestion mode, unspecific; maximum peptide mass (daltons), 10,000; minimum peptide length for unspecific search, 4; maximum peptide length for unspecific search, 50. On completion of the run, the evidence text file was used to map candidate peptides with MSMS scan numbers. The fragments were manually verified and annotated using the Xcalibur Qual Browser software (Thermo Fisher Scientific).

Data Availability

The metagenomic sequencing project of *M. hentscheli* has been deposited at the National Center for Biotechnology Information under BioProject ID PRJNA603662. Raw reads have been deposited at the Sequence Read Archive (accession no. PRJNA603662). Accession numbers of the individual bins are SAMN14054217–SAMN14054250. Information on the mycalamide, pateamine, and peloruside pathways was uploaded to the Minimum Biosynthetic

Information about a Biosynthetic Gene Cluster (MIBiG) database (IDs BGC0002055, BGC0002057, and BGC0002056, respectively).

Acknowledgements

This project has received funding from ETH Research Grant ETH-26 17-1, Swiss National Science Foundation Grants 205321 and 205320, and the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme Grant 742739 and the Helmut Horten Foundation. J.P. is grateful for an Investigator Grant of the Gordon and Betty Moore Foundation. P.N. was supported by a Swiss Government Excellence Scholarship. We thank the Functional Genomics Center Zurich for Illumina sequencing. We also thank M. Korneli for the construction of the pCDFBAD-Myc-HisA expression vector and A. Bhushan for providing the pCDFBAD-*aerD* plasmid.

References

1. J. L. Cleary, A. R. Condren, K. E. Zink, L. M. Sanchez, Calling all hosts: bacterial communication in situ. *Chem.* **2**, 334-358 (2017).
2. M. R. Wilson, L. Zha, E. P. Balskus, Natural product discovery from the human microbiome. *J. Biol. Chem.* **292**, 8546-8552 (2017).
3. L. V. Flórez, P. H. W. Biedermann, T. Engl, M. Kaltenpoth, Defensive symbioses of animals with prokaryotic and eukaryotic microorganisms. *Nat. Prod. Rep.* **32**, 904-936 (2015).
4. J. A. Zumberge *et al.*, Demosponge steroid biomarker 26-methylstigmastane provides evidence for Neoproterozoic animals. *Nat. Ecol. Evol.* **2**, 1709-1714 (2018).
5. M. F. Mehbub, J. Lei, C. Franco, W. Zhang, Marine sponge derived natural products between 2001 and 2010: trends and opportunities for discovery of bioactives. *Mar. Drugs* **12**, 4539-4577 (2014).
6. U. Hentschel, J. Piel, S. M. Degnan, M. W. Taylor, Genomic insights into the marine sponge microbiome. *Nat. Rev. Microbiol.* **10**, 641-654 (2012).
7. J. Tout *et al.*, Redefining the sponge-symbiont acquisition paradigm: sponge microbes exhibit chemotaxis towards host-derived compounds. *Environ. Microbiol. Rep.* **9**, 750-755 (2017).
8. M. W. Taylor *et al.*, 'Sponge-specific' bacteria are widespread (but rare) in diverse marine environments. *ISME J.* **7**, 438-443 (2013).
9. N. S. Webster, T. Thomas, The sponge hologenome. *mBio* **7**, e00135-00116 (2016).
10. M. C. Wilson *et al.*, An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58-62 (2014).
11. C. J. Wegerski, J. Hammond, K. Tenney, T. Matainaho, P. Crews, A serendipitous discovery of isomotuporin-containing sponge populations of *Theonella swinhoei*. *J. Nat. Prod.* **70**, 89-94 (2007).
12. G. Lackner, E. E. Peters, E. J. N. Helfrich, J. Piel, Insights into the lifestyle of uncultured bacterial natural product factories associated with marine sponges. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E347-E356 (2017).
13. M. J. Helf, A. Jud, J. Piel, Enzyme from an uncultivated sponge bacterium catalyzes *S*-methylation in a ribosomal peptide. *ChemBioChem* **18**, 444-450 (2017).
14. T. Mori *et al.*, Single-bacterial genomics validates rich and varied specialized metabolism of uncultivated *Entotheonella* sponge symbionts. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 1718-1723 (2018).
15. T. Wakimoto *et al.*, Calyculin biogenesis from a pyrophosphate protoxin produced by a sponge symbiont. *Nat. Chem. Biol.* **10**, 648-655 (2014).
16. Y. Nakashima, Y. Egami, M. Kimura, T. Wakimoto, I. Abe, Metagenomic analysis of the sponge *Discodermia* reveals the production of the cyanobacterial natural product kasumigamide by 'Entotheonella'. *PLoS One* **11**, e0164468 (2016).
17. E. W. Schmidt, A. Y. Obraztsova, S. K. Davidson, D. J. Faulkner, M. G. Haygood, Identification of the antifungal peptide-containing symbiont of the marine sponge *Theonella swinhoei* as a novel δ-proteobacterium, "*Candidatus* Entotheonella palauensis". *Mar. Biol.* **136**, 969-977 (2000).

18. M. D. Unson, N. D. Holland, D. J. Faulkner, A brominated secondary metabolite synthesized by the cyanobacterial symbiont of a marine sponge and accumulation of the crystalline metabolite in the sponge tissue. *Mar. Biol.* **119**, 1-11 (1994).
19. P. M. Flatt *et al.*, Identification of the cellular site of polychlorinated peptide biosynthesis in the marine sponge *Dysidea* (*Lamellodysidea*) *herbacea* and symbiotic cyanobacterium *Oscillatoria spongeliae* by CARD-FISH analysis. *Mar. Biol.* **147**, 761-774 (2005).
20. V. Agarwal *et al.*, Metagenomic discovery of polybrominated diphenyl ether biosynthesis by marine sponges. *Nat. Chem. Biol.* **13**, 537-543 (2017).
21. W. H. Gerwick, B. S. Moore, Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. *Chem. Biol.* **19**, 85-98 (2012).
22. L. J. Habener, J. N. A. Hooper, A. R. Carroll, Chemical and biological aspects of marine sponges from the family Mycalidae. *Planta Med.* **82**, 816-831 (2016).
23. P. R. Bergquist, P. J. Fromont, "The Marine Fauna of New Zealand: Porifera, Demospongiae, Part 4 (Poecilosclerida)" in New Zealand Oceanographic Institute Memoir 96. (New Zealand Oceanographic Institute, 1988), pp. 1-197.
24. S. A. Dyshlovoy *et al.*, Mycalamide A shows cytotoxic properties and prevents EGF-induced neoplastic transformation through inhibition of nuclear factors. *Mar. Drugs* **10**, 1212-1224 (2012).
25. N. B. Perry, J. W. Blunt, M. H. G. Munro, L. K. Pannell, Mycalamide A, an antiviral compound from a New Zealand sponge of the genus *Mycale*. *J. Am. Chem. Soc.* **110**, 4850-4851 (1988).
26. W.-K. Low *et al.*, Inhibition of eukaryotic translation initiation by the marine natural product pateamine A. *Mol. Cell* **20**, 709-722 (2005).
27. P. T. Northcote, J. W. Blunt, M. H. G. Munro, Pateamine: a potent cytotoxin from the New Zealand marine sponge, *Mycale* sp. *Tetrahedron Lett.* **32**, 6411-6414 (1991).
28. K. A. Hood *et al.*, Peloruside A, a novel antimitotic agent with paclitaxel-like microtubule-stabilizing activity. *Cancer Res.* **62**, 3356-3360 (2002).
29. L. M. West, P. T. Northcote, C. N. Battershill, Peloruside A: a potent cytotoxic macrolide isolated from the New Zealand marine sponge *Mycale* sp. *J. Org. Chem.* **65**, 445-449 (2000).
30. T. N. Gaitanos *et al.*, Peloruside A does not bind to the taxoid site on β-tubulin and retains its activity in multidrug-resistant cell lines. *Cancer Res.* **64**, 5063-5067 (2004).
31. M. J. Page, S. J. Handley, P. T. Northcote, D. Cairney, R. C. Willan, Successes and pitfalls of the aquaculture of the sponge *Mycale hentscheli*. *Aquaculture* **312**, 52-61 (2011).
32. M. J. Page, L. West, P. Northcote, C. Battershill, M. Kelly, Spatial and temporal variability of cytotoxic metabolites in populations of the New Zealand sponge *Mycale hentscheli*. *J. Chem. Ecol.* **31**, 1161-1174 (2005).
33. J. Piel, A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of *Paederus* beetles. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 14002-14007 (2002).
34. A. Nakabachi *et al.*, Defensive bacteriome symbiont with a drastically reduced genome. *Curr. Biol.* **23**, 1478-1484 (2013).
35. A. Kampa *et al.*, Metagenomic natural product discovery in lichen provides evidence for a family of biosynthetic pathways in diverse symbioses. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E3129-E3137 (2013).
36. J. Piel *et al.*, Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge *Theonella swinhoei*. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16222-16227 (2004).
37. K. M. Fisch *et al.*, Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting. *Nat. Chem. Biol.* **5**, 494-501 (2009).
38. C. Schleissner *et al.*, Bacterial production of a pederin analogue by a free-living marine alphaproteobacterium. *J. Nat. Prod.* **80**, 2170-2173 (2017).
39. A. Kust *et al.*, Discovery of a pederin family compound in a nonsymbiotic bloom-forming cyanobacterium. *ACS Chem. Biol.* **13**, 1123-1129 (2018).
40. J. Piel, G. Wen, M. Platzer, D. Hui, Unprecedented diversity of catalytic domains in the first four modules of the putative pederin polyketide synthase. *ChemBioChem* **5**, 93-98 (2004).
41. E. J. N. Helfrich, J. Piel, Biosynthesis of polyketides by *trans*-AT polyketide synthases. *Nat. Prod. Rep.* **33**, 231-316 (2016).
42. C. T. Calderone, Isoprenoid-like alkylations in polyketide biosynthesis. *Nat. Prod. Rep.* **25**, 845-853 (2008).
43. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824-834 (2017).

44. K. Blin *et al.*, antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* **45**, W36-W41 (2017).
45. T. Nguyen *et al.*, Exploiting the mosaic structure of *trans*-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat. Biotechnol.* **26**, 225-233 (2008).
46. T. Thomas *et al.*, Diversity, structure and convergent evolution of the global sponge microbiome. *Nat. Commun.* **7**, 11870 (2016).
47. U. Hentschel *et al.*, Molecular evidence for a uniform microbial community in sponges from different oceans. *Appl. Environ. Microbiol.* **68**, 4431-4440 (2002).
48. M. D. Tianero, J. N. Balaich, M. S. Donia, Localized production of defence chemicals by intracellular symbionts of *Haliclona* sponges. *Nat. Microbiol.* **4**, 1149-1159 (2019).
49. J. C. Kwan *et al.*, Genome streamlining and chemical defense in a coral reef symbiosis. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 20655-20660 (2012).
50. D. H. Parks *et al.*, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996-1004 (2018).
51. R. A. Meoded *et al.*, A polyketide synthase component for oxygen insertion into polyketide backbones. *Angew. Chem. Int. Ed. Engl.* **57**, 11644-11648 (2018).
52. C. Quast *et al.*, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-D596 (2013).
53. R. M. Tian *et al.*, The deep-sea glass sponge *Lophophysema eversa* harbours potential symbionts responsible for the nutrient conversions of carbon, nitrogen and sulfur. *Environ. Microbiol.* **18**, 2481-2494 (2016).
54. E. J. N. Helfrich *et al.*, Automated structure prediction of *trans*-acyltransferase polyketide synthase products. *Nat. Chem. Biol.* **15**, 813-821 (2019).
55. K. Bloudoffa, C. D. Fageb, M. A. Marahiel, T. M. Schmeinga, Structural and mutational analysis of the nonribosomal peptide synthetase heterocyclization domain provides insight into catalysis. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 95-100 (2017).
56. J. B. Biggins, M. A. Ternei, S. F. Brady, Malleilactone, a polyketide synthase-derived virulence factor encoded by the cryptic secondary metabolome of *Burkholderia pseudomallei* group pathogens. *J. Am. Chem. Soc.* **134**, 13192-13195 (2012).
57. J. Franke, K. Ishida, C. Hertweck, Genomics-driven discovery of burkholderic acid, a noncanonical, cryptic polyketide from human pathogenic *Burkholderia* species. *Angew. Chem. Int. Ed. Engl.* **51**, 11611-11615 (2012).
58. S. Spring *et al.*, Characterization of the first cultured representative of *Verrucomicrobia* subdivision 5 indicates the proposal of a novel phylum. *ISME J.* **10**, 2801-2816 (2016).
59. A. S. Eustáquio, J. E. Janso, A. S. Ratnayake, C. J. O'Donnell, F. E. Koehn, Spliceostatin hemiketal biosynthesis in *Burkholderia* spp. is catalyzed by an iron/α-ketoglutarate–dependent dioxygenase. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E3376-E3385 (2014).
60. X. Liu *et al.*, Genomics-guided discovery of thailanstatins A, B, and C as pre-mRNA splicing inhibitors and antiproliferative agents from *Burkholderia thailandensis* MSMB43. *J. Nat. Prod.* **76**, 685-693 (2013).
61. G. Pan *et al.*, Discovery of the leinamycin family of natural products by mining actinobacterial genomes. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E11131-E11140 (2017).
62. M. F. Freeman *et al.*, Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. *Science* **338**, 387-390 (2012).
63. T. Hamada *et al.*, Solution structure of polytheonamide B, a highly cytotoxic nonribosomal polypeptide from marine sponge. *J. Am. Chem. Soc.* **132**, 12941-12945 (2010).
64. M. F. Freeman, M. J. Helf, A. Bhushan, B. I. Morinaka, J. Piel, Seven enzymes create extraordinary molecular complexity in an uncultivated bacterium. *Nat. Chem.* **9**, 387-395 (2017).
65. B. I. Morinaka *et al.*, Radical *S*-adenosyl methionine epimerases: regioselective introduction of diverse D-amino acid patterns into peptide natural products. *Angew. Chem. Int. Ed. Engl.* **53**, 8503-8507 (2014).
66. A. Bhushan, P. J. Egli, E. E. Peters, M. F. Freeman, J. Piel, Genome mining- and synthetic biology-enabled production of hypermodified peptides. *Nat. Chem.* **11**, 931-939 (2019).
67. B. I. Morinaka, M. Verest, M. F. Freeman, M. Gugger, J. Piel, An orthogonal D_2O-based induction system that provides insights into D-amino acid pattern formation by radical *S*-adenosylmethionine peptide epimerases. *Angew. Chem. Int. Ed. Engl.* **56**, 762-766 (2017).
68. A. J. Pinto *et al.*, Metagenomic evidence for the presence of comammox *Nitrospira*-like bacteria in a drinking water system. *mSphere* **1**, e00054-15 (2016).

69. R. W. M. Van Soest *et al.*, Global diversity of sponges (Porifera). *PLoS One* **7**, e35105 (2012).
70. L. Pita, L. Rix, B. M. Slaby, A. Franke, U. Hentschel, The sponge holobiont in a changing ocean: from microbes to ecosystems. *Microbiome* **6**, 46 (2018).
71. A. Kanakkanthara, P. T. Northcote, J. H. Miller, Peloruside A: a lead non-taxoid-site microtubule-stabilizing agent with potential activity against cancer, neurodegeneration, and autoimmune disease. *Nat. Prod. Rep.* **33**, 549-561 (2016).
72. M. S. Donia, W. F. Fricke, J. Ravel, E. W. Schmidt, Variation in tropical reef symbiont metagenomes defined by secondary metabolism. *PLoS One* **6**, e17897 (2011).
73. T. E. Smith *et al.*, Accessing chemical diversity from the uncultivated symbionts of small marine animals. *Nat. Chem. Biol.* **14**, 179-185 (2018).
74. R. Ueoka *et al.*, Metabolic and evolutionary origin of actin-inhibiting polyketides from diverse organisms. *Nat. Chem. Biol.* **11**, 705-712 (2015).
75. Z. Lin, J. P. Torres, M. D. Tianero, J. C. Kwan, E. W. Schmidt, Origin of chemical diversity in *Prochloron*-tunicate symbiosis. *Appl. Environ. Microbiol.* **82**, 3450-3460 (2016).
76. R. L. L. Kellner, K. Dettner, Differential efficacy of toxic pederin in deterring potential arthropod predators of *Paederus* (Coleoptera: Staphylinidae) offspring. *Oecologia* **107**, 293-300 (1996).
77. T. Yamada, M. Hamada, P. Floreancig, A. Nakabachi, Diaphorin, a polyketide synthesized by an intracellular symbiont of the Asian citrus psyllid, is potentially harmful for biological control agents. *PLoS One* **14**, e0216319 (2019).
78. M. C. Wilson, J. Piel, Metagenomic approaches for exploiting uncultivated bacteria as a resource for novel biosynthetic enzymology. *Chem. Biol.* **20**, 636-647 (2013).
79. A. Crits-Christoph, S. Diamond, C. N. Butterfield, B. C. Thomas, J. F. Banfield, Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature* **558**, 440-444 (2018).
80. G. E. Lim, M. G. Haygood, "*Candidatus* Endobugula glebosa," a specific bacterial symbiont of the marine bryozoan *Bugula simplex*. *Appl. Environ. Microbiol.* **70**, 4921-4929 (2004).
81. J. Lopera, I. J. Miller, K. L. McPhail, J. C. Kwan, Increased biosynthetic gene dosage in a genome-reduced defensive bacterial symbiont. *mSystems* **2**, e00096-00017 (2017).
82. J. Zan *et al.*, A microbial factory for defensive kahalalides in a tripartite marine symbiosis. *Science* **364**, eaaw6732 (2019).
83. S. A. Anderson, P. T. Northcote, M. J. Page, Spatial and temporal variability of the bacterial community in different chemotypes of the New Zealand marine sponge *Mycale hentscheli*. *FEMS Microbiol. Ecol.* **72**, 328-342 (2010).
84. C. Gurgui, J. Piel, Metagenomic approaches to identify and isolate bioactive natural products from microbiota of marine sponges. *Methods Mol. Biol.* **668**, 247-264 (2010).
85. D. Li, C. M. Liu, R. Luo, K. Sadakane, T. W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674-1676 (2015).
86. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
87. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
88. D. D. Kang, J. Froula, R. Egan, Z. Wang, MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
89. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043-1055 (2015).
90. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256-W259 (2019).
91. T. Brettin *et al.*, RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* **5**, 8365 (2015).
92. J. Huerta-Cepas *et al.*, Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115-2122 (2017).
93. B. A. Pfeifer, S. J. Admiraal, H. Gramajo, D. E. Cane, C. Khosla, Biosynthesis of complex polyketides in a metabolically engineered strain of *E. coli*. *Science* **291**, 1790-1792 (2001).
94. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367-1372 (2008).

Supplementary Information

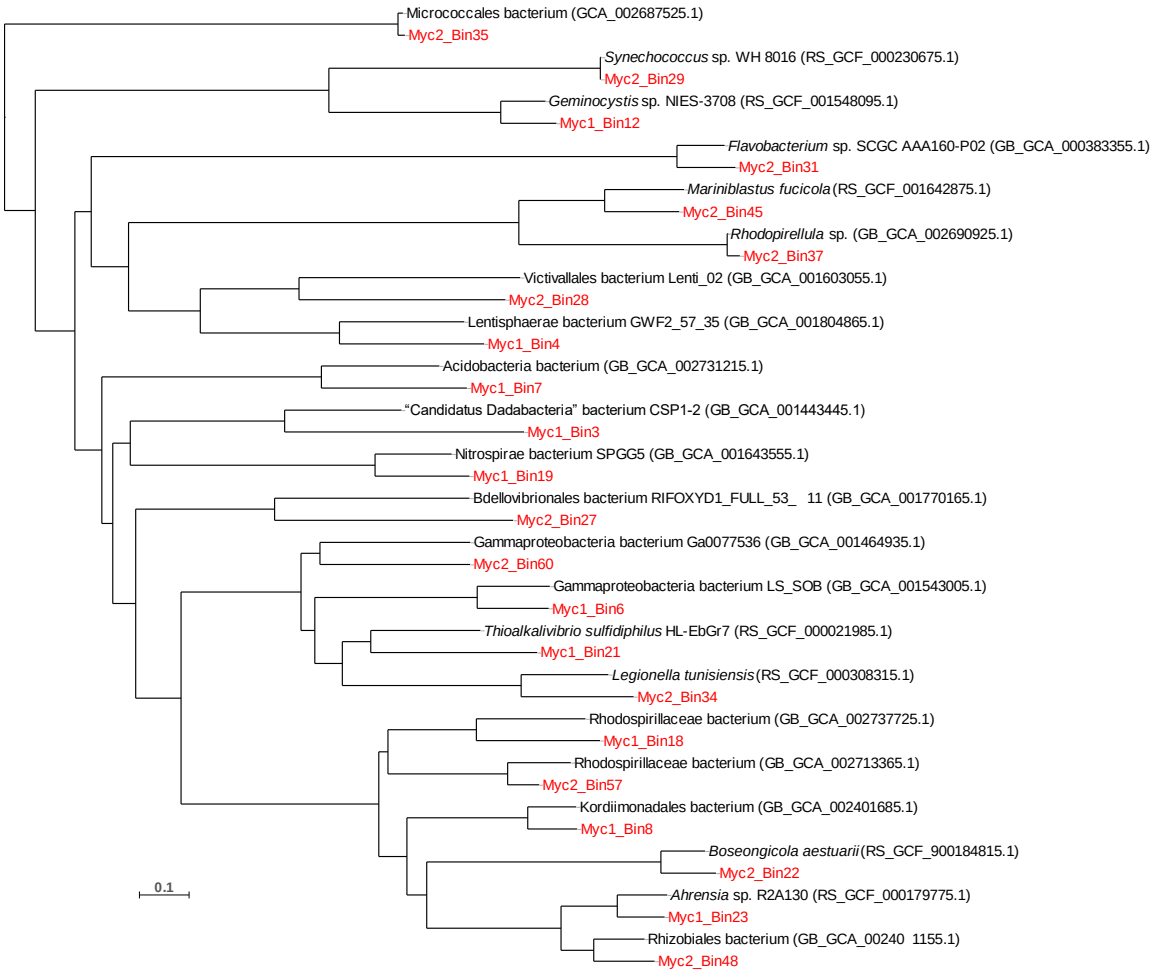


Figure S1. Nearest-neighbor tree for the 22 *Mycale* bins. Whole genome-based phylogenetic tree including the nearest neighbor of each of the 22 Mycale bins with >60% estimated genome completeness derived from the larger tree (Fig. 2) generated with the GTDB-Tk workflow (https://github.com/Ecogenomics/GtdbTk). Mycale bins are shown in red. Genome accession numbers of nearest neighbors are specified in parentheses. The scale bar represents 10% estimated sequence divergence.

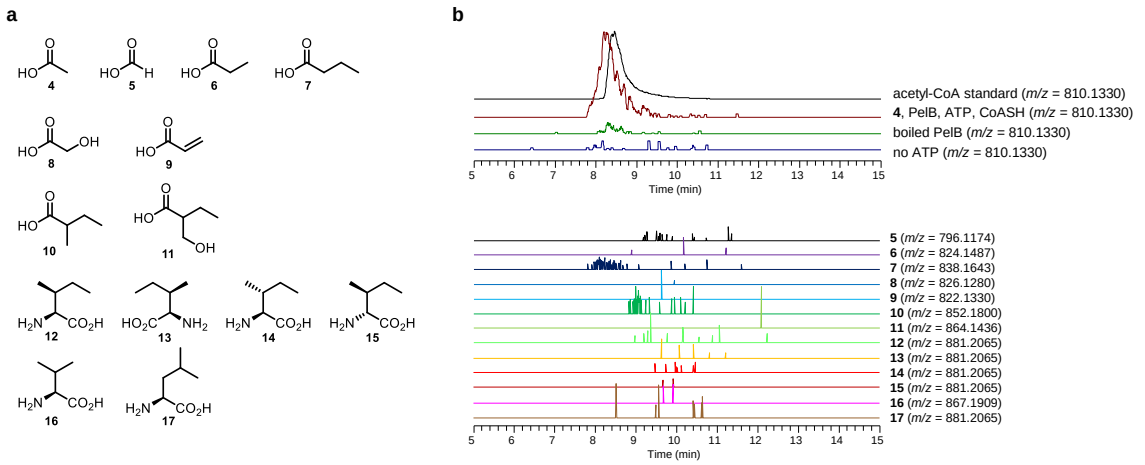


Figure S2. In vitro assays with the acyl-CoA/ACP ligase PelB. a) carboxylic acids tested as substrates for PelB. **b)** Extracted ion chromatograms of the $[M+H]^+$ ions of the corresponding CoA-adducts (5 ppm within calculated values given in parentheses). For the PelB assay with acetic acid as substrate, control assays without ATP and boiled enzyme as well as an acetyl-CoA standard are shown. All chromatograms have a fixed scale except for the acetyl-CoA standard which was scaled down for visualization purposes.

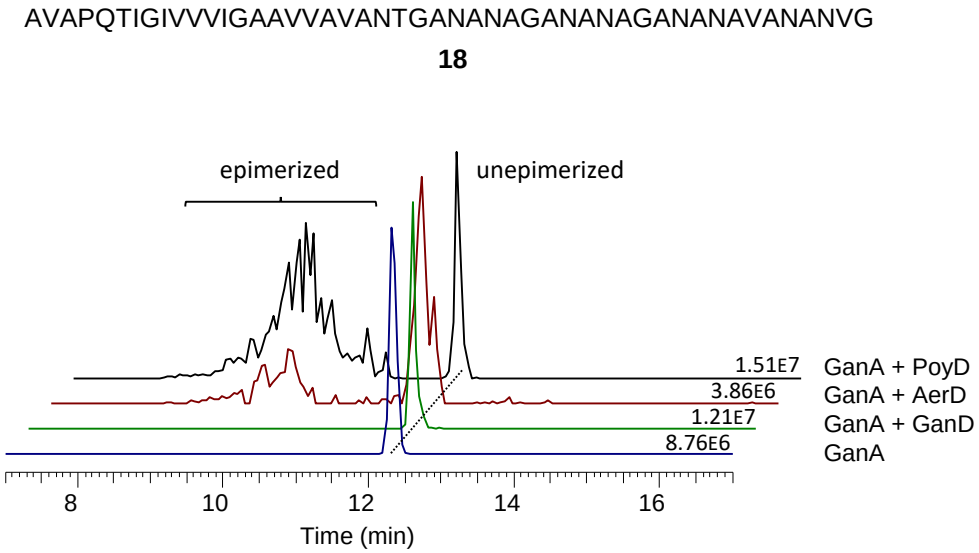


Figure S3. Extracted ion chromatograms of GluC-digested GanA precursor co-expressed with epimerases. Extracted ion chromatograms (8 ppm within calculated *m/z* of **18** = 1064.0651 $[M+4H]^{4+}$) of the GanA precursor peptide alone, and co-expressed with one of the epimerases GanD, AerD from M. aerodenitrificans, and PoyD from 'Entotheonella' digested in vitro with endoproteinase GluC. Respective base peak intensities are displayed. Differently epimerized derivatives are highlighted with a curly bracket.

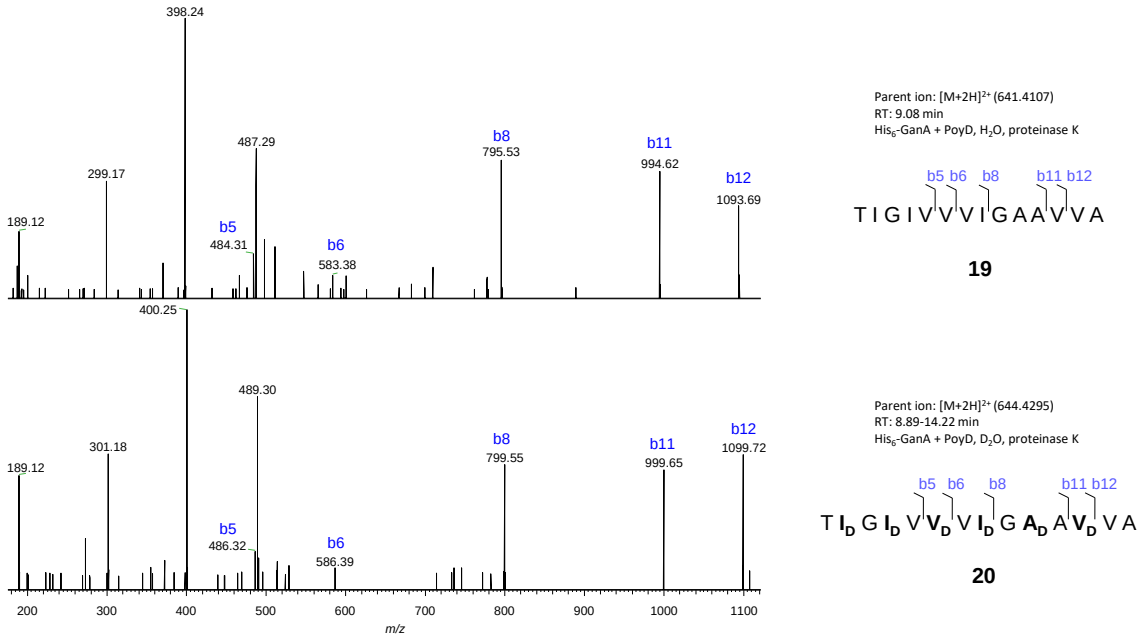


Figure S4. Characterization of selectively labeled products from co-expression of GanA and PoyD using ODIS.[1] MS^2 spectra derived from proteinase K-digested parent ions 19 and 20 acquired by LC-MS/MS of ODIS experiment. Bold residues indicate epimerized residues. Subscript "D" indicates selective labeling with deuterium at corresponding epimerized residue.

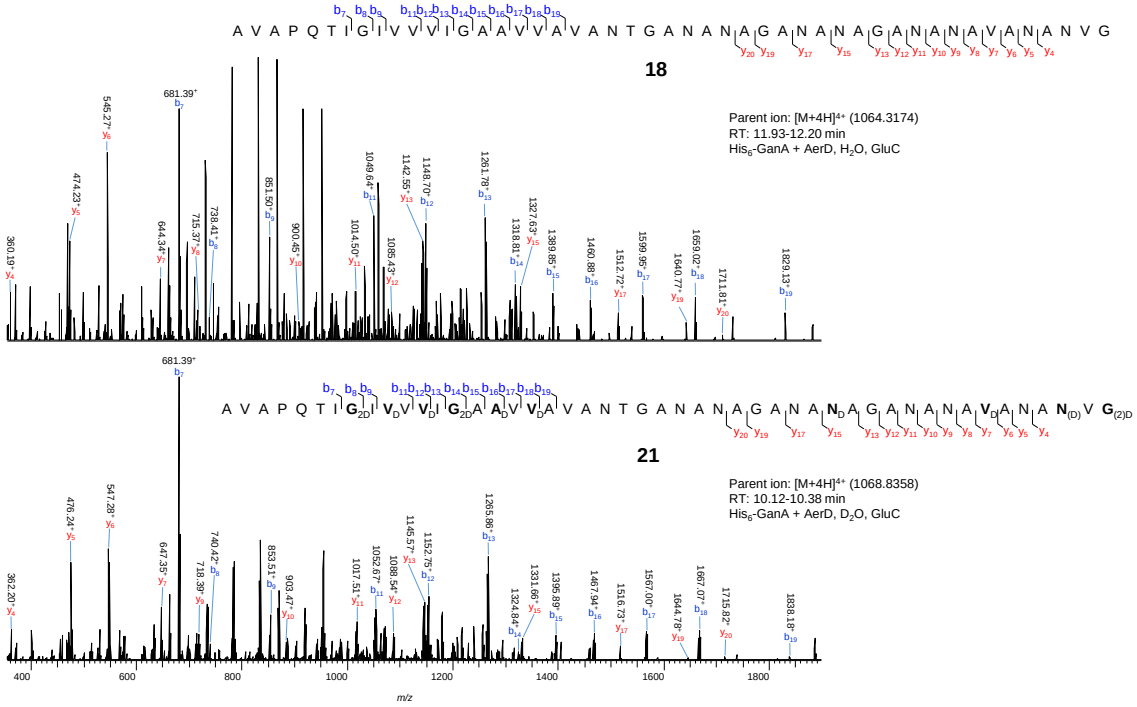


Figure S5. Characterization of selectively labeled products from co-expression of GanA and AerD using ODIS. MS^2 spectra derived from endoproteinase GluC-digested parent ions **18** and **21** acquired by LC-MS/MS of ODIS experiment. Bold residues indicate epimerized residues. Subscript "D" indicates selective labeling with deuterium at corresponding epimerized residue. The "D" and "2" in parentheses indicate that the deuterium incorporation could not be unambiguously localized.

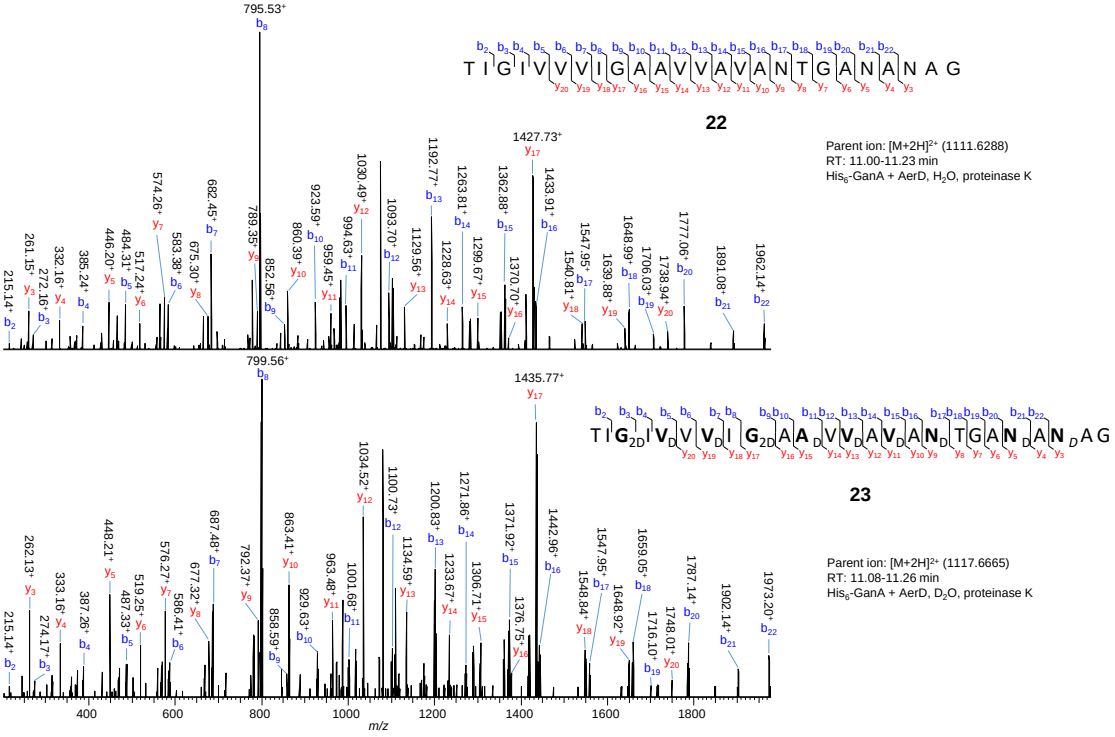


Figure S6: Characterization of selectively labeled products from co-expression of GanA and AerD using ODIS. MS^2 spectra derived from proteinase K-digested parent ions **22** and **23** acquired by LC-MS/MS of ODIS experiment. Bold residues indicate epimerized residues. Subscript "D" indicates selective labeling with deuterium at corresponding epimerized residue.

Table S1. *Mycale hentscheli* sponge samples from New Zealand used in this study and concentrations of the three polyketides mycalamide A, pateamine A, and peloruside A determined as described previously.[2]

Sample	Site	Reef	Latitude	Longitude	Year	Mycalamide A [µg/g]	Pateamine A [µg/g]	Peloruside A [µg/g]
1MJP40-24 (Myc1)	Pelorus Sound	114	-41.083516°	173.947246°	2005	173	386	16
1MJP3-79.12 (Myc2)	Pelorus Sound	114	-41.083516°	173.947246°	2006	98	55	12

Table S2. Comparison of the Illumina assemblies for the two sponge specimens.

	1MJP40-24 (Myc1)	1MJP3-79.12 (Myc2)
No. Scaffolds	369,503	714,954
No. Scaffolds (≥1500 bp)	24,860	44,234
Size (bp)	143,972,757	221,730,805
N50	10,906	7,948
No. Bins	23	68
No. Bins (>60% complete)	14	20

Table S3. Contigs >1,500 bp of the Myc1 (1MJP40-24) sponge sample harboring biosynthetic genes identified by antiSMASH[3] and manual analysis. Identifiers are given for gene clusters discussed in the main part of the paper. Contigs without assigned bin number are part of the unbinned fraction.

Cluster identifier	Contig name	Gene cluster type (antiSMASH notation)	Bin number
	NODE_980_length_21656_cov_9.627008	arylpolyene	6
	NODE_1269_length_17690_cov_5.632946	arylpolyene	20
	NODE_4135_length_6707_cov_14.709561	arylpolyene	-
	NODE_4718_length_5986_cov_12.015343	arylpolyene	9
	NODE_6263_length_4666_cov_8.875732	arylpolyene	6
	NODE_22895_length_1603_cov_47.178295	arylpolyene	-
	NODE_22897_length_1603_cov_29.748062	arylpolyene	-
	NODE_24585_length_1515_cov_1.981507	arylpolyene	-
	NODE_3_length_610169_cov_14.127476	bacteriocin	18
	NODE_86_length_92752_cov_8.140533	bacteriocin	8
	NODE_105_length_82472_cov_8.375578	bacteriocin	8
	NODE_122_length_73921_cov_73.430266	bacteriocin	19
	NODE_145_length_66988_cov_9.597971	bacteriocin	21
	NODE_148_length_65867_cov_72.858840	bacteriocin	19
	NODE_196_length_57705_cov_8.478786	bacteriocin	21
	NODE_2765_length_9587_cov_7.110890	ectoine	6
	NODE_12083_length_2660_cov_4.519770	ectoine	20

	NODE_21249_length_1698_cov_2.477176	ectoine	-
	NODE_1_length_1714095_cov_14.104484	hserlactone	18
	NODE_3_length_610169_cov_14.127476	hserlactone	18
	NODE_65_length_105797_cov_12.695750	hserlactone	18
	NODE_67_length_104018_cov_13.163876	hserlactone	18
	NODE_498_length_34250_cov_8.202632	hserlactone	8
	NODE_1121_length_19538_cov_48.126726	hserlactone	22
	NODE_81_length_95316_cov_9.455370	lassopeptide	7
	NODE_20061_length_1777_cov_2.254936	linaridin	-
	NODE_13_length_326697_cov_13.435979	nrps	18
	NODE_289_length_46986_cov_16.436257	nrps	17
	NODE_1375_length_16750_cov_12.917580	nrps	17
	NODE_1963_length_12658_cov_22.384353	nrps	12
	NODE_6065_length_4782_cov_7.823355	nrps	12
	NODE_22703_length_1614_cov_2.932008	nucleoside	-
	NODE_136_length_69545_cov_8.739013	other	7
Pam	NODE_1210_length_18381_cov_25.267980	other	4
	NODE_1303_length_17308_cov_24.639541	other	12
	NODE_5982_length_4839_cov_5.027592	other	20
	NODE_9698_length_3188_cov_8.315991	other	5
	NODE_17554_length_1969_cov_3.873041	other	-
Pam	NODE_1097_length_19903_cov_16.716999	otherks	4
	NODE_140_length_68311_cov_8.419113	phosphonate	8
	NODE_374_length_41034_cov_72.427560	phosphonate	19
	NODE_5009_length_5683_cov_5.643746	phosphonate	20
12	NODE_255_length_50741_cov_14.492542	t1pks	23
6	NODE_2959_length_9002_cov_16.348832	t1pks	4
7	NODE_4554_length_6159_cov_7.420872	t1pks	6
8	NODE_6368_length_4600_cov_6.631463	t1pks	6
	NODE_3_length_610169_cov_14.127476	t3pks	18
	NODE_16_length_239453_cov_10.755725	t3pks	20
	NODE_251_length_51268_cov_9.132349	t3pks	7
	NODE_9218_length_3327_cov_4.218521	t3pks	20
	NODE_22219_length_1642_cov_2.862634	t3pks	-
	NODE_4_length_588314_cov_13.995148	terpene	18
	NODE_16_length_239453_cov_10.755725	terpene	20
	NODE_46_length_137014_cov_73.316138	terpene	19
	NODE_281_length_47513_cov_18.593767	terpene	22
	NODE_451_length_36715_cov_12.748854	terpene	23
	NODE_488_length_34806_cov_8.279359	terpene	8
	NODE_565_length_31087_cov_33.284738	terpene	5
	NODE_572_length_30918_cov_7.619771	terpene	12
	NODE_636_length_29217_cov_73.102428	terpene	19
	NODE_859_length_24051_cov_8.858851	terpene	7
	NODE_878_length_23616_cov_73.612495	terpene	19

	NODE_888_length_23505_cov_15.460213	terpene	5
	NODE_1304_length_17302_cov_5.656230	terpene	11
	NODE_1641_length_14647_cov_4.902755	terpene	20
	NODE_5248_length_5453_cov_5.994628	terpene	20
	NODE_5572_length_5159_cov_7.773511	terpene	5
	NODE_7929_length_3789_cov_2.726031	terpene	5
	NODE_9380_length_3279_cov_2.924938	terpene	5
	NODE_11215_length_2833_cov_2.242981	terpene	5
Pam	NODE_440_length_37488_cov_16.320199	transatpks	4
1	NODE_496_length_34342_cov_16.206113	transatpks	4
3	NODE_879_length_23615_cov_16.784338	transatpks	4
Pam	NODE_977_length_21693_cov_23.213559	transatpks	4
2	NODE_1316_length_17179_cov_15.121584	transatpks	4
Pam	NODE_1511_length_15590_cov_23.193949	transatpks	4
11	NODE_1688_length_14320_cov_12.264353	transatpks	17
4	NODE_2456_length_10540_cov_17.949070	transatpks	4
5	NODE_3415_length_7985_cov_13.492560	transatpks	4
	NODE_5751_length_5026_cov_14.986924	transatpks	4
Pam	NODE_8035_length_3745_cov_30.982385	transatpks	4
	NODE_14954_length_2234_cov_18.107848	transatpks	-
Myc	NODE_29_length_165452_cov_18.316300	transatpks-otherks-nrps	6

Table S4. Contigs >1,500 bp of the Myc2 (1MJP3-79.12) sponge sample harboring biosynthetic genes identified by antiSMASH and manual analysis. Identifiers are given for gene clusters discussed in the main part of the paper. Contigs without assigned bin number are part of the unbinned fraction.

Cluster identifier	**Contig name**	**Gene cluster type (antiSMASH notation)**	**Bin number**
	NODE_43962_length_1506_cov_8.706409	arylpolyene	-
	NODE_17594_length_2958_cov_3.170858	arylpolyene	-
	NODE_12853_length_3760_cov_22.946019	arylpolyene	-
	NODE_10720_length_4342_cov_6.082109	arylpolyene	28
	NODE_6932_length_6106_cov_4.517435	arylpolyene	28
	NODE_1962_length_16680_cov_7.458827	arylpolyene	-
	NODE_1902_length_17034_cov_8.845456	arylpolyene	3
	NODE_995_length_25933_cov_6.137839	arylpolyene	45
	NODE_583_length_34253_cov_9.732294	arylpolyene	-
	NODE_18340_length_2865_cov_1.998932	bacteriocin	-
	NODE_18311_length_2868_cov_3.401706	bacteriocin	14
	NODE_7408_length_5795_cov_3.669686	bacteriocin	14
	NODE_2660_length_13296_cov_3.305566	bacteriocin	17
	NODE_2452_length_14216_cov_7.548619	bacteriocin	29
	NODE_2175_length_15607_cov_4.388310	bacteriocin	37

	NODE_2106_length_15911_cov_7.542886	bacteriocin	29
	NODE_1864_length_17203_cov_11.377945	bacteriocin	60
	NODE_1824_length_17419_cov_8.633322	bacteriocin	29
	NODE_1356_length_21694_cov_8.579047	bacteriocin	29
	NODE_1108_length_24457_cov_4.778953	bacteriocin	30
	NODE_745_length_30512_cov_31.023476	bacteriocin	47
	NODE_709_length_31279_cov_33.638675	bacteriocin	27
	NODE_195_length_57602_cov_4.929449	bacteriocin	-
	NODE_149_length_64650_cov_13.068086	bacteriocin	60
	NODE_1_length_3733361_cov_37.618454	bacteriocin	-
Gan	NODE_2229_length_15314_cov_14.643161	bacteriocin-proteusin	60
	NODE_23329_length_2384_cov_3.100902	ectoine	-
	NODE_1904_length_17018_cov_11.317102	ectoine	60
	NODE_1270_length_22674_cov_7.929219	ectoine	3
	NODE_32973_length_1845_cov_3.203352	hserlactone	-
	NODE_7177_length_5939_cov_3.915194	hserlactone	10
	NODE_1606_length_19122_cov_101.403105	hserlactone	22
	NODE_1_length_3733361_cov_37.618454	hserlactone	-
	NODE_1_length_3733361_cov_37.618454	hserlactone	-
	NODE_2481_length_14065_cov_5.059101	lassopeptide	34
	NODE_3_length_1469305_cov_73.933105	lassopeptide	56
	NODE_1947_length_16743_cov_74.049856	nrps	31
	NODE_820_length_29020_cov_5.043811	nrps	34
	NODE_298_length_46985_cov_75.366290	nrps	31
	NODE_1_length_3733361_cov_37.618454	nrps	-
	NODE_18727_length_2814_cov_2.812613	nucleoside	-
16	NODE_29016_length_2024_cov_2.537836	other	-
	NODE_17321_length_2995_cov_2.934694	other	19
	NODE_14728_length_3389_cov_5.613077	other	28
	NODE_14478_length_3432_cov_4.159609	other	-
	NODE_9845_length_4644_cov_10.303988	other	63
	NODE_1539_length_19657_cov_6.850474	other	45
9	NODE_527_length_36114_cov_11.534901	other	60
	NODE_2_length_1981894_cov_73.738136	other	51
	NODE_26907_length_2141_cov_1.852828	otherks	-
	NODE_7079_length_6000_cov_3.144491	otherks	17
	NODE_5485_length_7363_cov_4.546661	otherks	28
Pam	NODE_1517_length_19903_cov_13.762898	otherks	41
	NODE_2220_length_15364_cov_4.430531	phosphonate	30
	NODE_291_length_47628_cov_8.158598	phosphonate	45
6	NODE_4186_length_9227_cov_24.121784	t1pks	-
7	NODE_1913_length_16979_cov_7.342177	t1pks	3
8	NODE_991_length_25964_cov_8.672430	t1pks	3
12	NODE_257_length_50741_cov_71.127057	t1pks	52
10	NODE_99_length_76726_cov_14.086499	t1pks	60

	NODE_2_length_1981894_cov_73.738136	t1pks	51
13	NODE_121_length_72259_cov_5.286009	t1pks-nrps	34
	NODE_2015_length_16381_cov_5.356180	t1pks-otherks	28
	NODE_880_length_27655_cov_10.500942	t1pks-otherks	63
	NODE_20386_length_2642_cov_2.906069	t3pks	-
	NODE_12226_length_3911_cov_3.582728	t3pks	37
	NODE_7887_length_5533_cov_3.798101	t3pks	37
	NODE_7486_length_5752_cov_4.155520	t3pks	63
	NODE_5498_length_7347_cov_10.459133	t3pks	63
	NODE_2732_length_12996_cov_6.738042	t3pks	29
	NODE_1047_length_25099_cov_4.419302	t3pks	37
	NODE_442_length_39856_cov_12.172282	t3pks	35
	NODE_16_length_239453_cov_15.339076	t3pks	57
	NODE_3_length_1469305_cov_73.933105	t3pks	56
	NODE_1_length_3733361_cov_37.618454	t3pks	1
	NODE_39614_length_1622_cov_1.920230	terpene	-
	NODE_34922_length_1772_cov_4.343623	terpene	-
	NODE_24483_length_2301_cov_3.940338	terpene	-
	NODE_24204_length_2320_cov_1.943929	terpene	-
	NODE_20193_length_2660_cov_7.709789	terpene	-
	NODE_19772_length_2701_cov_2.294029	terpene	-
	NODE_17436_length_2980_cov_1.471111	terpene	-
	NODE_16060_length_3172_cov_4.282964	terpene	29
	NODE_12122_length_3939_cov_4.719361	terpene	-
	NODE_8934_length_5015_cov_4.992742	terpene	68
	NODE_7009_length_6050_cov_10.683236	terpene	63
	NODE_3593_length_10446_cov_23.136176	terpene	-
	NODE_3182_length_11551_cov_4.211813	terpene	37
	NODE_1826_length_17399_cov_4.594730	terpene	37
	NODE_1667_length_18590_cov_12.018559	terpene	60
	NODE_1635_length_18878_cov_11.428040	terpene	35
	NODE_1194_length_23505_cov_72.248913	terpene	31
	NODE_1075_length_24758_cov_13.654333	terpene	48
	NODE_809_length_29220_cov_4.705880	terpene	30
	NODE_557_length_34973_cov_31.553640	terpene	23
	NODE_533_length_35899_cov_7.499191	terpene	45
	NODE_519_length_36396_cov_7.466663	terpene	45
	NODE_383_length_42525_cov_71.444455	terpene	52
	NODE_330_length_44995_cov_12.613529	terpene	60
	NODE_202_length_57047_cov_5.102278	terpene	30
	NODE_194_length_57689_cov_37.466079	terpene	22
	NODE_141_length_67369_cov_4.578230	terpene	37
	NODE_101_length_75955_cov_7.041094	terpene	29
	NODE_16_length_239453_cov_15.339076	terpene	57
	NODE_3_length_1469305_cov_73.933105	terpene	56

	NODE_1_length_3733361_cov_37.618454	terpene	-
16	NODE_31287_length_1917_cov_2.799141	transatpks	-
	NODE_25447_length_2234_cov_13.353832	transatpks	-
	NODE_18683_length_2820_cov_3.684991	transatpks	-
	NODE_12165_length_3925_cov_31.710853	transatpks	-
	NODE_12107_length_3942_cov_5.112941	transatpks	-
	NODE_10909_length_4281_cov_4.721959	transatpks	-
16	NODE_10284_length_4488_cov_3.199413	transatpks	-
	NODE_8487_length_5223_cov_21.473104	transatpks	41
	NODE_7922_length_5510_cov_4.827864	transatpks	-
18	NODE_6931_length_6108_cov_4.086569	transatpks	-
	NODE_5918_length_6905_cov_3.610365	transatpks	-
5	NODE_5020_length_7956_cov_11.372231	transatpks	-
	NODE_4437_length_8784_cov_5.349410	transatpks	-
Pam	NODE_3848_length_9872_cov_27.763675	transatpks	-
4	NODE_3575_length_10489_cov_15.144048	transatpks	41
Pam	NODE_3490_length_10716_cov_29.356158	transatpks	-
15	NODE_2648_length_13333_cov_3.928077	transatpks	-
11	NODE_2430_length_14320_cov_72.084683	transatpks	31
14	NODE_2228_length_15326_cov_4.033397	transatpks	-
Pam	NODE_2184_length_15539_cov_18.456923	transatpks	41
2	NODE_1870_length_17186_cov_13.527815	transatpks	41
3	NODE_1208_length_23410_cov_12.709099	transatpks	41
14	NODE_1029_length_25397_cov_4.564202	transatpks	-
1	NODE_573_length_34442_cov_13.691482	transatpks	41
17	NODE_1866_length_17199_cov_5.518082	transatpks-nrps	-
Pam	NODE_505_length_37423_cov_14.324984	transatpks-nrps	41
Pel	NODE_242_length_53043_cov_9.005681	transatpks-nrps	-
Myc	NODE_64_length_91913_cov_18.309216	transatpks-otherks-nrps	-

Table S5. Genome properties of the 34 bins with >60% estimated genome completeness. M1 bins originate from sponge sample Myc1, M2 from sponge sample Myc2. In case the bin was present in both sponge samples, the corresponding bin number is listed in the last column. Abbreviations, comp: completeness, corr: corresponding, est: estimated.

Ac-cession	**Bin no.**	**Phylogeny**	**% est. comp.**	**Contig no.**	**Est. size (Mb)**	**% GC**	**N50**	**CDS no.**	**Corr. bin no.**
SAMN14054217	M1 bin3	*Candidatus* Dadabacteria	94	8	1.68	50	506,424	1,635	M2 bin8
SAMN14054218	M1 bin4	*Kiritimatiellae*	93	342	2.71	44	10,705	2,936	M2 bin41
SAMN14054219	M1 bin6	γ-Proteo-bacteria > UBA10353	96	374	4.17	46	16,998	4,155	M2 bin3
SAMN14054220	M1 bin7	Acidobac-teria > UBA890	92	144	4.82	64	49,606	4,286	M2 bin51
SAMN14054221	M1 bin8	α-Proteo-bacteria > Sphingo-monadales	96	160	4.01	51	38,060	3,844	
SAMN14054222	M1 bin10	Archaea	66	38	1.36	37	67,476	1,697	M2 bin49
SAMN14054223	M1 bin11	α-Proteo-bacteria > Rhizobiales	80	396	2.19	45	6,085	2,460	M2 bin48
SAMN14054224	M1 bin12	Cyano-bacteria	70	206	2.58	34	16,727	2,578	
SAMN14054225	M1 bin18	α-Proteo-bacteria > Rhodo-spirillales	100	101	6.98	55	231,845	6,941	M2 bin10
SAMN14054226	M1 bin19	Nitrospirae	96	74	2.60	48	65,867	2,582	M2 bin30
SAMN14054227	M1 bin20	α-Proteo-bacteria	100	970	9.58	60	13,622	9,924	M2 bin57
SAMN14054228	M1 bin21	γ-Proteo-bacteria > Thiotrichales	94	56	2.61	40	105,978	2,719	M2 bin14
SAMN14054229	M1 bin22	α-Proteo-bacteria > Rhodo-bacterales	95	201	3.65	62	26,418	3,823	M2 bin22
SAMN14054230	M1 bin23	α-Proteo-bacteria > Rhizobiales	98	260	4.71	53	30,778	4,569	M2 bin52
SAMN14054231	M2 bin3	γ-Proteo-bacteria	90	325	3.95	46	19,417	3,877	M1 bin6
SAMN14054232	M2 bin8	*Candidatus* Dadabacteria	94	4	1.18	50	759,554	1,151	M1 bin3
SAMN14054233	M2 bin10	α-Proteo-bacteria > Rhodo-spirillales	76	457	2.79	55	8,059	3,161	M1 bin18

SAMN14054234	M2 bin22	α-Proteobacteria > Rhodobacterales	96	165	3.58	61	29,577	3,787	M1 bin22
SAMN14054235	M2 bin27	Oligoflexia	84	52	1.82	45	50,621	1,664	
SAMN14054236	M2 bin28	Verrucomicrobia > Lentisphaeria	64	354	1.6	40	4,991	1,525	
SAMN14054237	M2 bin29	Cyanobacteria > Synechococcales	86	170	2.02	56	17,643	2,308	
SAMN14054238	M2 bin30	Nitrospirae	81	309	2.47	48	10,754	2,612	M1 bin19
SAMN14054239	M2 bin31	Bacteroidetes > *Flavobacteriaceae*	91	238	3.11	34	18,194	3,084	M1 bin17
SAMN14054240	M2 bin34	γ-Proteobacteria > *Legionellaceae*	82	142	1.61	41	17,423	1,679	
SAMN14054241	M2 bin35	Actinobacteria > Nanopelagicales	81	183	1.55	63	12,694	1,668	
SAMN14054242	M2 bin37	Planctomycetes > *Pirellulaceae*	83	859	6.71	51	11,312	5,627	
SAMN14054243	M2 bin41	*Kiritimatiellae*	77	222	2.03	44	11,296	2,066	M1 bin4
SAMN14054244	M2 bin45	Planctomycetes > *Pirellulaceae*	88	540	4.76	48	13,365	3,823	
SAMN14054245	M2 bin48	α-Proteobacteria > Rhizobiales	82	294	1.91	45	7,440	2,057	M1 bin11
SAMN14054246	M2 bin49	Archaea	66	48	1.21	37	43,595	1,564	M1 bin10
SAMN14054247	M2 bin51	Acidobacteria > UBA890	67	4	2.91	64	1,962,698	2,542	M1 bin7
SAMN14054248	M2 bin52	α-Proteobacteria > Rhizobiales	97	181	4.20	53	33,351	4,019	M1 bin23
SAMN14054249	M2 bin57	α-Proteobacteria > UBA6615	100	30	3.90	60	225,171	3,721	M1 bin20
SAMN14054250	M2 bin60	γ-Proteobacteria	90	274	5.83	58	30,432	5,977	

Table S6. Presence/absence of core metabolism and amino acid biosynthesis pathways across the *Mycale* bins with >90% estimated genome completeness and '*Candidatus* Entotheonella factor' (GCA_000522425) and a '*Candidatus* Poribacterium' bacterium (GCA_002726605). Complete pathways are shown in green, partial pathways in green stripes, and absent pathways in red crosshatches.

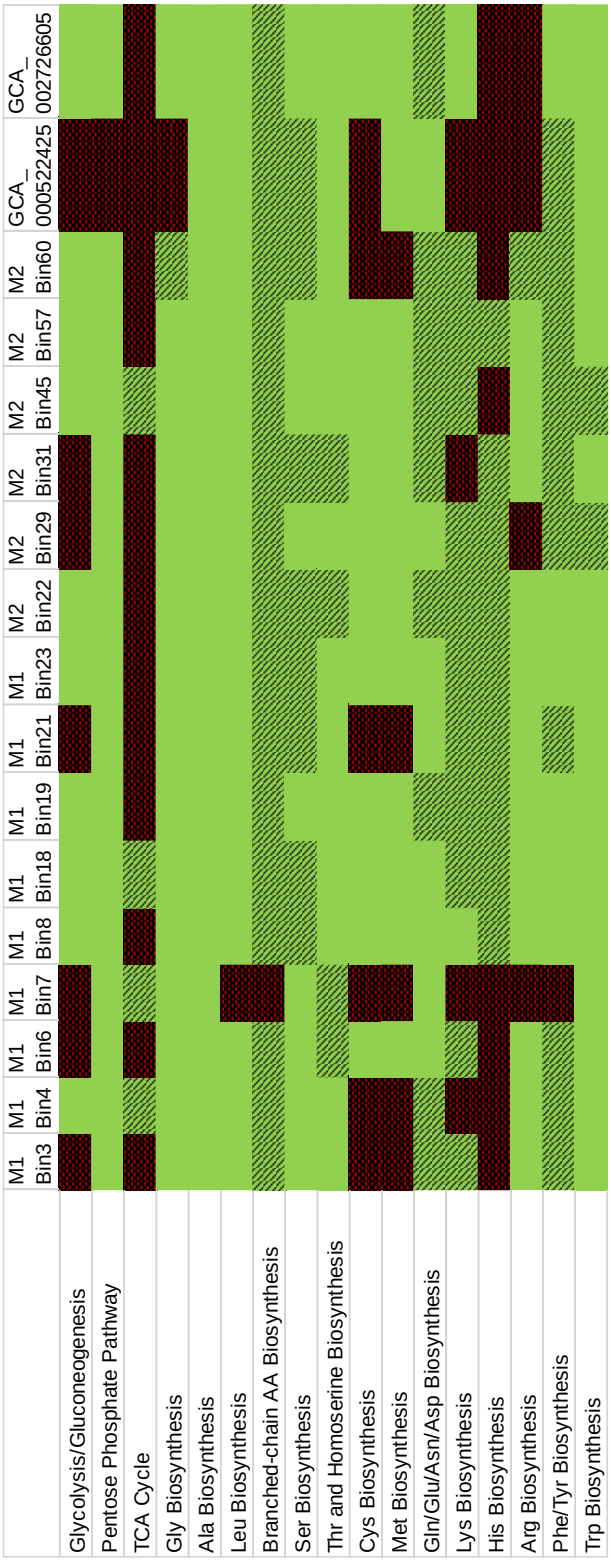


Table S7. Selected COG/TIGRFAM entries commonly associated with symbiosis[4-7] across the *Mycale* bins with >90% estimated genome completeness and '*Candidatus* Entotheonella factor' (GCA_000522425) and a *'Candidatus* Poribacterium' bacterium (GCA_002726605).

	M1 Bin3	M1 Bin4	M1 Bin6	M1 Bin7	M1 Bin8	M1 Bin18	M1 Bin19	M1 Bin21	M1 Bin23	M2 Bin22	M2 Bin29	M2 Bin31	M2 Bin45	M2 Bin57	M2 Bin60	GCA_ 000522425	GCA_ 002726605	Description
COG5461	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	Type IV pilus biogenesis protein
COG0666	4	0	2	1	0	3	0	1	4	1	1	0	1	2	3	1	5	Ankyrin repeats
COG0457	4	3	6	5	8	32	4	2	3	3	1	2	8	12	11	15	18	Tetratricopeptide repeats
COG0790	1	0	3	1	1	3	6	2	4	1	0	0	0	18	14	4	2	Tetratricopeptide repeats
COG4886	0	0	1	0	0	0	0	0	0	0	0	1	2	0	0	3	8	Leucine-rich repeat (LRR) protein
COG2319	0	0	3	13	0	0	3	0	2	1	0	0	2	0	2	11	1	WD40 repeats
COG1520	0	0	1	1	2	3	0	1	0	1	0	0	4	2	1	2	2	WD40 repeats
COG5424	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	pyrrolo-quinoline quinone repeat (PQQ)
COG1587	1	1	1	0	1	2	1	1	0	0	1	0	2	1	1	3	2	uroporphyrinogen-III synthase
COG2109	0	0	1	0	0	1	1	0	1	1	3	0	0	1	1	1	0	ATP:corrinoid adenosyltransferase
COG0863	6	3	0	0	1	2	0	1	2	1	0	12	3	1	9	1	1	DNA modification methylase
COG3093	2	0	0	1	0	0	4	0	0	1	0	1	0	0	6	4	0	Plasmid maintenance system antidote protein VapI
COG4634	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Predicted nuclease, potential toxin-antitoxin system component
COG2810	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Predicted type IV restriction endonuclease
COG2141	6	3	6	1	0	7	0	0	3	3	0	0	0	21	25	122	1	Flavin-dependent oxidoreductase, luciferase family
COG3210	0	0	0	0	1	0	0	0	0	0	3	0	0	0	1	5	0	large exoprotein involved in adhesion
COG3979	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	Chitinase
COG1518	1	0	2	0	0	0	0	0	0	0	0	1	0	0	2	0	0	CRISPR/Cas
COG3513	1	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	CRISPR/Cas
COG2189	2	0	2	0	0	0	0	0	2	0	0	3	0	0	3	0	1	Transposase
COG3436	0	3	4	0	0	0	0	0	1	0	0	1	0	0	1	6	0	Transposase
COG3039	0	1	2	0	0	0	0	0	0	0	0	3	0	0	7	1	0	Transposase
TIGR03558	0	0	0	0	0	1	0	0	0	0	0	0	0	2	1	2	0	luciferase family oxidoreductase, group 1
TIGR03619	0	0	0	0	0	0	0	0	0	1	0	0	0	0	10	34	0	probable F420-dependent oxidoreductase, Rv2161c family
TIGR03841	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	probable F420-dependent oxidoreductase, Rv3093c family
TIGR03854	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	probable F420-dependent oxidoreductase

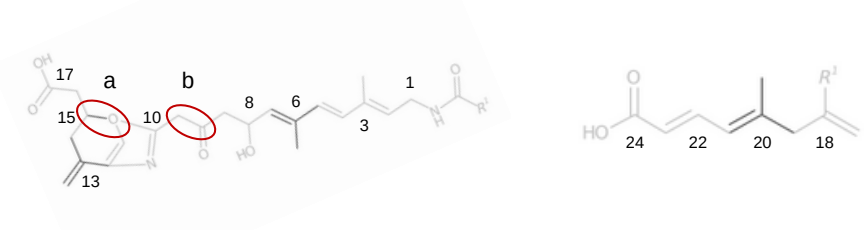
Table S8. Proteins encoded in the proposed mycalamide BGC (*myc*) and their putative functions. *Methylobacter tundripaludum* contains a biosynthetic gene cluster for an as-yet uncharacterized pederin-type compound. Abbreviations, aa: amino acid, I: identity, S: similarity.

Protein	**Size, aa**	**Proposed function**	**Closest homolog, protein (origin)**	**S/I, %/%**	**GenBank accession number**	**Pederin cluster homolog**
MycN	79	ACP	Acyl carrier protein (*Methylocucumis oryzae*)	80/62	WP_045778439	PedN
MycM	415	KS	Polyketide beta-ketoacyl:ACP synthase (*Sorangium cellulosum*)	73/58	WP_012235813	PedM
MycP	419	HMGS	3-Hydroxy-3-methylglutaryl-ACP synthase (*Bacillus atrophaeus*)	86/74	WP_013390523	PedP
MycL	260	ECH	Enoyl-CoA hydratase/isomerase (*Paenibacillus* sp. 1-49)	70/54	WP_025683544	PedL
MycC	332	AT	AT domain-containing protein (*Azospirillum brasilense*)	63/43	WP_114857239	PedC
MycQ	346	unknown	Prohibitin family protein (*Methylobacter tundripaludum*)	75/54	WP_104425067	
MycR	553	transporter	Putative ATP-binding cassette transporter (*M. tundripaludum*)	67/47	PPK66286	
MycS	289	unknown	DUF697 domain-containing protein (*Algicola sagamiensis*)	68/52	WP_018694097	
MycD	375	AT	ACP S-malonyltransferase (*M. tundripaludum*)	68/52	WP_104425065	PedD
MycB	452	ER	PfaD family polyunsaturated fatty acid/polyketide biosynthesis protein (*M. tundripaludum*)	78/64	WP_104425071	PedB
MycI	5815	PKS	OnnB ('Entotheonella factor')	59/43	AAV97870	PedI
MycJ	375	monooxy-genase	LLM class flavin-dependent oxidoreductase (*M. tundripaludum*)	82/68	WP_104425061	PedJ
MycK1	345	unknown	Hypothetical protein (*M. tundripaludum*)	78/66	WP_104425072	PedK
MycA	334	MT	SAM-dependent MT (*M. tundripaludum*)	80/64	WP_104425074	PedA
MycT	242	dioxygenase	Phytanoyl-CoA dioxygenase (*M. tundripaludum*)	66/48	WP_104425062	
MycK2	339	unknown	Hypothetical protein (*M. tundripaludum*)	79/64	WP_104425075	PedK
MycE	269	MT	Methyltransferase domain-containing protein (*M. tundripaludum*)	75/60	WP_104425076	PedE

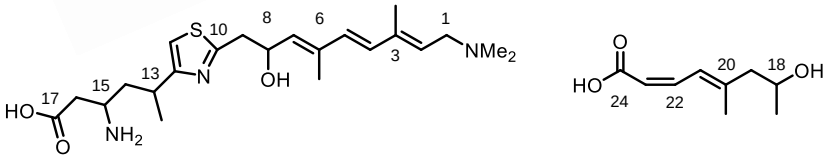
MycF	9241	PKS	OnnI ('E. factor')	61/45	AAV97877	PedF
MycG	433	monooxy-genase	Monooxygenase (*M. tundripaludum*)	87/80	WP_104425078	PedG
MycH	6918	PKS	Non-ribosomal peptide synthetase (*M. tundripaludum*)	61/44	WP_104424071	PedH

Table S9. TransATor-based[8] structure prediction (moieties in gray have a lower confidence than the ones in black) and A/KS domain substrate prediction of the *pam* BGC. The structures are mostly in agreement except for two major differences: a) The bond from C15 to the oxazole moiety (thiazole in pateamine) was wrongly introduced by TransATor because the software was unable to predict the unprecedented *trans*-AT PKS functionality installing the amine moiety. A pyran ring was predicted instead as a low-confidence hit for KS9 and wrongly connected to the oxazole. b) The prediction contains an additional C_2 unit compared to the linearized molecule because KS5 was not recognized as a non-elongating KS due to lack of a downstream KS (replaced by a thiazole-incorporating NRPS module).

TransATor prediction



linearized pateamine



A/KS No.	**Predicted specificity**
A1	Glycine
KS1	Clade_10 amino acids (glycine)
KS2	Clade_55 α Me double bonds (*E*-configured)
KS3	Clade_99 double bonds (e-configured)
KS4	Clade_73 exomethylene
KS5	Clade_62 β D-OH (some with α L-Me)
A2	Serine
A3	Cysteine
KS6	Clade_36 non-elongating (oxazole/thiazole rings)
KS7	Clade_5 amino acids (oxazole/thiazole)
KS8	Clade_14 exomethyl/exoester
KS9	Clade_26 pyran/furan rings
KS10	Clade_27 acetyl starter
KS11	Clade_14 exomethyl/exoester
KS12	Clade_73 exomethylene
KS13	Clade_31 non-elongating (bimodule β D-OH)
KS14	Clade_51 double bonds (*Z*-configured)

Table S10. Proteins encoded in the proposed pateamine BGC (*pam*) and their putative functions.

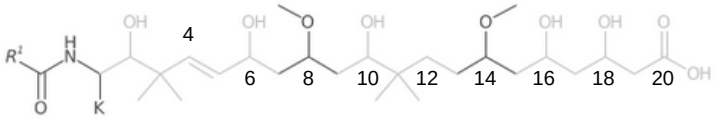
Protein	**Size, aa**	**Proposed function**	**Closest homolog, protein (origin)**	**S/I, %/%**	**GenBank accession number**
PamA	5652	PKS	PKS (*Nostoc* sp. UHCC 0450)	65/48	ATX68127
PamB	3384	PKS	SDR family NAD(P)-dependent oxidoreductase (*Clostridium formicaceticum*)	69/52	WP_070967484
PamC	7556	PKS	Hypothetical protein JL50_14400, partial (*Peptococcaceae* bacterium BICA1-7)	69/52	KJS65997
PamD	6581	PKS	KR domain-containing protein (*Brevibacillus laterosporus*)	71/55	WP_104149413
PamE	84	ACP	ACP (*Leptolyngbya* sp. PCC 7375)	81/64	WP_006512953
PamF	416	KS	KS (*Leptolyngbya* sp. PCC 7375)	76/62	WP_006512954
PamG	420	HMGS homolog	HMGS homolog (*Dendrosporobacter quercicolus*)	87/78	WP_092069005
PamH	259	ECH	ECH (*Leptolyngbya* sp. PCC 7375)	87/72	WP_006512938
PamI	249	ECH	ECH (*Desulfosporosinus* sp. SbF1)	87/72	WP_106798973
PamK	1018	unknown	Hypothetical protein ACD_62C00445G0006 (uncultured bacterium)	46/28	EKD50720
PamL	455	AT-ER	AT (*Chitinophaga* sp. CF118)	80/66	WP_090106938
	592	(kinase pseudogene)	AarF/ABC1/UbiB kinase family protein (*Leptolyngbya* sp. PCC 7375)	68/47	WP_006512948
PamM	315	AT	AT (*Leptolyngbya* sp. ISBN3-Nov-94-8)	65/46	AMH40435

Table S11. Proteins encoded in the proposed peloruside BGC (*pel*) and their putative functions.

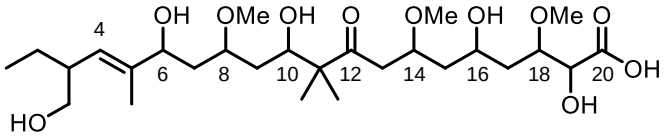
Protein	**Size, aa**	**Proposed function**	**Closest homolog, protein (origin)**	**S/I, %/%**	**GenBank accession number**
PelA	133	ACP	Hypothetical protein (*Corallococcus* sp. CA051B)	67/46	WP_120641751
PelB	453	acyl-CoA/ACP ligase	Long-chain fatty acid--CoA ligase ('*Candidatus* Parcubacteria' bacterium)	71/53	RME30562
PelC	7959	PKS	PKS/NRPS (*Dyella* sp. OK004)	58/41	WP_090451852
PelD	8369	PKS	PKS (*Scytonema* sp. PCC 10023)	59/42	AKQ22650
PelE	269	hydroxylase	Hypothetical protein C6503_23180 ('*Candidatus* Poribacteria' bacterium)	68/55	RKU08362
PelF	287	O-methyl-transferase (OMT)	OMT (*Mycolicibacter* sp.)	50/32	WP_082952520
PelG	198	unknown	Hypothetical protein X764_08200 (*Mesorhizobium* sp. LSHC440A00)	57/41	ESX43285

Table S12. TransATor-based[8] structure prediction (moieties in gray have a lower confidence than the ones in black) and A/KS domain substrate prediction of the *pel* BGC. The structures are mostly in agreement except for the starter unit (C1-C4).

TransATor prediction



linearized peloruside



A/KS No.	**Predicted specificity**
KS1	Clade_28 amino acids
KS2	Clade_112 α L-(di)Me β OH
KS3	Clade_88 non-elongating
KS4	Clade_99 double bonds (*E*-configured)
KS5	Clade_102 β D-OH
KS6	Clade_81 non-elongating (β D-OH)
KS7	Clade_103 β OMe
KS8	Clade_111 α L-(di)Me β OH
KS9	Clade_25 completely reduced
KS10	Clade_75 non-elongating (β L-OH)
KS11	Clade_103 β OMe
KS12	Clade_102 β D-OH
KS13	Clade_93 non-elongating (β OH)

Table S13. Proteins encoded in the proposed gananamide BGC (*gan*) and their putative functions.

Protein	**Size, aa**	**Proposed function**	**Closest homolog, protein (origin)**	**S/I, %/%**	**GenBank accession number**
Ψ		*C*-methyl-transferase	Magnesium-protoporphyrin IX monomethyl ester (oxidative) cyclase (*Microvirgula* sp. AG722)	62/48	RAS15729
GanA	147	Precursor peptide	Hypothetical protein VR70_09970 (*Rhodospirillaceae* bacterium BRH_c57)	78/61	KJS38424
GanB	683	*C*-methyl-transferase	RiPP maturation radical SAM protein 1 (*Nostoc* sp. PCC 7120)	58/41	WP_010996184
GanD	494	Epimerase	Aeronamide epimerase AerD (*Microvirgula aerodenitrificans*)	69/53	WP_107890183
GanE	410	*N*-methyl-transferase	Hypothetical protein VR70_09945 (*Rhodospirillaceae* bacterium BRH_c57)	63/49	KJS38421
GanF	595	Dehydratase	Type 2 lantipeptide synthetase LanM (*Cystobacter fuscus*)	47/37	WP_095985948
GanG	168	Unknown	Conserved hypothetical protein (uncultured archaeon GZfos12E1)	55/39	AAU85410
GanH	345	Protease	Hypothetical protein VR70_09935 (*Rhodospirillaceae* bacterium BRH_c57)	74/58	KJS38419
GanI	438	Protease	Hypothetical protein VR70_09935 (*Rhodospirillaceae* bacterium BRH_c57)	64/46	KJS38419
GanK	86	Unknown	DUF2958 domain-containing protein [*Coprobacillus* sp. AF36-10BH]	39/28	WP_117851561
GanL	677	*C*-methyl-transferase	RiPP maturation radical SAM protein 1 (*Trichormus variabilis*)	59/43	WP_011317980

Table S14. Primers used in this study.

No.	Primer Name	Primer Sequence
1	pBAD_F	5'-GGG CGG ATG AGA TCC GAG CTC GAG ATC TG-3'
2	pBAD_R	5'-GTA CTG CAT TCA TGG TTA ATT CCT CCT GTT AG-3'
3	GanD_F	5'-ATT AAC CAT GAA TGC AGT ACT TGA GTC CGA G-3'
4	GanD_R	5'-AGC TCG GAT CTC ATC CGC CCT TCG ATG C-3'
5	pCDF_F	5'-CGT CGG CTA GAG CCA GGA TCC GAA TTC G-3'
6	pCDF_R	5'-TTT GGT CAG CGT GGT GAT GAT GGT GAT G-3'
7	GanA_F	5'-TCA TCA CCA CGC TGA CCA AAG ACC TGC ATT C-3'
8	GanA_R	5'-GAT CCT GGC TCT AGC CGA CGT TGG CGT TG-3'
9	nMT_F	5'-TAT ACA TAT GCT GGA GCC GGA GCC GA-3'
10	nMT_R	5'-TAT ACT CGA GTT AGT TTT GCA GCA AGA TGG-3'
11	BamHI-GanA_F	5'-TGA GGA TCC ATG GCT GAC CAA AGA CCT-3'
12	HindIII-GanA_R	5'-CGT AAG CTT CTA GCC GAC GTT GGC GTT-3'
13	pACYC_F	5'-CGT CGG CTA GGA ATT CGA GCT CGG CGC G-3'
14	pACYC_R	5'-GGT CAG CCA TGT GGT GAT GAT GGT GAT GGC-3'
15	GanA_F	5'-TCA TCA CCA CAT GGC TGA CCA AAG ACC TG-3'
16	GanA_R	5'-GCT CGA ATT CCT AGC CGA CGT TGG CGT TG-3'
17	ganF_FW_NdeI	5'-GGA ATT CCA TAT GGT GAA TGC ATT GCA CGC CCG G-3'
18	ganF_RV_XhoI	5'-CCG CTC GAG TCA TGT CGT AAT CCT GCC GGA TAC-3'
19	nMT-Gib_F	5'-ATA TAC CAT GCT GGA GCC GGA GCC GAC A-3'
20	nMT-Gib_R	5'-TGA TGA TGA TGA TGA TGT TAG TTT TGC AGC AAG ATG GAC-3'
21	pET28b_F	5'-TGC AAA ACT AAC ATC ATC ATC ATC ATC ACA G-3'
22	pET28b_R	5'-CCG GCT CCA GCA TGG TAT ATC TCC TTC TTA AAG-3'
23	poyA-lead_F	5'-TAT GGA TCC TAT GGC AGA CAG CGA CAA C-3'
24	poyA-lead_R	5'-TAC CGA TAG TGC CAC CCG CTG CTT GAT C-3'
25	ganA-core_F	5'-AGC GGG TGG CAC TAT CGG TAT CGT AGT GGT GAT TG-3'
26	ganA-core_R	5'-TAT AAG CTT CTA GCC GAC GTT GGC GTT G-3'
27	ganA-lead_F	5'-TAT GGA TCC TAT GGC TGA CCA AAG ACC TGC-3'
28	ganA-lead_R	5'-CAA TAC CCG TCT GCG GAG CGA CCG CTT C-3'
29	poyA-core_F	5'-CGC TCC GCA GAC GGG TAT TGG TGT CGT TG-3'
30	poyA-core_R	5'-TAT AAG CTT TTA GGT GGT CTG ATT CAT GTT C-3'
31	Pel-TE_F	5'-TCC AAC CAT CCA GGC TCT-3'
32	Pel-TE_R	5'-CGC TGA TAA CGA CCT CTC-3'
33	EcoRI-PelB_F	5'-TCA AGA ATT CAT GAA TCA TTG GTT ATT CAA G-3'
34	HindIII-PelB_R	5'-TGC AAG CTT TCA AAC ACG TCC TGA AAC-3'
35	EcoRI-PelA_F	5'-TCA AGA ATT CTT GCG GAA GAT AGA ATG G-3'
36	HindIII-PelA_R	5'-TGC AAG CTT TCA TTT TTT TCT CGT GAC TC-3'
37	Pel-mod1_F	5'-GCA TGT GAG GTA AAA AAA GGA G-3'
38	Pel-mod1_R	5'-AAT CGG CAG TCA TTG AGT-3'
39	Pel-start_R	5'-ATC CAT CCA AAC CCT ACC C-3'
40	Pel-end_F1	5'-GCG AGC AGG AGG TAG TAC A-3'
41	Pel-end_F2	5'-CCA AAC GAG AGC CAT TAT CAC-3'
42	Pel-OXY_F	5'-CGG ACA AAC TTT ACT CAC CA-3'
43	Pel-DUF_R	5'-TCC CAC AAC TGC ACA CAA-3'

44	Pel-OMT_R	5'-ATC GGT AAT CTA AGC ACT CA-3'
45	N1291_R	5'-GAA AGT GGG TGA AAG GGA GT-3'
46	Pat-ACP_F	5'-ACA TAG ACT CAG GCA GCA-3'
47	Pat-ACP_R	5'-GTC AAC ACT CTT CGG TCA-3'
48	Pat-CC_F	5'-TGG TTG CAG ATC AGT TGG-3'
49	Pat-AA_R	5'-GTC AAC ACT CTT CGG TCA-3'
50	Pat-AA_F	5'-GGA AAT AGA ACA TAG ACT CAG G-3'
51	Pat-KSoxa_R	5'-ATG TGG TCA AAG GTG AGA A-3'
52	Pat-end_F	5'-AGA GCT TGG GAT AGT GGA GA-3'
53	Pat-end_R	5'-GAG CGG GAC GAA GAA GGA T-3'
54	Pat-KS8_F	5'-CAC GCA CCC ATA ACT CAC-3'
55	Pat-Ksend_R	5'-ATC CCT TCT GCC TCC TTT-3'
56	Pat-2898_F	5'-TTG TTG AAA TGG GGT TGG A-3'
57	Pat-42591-start_R	5'-ATT AGG TTT GGA TGC AGG A-3'
58	Pat-910529-end_F	5'-TTT CGT ATT TAC CGG CTC CA-3'
59	Pat-42591-end_F	5'-AAA TAA CCG TAA CCC CGA A-3'
60	Pat-1730153-start_R	5'-GCT TGA GTC TGG GTT GTG-3'
61	Pat1_F	5'-CCC CAA AGA AGA GCT CAC A-3'
62	Pat2_R	5'-GCG TTT GGA GTG TGG TGG A-3'
63	Pat2_F	5'-AAT GCA CAA CCC GAG AAC-3'
64	Pat3_R	5'-CAA TCA ACT CAT GCA GAC AC-3'
65	Pat3_F	5'-AGA GTG TTG ACC ATT GGG A-3'
66	Pat4_R	5'-CAA ACA GGT CAT CGC TCA-3'
67	Pat4_F	5'-CGT GGG TGT GGA ATG GAT-3'
68	11306_R	5'-GAG GGC AGC AAT GAG AAG GA-3'
69	Pat-ver1_F	5'-GGA GGG TAT GAT GCA GGA G-3'
70	Pat-ver2_R	5'-GGT CCG CAG CAT AGT TCA-3'
71	Pat-AA_F	5'-AAG CTG ACA CAA AAG AAG G-3'
72	Pat-AA_R	5'-GGA AGA GGC ATG ACA AGA-3'
73	Pat-2910_F	5'-TAC AGA GAT ATT GGA TGC GA-3'
74	Pat-2911_R	5'-GAG GGA TAG AGA TAT TTG GG-3'
75	Pat-ver3_F	5'-TTG TTG AAA TGG GGT TGG-3'
76	Pat-ver4_R	5'-ATC TCT CCG TTT TCT TCC-3'
77	Pat_ver4_F	5'-CCC TCC GTG CTT TAT CTC-3'
78	Pat-end_R	5'-TCT CCC CGT TCT CTT CCT-3'
79	Pat-KR_F	5'-TCG CTT GGG ATT TGA AGT-3'
80	Pat-Ksexometh_R	5'-AAA CCA CCC TTT GTA CAG-3'
81	Pat-16S_F	5'-TGT GAT TTG CCC ATC CTG CT-3'
82	Pat-16S_R	5'-CGA AGC AAA ACG GAC CAC TG-3'
83	Myc-16S_F	5'-AAC ACT GGC TCG CTT CAC AT-3'
84	Myc-16S_R	5'-TGC GTA TAA TTC GCG CTC CT-3'
85	Gan-16S_F1	5'-GGT TGA CGG ACC TTT AGC GA-3'

86	Gan-16S_F2	5'-AAT TTG TGT GGG CGC TTA CG-3'
87	U1492_R	5'-GGT TAC CTT GTT ACG ACT T-3'
88	Bin45-16S_F1	5'-TCG GGT AGG ATT TTG TTC TTT G-3'
89	Bin45-16S_F2	5'-TCT TCC TGC TCG TTT TTT GTT C-3'
90	Mcz1_F	5'-GAA CGA CAC AAA CAA CAA GA-3'
91	Mcz1_R	5'-AAG ACG TTG TGA GAA ATG G-3'
92	Mcz2_F	5'-CGG ACG CAA AGA ACT GAA-3'
93	Mcz2_R	5'-GGA AAG GGA GAG GCT AGT-3'
94	Mcz3_F	5'-AGA AGG AGG GAT GGA CAG-3'
95	Mcz3_R	5'-AAA GGT TGC GAG CGG ATA-3'
96	Mcz4_F	5'-CAT ATC CGC TCG CAA CCT-3'
97	Mcz4_R	5'-AGA GGA CGC GAT CAG GAA-3'
98	Mcz5_F	5'-GGT TTG GAG CTA GTC GTT-3'
99	Mcz5_R	5'-ACT GGG TAT TGG TTT TGG-3'
100	Mcz6_F	5'-AAA CCA GAT CTC AAC CAC-3'
101	Mcz6_R	5'-CCT TCA GCA TTG ACA AAC-3'
102	Mcz7_F	5'-CAG TAC ATC CCG TGC GAA-3'
103	Mcz7_R	5'-GTT GCG TGA ATT GGG TCT-3'
104	Mcz8_F	5'-TCT CTC GCT GTT CAC TCT-3'
105	Mcz8_R	5'-CAA TGC CTA CCT GAC GAT-3'
106	NR-PKS-1_F	5'-TTG AGC GCA TAT CGA GTT-3'
107	NR-PKS-1_R	5'-GAT CGC TGA TTG GTA GGT-3'
108	NR-PKS-2_F	5'-CCA GCT ATC AAA TTC TCC T-3'
109	NR-PKS-2_R	5'-GTG TGA TAA ATG CGA TGA G-3'
110	NR-PKS-3_F	5'-ACC ATT ACC TTG CCA CCC-3'
111	NR-PKS-3_R	5'-AGC ATC AGC GAC ATC CAC-3'
112	NR-PKS-3_F2	5'-ACC GGG TTC GTT TTC TAC CC-3'
113	NR-PKS-3_R2	5'-CGC CTA GAC CTC CCG TAA TG-3'
114	NR-PKS-4_F	5'-TGG TTG GTG TTC GAT GAT GG-3'
115	NR-PKS-4_R	5'-GCG GGG ATT TGT ATT TGG T-3'
116	Hib1_F	5'-GGA ATA AGC AGG ATA AGA AGG-3'
117	Hib1_R	5'-TCA CGA ATT GAA GGC TGT T-3'
118	Hib2_F	5'-AAT CCC TCT ATC CGC TAA-3'
119	Hib2_R	5'-TCC AAA CGA AGA GAC TCC-3'
120	Hib3_F	5'-CGC CGT TTA TTG TCC AAC A-3'
121	Hib3_R	5'-CTA TTT CTT CGG TTG GCT-3'
122	Hib4_F	5'-GGC GCC TCG ATT AAC AGA-3'
123	Hib4_R	5'-GGA TGC TCG CTA TAC GGA-3'

Supplementary References

1. B. I. Morinaka, M. Verest, M. F. Freeman, M. Gugger, J. Piel, An orthogonal D_2O-based induction system that provides insights into D-amino acid pattern formation by radical *S*-adenosylmethionine peptide epimerases. *Angew. Chem. Int. Ed. Engl.* **56**, 762-766 (2017).
2. M. J. Page, L. West, P. Northcote, C. Battershill, M. Kelly, Spatial and temporal variability of cytotoxic metabolites in populations of the New Zealand sponge *Mycale hentscheli*. *J. Chem. Ecol.* **31**, 1161-1174 (2005).
3. K. Blin *et al.*, antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* **45**, W36-W41 (2017).
4. E. Karimi *et al.*, Genomic blueprints of sponge-prokaryote symbiosis are shared by low abundant and cultivatable *Alphaproteobacteria*. *Sci. Rep.* **9**, 1999 (2019).
5. G. Lackner, E. E. Peters, E. J. N. Helfrich, J. Piel, Insights into the lifestyle of uncultured bacterial natural product factories associated with marine sponges. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E347-E356 (2017).
6. T. Thomas *et al.*, Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *ISME J.* **4**, 1557-1567 (2010).
7. M. Moreno-Pino, A. Cristi, J. F. Gillooly, N. Trefault, Characterizing the microbiomes of Antarctic sponges: a functional metagenomic approach. *Sci. Rep.* **10**, 645 (2020).
8. E. J. N. Helfrich *et al.*, Automated structure prediction of *trans*-acyltransferase polyketide synthase products. *Nat. Chem. Biol.* **15**, 813-821 (2019).

Chapter III

Manuscript in preparation

Characterization of an Orphan NRPS-PKS Hybrid Cluster in the Gananamide Producer "*Ca.* Caria hoplita" Reveals Homologous Clusters in Bacteria of the *Rhodobacteraceae* Family

Michael Rust[1], Pakjira Nanudorn[1], Mariella Greutmann[1], Tomas Kündig[1], Jörn Piel[1*]

[1] Institute of Microbiology, ETH Zurich, 8093 Zurich, Switzerland.

Author Contributions

MR and JP designed research. MR and PN established the transformation protocol and performed knockout studies. MR, PN, MG, and TK performed cloning and expression experiments. MR performed GUS assays and HPLC-MS analysis. MR and JP wrote the manuscript with contributions from all authors.

Abstract

Bacterial sponge symbionts harbor a plethora of biosynthetic gene clusters (BGCs) without assigned products. Typically, these orphan pathways are studied by generating knockout mutants. However, many symbionts are either recalcitrant to cultivation or the biosynthetic pathways are inactive under laboratory conditions, restricting the analysis to BGC expression in heterologous hosts. Here, we pursued an alternative strategy for a small hybrid NRPS/PKS cluster linked to the "*Ca.* Caria hoplita" bin of the *Mycale hentscheli* metagenome. Searching sequenced genomes for similar pathways revealed homologous clusters in bacteria of the *Rhodobacteraceae* family. The cultivated free-living alpha-proteobacterium *Pseudophaeobacter arcticus* harbors a BGC with identical module architecture, suggesting the production of highly similar metabolites. We established a transformation protocol for *P. arcticus* that allowed β-glucuronidase (GUS) reporter assays to analyze the promoter activity of the *P. arcticus* cluster. Testing different expression media revealed conditions under which the cluster is active. The characterization of the natural product in *P. arcticus* will facilitate the isolation of similar compounds from sponge extracts and open avenues for studying the ecological function of these metabolites.

Introduction

Polyketides and nonribosomal peptides represent two major classes of natural products. They are biosynthesized by large multimodular enzyme machineries called polyketide synthases (PKSs) and nonribosomal peptide synthetases (NRPSs), respectively.[1] Generally, each module consists of multiple enzymatic domains and incorporates one building block into the polymeric chain. In both systems, the growing chain is covalently tethered to carrier proteins, acyl carrier proteins (ACPs) in PKSs and peptidyl carrier proteins (PCPs) in NRPSs. A thioesterase (TE) usually located at the C-terminus of the terminal PKS or NRPS catalyzes the release of the polyketide or peptide natural product.[2]

Traditional activity-based screenings of metabolically talented bacteria such as actinomycetes and myxobacteria have yielded numerous bioactive polyketides and peptides, some of which are used in human and veterinary medicine.[3, 4] The high rediscovery rate in well-known producers and the rapidly growing amount of sequence data has shifted the focus to a more global analysis of genes encoding the biosynthesis of natural products.[5] Sequences from environmental microbiomes have been a particularly promising source of novel biosynthetic gene clusters (BGCs).[6] However, only a few BGCs have been linked to the production of a specific metabolite because the producing organisms are often not cultivated/cultivable or unamenable to genetic manipulation, restricting the study of pathways to heterologous hosts.[7]

In our quest to characterize a small hybrid NRPS-PKS cluster from the metagenome of the sponge *Mycale hentscheli*, we identified a homologous cluster with identical module architecture in the free-living bacterium *Pseudophaeobacter arcticus*. *In silico* analysis of biosynthetic enzymes suggested the production of highly similar compounds. Promoter activity studies revealed culturing conditions under which the cluster is active in *P. arcticus* and paved the way for comparative metabolomics experiments to identify the natural product.

Results

In Silico Analysis of the NRPS-PKS Hybrid Cluster

Mining the metagenome of *M. hentscheli* revealed numerous BGCs that could not be assigned to known metabolites. Among them was the small (20 kb) and complete NRPS/*cis*-AT PKS cluster (*cah*) in the "*Ca.* Caria hoplita" bin, the gammaproteobacterial strain that harbors the gananamide pathway.[8] The cluster consists of five unidirectional genes, each encoding a single module (**Fig. 1, Table S1**). The first gene encodes a fatty acyl-AMP ligase (FAAL) fused to an acyl-CoA dehydrogenase. The recently developed AdenylPred tool[9] predicts a substrate length of C_{13}-C_{17} for the FAAL, suggesting an α,β-unsaturated fatty acid thioester bound to the ACP. Prediction of adenylation domain (A) specificities for the three NRPS modules only provided a single amino acid hit for the second A domain, which is predicted to incorporate tyrosine. The methyltransferase domain encoded within this module suggests an *N*-methylated amide bond. The acyltransferase (AT) of the PKS module is predicted to load malonyl-CoA, which together with the ketoreductase (KR) function yields a β-hydroxyacyl thioester. The terminal NRPS module is predicted to incorporate a hydrophobic-aliphatic amino acid and contains a TE that likely catalyzes product release. Interestingly, antiSMASH[10] did not identify similar clusters in public datasets, suggesting a novel natural product. Overall, the *cah* cluster encodes the production of a modified lipopeptide containing at least one tyrosine moiety.

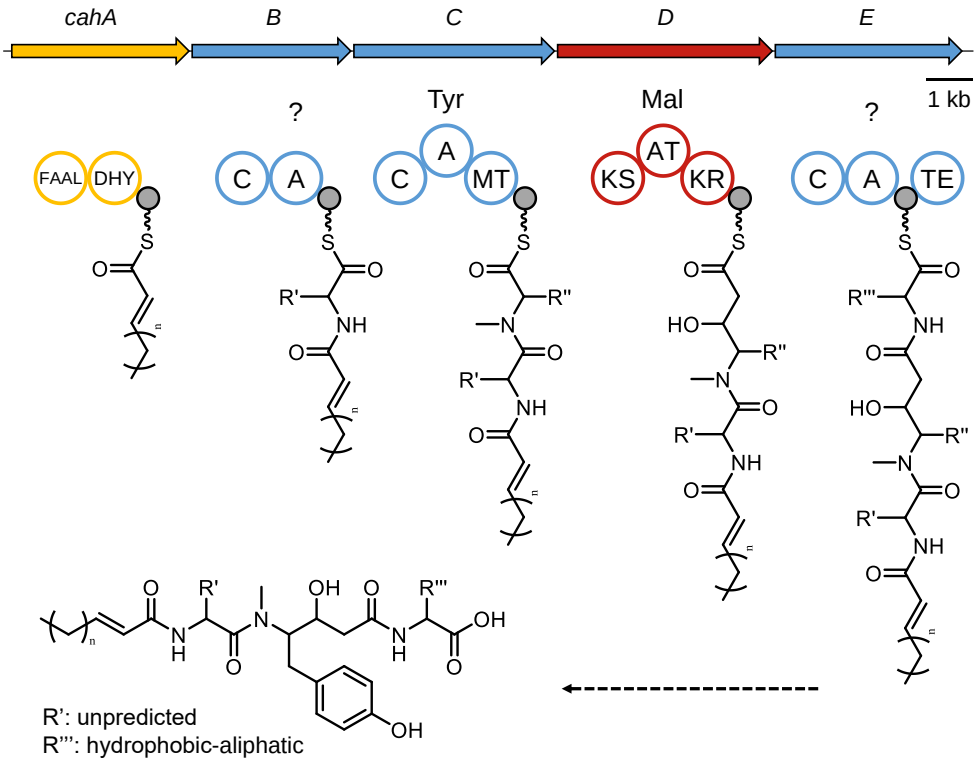


Figure 1. Gene cluster and biosynthetic model for the NRPS/PKS hybrid cluster (*cah*) in "*Ca.* Caria hoplita". Genes are colored according to their proposed function. Predicted substrates are depicted above adenylation (A) and acyltransferase (AT) domains. C, condensation domain; DHY, dehydrogenase; FAAL, fatty acyl-AMP ligase; KR, ketoreductase; KS, ketosynthase; MT, methyltransferase; TE, thioesterase.

Identification of Homologous Clusters in Bacteria of the *Rhodobacteraceae* Family

BLASTx searches of the individual *cah* genes revealed proteins with similar architectures in bacteria of the *Rhodobacteraceae* family (**Fig. S1**). Among them was a cluster (*pba*) with the identical module architecture and an overall amino acid similarity of >50% in *P. arcticus*, a Gram-negative psychrophilic bacterium isolated from marine sediment in the Arctic ocean (**Fig. 2, Table S2**).[11] The presence of a transposase gene directly upstream of both clusters provides evidence for a potential horizontal gene transfer event.[12] Interestingly, the phosphopantetheinyl transferase (PPTase) encoded 10 kb upstream of the *pba* cluster exhibits a high sequence identity (47%) to the PPTase (CahF) in the "*Ca.* Caria hoplita" bin (**Table S2**). The substrate predictions for the FAAL, A, and KS domains of the *pba* cluster are identical to the ones for the *cah* cluster. We therefore concluded that the clusters encode production of highly similar compounds. Because *P. arcticus* was available from strain collections and not known as natural product source, we decided to first characterize the *pba* cluster. The identification of a compound in *P. arcticus* will greatly facilitate the identification of similar compounds in sponge extracts and other *Rhodobacteraceae* members.

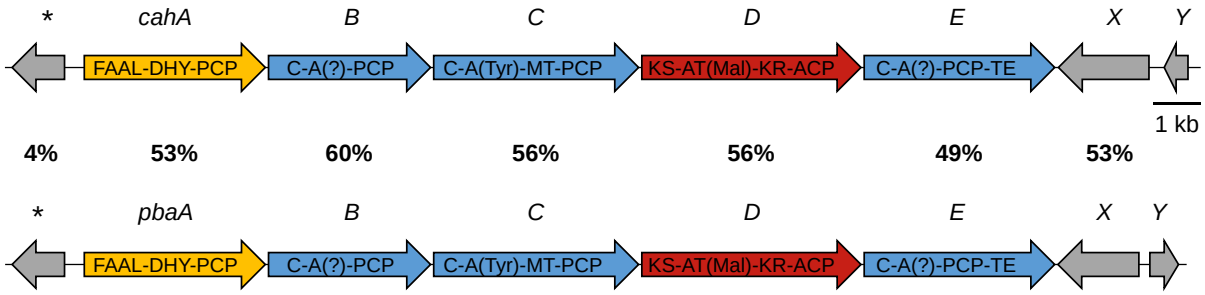


Figure 2. Comparative gene maps of the *cah* ("*Ca.* C. hoplita") and *pba* (*P. arcticus*) clusters. Amino acid identities of the homologs are given above the *pba* genes. Asterisks denote a transposase gene. *CahX* and *pbaX* encode an alkaline phosphatase, *cahY* (NADP Rossmann family) and *pbaY* (OsmY superfamily) are not related.

GUS Reporter Assays Reveal *Pba* Promoter Activity

To verify whether the *pba* pathway in *P. arcticus* is intact, we amplified the complete cluster in four parts by PCR (**Table S3**). Since a metabolic comparison of the wild-type strain to a strain with an inactivated *pba* cluster would facilitate the detection of the corresponding natural product, we attempted to generate a knockout mutant. We established an electroporation-based transformation protocol for *P. arcticus*, in which sufficient washing of the cells proved to be crucial for efficient DNA uptake. *P. arcticus* was transformed with various plasmids harboring different origins of replication and resistance genes (**Table S4**). However, experiments to knock out a part of *pbaA* using the suicide plasmid pEX18Gm[13] failed (**Table S5**). Although we continuously observed single recombinant mutants, SacB-based counterselection did not yield any double-crossover mutants.

To address whether the cluster is active under laboratory growth conditions, we screened for potential promoters[14] of the *pba* cluster and identified a candidate upstream of *pbaA*. The 194 bp intergenic region between the transposase gene and *pbaA* was cloned into the

pSEVA438 plasmid harboring the *gusA* reporter gene by Gibson assembly (**Table S6**). GUS reporter assays revealed that the *pba* promoter is active in medium 1 and to a lesser extent in medium 2, but not in media 3+4 (**Fig. 3**). These results provided an ideal setup for comparative analyses of culture extracts. *P. arcticus* cultures were grown in all four media and analyzed by high-performance liquid chromatography–mass spectrometry (HPLC-MS). Manual inspection of the chromatograms did not reveal significant peak differences between the active and inactive conditions (**Fig. S2**). However, natural products are often produced in low amounts that are difficult to detect.[15] Experiments using different extraction conditions coupled to comparative metabolomics workflows are currently underway to identify the natural product. Interestingly, GUS assays with cultures harboring the inducible XylS/*Pm* regulator/promoter system revealed that the *Pm* promoter is active in *P. arcticus*, even without induction (**Fig. S3**). Therefore, we attempted to clone the *cah* and *pba* clusters into expression plasmids for heterologous expression experiments.

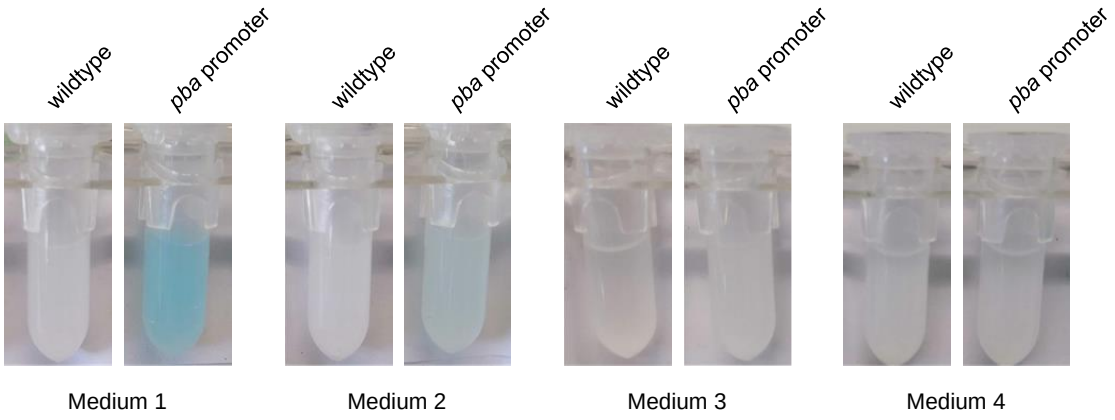


Figure 3. GUS reporter assays in *P. arcticus*. Comparison of the wild type and a strain harboring the plasmid with the *pba* promoter upstream of the *gusA* reporter gene in four different media.

Cloning the *Cah* and *Pba* Clusters for Heterologous Expression Studies

The small size of the *cah* and *pba* clusters and the activity of the XylS/*Pm* regulator/promoter system in *P. arcticus* and other strains render these systems compatible with plasmid-based homologous and heterologous expression. We therefore aimed to clone the clusters into pAMK159 (see Chapter V) using ExoCET, a method based on *in vitro* exonuclease assembly and *in vivo* RecET recombination.[16] The pAMK159 plasmid was amplified using primers with 40 bp homology arms to the start and the end of the *pba* cluster. The cluster was amplified with two primer sets to generate 10 kb fragments with an overlap of 80 bp (**Table S7**). Electroporation of the three exonuclease-treated fragments into the induced recombineering strain yielded the desired construct pAMK159-*pba*, which was verified by PCR and end-sequencing. We experienced that the amplification of long fragments from complex metagenome templates is often not feasible. We therefore amplified the *cah* cluster from the *M. hentscheli* metagenome in four overlapping parts (**Table S8**) and experiments to assemble them with the expression vector by ExoCET are currently underway. These experiments will pave the way towards homologous and heterologous expression studies, potentially linking lipopeptide natural products to a BGC family from different environments.

Discussion

The wealth of BGCs discovered through metagenome mining emphasizes the potential to find chemical novelty in environmental communities.[6] However, only a minor fraction of clusters has been characterized and linked to specific metabolites.[17] The large size of BGCs, combined with their presence in genomes of uncultivated strains, renders many pathways challenging to study. Here, we analyzed a small NRPS/*cis*-AT PKS hybrid cluster (*cah*) identified in the metagenome of the chemically rich sponge *M. hentscheli*.[8] Analysis of the biosynthetic genes revealed highly similar clusters in strains of the *Rhodobacteraceae* family. Members of the *Roseobacter* clade within this family play important roles in algal-bacterial symbioses, in which they biosynthesize roseobacticides[18] and roseochelins[19] in response to algal stimuli. Some *Roseobacter* strains also harbor NRPS and PKS genes, however, no natural products have been isolated or linked to these pathways.[20]

Among the clusters like *cah* was a pathway (*pba*) with identical gene and module architecture from the psychrophilic alpha-proteobacterium *P. arcticus*, suggesting the production of related metabolites. We identified culturing conditions under which the *pba* promoter is active and comparative metabolomics experiments to identify the product are underway. The identification of the natural product in *P. arcticus* will open avenues for the identification of related compounds in the sponge and other members of the *Rhodobacteraceae* family. This approach holds great potential for discovering new chemical scaffolds and study their ecological role in free-living and symbiotic bacteria. Recently, a similar approach led to the HPLC-MS–guided isolation of the first anoxically biosynthesized NRPS/PKS-based hybrid natural product barnesin A.[21]

Interestingly, the *cah* cluster was assigned to the "*Ca.* Caria hoplita" bin that also encodes production of the orphan natural product gananamide, a ribosomally synthesized and posttranslationally modified peptide (RiPP).[8] Since horizontal gene transfer might explain the presence of the NRPS/PKS hybrid cluster in various *Rhodobacteraceae* strains, members of this family could represent good candidates for finding novel RiPP pathways. Overall, the study demonstrates that a detailed *in silico* analysis of orphan BGCs can lead to the identification of homologous pathways in cultivated strains, thereby allowing for a targeted and more efficient isolation approach.

Materials and Methods

DNA Isolation

Metagenomic DNA from *M. hentscheli* was isolated as described previously.[8] Genomic DNA from *P. arcticus* (DSM 23566) was isolated with the Nucleospin Microbial DNA kit (Macherey-Nagel) according to the manufacturer's protocol.

Amplification of DNA

PCRs were performed with Q5 High-Fidelity DNA Polymerase (NEB). A typical PCR (25 µL) contained 1× Q5 reaction buffer, 200 µM deoxyribose nucleoside triphosphates (dNTPs), 0.5 µM of each primer, 20 to 50 ng template DNA, and 0.5 U Q5 High-Fidelity DNA Polymerase. For the amplification of long fragments and vector backbones, 5% DMSO or 1 M betaine were included in the PCR. The reaction was heated to 98 °C for 30 s followed by 30 cycles of 98 °C for 10 s, 62 °C for 20 s, and 72 °C for 30 s per kilobase DNA target sequence. At the end, a final incubation at 72 °C for 2 min was performed. For Gibson assembly fragments, a two-step PCR

was performed with an annealing temperature for the 3' binding part of the primers for the first 5 cycles, followed by 30 cycles with an annealing temperature of the full-length primers.

Gibson Assembly

Approximately 100 ng of PCR-amplified vector DNA was mixed with three times molar excess of insert DNA fragments and 10 µL Gibson assembly master mix (NEB) in a total volume of 20 µL. The reaction mixture was incubated at 50 °C for 1 h. Chemically competent *E. coli* DH5α cells were transformed with 5 µL of the reaction mixture.

ExoCET Cloning

ExoCET assemblies were performed as described previously[16] with slight modifications. 100 ng of each PCR-amplified fragment were mixed with 1× NEBuffer 2.1 and 0.2 µL of T4 DNA polymerase (NEB) in a total reaction volume of 20 µL and incubated at 25 °C for 1 h, 75 °C for 20 min, and 50 °C for 30 min in a thermocycler. The reaction mixture was dialyzed against water on Millipore Membrane Filters (Merck-Millipore) at room temperature for 30 min. 280 µL of an overnight culture (*E. coli* GBdir-gyrA462 harboring the pSC101-BAD-ETgA-tet plasmid[22]) were added to 10 mL LB medium and incubated at 30 °C and 200 rpm for 2 h. The culture was induced with 250 µL of 10% (w/v) L-arabinose and incubated at 37 °C and 200 rpm for 40 min. The culture was cooled on ice and the cells were washed two times with ice-cold water. Cells were resuspended in 20 µL of water and 5 µL of the desalted exonuclease reaction were added. The mixture was transferred to a precooled 1 mm cuvette and electroporated at 1.35 kV.

Transformation of *P. arcticus*

A preculture of *P. arcticus* was diluted 1:50 in marine broth (BD Difco) and cells were grown at 16 °C and 200 rpm to an OD_{600} of 0.6 to 1.0 for approximately 48 h. Cultures were cooled on ice and centrifuged at 5,000× g and 4 °C for 15 min. The supernatant was discarded, and the cells were washed five times with ice-cold 10% glycerol. The cells were resuspended in ice-cold 10% glycerol and 80 µL aliquots were immediately used for electroporation.

Electrocompetent cells were mixed with 2 µL of plasmid DNA and the mixture was transferred to a precooled 2 mm cuvette. The cells were electroporated at 2.5 kV and recovered in marine broth containing 50% of salts (peptone 5 g/L, yeast extract 1 g/L, Instant Ocean sea salts 16 g/L) overnight. The cells were plated on marine broth agar containing 50% of salts and incubated at room temperature for several days.

GUS Reporter Assays

P. arcticus cultures harboring the *gusA* plasmid were grown to an OD_{600} of 0.6, centrifuged at 5,000× g for 10 min and washed with 0.9% NaCl (w/v). The cells were resuspended in 200 µL 0.9% NaCl and mixed with 800 µL of GUS assay buffer (50 mM NaH_2PO_4, 10 mM β-mercaptoethanol, 2 mM EDTA, 0.1% sarkosyl, and 0.1% Triton X-100) and three drops of chloroform. The mixture was vortexed for 10 s and 125 µL of X-Gluc (4 mg/mL) were added. The mixture was vortexed for 10 s and incubated at 37 °C. A blue color appeared in β-glucuronidase-positive samples after 15 to 30 min. Medium 1: marine broth (BD Difco) 18.7 g/L; medium 2: peptone 5 g/L, yeast extract 1 g/L, Instant Ocean sea salts 16 g/L; medium 3: tryptone 20 g/L, yeast extract 5 g/L, dextrose 4 g/L, maltose 4 g/L, $CaCO_3$ 4 g/L, Instant Ocean sea salts 16 g/L; medium 4: same as medium 3 without $CaCO_3$.

Extraction and HPLC-MS Analysis

Cultures were centrifuged at 10,000× g for 5 min. The supernatant was extracted with 2 volumes of ethyl acetate and the pellet was extracted with 5 mL of acetone:methanol (1:1). The organic phase was dried under reduced pressure. Extracts were dissolved in acetonitrile:water (1:1) and centrifuged at 20,000 × g for 10 min. The supernatant was analyzed by HPLC-MS on a Phenomenex Kinetex 2.6 µm XB-C18 100 Å (150 × 2.1 mm) column. The column was heated to 27 °C and the solvents used were water with 0.1% (v/v) formic acid (solvent A) and acetonitrile with 0.1% (v/v) formic acid (solvent B). A flow rate of 0.8 mL/min with solvent B at 5% from 0 to 2 min, 5% to 98% from 2 to 12 min, 98% from 12 to 15 min, 98% to 5% from 15 to 17 min, and 5% from 17 to 19 min was used. ESI-MS was performed in positive ion mode, with a spray voltage of 3,500 V, a capillary temperature of 280 °C, probe heater temperature of 475 °C and an S-Lens RF level of 50. Full MS was performed at a resolution of 140,000 (AGC target 1e6, maximum IT 150 ms, range 300–800 *m/z*). Data-dependent tandem mass spectrometry (MSMS) was performed at a resolution of 17,500 (AGC target between 2e5 and 1e6, maximum IT between 100 and 500 ms, isolation windows in the range of 1.1 to 2.2 *m/z*) using a normalized collision energy (NCE) of 30.

Acknowledgements

This project has received funding from ETH Research Grant ETH-26 17-1, Swiss National Science Foundation Grants 205321 and 205320, and the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme Grant 742739 and the Helmut Horten Foundation. J.P. is grateful for an Investigator Grant of the Gordon and Betty Moore Foundation. P.N. was supported by a Swiss Government Excellence Scholarship.

References

1. D. E. Cane, C. T. Walsh, The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases. *Chem. Biol.* **6**, R319-R325 (1999).
2. M. A. Fischbach, C. T. Walsh, Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.* **106**, 3468-3496 (2006).
3. K. S. Lam, Discovery of novel metabolites from marine actinomycetes. *Curr. Opin. Microbiol.* **9**, 245-251 (2006).
4. K. J. Weissman, R. Müller, Myxobacterial secondary metabolites: bioactivities and modes-of-action. *Nat. Prod. Rep.* **27**, 1276-1295 (2010).
5. N. Ziemert, M. Alanjary, T. Weber, The evolution of genome mining in microbes – a review. *Nat. Prod. Rep.* **33**, 988-1005 (2016).
6. A. Milshteyn, J. S. Schneider, S. F. Brady, Mining the metabiome: identifying novel natural products from microbial communities. *Chem. Biol.* **21**, 1211-1223 (2014).
7. A. G. Atanasov *et al.*, Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discovery* **20**, 200-216 (2021).
8. M. Rust *et al.*, A multiproducer microbiome generates chemical diversity in the marine sponge *Mycale hentscheli*. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9508-9518 (2020).
9. S. L. Robinson *et al.*, Global analysis of adenylate-forming enzymes reveals β-lactone biosynthesis pathway in pathogenic *Nocardia*. *J. Biol. Chem.* **295**, 14826-14839 (2020).
10. K. Blin *et al.*, antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81-W87 (2019).
11. D.-C. Zhang *et al.*, *Phaeobacter arcticus* sp. nov., a psychrophilic bacterium isolated from the Arctic. *Int. J. Syst. Evol. Microbiol.* **58**, 1384-1387 (2008).
12. P. R. Jensen, Natural products and the gene cluster revolution. *Trends Microbiol.* **24**, 968-977 (2016).

13. T. T. Hoang, R. R. Karkhoff-Schweizer, A. J. Kutchma, H. P. Schweizer, A broad-host-range Flp-*FRT* recombination system for site-specific excision of chromosomally-located DNA sequences: application for isolation of unmarked *Pseudomonas aeruginosa* mutants. *Gene* **212**, 77-86 (1998).
14. R. K. Umarov, V. V. Solovyev, Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS One* **12**, e0171410 (2017).
15. B. C. Covington, J. A. McLean, B. O. Bachmann, Comparative mass spectrometry-based metabolomics strategies for the investigation of microbial secondary metabolites. *Nat. Prod. Rep.* **34**, 6-24 (2017).
16. H. Wang *et al.*, ExoCET: exonuclease *in vitro* assembly combined with RecET recombination for highly efficient direct DNA cloning from complex genomes. *Nucleic Acids Res.* **46**, e28 (2018).
17. R. V. O'Brien, R. W. Davis, C. Khosla, M. E. Hillenmeyer, Computational identification and analysis of orphan assembly-line polyketide synthases. *J. Antibiot.* **67**, 89-97 (2014).
18. M. R. Seyedsayamdost, G. Carr, R. Kolter, J. Clardy, Roseobacticides: small molecule modulators of an algal-bacterial symbiosis. *J. Am. Chem. Soc.* **133**, 18343-18349 (2011).
19. R. Wang, M. R. Seyedsayamdost, Roseochelin B, an algaecidal natural product synthesized by the *Roseobacter Phaeobacter inhibens* in response to algal sinapic acid. *Org. Lett.* **19**, 5138-5141 (2017).
20. T. Martens *et al.*, Bacteria of the *Roseobacter* clade show potential for secondary metabolite production. *Microb. Ecol.* **54**, 31-42 (2007).
21. M. Rischer *et al.*, Biosynthesis, synthesis, and activities of barnesin A, a NRPS-PKS hybrid produced by an anaerobic epsilonproteobacterium. *ACS Chem. Biol.* **13**, 1990-1995 (2018).
22. H. Wang *et al.*, RecET direct cloning and Redαβ recombineering of biosynthetic gene clusters, large operons or single genes for heterologous expression. *Nat. Protoc.* **11**, 1175-1190 (2016).

Supplementary Information

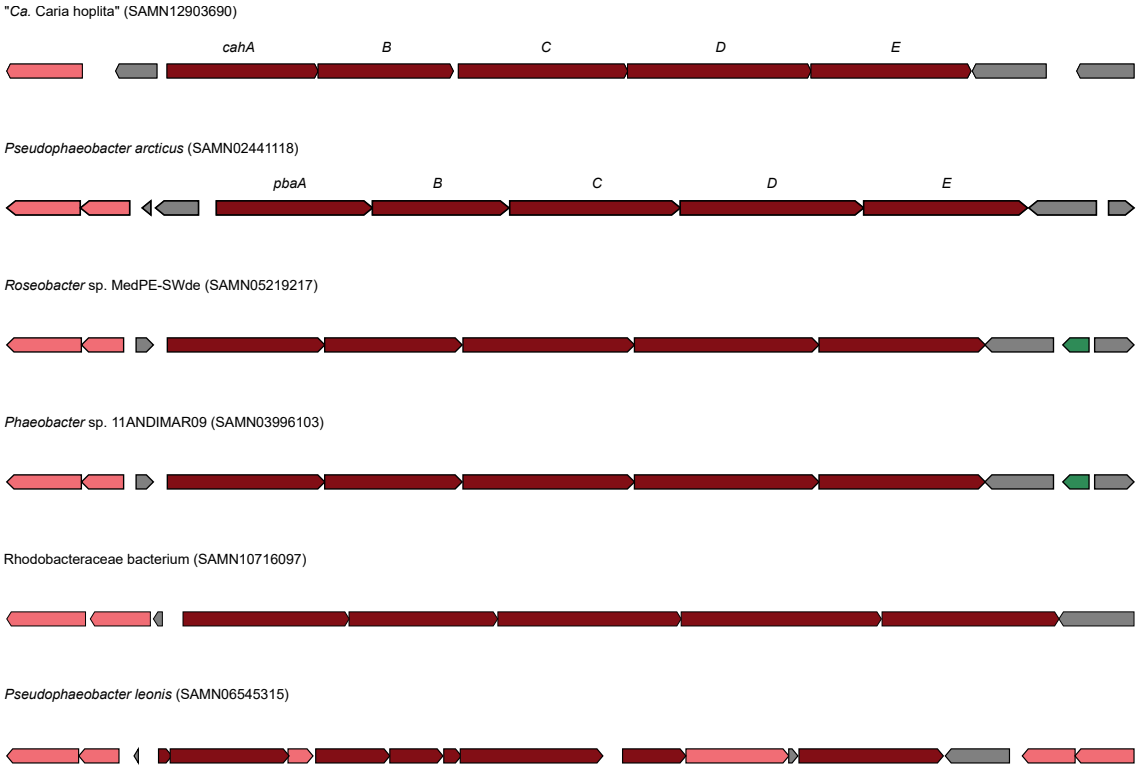


Figure S1. Comparative gene maps of the *cah* cluster and homologous clusters in bacteria of the *Rhodobacteraceae* family.

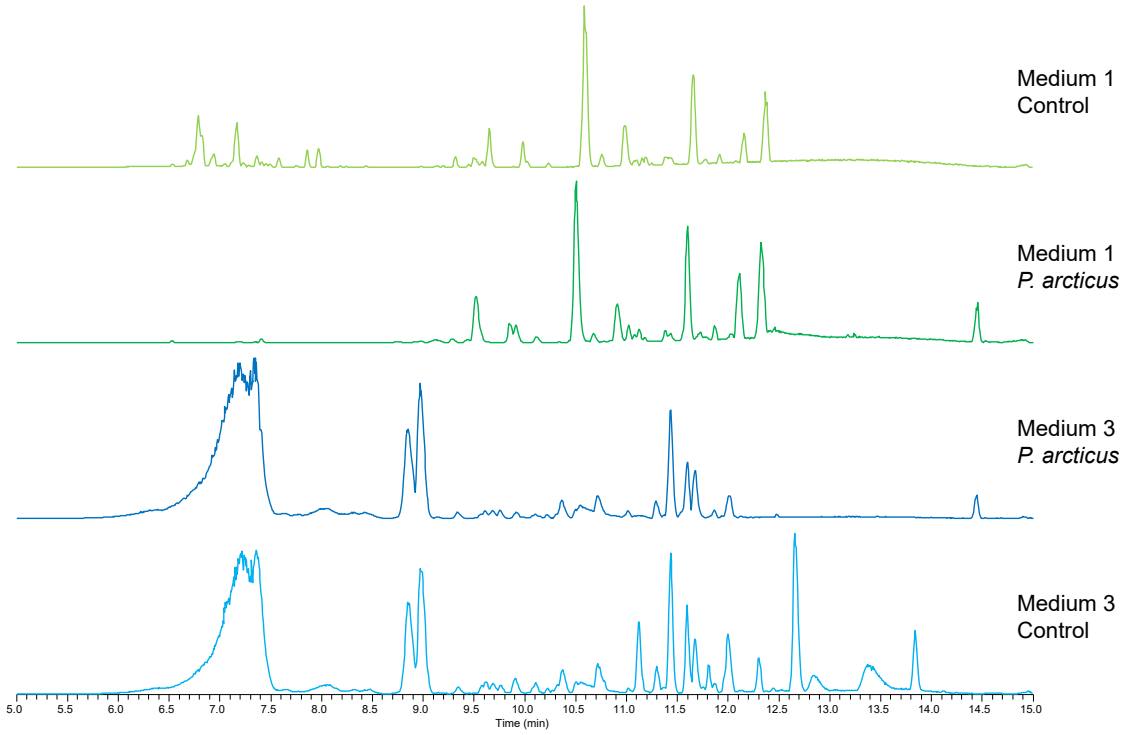


Figure S2. Total ion chromatograms (TICs) for *P. arcticus* cultures and media controls.

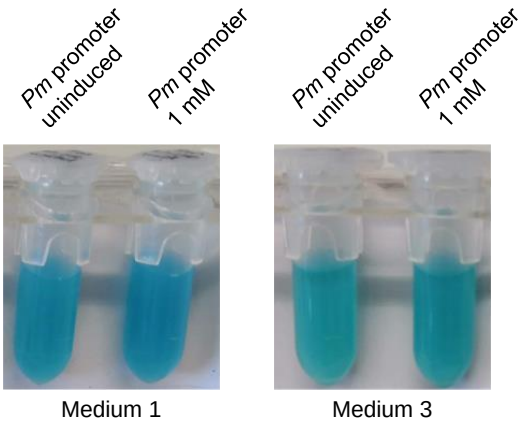


Figure S3. GUS reporter assays for *P. arcticus* harboring the *Pm* promoter construct. Uninduced and induced (1 mM *m*-toluic acid) *P. arcticus* cultures with the pSEVA438 plasmid harboring the *Pm* promoter upstream of the *gusA* reporter gene.

Table S1. Proteins encoded in the NRPS/PKS hybrid cluster (*cah*) of "*Ca.* Caria hoplita" and their putative functions.

Protein	**Size, aa**	**Proposed function**	**Closest homolog, protein (origin)**	**S/I, %/%**	**GenBank accession number**
CahA	1,275	Fatty acyl-AMP ligase (FAAL) and acyl-CoA dehydrogenase	hypothetical protein EP318_04530 (*Rhodobacteraceae* bacterium)	68/56	TNF22211.1
CahB	1,141	NRPS	NRPS (*Pseudophaeobacter arcticus*)	71/60	WP_027238316.1
CahC	1,428	NRPS	NRPS (*Pseudophaeobacter arcticus*)	69/56	WP_027238317.1
CahD	1,548	PKS	hypothetical protein BM560_12700 (*Roseobacter* sp. MedPE-SWde)	68/56	OIQ40355.1
CahE	1,349	NRPS	NRPS (*Pseudophaeobacter* sp. EL27)	62/49	WP_122077281.1

Table S2. Comparison of the *cah* cluster from "*Ca.* Caria hoplita" with the homologous *pba* cluster in *P. arcticus*.

Protein	**Closest homolog in *P. arcticus***	**Query Cover**	**Percent Identity**	**GenBank accession number**
CahA	AMP-binding protein	98%	53%	WP_027238315.1
CahB	Non-ribosomal peptide synthetase	98%	60%	WP_027238316.1
CahC	Non-ribosomal peptide synthetase	97%	56%	WP_027238317.1
CahD	Type I polyketide synthase	99%	56%	WP_027238318.1
CahE	Non-ribosomal peptide synthetase	99%	49%	WP_027238319.1
CahF	Phosphopantetheinyl transferase	84%	47%	WP_161631270.1

Table S3. Primers used to screen for the presence of the *pba* cluster in *P. arcticus*.

Name	**Sequence (5'-3')**
PA_pba1_F	TGGAGTCATGAACATGGATTTTGAAAAAGGCTTGATCAG
PA_pba1_R	CATGGCTGGCAAATGTGGCCGTCTCGTCGAA
PA_pba2_F	CGAGACGGCCACATTTGCCAGCCATGTTGCTG
PA_pba2_R	GCCGTCGCACCATATCGCCGCTGCGGTAAAG
PA_pba3_F	CCGCAGCGGCGATATGGTGCGACGGCTGGAG
PA_pba3_R	GCAATACGGGTTGGATCCTCTGTATCCAGATCCAGAACC
PA_pba4_F	GATACAGAGGATCCAACCCGTATTGCCAGCC
PA_pba4_R	CTATAAACGCAGATTACCCCTGTTGCTCAAGAAC

Table S4. Plasmids used to transform *P. arcticus*.

Plasmid name	**Origin of replication**	**Resistance**
pEX18Gm	ColE1	Gentamycin
pSEVA438	pBBR1	Streptomycin
pAMK159 (see chapter V)	RK2	Kanamycin

Table S5. Primers used to amplify the fragments of the suicide plasmid by overlap-extension PCR, and primers used to screen for deletion mutants.

Name	**Sequence (5'-3')**
FAAL_UP_F	CGGAATTCCTATATGGCGAGGTTCACGCC
FAAL_UP_R	CTGGAATTGTCGCCACAATCTGCGGCAAC
FAAL_DOWN_F	GATTGTGGCGACAATTCCAGATATCGATGCGC
FAAL_DOWN_R	GGGGTACCCCGGTGATATGGTGCATCACCA
FAAL_mut_F	TTGCGTCAGGCGTGGAAGAC
FAAL_mut_R	CTGCATACAGCTGCTGCAGC
PA_genome_F	GAAATTCCCACCGCCTCTGC
PA_genome_R	CCTTCAGCGGCAGCATATTG

Table S6. Primers used to amplify pSEVA438, the *gusA* gene, and the *pba* promoter region for Gibson assemblies.

Name	**Sequence (5'-3')**
p438_F	CAAACAATGATACCCGGGGATCCTCTAG
p438_R	TGGTCGGCGGGCACTCCTGTATCCGCTTC
pba_prom_F	ACAGGAGTGCCCGCCGACCATGCCACAA
pba_prom_R	GACGGACCATGGCATCGAAAATCTACAAGAAAACTATCGG
gusA_F	TTTCGATGCCATGGTCCGTCCTGTAGAAAC
gusA_R	TCCCCGGGTATCATTGTTTGCCTCCCTG
p438_gusA_F	CAAACAATGATACCCGGGGATCCTCTAG
p438_gusA_R	GGTCATGACTCCATTATTATTGTTTC
gusA_p438_F	ATAATAATGGAGTCATGACCATGGTCCGTCCTGTAGAAAC
gusA_p438_R	TCCCCGGGTATCATTGTTTGCCTCCCTG

TableS7. Primers used to amplify the *pba* cluster of *P. arcticus* and pAMK159 for ExoCET cloning.

Name	**Sequence (5'-3')**
p159_pba_F	TTCTTGAGCAACAGGGGTAATCTGCGTTTATAGCTCTTC
p159_pba_R	AAGCCTTTTTCAAAATCCATGTTCATGACTCCATTATTATTG
PA_pba1_F	ATAATAATGGAGTCATGAACATGGATTTTGAAAAAGGCTTGATCAG
PA_pba1_R	CGTCATAGACCAGAAATTCAGAAATCGACGGCCACCAC
PA_pba2_F	GAGTGGTGGCCGTCGATTTCTGAATTTCTGGTCTATGACG
PA_pba2_R	TGAAGAGCTATAAACGCAGATTACCCCTGTTGCTCAAG

TableS8. Primers used to amplify the *cah* cluster of "*Ca.* Caria hoplita" and pAMK159 for ExoCET cloning.

Name	**Sequence (5'-3')**
p159_cah1_F	ACCGGAACTGGCGCACCGCCTTCGTTCACTGTTCGACTGATCTGCGTTTATAGCTCTTC
p159_cah1_R	GGCGCCGGCGACTCAGTGCTCCAATCAACTCGTCAATGTTCATGTTCATGACTCCATT
MH_cah1_F	AACATTGACGAGTTGATTGG
MH_cah1_R	CAGCACGAAGGTGGTTTG
MH_cah2_F	CGTTTTCAATGATTGCGG
MH_cah2_R	TCGGGTAGAAAACGAACC
MH_cah3_F	TTGTGCCGCATCCTTTCG
MH_cah3_R	GACTTCATCGAAGCTGCC
MH_cah4_F	GGAGCAAGGGTTCTTTGC
MH_cah4_R	TCAGTCGAACAGTGAACG

Chapter IV

Manuscript in preparation

Indirect Sequence Capture and Long-Read Sequencing Link Biosynthetic Gene Clusters to Bacterial Producers in a Sponge Metagenome

Michael Rust[1], Charlotte Carlström[1], Jeremy G. Owen[2], Robert A. Keyzers[3], Michael J. Page[4], Christopher M. Field[1], Salome A. Hegi[1], Shinichi Sunagawa[1], Jörn Piel[1]

[1] Institute of Microbiology, ETH Zurich, 8093 Zurich, Switzerland.
[2] School of Biological Sciences, Victoria University of Wellington, 6012 Wellington, New Zealand.
[3] School of Chemical and Physical Sciences, Victoria University of Wellington, 6140 Wellington, New Zealand.
[4] Aquaculture, Coasts and Oceans, National Institute of Water and Atmospheric Research Ltd., 7010 Nelson, New Zealand.

Author Contributions

MR, CC, SS, and JP designed research. MR, JGO, RAK, and MJP collected samples. MR, CC, and SAH performed experiments. MR, CC, CMF, and SS performed sequencing and data analysis. MR and JP wrote the manuscript with contributions from all authors.

Abstract

Marine sponges are a rich source of bioactive natural products with therapeutic potential. Pelorusides isolated from the sponge *Mycale hentscheli* are potent microtubule stabilizers with substantial promise for clinical development. However, challenging chemical syntheses and the short supply of sponge material have restricted clinical trials. The identification of the biosynthetic pathway in the metagenome of *M. hentscheli* opened avenues for biotechnological production efforts. However, the biosynthetic gene cluster was not assigned to a producer genome, complicating targeted isolation approaches and heterologous expression studies. Here, we tested long-read sequencing and a targeted enrichment strategy to gain insights into producers in *M. hentscheli*. The enrichment method relies on the PCR-based amplification of a short DNA sequence inside emulsion droplets. Droplets are sorted and the enriched DNA is amplified prior to sequencing. Our results putatively linked the peloruside pathway to the mycalamide producer "*Ca.* Entomycale ignis", a member of an uncultivated gammaproteobacterial taxon that was previously not known as source of natural products. The strategy was expanded to two other unassigned clusters and putatively linked a large orphan polyketide pathway to a member of the Verrucomicrobiota phylum.

Introduction

Marine natural products exhibit a wide range of biological activities that make them exceptional candidates for therapeutic applications.[1] Many marine-derived compounds have entered clinical trials for the treatment of cancer.[2] The polyketide peloruside A isolated from the marine sponge *Mycale hentscheli*[3] has gained substantial attention owing to its activity against diverse cancer cell types and effects in neurological and immune disorders.[4] Interestingly, peloruside A inhibits cancer cells that are resistant to paclitaxel, an important anticancer drug used for the treatment of various solid tumors.[5] Competition assays revealed that peloruside A binds to a different non-taxoid site on β-tubulin[6] that is only shared with laulimalide, a macrolide isolated from other sponges (**Fig. 1**).[7]

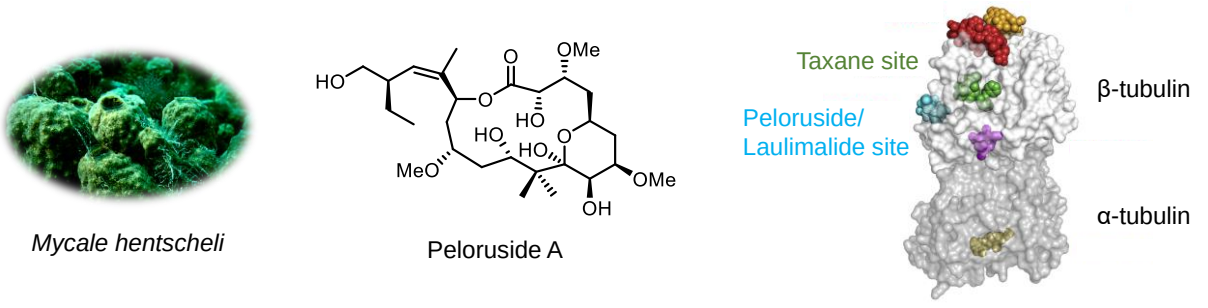


Figure 1. Peloruside A isolated from the sponge *Mycale hentscheli* and its external binding site on the β-tubulin subunit. The peloruside/laulimalide binding site is shown in blue relative to the taxane binding site in green (adapted from Steinmetz and Prota, *Trends Cell Biol.* 2018[8]).

Major challenges in the drug development of sponge-derived metabolites are the limited natural supply and difficult chemical syntheses.[9] Numerous natural products originally isolated from sponges have been assigned to bacterial symbionts,[10-13] providing alternative strategies to secure sufficient material for clinical trials. Genome data of the producing organisms provides crucial information about metabolic requirements or suitable host systems.[14] However, biosynthetic gene clusters (BGCs) of large modular pathways such as polyketide synthases (PKSs) are often scattered between multiple contigs in fragmented assemblies, and therefore not linked to specific microbial genomes in metagenomic datasets.[15] The combination of short- and long-read sequencing has been shown to greatly improve the quality of metagenome-assembled genomes (MAGs).[16] However, low-abundance organisms might remain unnoticed in such datasets or require extensive sequencing to achieve sufficient coverage.[17]

Here, we performed long-read sequencing and applied a new indirect sequence capture technology (Xdrop™)[18] for the enrichment of long DNA fragments to link unbinned polyketide BGCs to microbial producers in the metagenome of *M. hentscheli* (**Fig. 2**). The method relies on partitioning of high-molecular-weight (HMW) DNA together with PCR reagents into millions of double emulsion droplets. A short (150 bp) detection sequence is amplified by PCR inside droplets harboring the region of interest and a DNA-intercalating fluorescent dye stains the droplets in which product amplification occurred, facilitating cell sorting via standard flow cytometry. Subsequently, the enriched DNA containing the region of interest (ROI) is released and partitioned into emulsion droplets for single molecule multiple displacement amplification (dMDA). The amplified enriched DNA is compatible with short- and long-read sequencing, and PCR amplicons are rarely detected in downstream sequencing because the MDA reaction is most efficient for fragments larger than five kilobases. We targeted the peloruside (*pel*) BGC and two other unbinned polyketide pathways from the metagenome of *M. hentscheli* and identified putative links to microbial genomes for the peloruside cluster and a large orphan polyketide pathway.

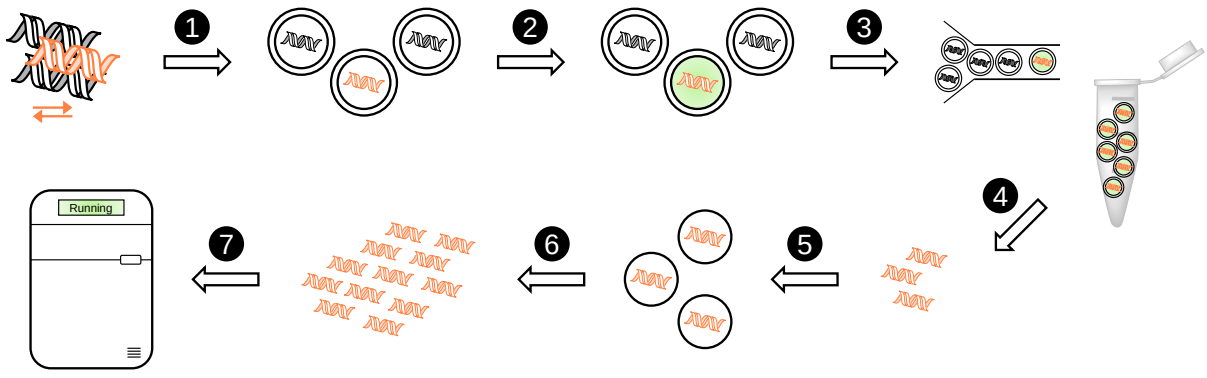


Figure 2. Indirect sequence capture and multiple displacement amplification workflow. 1) DNA is mixed with detection primers and partitioned into double emulsion droplets using the XDrop™ instrument. 2) In-droplet PCR amplifies the detection sequence in droplets containing the region of interest (ROI) and an intercalating dye stains positive droplets. 3) Positive droplets are separated from negative droplets and collected by fluorescence-activated cell sorting. 4) DNA enriched for the ROI is released from double emulsion droplets. 5) Each DNA fragment is partitioned into single emulsion droplets for multiple displacement amplification (MDA). 6) Amplified DNA fragments are released from the single emulsion droplets. 7) The enriched DNA is subjected to short- and long-read sequencing.

Results

Unbinned Clusters in the *M. hentscheli* Metagenome

We started our quest to link unbinned BGCs in the metagenome of *M. hentscheli* to microbial genomes by selecting candidate pathways for the indirect sequence capture technology. From the clusters identified for the three known polyketides mycalamide, pateamine, and peloruside, only the latter was not assigned to a microbial producer in the binning process.[12, 13] The identification of the peloruside producer is relevant in several ways. First, it will reveal whether peloruside is generated by yet another of the numerous symbionts harboring only few biosynthetic pathways or a biosynthetically more decorated strain. Second, it would allow for a targeted isolation of the producer from *M. hentscheli* or alternative sources and enable studies about the ecological function of the compound. Last, it would provide valuable information for heterologous expression studies regarding host selection and metabolic requirements. The relatively small size of the six-gene *pel* cluster and the pharmacological importance of peloruside[4] render this pathway a promising system for biotechnological production. The second pathway is the *mcz* cluster putatively involved in the biosynthesis of mycalazols and mycalazals (see Chapter V). Since these compounds were isolated from multiple *Mycale* species,[19] information about the producing organism(s) will provide valuable insights about the evolution of the pathway and the ecological relevance of the compounds. The third BGC (*ver*) encodes a huge and seemingly complete *trans*-AT PKS consisting of three PKS proteins, a methyltransferase, and a glycosyltransferase. The identification of a producer will facilitate product isolation for a pathway that is currently too complex for cloning and heterologous expression systems (**Fig. 3**).

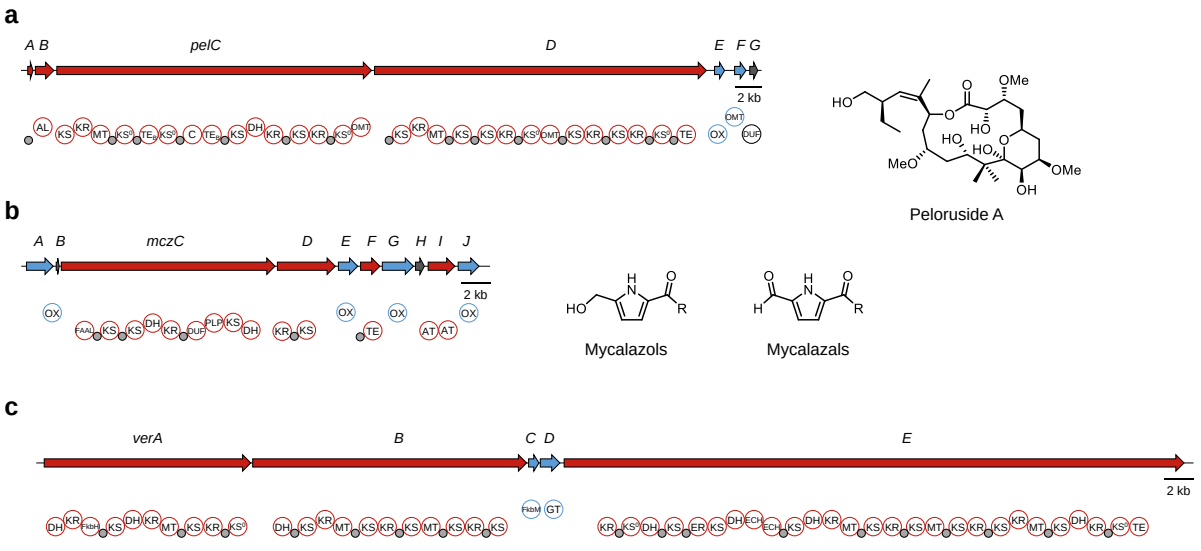


Figure 3. Unbinned BGCs from the *M. hentscheli* metagenome targeted for indirect sequence capture. Core PKS genes are shown in red, tailoring biosynthetic genes in blue, and genes with unknown function in gray. a) *pel* cluster responsible for peloruside biosynthesis. b) *mcz* cluster hypothesized to encode production of mycalazols and mycalazals. c) *ver* cluster encoding a large *trans*-AT PKS with unknown product.

Indirect Sequence Capture and Multiple Displacement Amplification Enrich Regions of Interest

Two primer sets targeting both ends of the three clusters (*pel*, *mcz*, *ver*) were designed and evaluated by qPCR (**Table S1**). The first primer set (dPCR primers) is used to amplify the detection sequence in droplets, while the second set (qPCR primers) facilitates calculation of the enrichment factor compared to the original sample after MDA. The designed primers include non-overlapping amplicons and are placed within 2 kb from each other for optimal amplification. All primers showed an efficiency between 90% and 110% and single peak amplicon melting profiles indicating a high specificity. This is crucial for generating a strong signal in positive droplets containing the target sequence, providing a means for separation from the negative population into distinct clusters (**Fig. S1**). The optimal amount of DNA was determined from the threshold cycle (C_t) values obtained in the qPCRs with varying input concentrations. Droplet generation and sorting were successful for all samples with a typical outcome of 200 to 800 positive droplets from 2 ng input material. The DNA was isolated and amplified in single emulsion droplets by dMDA, yielding 400 to 1,000 ng enriched DNA per sample. Evaluation by qPCR estimated an enrichment of 60- to 600-fold for the different samples, reaching the cutoff criterium of 100-fold for most but not all the samples. The low enrichment factors obtained for both samples and the positive control in some workflows suggested that individual experimental steps rather than the sample or primers were the cause for reduced enrichment.

After enrichment, we proceeded with long-read sequencing (Oxford Nanopore Technologies, ONT). Although we observed an enrichment for the targeted parts of the clusters (**Fig. S2**), chimeric reads consisting of concatemers of inverted repeats or the same sequence prevented the extension of the BGC contigs. We are currently pursuing the resolution of this issue by applying strategies for splitting palindromic reads, including tools such as Pacasus,[20] and complementing the analysis with short-read sequencing data.

Long-Read Sequencing Extends Biosynthetic Contigs

Based on the results that both ends of the three unbinned clusters were amplified with relatively low C_t values during primer validation, we hypothesized that long-read sequencing of the original HMW DNA facilitates contig mapping and might be sufficient for BGC extension. We therefore subjected the original HMW DNA used for the Xdrop™ workflow to ONT long-read sequencing and generated approximately 310,000 reads longer than 5 kb. The three unbinned clusters (*pel*, *mcz*, *ver*) were largely covered by the ONT reads (**Fig. S3**). We manually inspected reads mapped to the ends of the clusters and mapped them back to the Illumina dataset. One of the ONT reads harboring the C-terminal part of the *pel* cluster spanned two Illumina contigs, the latter of which was assigned to the bin "*Ca.* Entomycale ignis" harboring the mycalamide pathway. The connection between the two contigs was verified by PCR (**Fig. 4** and **Table S2**). Since the *pel* pathway does not contain an acyltransferase (AT) required for the loading of the acyl carrier proteins (ACPs), the AT encoded in the mycalamide cluster might be shared between the two pathways.

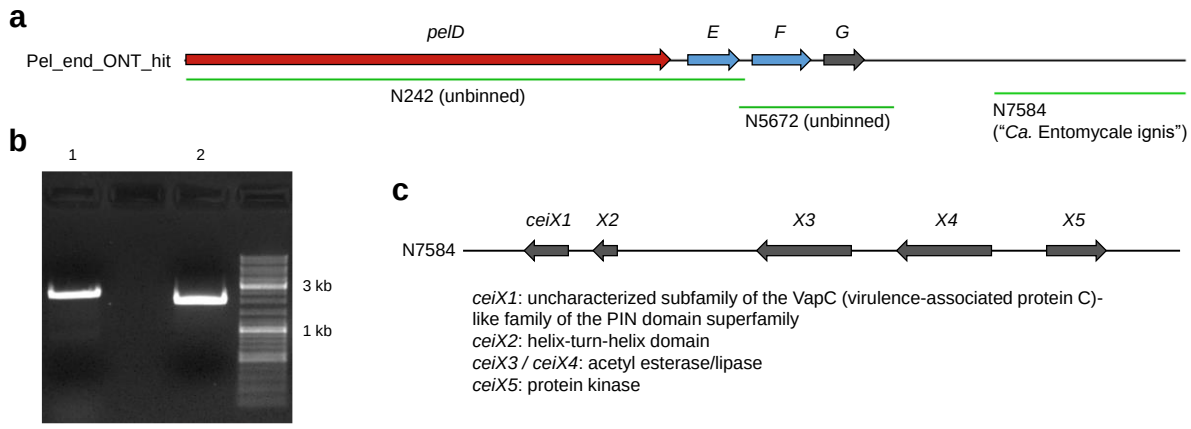


Figure 4. Extension of the *pel* contig by long-read sequencing. a) Mapping of Illumina contigs to the ONT read containing the terminal part of the *pel* cluster. b) PCR verification of the connection between N5672 and N7584 (expected fragment sizes: 1, 2.3 kb; 2, 1.9 kb). c) Genes encoded on the N7584 contig and their predicted functions.

We performed the same mapping strategy for the *mcz* and the *ver* clusters. While the *mcz* cluster was extended to another unbinned contig (**Fig. S4**), we identified connections to binned contigs for both ends of the *ver* cluster (**Fig. S5**). Both contigs mapping to the *ver* cluster were assigned to the same bin (MH2_bin28), which is phylogenetically closest to bacteria of the Lentisphaeria class (phylum Verrucomicrobiota). Interestingly, this bin harbors no other polyketide BGC, but contains a contig that encodes an AT and a β-branching cassette (**Fig. S6**). This could provide the missing AT function in the *ver* pathway and suggests a product with β-branching. The results demonstrate that long-read sequencing is a promising strategy for linking unbinned BGCs to microbial genomes. Future efforts will focus on subjecting the metagenomic DNA to additional rounds of long-read sequencing and indirect sequence capture workflows to map the three pathways to specific genomes.

Discussion

Metagenome mining has revealed an enormous reservoir of orphan BGCs showcasing the untapped potential for expanding our knowledge of chemical space in environmental communities. However, the complexity of metagenomic datasets renders the analysis challenging.[21] The large size and the repetitiveness of multimodular pathways complicate the assembly process, often leading to short fragmented contigs.[15] Sequences consisting of purely biosynthetic genes are difficult to assign to individual members within complex microbiomes,[22] but long-read sequencing can greatly reduce contig number and improve the completeness of MAGs.[23] However, the lower sequencing depth compared to short-read sequencing methods limits the high-quality genomes to the abundant microbiome members. Current target enrichment methods were mainly developed for short-read sequencing technologies and require relatively large amounts of DNA.[24] Therefore, a targeted method to enrich DNA fragments from limited input amounts prior to sequencing is highly desirable. Here, we applied an indirect sequence capture method (Xdrop™)[18] to enrich DNA that contains polyketide BGCs from a sponge metagenome.

The sponge *M. hentscheli* harbors numerous orphan BGCs for which the products are unknown, suggesting a greater chemical diversity than previously expected.[12, 13] Its

microbiome is particularly rich in a distinct family of PKSs, termed *trans*-acyltransferase (*trans*-AT) PKSs, that are often found in uncultivated symbiotic bacteria.[25] The unprecedented number of module variants in *trans*-AT PKSs translates into an astonishing diversity of chemical modifications in their products.[26] We targeted three unbinned *trans*-AT PKS pathways from the *M. hentscheli* metagenome using the indirect sequence capture technology. Due to the pharmacological relevance of peloruside,[4] the assignment of the *pel* pathway to a producer is of great importance, and would not only allow for targeted isolation and cultivation approaches but also facilitate heterologous expression studies. The putative assignment of the peloruside cluster to the mycalamide producer "*Ca.* Entomycale ignis" could explain the missing AT function in the *pel* pathway and suggests a metabolically talented producer. The affiliation of the symbiont with the marine group UBA10353[13] raises intriguing questions about the biosynthetic potential of this uncultivated gammaproteobacterial taxon that is not known as source of other natural products. Since distinct *M. hentscheli* chemotypes exist that differ in the presence or absence of mycalamides, pateamines, and pelorusides,[27] "*Ca.* Entomycale ignis" could provide a model system for studying the co-regulation of polyketide biosynthesis. The preliminary assignment of the orphan *ver* cluster to a bin belonging to the Lentisphaeria class of bacteria adds another PKS-encoding member to the multiproducer consortium in *M. hentscheli*.

Overall, the indirect sequence capture technology provides a promising tool for natural product research. The short detection sequence allows for an enrichment based on individual enzymatic domains in fragmented datasets leading to the identification of complete pathways. Links between BGCs and genomes inform about the lifestyle of producers, thereby facilitating targeted isolation and heterologous expression studies. In addition, they enable experiments to study the evolutionary history of similar pathways and provide insights into the ecological function of the encoded compounds.

Materials and Methods

DNA Isolation

Metagenomic DNA was isolated as described previously.[13] To improve the quality of the metagenomic DNA, samples were re-purified with the Nucleobond HMW DNA kit (Macherey-Nagel) according to the manufacturer's instructions.

Primer Design and Testing

Non-overlapping dPCR and qPCR primer sets were designed with the "Primer design tool" available at the Samplix homepage using 2 kb of both ends of each BGC contig as input sequences. The qPCR primer sets were used to validate the enrichment at the end of the Xdrop™ workflow. All primer sets were tested with a representative HMW DNA sample and the 2× FastStart SYBR Green Master mix (Roche). Efficiency was calculated based on 5-fold dilution series of input DNA (10 ng, 2 ng, 0.4 ng) in duplicate measurements. To evaluate specificity, amplicon melting profiles were generated and evaluated for single peak amplicon melting profiles to ensure absence of primer dimer formation or off-target amplification.

Target Detection by dPCR

Double emulsion droplets were generated in the Xdrop™ instrument by using the dPCR cartridge and the dPCR kit (Samplix). The dPCR cartridge was loaded with the 1× dPCR buffer, dPCR mix including dPCR primers and 1 to 3 ng HMW DNA, and dPCR oil in the respective

order. The cartridge was sealed with the rubber gasket following manufacturer's recommendations. After droplet generation, droplets were collected, aliquoted in four 0.2 mL tubes, and subjected to the following program in a thermocycler with slow ramping rates at all stages (1 °C/s): 30 °C for 5 s, 94 °C for 3 min, 40 cycles of 94 °C for 3 s and 60 °C for 30 s. Droplets were stored at 4 °C overnight.

Fluorescence-Activated Cell Sorting

Droplets were stained with an intercalating fluorescent dye (Samplix) and sorted on a FACSAria III (BD Biosciences) equipped with a 100 µm nozzle. The cell sorter control kit (Samplix), consisting of ready-made dPCR droplets with a large population of positive droplets, was used to preset the sorting gates using a sample pressure set to aim for approximately 4,000 to 6,000 events/s. Positive droplets were gated and sorted out into 15 µL of PCR grade water in a 1.5 ml DNA LoBind tube (Eppendorf). The positively sorted dPCR droplets were kept at 4°C before dMDA amplification on the same day.

Droplet Multiple Displacement Amplification

After sorting, the DNA was released from the positive droplets by adding Break solution and Break color (Samplix) to each sample. After brief centrifugation, the clear bottom phase was carefully pipetted off. The upper phase containing the enriched DNA was used to set up dMDA reactions using the dMDA kit (Samplix) and following manufacturer's recommendations. To monitor contamination, we included as negative controls both aliquots of sheath fluid from the flow cytometer and PCR grade water. As a positive control 1 pg of metagenomic sponge DNA was used. The samples and the dMDA reagents were mixed and injected into the dMDA cartridge followed by overlaying with dMDA oil. The loading of reagents in the dMDA cartridge carefully followed manufacturer's recommendations. The gasket-sealed cartridge was inserted into the Xdrop™ instrument for single emulsion droplet generation. Droplets were collected and transferred into a 0.2 mL tube and incubated in a thermal cycler for 16 h at 30°C and at 65 °C for 10 min to terminate the reaction. After dMDA, the amplified DNA was harvested by adding Break solution and Break color (Samplix) to each sample, followed by pipetting and discarding the clear bottom phase.

Evaluation of Enrichment

The total amount of enriched DNA released from the dMDA droplets was measured with a Quantus™ Fluorometer (Promega Inc.). Fold enrichment of target DNA was assessed by qPCR using the 2× FastStart SYBR Green Master mix (Roche) and the validation primers (qPCR), not overlapping with the detection sequence. The fold enrichment was calculated with the online "Enrichment calculator" available at the Samplix homepage.

Long-Read Sequencing

For long-read sequencing of enriched DNA from the Xdrop™ workflow, 1.5 µg of enriched DNA were debranched with 15 units of T7 endonuclease I (NEB) by incubation at 37 °C for 15 min. The DNA was repaired and end-prepped using the NEBNext FFPE DNA Repair Mix (NEB) and the NEBNext End repair/dA-Tailing Module (NEB) according to the manufacturer's instructions. Adapter ligation was performed using the Ligation Sequencing kit (ONT) with the Long Fragment buffer according to the manufacturer's instructions and Samplix recommendations. Libraries were quantified with a Quantus™ Fluorometer (Promega Inc.). The MinION flowcell (R9.4.1, ONT) was loaded with 5-50 fmol DNA and run using standard settings

for 16 h. Basecalling of native reads was performed using Guppy (v.3.4.5) with high accuracy and quality filtering (ONT).

For long-read sequencing of the original metagenomic DNA sample, 1 µg of HMW DNA was treated as described above without the T7 endonuclease I step.

Acknowledgements

This project has received funding from ETH Research Grant ETH-26 17-1, Swiss National Science Foundation Grants 205321 and 205320, and the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme Grant 742739 and the Helmut Horten Foundation. J.P. is grateful for an Investigator Grant of the Gordon and Betty Moore Foundation. We thank the Flow Cytometry Core Facility at ETH Zurich for droplet sorting.

References

1. W.-Y. Lu, H.-J. Li, Q.-Y. Li, Y.-C. Wu, Application of marine natural products in drug research. *Bioorg. Med. Chem.* **35**, 116058 (2021).
2. S. A. Dyshlovoy, F. Honecker, Marine compounds and cancer: 2017 updates. *Mar. Drugs* **16**, 41 (2018).
3. L. M. West, P. T. Northcote, C. N. Battershill, Peloruside A: a potent cytotoxic macrolide isolated from the New Zealand marine sponge *Mycale* sp. *J. Org. Chem.* **65**, 445-449 (2000).
4. A. Kanakkanthara, P. T. Northcote, J. H. Miller, Peloruside A: a lead non-taxoid-site microtubule-stabilizing agent with potential activity against cancer, neurodegeneration, and autoimmune disease. *Nat. Prod. Rep.* **33**, 549-561 (2016).
5. E. K. Rowinsky, R. C. Donehower, Paclitaxel (Taxol). *N. Engl. J. Med.* **332**, 1004-1014 (1995).
6. T. N. Gaitanos *et al.*, Peloruside A does not bind to the taxoid site on β-tubulin and retains its activity in multidrug-resistant cell lines. *Cancer Res.* **64**, 5063-5067 (2004).
7. S. L. Mooberry, G. Tien, A. H. Hernandez, A. Plubrukarn, B. S. Davidson, Laulimalide and isolaulimalide, new paclitaxel-like microtubule-stabilizing agents. *Cancer Res.* **59**, 653-660 (1999).
8. M. O. Steinmetz, A. E. Prota, Microtubule-targeting agents: strategies to hijack the cytoskeleton. *Trends Cell Biol.* **28**, 776-792 (2018).
9. M. J. Page, S. J. Handley, P. T. Northcote, D. Cairney, R. C. Willan, Successes and pitfalls of the aquaculture of the sponge *Mycale hentscheli*. *Aquaculture* **312**, 52-61 (2011).
10. M. C. Wilson *et al.*, An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58-62 (2014).
11. M. D. Tianero, J. N. Balaich, M. S. Donia, Localized production of defence chemicals by intracellular symbionts of *Haliclona* sponges. *Nat. Microbiol.* **4**, 1149-1159 (2019).
12. M. A. Storey *et al.*, Metagenomic exploration of the marine sponge *Mycale hentscheli* uncovers multiple polyketide-producing bacterial symbionts. *mBio* **11**, e02997-02919 (2020).
13. M. Rust *et al.*, A multiproducer microbiome generates chemical diversity in the marine sponge *Mycale hentscheli*. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9508-9518 (2020).
14. L. Huo *et al.*, Heterologous expression of bacterial natural product biosynthetic pathways. *Nat. Prod. Rep.* **36**, 1412-1436 (2019).
15. D. Meleshko *et al.*, BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Res.* **29**, 1352-1362 (2019).
16. E. L. Moss, D. G. Maghini, A. S. Bhatt, Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* **38**, 701-707 (2020).
17. V. Libis *et al.*, Uncovering the biosynthetic potential of rare metagenomic DNA using co-occurrence network analysis of targeted sequences. *Nat. Commun.* **10**, 3848 (2019).
18. E. B. Madsen, I. Höijer, T. Kvist, A. Ameur, M. J. Mikkelsen, Xdrop: targeted sequencing of long DNA molecules from low input samples using droplet sorting. *Hum. Mutat.* **41**, 1671-1679 (2020).
19. S.-C. Mao *et al.*, Lipophilic 2,5-disubstituted pyrroles from the marine sponge *Mycale* sp. inhibit mitochondrial respiration and HIF-1 activation. *J. Nat. Prod.* **72**, 1927-1936 (2009).

20. S. Warris *et al.*, Correcting palindromes in long reads after whole-genome amplification. *BMC Genomics* **19**, 798 (2018).
21. N. Ziemert, M. Alanjary, T. Weber, The evolution of genome mining in microbes – a review. *Nat. Prod. Rep.* **33**, 988-1005 (2016).
22. I. J. Miller, M. G. Chevrette, J. C. Kwan, Interpreting microbial biosynthesis in the genomic age: biological and practical considerations. *Mar. Drugs* **15**, 165 (2017).
23. S. Goldstein, L. Beka, J. Graf, J. L. Klassen, Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics* **20**, 23 (2019).
24. L. Mamanova *et al.*, Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111-118 (2010).
25. E. J. N. Helfrich, J. Piel, Biosynthesis of polyketides by *trans*-AT polyketide synthases. *Nat. Prod. Rep.* **33**, 231-316 (2016).
26. E. J. N. Helfrich *et al.*, Evolution of combinatorial diversity in *trans*-acyltransferase polyketide synthase assembly lines across bacteria. *Nat. Commun.* **12**, 1422 (2021).
27. M. J. Page, L. West, P. Northcote, C. Battershill, M. Kelly, Spatial and temporal variability of cytotoxic metabolites in populations of the New Zealand sponge *Mycale hentscheli*. *J. Chem. Ecol.* **31**, 1161-1174 (2005).

Supplementary Information

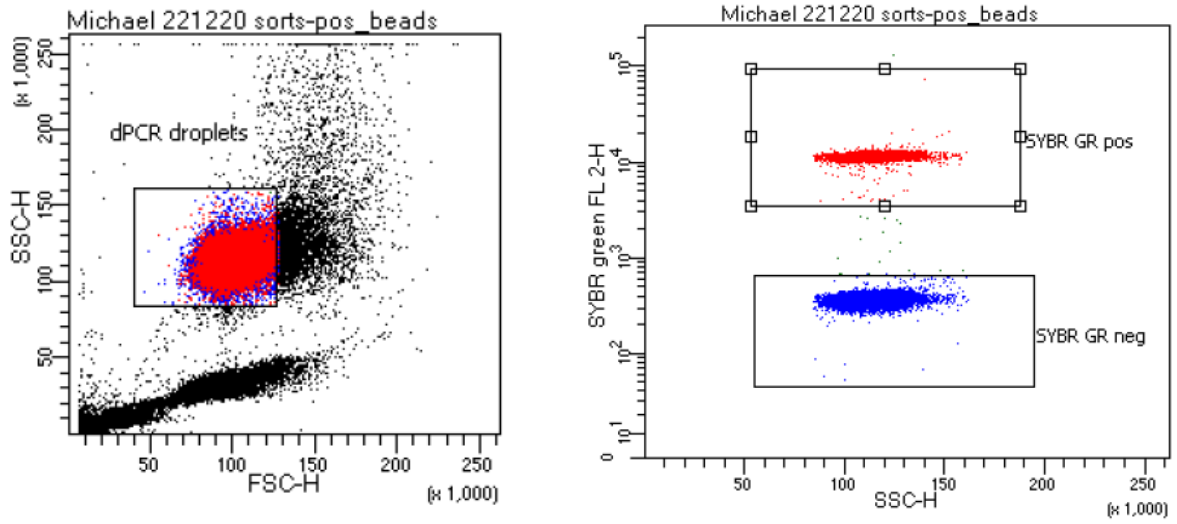


Figure S1. Sorting of dPCR droplets. Plotting of the forward scatter (FSC) against the side scatter (SSC) allows the identification of double emulsion dPCR droplets (left). After gating the dPCR droplets, positive droplets (red) are identified and collected based on the green fluorescent signal versus the SSC.

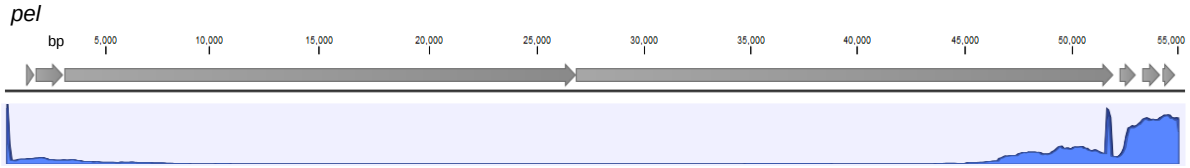


Figure S2. Mapping of the enriched ONT reads to the *pel* cluster. Coverage graph of ONT reads from the enriched HMW DNA sample mapped to the *pel* cluster.

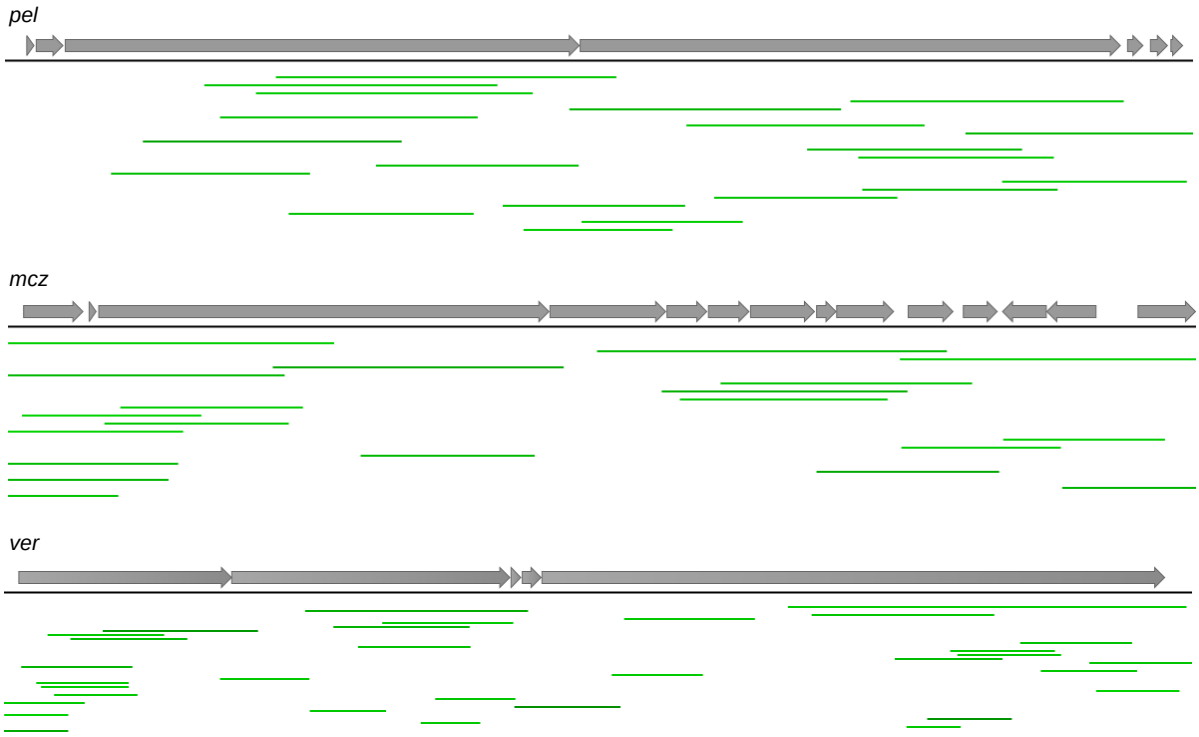


Figure S3. Coverage of the unbinned clusters in the ONT sequencing dataset. ONT reads (green) were mapped against the three unbinned clusters (*pel*, *mcz*, *ver*).

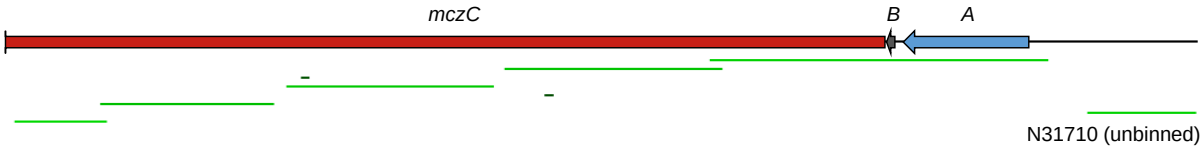


Figure S4. Extension of the *mcz* cluster to an unbinned contig. Illumina reads (green) were mapped to the ONT read containing the start of the *mcz* cluster.

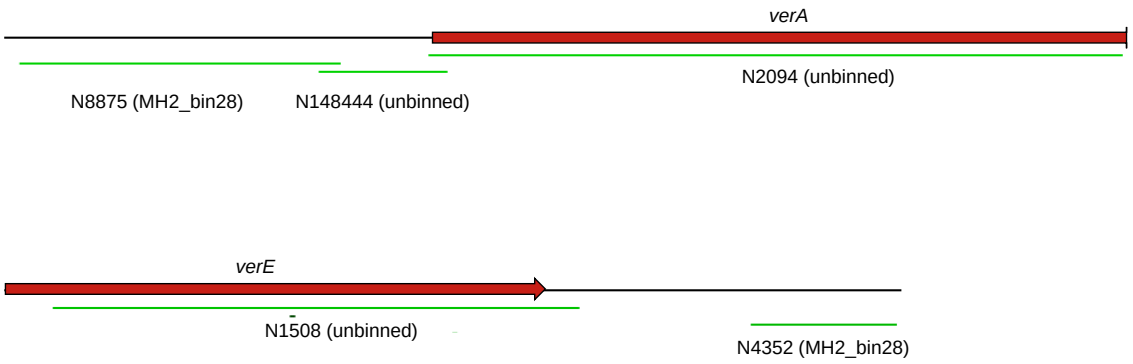


Figure S5. Extension of the *ver* cluster. Illumina reads (green) mapped to ONT reads containing the start (top) and the end (bottom) of the *ver* cluster.

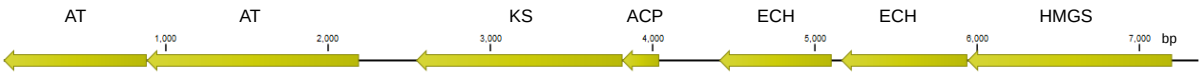


Figure S6. Contig in the Lentisphaeria bin encoding ATs and the β-branching cassette. Gene map of NODE_5485 encoding enzymes for β-branching in MH2_bin28. ACP, acyl carrier protein; AT, acyltransferase; ECH, enoyl-CoA hydratase; HMGS, 3-hydroxy-3-methylglutaryl-CoA synthase homolog; KS, ketosynthase.

Table S1. Detection primers and qPCR verification primers for the *pel*, *mcz*, and *ver* clusters.

Name	**Sequence (5'-3')**
Pel_dPCR1_F	CGTTCAACCGGGTTCTCGAT
Pel_dPCR1_R	TCCGGATGCAGGCACAATAG
Pel_qPCR1_F	TCGATATTGGCGAGGTTGAGG
Pel_qPCR1_R	ATTGTCGCAGGTCGCAGATC
Pel_dPCR2_F	GTCCTGCAAAGCGTTATGGC
Pel_dPCR2_R	GTCGGCTAACCCTTTGCAAA
Pel_qPCR2_F	GCCATCGTTGTCAGACCCA
Pel_qPCR2_R	ATGCACCAAGCTTCTGTCGT
Mcz_dPCR1_F	TCCTTCTCCAGATGGTCGAGT
Mcz_dPCR1_R	GGCGAAAACCTTCCAGGCTA
Mcz_qPCR1_F	TCGATAGCGGTATCCATTCCC
Mcz_qPCR1_R	ACAATTGCATTACGCGACTCG
Mcz_dPCR2_F	CGCTATCAAGAACGCCAGGA
Mcz_dPCR2_R	CCAGGCACCGATCACTAAGG
Mcz_qPCR2_F	CACTAGCCAGGTGATTGCGA
Mcz_qPCR2_R	AGGGCAGTTGTTTCGGGAAT
Ver_dPCR1_F	GGGGAAATCGCATCTCAGGT
Ver_dPCR1_R	CGCTCCTGACAAACGCACTA
Ver_qPCR1_F	GGGCTACATTTGTCCAGGCT
Ver_qPCR1_R	GCTTCAGGGATTGGGCTTCT

Ver_dPCR2_F	AAAATCGGTTGCCCAGAGGA
Ver_dPCR2_R	GCCGTTGCAAATCTCGGTTA
Ver_qPCR2_F	TGAGTTTTTGGCCTCCCGTT
Ver_qPCR2_R	CGATACCTGCACCATCGTCA

Table S2. Primers for the verification of the connection between contigs N5672 and N7584.

Name	**Sequence (5'-3')**
N5672_F1	GCTTTAGTTTGGCCCATGA
N5672_F2	TGGGGTTTTATATTGAGCGT
N7585_R1	CGGTTTACGTCGGTTGTG
N7585_R2	GAAATGACCGAAAAAGAACGC

Chapter V

Cloning *Trans*-AT PKS Gene Clusters from Cultivated and Uncultivated Microbes for a Broad-Host-Range Expression Platform

Michael Rust[1], Andrew King[2], Amy E. Fraley[1], Franziska Hemmerling[1], Tomas Kündig[1], Mariella Greutmann[1], Roy A. Meoded[1], Hannah A. Minas[1], Mathijs F. J. Mabesoone[1], Christopher A. Voigt[2], Jörn Piel[1]

[1] Institute of Microbiology, ETH Zurich, 8093 Zurich, Switzerland.
[2] Synthetic Biology Center, Department of Biological Engineering, Massachusetts Institute of Technology, Boston, USA.

Author Contributions

MR and JP designed research. AK synthesized expression plasmids. RAM generated knockout mutants. MR, AEF, FH, TK, MG, HAM, and MFJM performed cloning and expression experiments. MR performed mass spectrometry experiments. MR and JP wrote the manuscript with contributions from all authors.

Abstract

Polyketide natural products exhibit extraordinary structural diversity and are among the most promising microbial compounds used in medicine. The modular architecture of their biosynthetic machineries has inspired endeavors to generate designer molecules by combinatorial biosynthesis. However, complex biosynthetic gene clusters (BGCs) and the large size of many proteins have complicated biotechnological production and engineering efforts. Here, we applied a recombination-based strategy to clone BGCs of a specific family of multimodular polyketide synthases (PKSs), termed *trans*-acyltransferase PKSs (*trans*-AT PKSs). These systems often originate in poorly studied bacteria, such as uncultivated symbionts, which therefore provide a great source to find potential chemical novelty. We targeted two small *trans*-AT PKSs with known products from cultivated sources that could serve as model systems to understand PKS function and expression. The cloning strategy was expanded to orphan *trans*-AT PKS clusters from a sponge metagenome. Finally, we implemented an inducible expression system in an engineered *Pseudomonas* strain and a *Serratia* knockout mutant that provides the basis for heterologous expression studies of *trans*-AT PKSs.

Introduction

Polyketides represent a large class of natural products with remarkable functional and structural diversity. Various pharmaceutical properties have rendered polyketides important players in the drug discovery process. They are used as antibiotics, antiparasitics, antifungals, immunosuppressants, and chemotherapy drugs.[1] Many bacterial polyketides are biosynthesized by large multimodular assembly lines termed type I polyketide synthases (PKSs).[2] Generally, each module incorporates one building block into the growing polyketide chain before passing it on to the next module. A module harbors a minimum of three essential catalytic domains; an acyl carrier protein (ACP) to which the growing polyketide chain is bound, an acyltransferase (AT) that selects acyl-CoA building blocks and loads them onto the ACPs, and a ketosynthase (KS) that catalyzes chain elongation in a Claisen-like condensation reaction.[3] In addition, modules can harbor reductive domains that act on the processed intermediates and alter the oxidation state of the β-carbon. The genes encoding the biosynthesis of polyketides usually cluster in microbial genomes.[4] These biosynthetic gene clusters (BGCs) often contain additional genes for post-PKS tailoring, regulation, resistance, and transport.[5] The close correlation between the module architecture of a PKS and the succession of chemical moieties in the polyketide backbone allows the prediction of the polyketide core structure from the arrangement of catalytic domains and *vice versa*. This relation is known as colinearity rule and has greatly facilitated the assignment of BGCs to polyketides.[6]

The clustered appearance of genes and the modular architecture of type I PKSs have provided exciting opportunities for the generation of designer polyketides in a Lego-like fashion.[7] However, several PKS characteristics render such approaches highly challenging. Many native producers are difficult to culture or genetically intractable, limiting PKS engineering to heterologous hosts.[8] The large size of the BGCs requires the transfer of numerous genes and long DNA fragments into a host strain,[9] and repetitive regions in PKS-

encoding genes further complicate the cloning process.[10] Even in case of successful cloning, the heterologous production of a desired compound requires concerted expression, proper folding, and correct posttranslational modifications.[8] Progress has mainly been achieved for few well-studied systems of the textbook class of PKSs, notably the erythromycin pathway,[11] and often relied on strains that are phylogenetically close to the native producing organism.[12] Interestingly, there is a second large family of assembly-line PKSs, termed *trans*-AT PKSs, that deviate from the textbook *cis*-AT PKSs in various aspects. The apparent difference from *cis*-AT PKSs is the missing AT domain within each module in *trans*-AT PKSs.[13] Instead, one or more freestanding ATs iteratively load the ACPs with the same substrate. In addition, numerous non-canonical biosynthetic features are known from *trans*-AT PKSs and the diversity of module variants greatly exceeds those found in *cis*-AT PKS systems.[14] In contrast to *cis*-AT PKSs, the various *trans*-acting enzymes prevent a colinearity-based prediction of structures in *trans*-AT systems. However, there is a close correlation between the KS sequence and the chemical moieties introduced by the upstream module in *trans*-AT PKSs.[15] This relation enabled the development of the TransATor tool, which predicts KS specificities and polyketide structures based on input PKS sequences.[16]

Initially overlooked due to their presence in poorly studied microbes, more sophisticated cultivation methods and metagenomic approaches have revealed that *trans*-AT PKSs are widespread in nature.[17] This provides exciting opportunities for the identification of new chemistry, but also renders these systems extremely challenging to study. Native producers are often unknown or taxonomically distant to typical host strains. For example, only two *trans*-AT PKSs, the nocardiosis-associated PKS from several *Nocardia* species and the chejuenolide PKS from *Hahella chejuensis*, have been successfully reconstituted in *Escherichia coli*.[18, 19] In contrast, improved bioinformatic tools and the vast amount of sequencing data have revealed numerous BGCs with no assigned product, so called orphan BGCs.[17] Therefore, the development of a broadly applicable host system for *trans*-AT PKSs would expand the biosynthetic understanding and open avenues for the characterization of pathways from uncultivated sources, potentially leading to novel chemistry and a sustainable production of high-value compounds. Here, we provide a basis for heterologous expression studies by cloning complete *trans*-AT PKS clusters into broad-host-range plasmids and establishing an inducible expression system in different hosts.

Results

Small *Trans*-AT PKSs from Cultivated Sources

We initiated our studies by mining genomes for *trans*-AT PKSs suitable for heterologous expression. Major challenges with these pathways include the large size of their BGCs and the origin in strains that are taxonomically distant from commonly used production hosts. Therefore, we initially targeted small *trans*-AT PKSs with known products from cultivated sources that are compatible with a plasmid-based expression. *Methylobacterium extorquens* AM1 harbors one of the smallest *trans*-AT PKS pathways (25 kb) consisting of only few PKS components and accessory genes. It is responsible for the production of toblerol-type polyketides (**Fig. 1a**).[20] The pathway contains various unusual enzyme components including a FAD-dependent oxygenase and a pyridoxal-phosphate-dependent enzyme. The relatively small size of the largest gene (6.8 kb) and the unidirectionality of all open reading frames in

the BGC rendered this a promising system for heterologous expression studies. However, due to the unusual architecture of the toblerol BGC (*tob*) and the number of generated products (toblerols A-H (**1-8**)), we targeted a second pathway with a more canonical organization and less components.

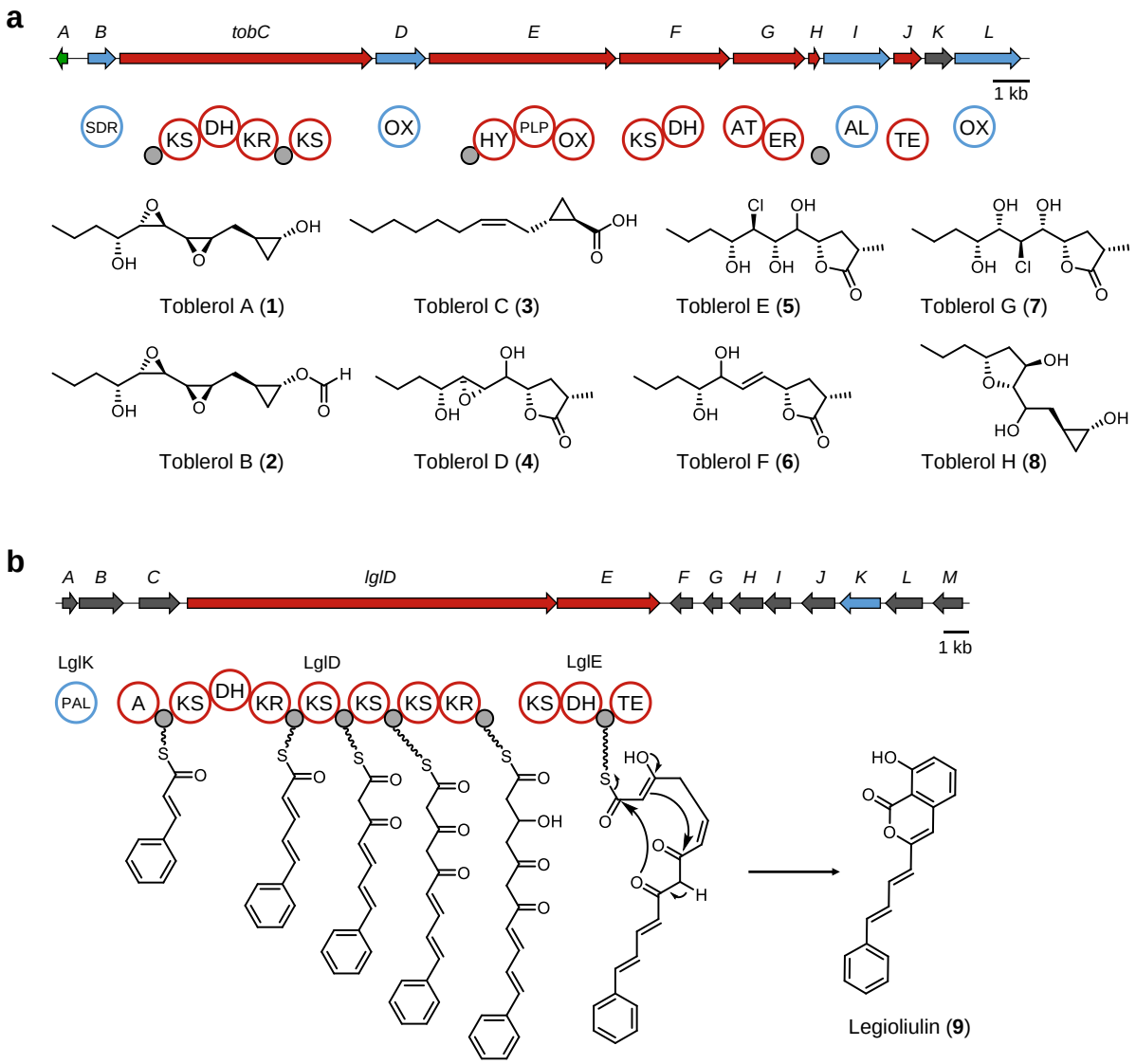


Figure 1. Toblerol and legioliulin biosynthesis. Core PKS genes are shown in red, tailoring biosynthetic genes in blue, regulatory genes in green, and genes with unknown function in gray. Biosynthetic intermediates are shown tethered to the carrier proteins (small gray circles). a) *tob* BGC and structures of toblerols A-H (1-8). b) *lgl* BGC and biosynthetic model for legioliulin (9). A, adenylation domain; AL, acyl-CoA/ACP ligase; AT, acyltransferase; DH, dehydratase; ER, enoylreductase; HY, hydrolase; KR, ketoreductase; KS, ketosynthase; OX, oxidoreductase; PAL, phenylalanine-ammonia lyase; PLP, pyridoxal-phosphate-dependent enzyme; SDR, short chain dehydratase reductase; TE, thioesterase (adapted from Ueoka *et al.*, *Angew. Chem. Int. Ed.* 2018[20] and Ahrendt *et al.*, *ChemBioChem* 2013[21]).

Legioliulin (**9**) is an isocoumarin compound isolated from several *Legionella* species that exhibits blue-white autofluorescence when exposed to long-wave UV light.[22] Transposon mutagenesis revealed the BGC responsible for legioliulin (*lgl*) production in *Legionella parisiensis* (**Fig. 1b**).[21] In contrast to the fragmented toblerol PKS, the pathway contains 13 genes out of which only three genes seem essential for legioliulin production. *LglK* encodes a phenylalanine-ammonia lyase that generates the cinnamic acid starter unit, which is loaded onto the first ACP by the adenylation domain of LglD. Five PKS modules subsequently perform extension cycles before the dehydratase (DH) of LglE initiates *cis*-double bond formation, followed by aldol condensation and water elimination to form the isocoumarin ring.[21] Genome mining revealed a phosphopantetheinyl transferase (PPTase) and two ATs, encoded distantly from the *lgl* cluster in the genome of *L. parisiensis*, expected to perform ACP activation and loading.[21] The autofluorescence of legioliulin has not only aided the identification of *Legionella* species,[23] but would also provide a rapid fluorescence-based readout of PKS productivity.

Capturing the *Tob* and *Lgl* Gene Clusters

Knockout mutants revealed the minimal gene cluster essential for toblerol production (*tobB*-*tobL*) in *M. extorquens* AM1.[20] In a first attempt, we divided the minimal cluster into three fragments and amplified them by PCR. Subsequent Gibson assembly did not yield the desired product and we decided to capture the *tob* cluster by construction of a fosmid library. Genomic DNA was digested with restriction enzymes that cut adjacent to but not within the *tob* cluster. After screening of approximately 300 clones by PCR (**Table S1**), we detected a positive clone from which both ends of the *tob* cluster could be amplified. The resulting plasmid (pCC1-*tob*) was isolated and verified by sequencing.

The same fosmid library approach was used to capture the *lgl* cluster. In addition to restriction enzymes, mechanical shearing was necessary to generate fragments of approximately 40 kb. Screening of 500 clones (**Table S2**) only yielded fosmids that harbored parts of the *lgl* cluster, and we continuously observed rearrangements during fosmid isolation. Therefore, we decided to amplify the two PKS genes *lglD* (14.5 kb) and *lglE* (4 kb) with a high-fidelity polymerase from genomic DNA and clone them into *E. coli*-based expression plasmids by restriction cloning (**Table S3**). Using this strategy, we generated the plasmids pET28b-*lglD* and pCDFDuet-*lglE*, for which we did not observe rearrangements.

Expression in *E. coli* BAP1

Initial expression studies with the pCC1-*tob* fosmid were performed in *E. coli* BAP1, a strain that harbors a chromosomal copy of the *sfp* PPTase gene under control of the T7 RNA polymerase promoter.[11] Analysis of culture extracts by high-performance liquid chromatography–mass spectrometry (HPLC-MS) did not result in the detection of any of the toblerol analogs **1-8**. Therefore, we decided to clone the *tob* cluster from the fosmid into a broad-host-range plasmid by recombineering.

Legioliulin expression tests were performed in *E. coli* BAP1 with pET28b-*lglD* and pCDFDuet-*lglE*. In addition, we cloned the genes encoding the two ATs and the PPTase as well as *lglK* from *L. parisiensis* into a third expression vector (**Table S4**). As previously observed,[21] cinnamic acid was formed in induced samples harboring the phenylalanine-ammonia lyase LglK (**Fig. S1**). However, legioliulin production was not detected by HPLC-MS analysis and we decided to expand the host range by transferring the *lgl* genes into broad-host-range vectors.

Recombineering Generates Broad-Host-Range Expression Plasmids

To expand the host range for the expression studies, we searched for a broad-host-range vector that harbors an inducible promoter and is capable of maintaining large DNA inserts. The RK2-based plasmid pJB861[24] contains the XylS/*Pm* regulator/promoter system, which is activated by low-cost benzoic acid derivatives.[25] We equipped pJB861 with sequences homologous to the terminal regions of the *tob* cluster (**Table S5**) and electroporated the linearized construct into *E. coli* DY380 harboring the pCC1-*tob* fosmid. *E. coli* DY380 contains a defective λ prophage with the recombination genes *exo*, *bet*, and *gam* under control of a temperature-sensitive repressor.[26] Successful recombination yielded pJB861-*tob*, which was verified by restriction digests and end-sequencing.

Due to the observed instability of fosmids containing parts of the *lgl* cluster, we synthesized a modified version of pJB861 by adding the *par* genes (*parA-E*) for increased plasmid stability.[27] Furthermore, we added the superfolder variant of the green fluorescent protein (sfGFP)[28] downstream of the *Pm* promoter to generate the plasmid pAMK159 (**Fig. S2**). We applied the ExoCET cloning method,[29] based on *in vitro* exonuclease assembly and *in vivo* RecET recombination, to clone the PKS genes *lglD* and *lglE* into pAMK159 (**Fig. 2**). The plasmid was amplified using primers with 40 bp homology arms to the start of *lglD* and the end of *lglE*. The PKS genes *lglD* and *lglE* were amplified with two primer sets to generate 9 kb-fragments with an overlap of 80 bp (**Fig. S3** and **Table S6**). Electroporation of the three exonuclease-treated fragments into the induced recombineering strain yielded the desired construct pAMK159-*lglDE*, which was verified by PCR and end-sequencing.

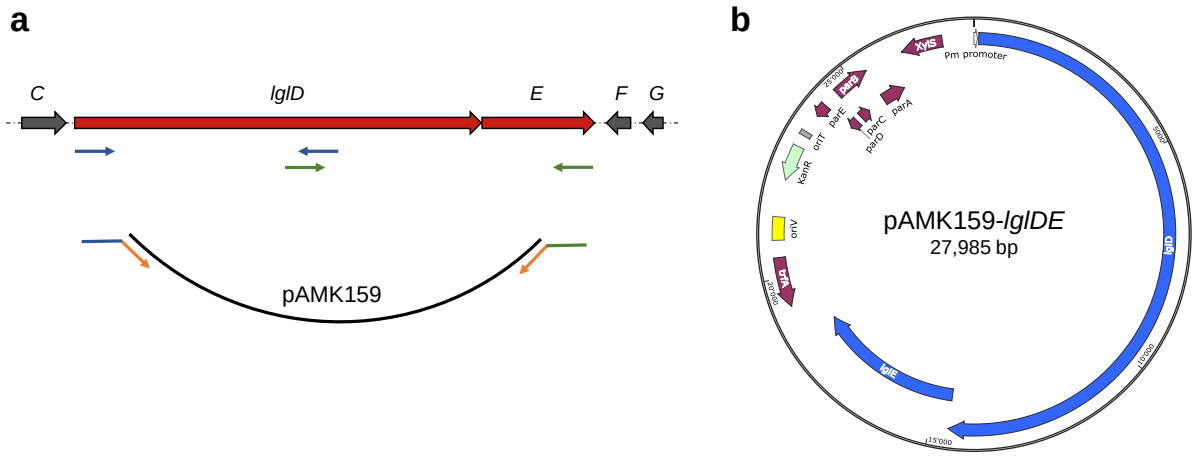


Figure 2. ExoCET cloning of legioliulin PKS genes. a) Binding sites of the two primer sets for the amplification of *lglDE* (blue and green) and the pAMK159 primers (orange) with overlaps to the start of *lglD* and the end of *lglE*. b) Vector map of the recombined plasmid harboring full-length *lglDE* downstream of the inducible *Pm* promoter.

Host Selection

Although *E. coli* has been used for the heterologous production of multiple polyketides and nonribosomal peptides, only two *trans*-AT PKSs have been successfully reconstituted.[18, 19] Therefore, we aimed at testing additional host strains with our model clusters. Members of the genus *Pseudomonas* have received much attention because of their metabolic diversity and tolerance to chemical stresses.[30] Among the pseudomonads, the non-pathogenic soil bacterium *Pseudomonas putida* has become a biotechnological workhorse.[31] Advanced techniques for cultivation and genetic manipulation have led to the production of numerous natural products in *P. putida*.[32] Noteworthy, the predominantly used *P. putida* strain KT2440 harbors a broad substrate range PPTase that is able to activate both ACPs and peptidyl carrier proteins (PCPs).[33, 34] Instead of the wild-type KT2440 strain, we used the engineered *P. putida* EM383 designed for improved functional expression of heterologous genes.[35] Although the clean background of *P. putida* in terms of natural product biosynthesis facilitates the detection of heterologously produced metabolites, the presence of an endogenous pathway in a host can be beneficial in terms of precursor supply, potential enzyme sharing, and correct protein folding . Therefore, we searched for an easily culturable strain that harbors a native *trans*-AT PKS. *Serratia plymuthica* 4Rx13 is a gammaproteobacterium that produces oocydin-type metabolites.[36] These metabolites are biosynthesized by a *trans*-AT PKS (*ooc*) with numerous non-canonical features including an enzyme that catalyzes oxygen insertion into the polyketide backbone.[37] The *ooc* pathway encodes two discrete ATs that could serve heterologous systems.[38] To avoid interference of the oocydin pathway with the heterologous system, we generated a knockout strain in which the PKS gene *oocL* was replaced with a chloramphenicol resistance gene (**Table S7**). This abolishes oocydin production but should leave *trans*-acting enzymes such as the ATs intact.

Functional Expression System in *Pseudomonas* and *Serratia*

To test transformation efficiency and inducible expression, *P. putida* EM383 and *S. plymuthica* 4Rx13Δ*oocL* were transformed with the pAMK159 plasmid harboring the sfGFP gene downstream of the *Pm* promoter. Both strains produced sfGFP upon induction with *m*-toluic acid, proving that the XylS/*Pm* system is functional in both strains (**Fig. 3**). We then attempted to transform the strains with the PKS constructs pJB861-*tob* and pAMK159-*lglDE*. While *Pseudomonas* efficiently took up both constructs, we observed a reduced transformation efficiency for larger constructs (>10 kb) in *Serratia* and were unable to introduce pAMK159-*lglDE*. Experiments to investigate toblerol and legioliulin production in *Pseudomonas* and *Serratia* are underway. A functional production system will guide the expression of orphan BGCs from uncultivated sources such as sponge metagenomes.

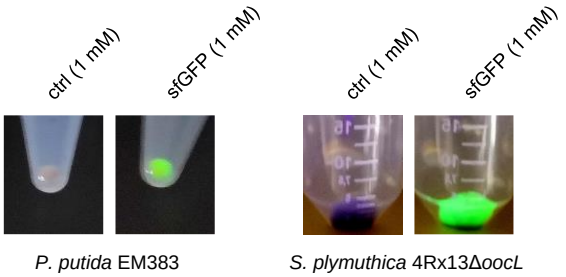


Figure 3. Expression in *Pseudomonas* and *Serratia*. Cultures harboring the pAMK159 plasmid with (sfGFP) and without (ctrl) the superfolder GFP were induced with 1 mM *m*-toluic acid.

Orphan *Trans*-AT PKS Gene Clusters from a Sponge Metagenome

To test whether the ExoCET cloning method is applicable to orphan BGCs from uncultivated sources, we mined the metagenome of the chemically rich sponge *Mycale hentscheli* for suitable *trans*-AT PKSs in terms of size and composition. A small (20 kb) and seemingly complete cluster (*hib*) in the bin of the pateamine producer "*Ca.* Patea custodiens" caught our attention (**Table S8**). The cluster starts with genes encoding an acyl-CoA/ACP ligase, a ketoreductase, and an ACP followed by two PKS genes. Interestingly, the second PKS (HibF) exhibits a module architecture almost identical to the last module of the psymberin PKS from the sponge *Psammocinia* aff. *bulbosa*.[39] Consistent with this observation, PsyD from the psymberin pathway is the closest BLAST hit to HibF. Psymberin is a cytotoxin with an isocoumarin moiety important for its toxicity.[40] We then searched the Mycalidae family for isocoumarin-containing compounds and noticed hiburipyranone (**10**), a cytotoxic compound with a 3,4-dihydroisocoumarin skeleton isolated from the sponge *Mycale adhaerens*.[41] This led us hypothesize that the small *hib* cluster in "*Ca.* Patea custodiens" might encode production of a hiburipyranone-like molecule (**Fig. 4a**). Loading of a C_4 substrate by the acyl-CoA/ACP ligase HibB followed by four PKS extension cycles is in good agreement with the C_{12} backbone of hiburipyranone. The lack of an AT in the *hib* cluster might be complemented by the AT of the pateamine pathway (PamM).[42] Also, there is no apparent halogenase encoded in the *hib* cluster that could introduce the bromine moiety of hiburipyranone. Therefore, we hypothesized that the halogenase is non-canonical, encoded distantly in the genome, or that the cluster encodes production of a non-brominated molecule. To investigate a potential role of the *hib* cluster in hiburipyranone biosynthesis, we aimed at cloning the complete cluster into pAMK159. We split the cluster in four parts of approximately 5 kb and amplified each fragment by PCR. We then used the ExoCET method to assemble all four products into pAMK159. Screening by PCR revealed the correct construct (pAMK159-*hib*) that was verified by Sanger sequencing (**Table S9**). Subsequently, *Pseudomonas* and *Serratia* were transformed with pAMK-*hib*. Due to the lack of an endogenous AT in the *Pseudomonas* strain, we additionally cloned the AT of the pateamine pathway (*pamM*) into two expression plasmids compatible with pAMK159 (**Table S10**). The *pamM* constructs can also be used in *Serratia* in case the native ATs are incapable of loading the *hib* ACPs. Expression tests with the *hib* cluster in *E. coli*, *Pseudomonas* and *Serratia* are underway and will potentially guide the product isolation from sponge extracts.

In the unbinned fraction of the *M. hentscheli* metagenome, we identified a highly aberrant but seemingly complete *trans*-AT PKS (**Fig. 4b**). Apart from canonical domains, the first PKS harbors a fatty acyl adenylate ligase and a module that contains a domain of unknown function (DUF2156) and a pyridoxal phosphate-dependent enzyme belonging to the aspartate aminotransferase superfamily. We hypothesized that this domain might introduce an amine via a PLP-dependent transamination reaction. Searching for natural products that harbor a fatty acyl chain and an amino group isolated from sponges of the Mycalidae family directed our attention to mycalazols and mycalazals. These compounds comprise a group of structurally related 2,5-disubstituted pyrroles that were first isolated from the sponge *Mycale micracanthoxea* and exhibited significant toxicity against several cancer cell lines.[43] Interestingly, the isolation of known as well as novel compounds of this family from a different *Mycale* species[44] suggests a widespread occurrence within the Mycalidae. To study individual enzymes of the putative mycalazol pathway, we subcloned the whole cluster (*mcz*) in eight fragments (**Table S11 and S12**). Experiments to assemble the unidirectional cluster into pAMK159 by ExoCET are currently underway. In addition, we identified a cluster with similar

architecture in a culturable *Enterovibrio* strain. Combined efforts of heterologous pathway expression and HPLC-MS analysis of *M. hentscheli* and *Enterovibrio* are underway and provide promising opportunities for the characterization of the cryptic *mcz* pathway.

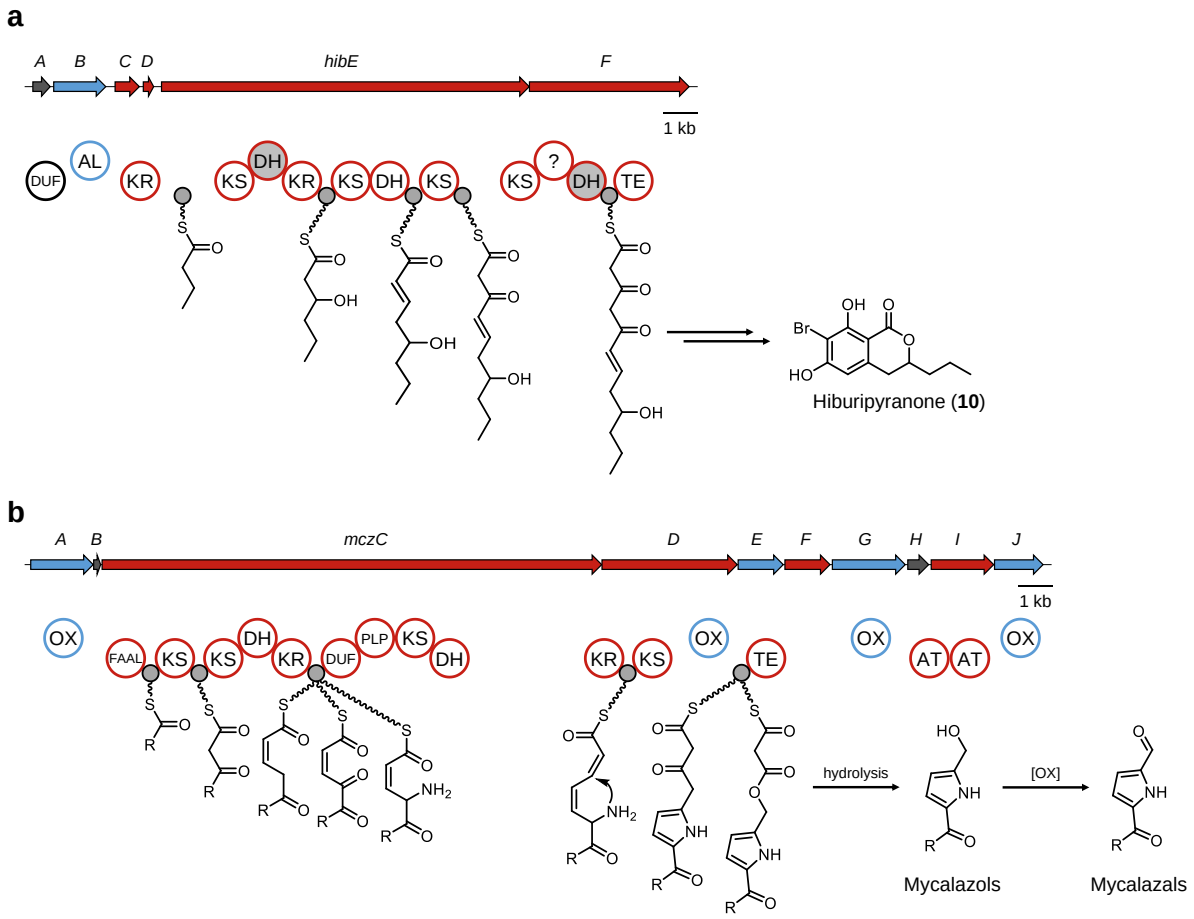


Figure 4. Putative gene clusters and biosynthetic models for hiburipyranone and mycalazols/mycalazals. Core PKS genes are shown in red, tailoring biosynthetic genes in blue, and genes with unknown function in gray. Biosynthetic intermediates are shown tethered to the carrier proteins (small gray circles). **a)** *hib* BGC and putative biosynthetic model for hiburipyranone. DH domains in gray are thought to be inactive. **b)** *mcz* BGC and putative biosynthetic model for mycalazols and mycalazals. DUF, domain of unknown function; FAAL, fatty acyl adenylate ligase.

To test whether the ExoCET method is applicable to larger *trans*-AT PKSs, we targeted the peloruside cluster (*pel*) from *M. hentscheli*.[42] The therapeutic relevance of peloruside[45] and the tentatively assigned identity of the producing organism (Chapter IV) prioritize the cluster for heterologous expression studies. The *pel* cluster (55 kb) consists of only two large PKS genes and few accessory genes. In a first step, we divided each of the two PKS genes *pelC* and *pelD* into two smaller fragments of about 13 kb with the aim to assemble these fragments into pAMK159 by ExoCET. Three out of the four constructs were successfully assembled and experiments to clone the missing part are underway (**Table S13**). Once all four constructs are assembled, we will perform additional rounds of ExoCET cloning to assemble full-length *pelC* and *pelD*. We therefore exchanged the resistance gene in pAMK159 (**Table S14**) to generate

two compatible plasmids for subsequent rounds of ExoCET cloning, with the goal to assemble both PKS genes (*pelDE*) in one expression vector. Overall, our results demonstrate that the ExoCET cloning method can be exploited to clone orphan *trans*-AT PKS clusters from metagenomes, thereby opening avenues for targeted product identification and synthetic biology approaches.

Discussion

Trans-AT PKSs are among the most complex biosynthetic enzymes and produce polyketides of remarkable structural diversity. Compared to textbook *cis*-AT PKSs, these systems harbor a variety of unorthodox biosynthetic components, leading to an astonishing diversity of module variants.[46] *Trans*-AT PKSs remained undiscovered in initial screening efforts because they often originate in less studied microbes. The increasing amount of sequencing data and cultivation-independent approaches have revealed that *trans*-AT PKSs constitute a major family of PKSs distributed over a wide phylogenetic range of bacteria.[17] Interestingly, symbiotic strains often harbor *trans*-AT PKSs, highlighting putative roles in microbe-host interactions.[47] The discrepancy between the number of identified BGCs and characterized *trans*-AT PKSs exemplifies the challenges of studying such pathways. Even for the more canonical *cis*-AT PKSs, expression of the best studied pathway (6-deoxyerythronolide B synthase, generating the precursor of the antibiotic erythromycin) required a highly engineered *E. coli* strain and substantial optimization procedures.[11] The production of the final compound erythromycin required the expression of 17 additional genes.[48]

Here, we provide a basis for heterologous expression studies of several *trans*-AT PKS systems. The toblerol and legioliulin pathways represent suitable model systems due to their relatively small BGCs and characterized products. The autofluorescence of legioliulin could aid the development of an optimized production system by methods such as fluorescence-activated cell sorting, allowing the evaluation of different expression conditions at single-cell resolution, or the selection of improved producers in mutagenesis studies. The generated pAMK159 plasmid can harbor large PKS genes and is compatible with a wide range of heterologous hosts. The sfGFP reporter allows for rapid testing of functional expression in new hosts. The inducible XylS/*Pm* regulator/promoter system is compatible with a broad range of bacteria and does not require expensive induction agents. We demonstrate that the ExoCET cloning strategy is applicable to *trans*-AT PKS BGCs from a sponge metagenome and outperforms previously limiting cloning strategies such as Gibson assembly, restriction cloning or pure RecET-based recombination,[49] particularly for multi-piece assemblies.[50]

Various bioactivities of *trans*-AT PKS-derived polyketides make them exceptional candidates for use in human and veterinary medicine. The relatively few characterized pathways stand in contrast to an impressive number of orphan BGCs, showcasing considerable potential for novel bioactive molecules. Despite improved bioinformatic tools to screen the fast growing sequence space for novel BGCs, *in vivo* or *in vitro* studies remain crucial for a profound understanding of *trans*-AT PKSs.[17] The wealth of biosynthetic components and the presence in taxonomically diverse bacteria complicates the development of a universal host system for *trans*-AT PKSs. However, a heterologous production platform that is applicable to a broad range of pathways from different sources could revolutionize the field in several ways. First, it would expand our limited understanding of how these multienzyme machines function in a cell. Second, it allows the characterization of orphan pathways from uncultivated sources. Third, a biotechnological production system provides

sustainable access to high-value products that are scarce in nature and often difficult to chemically synthesize. Last, these host systems will pave the way towards the generation of hybrid products by combinatorial biosynthesis. Since *trans*-AT PKSs seem to have evolved through extensive exchange and recombination of module series between pathways,[46] these systems hold great promise for rational engineering approaches to produce chimeric polyketides.

Materials and Methods

DNA Isolation

Metagenomic DNA from *M. hentscheli* was isolated as described previously.[42] For *M. extorquens* AM1 and *L. parisiensis*, 40 mL of mid to late exponential phase bacterial cells were resuspended in 4.5 mL of solution 1 (10% sucrose, 50 mM Tris-HCl (pH 8.0), 10 mM EDTA). 500 µL of 30 mg/mL lysozyme dissolved in solution 1 were added and the mixture was incubated on a shaking platform at 37 °C for 1 h. 100 µL of proteinase K (5 mg/mL) and 1.5 mL of SDS (3.3%) were added and the mixture was incubated at 37 °C for 1 h. The sticky lysate was transferred to a MaXtract High Density tube (Qiagen) by decantation and 5 mL of phenol:chloroform:isoamyl alcohol (25:24:1) were added. The mixture was shaken vigorously, vortexed for 30 s and centrifuged at 3,500 g for 5 min. The upper phase was transferred to a new falcon tube and 7 mL of isopropanol were added. The tube was inverted several times until the DNA became visible. The supernatant was carefully removed by pipetting and 10 mL of freshly prepared 70% ethanol were added. The supernatant was completely removed and 3 mL TE buffer containing RNase A (0.1 mg/mL) were added to the DNA pellet. The tube was incubated at 4 °C overnight to allow the genomic DNA to dissolve completely. The DNA was repurified by ethanol precipitation and resuspended in 500 µL TE buffer (0.2×) by incubation at 37 °C for 2 h.

Amplification of DNA

PCRs were performed with Q5 High-Fidelity DNA Polymerase (NEB). A typical PCR (25 µL) contained 1× Q5 reaction buffer, 200 µM deoxyribose nucleoside triphosphates (dNTPs), 0.5 µM of each primer, 20 to 50 ng template DNA, and 0.5 U Q5 High-Fidelity DNA Polymerase. For the amplification of long fragments and vector backbones, 5% DMSO or 1 M betaine were included in the PCR. The reaction was heated to 98 °C for 30 s followed by 30 cycles of 98 °C for 10 s, 62 °C for 20 s, and 72 °C for 30 s per kilobase DNA target sequence. At the end, a final incubation at 72 °C for 2 min was performed. For Gibson assembly fragments, a two-step PCR was performed with an annealing temperature for the 3’ binding part of the primers for the first 5 cycles, followed by 30 cycles with an annealing temperature of the full-length primers.

Gibson Assembly

Approximately 100 ng of PCR-amplified vector DNA was mixed with three times molar excess of insert DNA fragments and 10 µL Gibson assembly master mix (NEB) in a total volume of 20 µL. The reaction mixture was incubated at 50 °C for 1 h. Chemically competent *E. coli* DH5α cells were transformed with 5 µL of the reaction mixture.

Preparation of Fosmid Libraries
Fosmid libraries were constructed as described previously.[51] Genomic DNA was digested with restriction enzymes that cut adjacent to but not within the BGC of interest (*Nhe*I and *Sfi*I for *M. extorquens* AM1; *Avr*II, *Asc*I, *Nhe*I, *Not*I, and *Sfi*I for the *L. parisiensis*).

Restriction Cloning
PCR fragments were digested with restriction endonucleases (NEB) according to the manufacturer's instructions. Digested products were cleaned up and ligated using T4 DNA ligase (NEB) according to the manufacturer's instructions.

Expression in *E. coli* BAP1
Expression cultures were inoculated from overnight cultures in a 1:100 (vol/vol) dilution in Terrific Broth (TB) medium. Cultures were grown (37 °C, 200 rpm) to an OD_{600} of 0.6 to 1.0 and cooled on ice for 30 min. Gene expression was induced by adding 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG), and the cultures were incubated (16 °C, 200 rpm) overnight.

HPLC-MS Analysis
Cultures were extracted with ethyl acetate and the organic phase was dried under reduced pressure. Extracts were dissolved in acetonitrile:water (1:1) and centrifuged at 20,000× g for 10 min. The supernatant was analyzed by HPLC-MS on a Phenomenex Kinetex 2.6 µm XB-C18 100 Å (150 × 2.1 mm) column. The column was heated to 27 °C and the solvents used were water with 0.1% (v/v) formic acid (solvent A) and acetonitrile with 0.1% (v/v) formic acid (solvent B). A flow rate of 0.5 mL/min with solvent B at 5% from 0 to 2 min, 5% to 98% from 2 to 12 min, 98% from 12 to 15 min, 98% to 5% from 15 to 17 min, and 5% from 17 to 19 min was used. ESI-MS was performed in positive ion mode, with a spray voltage of 3,500 V, a capillary temperature of 280 °C, probe heater temperature of 475 °C and an S-Lens RF level of 50. Full MS was performed at a resolution of 140,000 (AGC target 1e6, maximum IT 150 ms, range 100 to 1,000 *m/z*).

λ/Red Recombination
Recombineering experiments were performed using *E. coli* DY380. A colony harboring the target fosmid was inoculated in 5 mL LB medium containing the respective antibiotics at 30 °C overnight. The culture was diluted 1:50 in 50 mL SOB medium containing the appropriate antibiotics and incubated at 30 °C to an OD_{600} of 0.6. The flask was transferred to a shaking waterbath at 42 °C for 15 min. The culture was cooled on ice and transferred into precooled 50 mL falcon tubes. The cells were centrifuged at 4 °C and 5000× g for 5 min. The cells were washed 3 times with ice-cold 10% glycerol and resuspended in 200 µL ice-cold 10% glycerol. 70 µL of cells were mixed with 100 ng of linearized DNA and the mixture was transferred to a pre-cooled 1 mm cuvette and electroporated at 1,800 V. The cells were recovered in 1 mL SOC medium at 30 °C for 1 h and plated on LB agar plates containing the respective antibiotics.

ExoCET Cloning
ExoCET assemblies were performed as described previously[29] with slight modifications. 100 ng of each PCR-amplified fragment was mixed with 1× NEBuffer 2.1 and 0.2 µL of T4 DNA polymerase (NEB) in a total reaction volume of 20 µL and incubated at 25 °C for 1 h, 75 °C for 20 min, and 50 °C for 30 min in a thermocycler. The reaction mixture was dialyzed against water on Millipore Membrane Filters (Merck-Millipore) at room temperature for 30 min. 280 µL

of an overnight culture (*E. coli* GBdir-gyrA462 harboring the pSC101-BAD-ETgA-tet plasmid[49]) were added to 10 mL LB medium and incubated at 30 °C and 200 rpm for 2 h. The culture was induced with 250 µL of 10% (w/v) L-arabinose and incubated at 37 °C and 200 rpm for 40 min. The culture was cooled on ice and the cells were washed two times with ice-cold water. Cells were resuspended in 20 µL of water and 5 µL of the desalted exonuclease reaction were added. The mixture was transferred to a precooled 1 mm cuvette and electroporated at 1.35 kV.

Generation of *S. plymuthica* 4Rx13Δ*oocL*

The knockout strain was generated as described previously[52] with the primers listed in Table S7.

Transformation of *P. putida* EM383

P. putida EM383 was transformed as described previously.[53] Cells were grown in LB to an OD_{600} of 0.6 to 0.8 and washed two times with 300 mM sucrose. Cell pellets were resuspended in 300 mM sucrose and 2 µL of plasmid were added. The mixture was transferred to a 2 mm cuvette and electroporated at 2.5 kV.

Transformation of *S. plymuthica*

S. plymuthica was transformed as described previously.[52] Cells were grown in LB to an OD_{600} of 0.6 to 0.8, cooled on ice and washed five times with ice-cold 10% glycerol. Cell pellets were resuspended in 10% glycerol and 2 µL of plasmid were added. The mixture was transferred to a precooled 2 mm cuvette and electroporated at 2.5 kV.

Gene Expression in *Pseudomonas* and *Serratia*

Cultures were grown in 20 mL LB medium at 30 °C and 200 rpm to an OD_{600} of 0.6 to 0.8 in 100 mL flasks. Gene expression was induced by adding 1 mM of *m*-toluic acid (in ethanol). Cultures were incubated at 30 °C and 200 rpm for 16 h. Cultures were centrifuged and expression of sfGFP was visualized under UV light.

Acknowledgements

This project has received funding from ETH Research Grant ETH-26 17-1, Swiss National Science Foundation Grants 205321 and 205320, and the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme Grant 742739 and the Helmut Horten Foundation. J.P. is grateful for an Investigator Grant of the Gordon and Betty Moore Foundation.

References

1. T. J. Buchholz, J. D. Kittendorf, D. H. Sherman, Polyketide biosynthesis: modular polyketide synthases. *Wiley Encycl. Chem. Biol.*, 1-14 (2008).
2. J. Staunton, K. J. Weissman, Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.* **18**, 380-416 (2001).
3. C. Hertweck, The biosynthetic logic of polyketide diversity. *Angew. Chem. Int. Ed. Engl.* **48**, 4688-4716 (2009).
4. J. J. J. van der Hooft *et al.*, Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.* **49**, 3297-3314 (2020).
5. M. A. Fischbach, C. T. Walsh, Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.* **106**, 3468-3496 (2006).
6. J. Piel, Biosynthesis of polyketides by *trans*-AT polyketide synthases. *Nat. Prod. Rep.* **27**, 996-1047 (2010).

7. D. H. Sherman, The Lego-ization of polyketide biosynthesis. *Nat. Biotechnol.* **23**, 1083-1084 (2005).
8. K. P. Yuet, C. Khosla, Challenges and opportunities for engineering assembly-line polyketide biosynthesis in *Escherichia coli*. *Metab. Eng. Commun.* **10**, e00106 (2020).
9. J. H. Kim *et al.*, Cloning large natural product gene clusters from the environment: piecing environmental DNA gene clusters back together with TAR. *Biopolymers* **93**, 833-844 (2010).
10. C. Song *et al.*, RedEx: a method for seamless DNA insertion and deletion in large multimodular polyketide synthase gene clusters. *Nucleic Acids Res.* **48**, e130 (2020).
11. B. A. Pfeifer, S. J. Admiraal, H. Gramajo, D. E. Cane, C. Khosla, Biosynthesis of complex polyketides in a metabolically engineered strain of *E. coli*. *Science* **291**, 1790-1792 (2001).
12. K. J. Weissman, Genetic engineering of modular PKSs: from combinatorial biosynthesis to synthetic biology. *Nat. Prod. Rep.* **33**, 203-230 (2016).
13. J. Piel, A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of *Paederus* beetles. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 14002-14007 (2002).
14. E. J. N. Helfrich, J. Piel, Biosynthesis of polyketides by *trans*-AT polyketide synthases. *Nat. Prod. Rep.* **33**, 231-316 (2016).
15. T. Nguyen *et al.*, Exploiting the mosaic structure of *trans*-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat. Biotechnol.* **26**, 225-233 (2008).
16. E. J. N. Helfrich *et al.*, Automated structure prediction of *trans*-acyltransferase polyketide synthase products. *Nat. Chem. Biol.* **15**, 813-821 (2019).
17. R. V. O'Brien, R. W. Davis, C. Khosla, M. E. Hillenmeyer, Computational identification and analysis of orphan assembly-line polyketide synthases. *J. Antibiot.* **67**, 89-97 (2014).
18. K. P. Yuet *et al.*, Complete reconstitution and deorphanization of the 3 MDa nocardiosis-associated polyketide synthase. *J. Am. Chem. Soc.* **142**, 5952-5957 (2020).
19. B. G. Ng, J. W. Han, D. W. Lee, G. J. Choi, B. S. Kim, The chejuenolide biosynthetic gene cluster harboring an iterative *trans*-AT PKS system in *Hahella chejuensis* strain MB-1084. *J. Antibiot.* **71**, 495-505 (2018).
20. R. Ueoka, M. Bortfeld-Miller, B. I. Morinaka, J. A. Vorholt, J. Piel, Toblerols: cyclopropanol-containing polyketide modulators of antibiosis in *Methylobacteria*. *Angew. Chem. Int. Ed. Engl.* **57**, 977-981 (2018).
21. T. Ahrendt *et al.*, Biosynthesis of the natural fluorophore legioliulin from *Legionella*. *ChemBioChem* **14**, 1415-1418 (2013).
22. J. Amemura-Maekawa *et al.*, Legioliulin, a new isocoumarin compound responsible for blue-white autofluorescence in *Legionella* (*Fluoribacter*) *dumoffii* under long-wavelength UV light. *Biochem. Biophys. Res. Commun.* **323**, 954-959 (2004).
23. R. M. Vickers, V. L. Yu, Clinical laboratory differentiation of *Legionellaceae* family members with pigment production and fluorescence on media supplemented with aromatic substrates. *J. Clin. Microbiol.* **19**, 583-587 (1984).
24. J. M. Blatny, T. Brautaset, H. C. Winther-Larsen, P. Karunakaran, S. Valla, Improved broad-host-range RK2 vectors useful for high and low regulated gene expression levels in Gram-negative bacteria. *Plasmid* **38**, 35-51 (1997).
25. A. Gawin, S. Valla, T. Brautaset, The XylS/*Pm* regulator/promoter system and its use in fundamental studies of bacterial gene expression, recombinant protein production and metabolic engineering. *Microb. Biotechnol.* **10**, 702-718 (2017).
26. E. C. Lee *et al.*, A highly efficient *Escherichia coli*-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA. *Genomics* **73**, 56-65 (2001).
27. C. L. Easter, H. Schwab, D. R. Helinski, Role of the *parCBA* operon of the broad-host-range plasmid RK2 in stable plasmid maintenance. *J. Bacteriol.* **180**, 6023-6030 (1998).
28. J.-D. Pédelacq, S. Cabantous, T. Tran, T. C. Terwilliger, G. S. Waldo, Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **24**, 79-88 (2006).
29. H. Wang *et al.*, ExoCET: exonuclease *in vitro* assembly combined with RecET recombination for highly efficient direct DNA cloning from complex genomes. *Nucleic Acids Res.* **46**, e28 (2018).
30. J. Kim, W. Park, Oxidative stress response in *Pseudomonas putida*. *Appl. Microbiol. Biotechnol.* **98**, 6933-6946 (2014).
31. P. I. Nikel, V. de Lorenzo, Robustness of *Pseudomonas putida* KT2440 as a host for ethanol biosynthesis. *New Biotechnol.* **31**, 562-571 (2014).

32. A. Loeschcke, S. Thies, *Pseudomonas putida*—a versatile host for the production of natural products. *Appl. Microbiol. Biotechnol.* **99**, 6197-6214 (2015).
33. F. Gross, D. Gottschalk, R. Müller, Posttranslational modification of myxobacterial carrier protein domains in *Pseudomonas* sp. by an intrinsic phosphopantetheinyl transferase. *Appl. Microbiol. Biotechnol.* **68**, 66-74 (2005).
34. Jeremy G. Owen, Janine N. Copp, David F. Ackerley, Rapid and flexible biochemical assays for evaluating 4'-phosphopantetheinyl transferase activity. *Biochem. J.* **436**, 709-717 (2011).
35. E. Martínez-García, P. I. Nikel, T. Aparicio, V. de Lorenzo, *Pseudomonas* 2.0: genetic upgrading of *P. putida* KT2440 as an enhanced host for heterologous gene expression. *Microb. Cell Fact.* **13**, 159 (2014).
36. M. A. Matilla, H. Stöckmann, F. J. Leeper, G. P. C. Salmond, Bacterial biosynthetic gene clusters encoding the anti-cancer haterumalide class of molecules: biogenesis of the broad spectrum antifungal and anti-oomycete compound, oocydin A. *J. Biol. Chem.* **287**, 39125-39138 (2012).
37. R. A. Meoded *et al.*, A polyketide synthase component for oxygen insertion into polyketide backbones. *Angew. Chem. Int. Ed. Engl.* **57**, 11644-11648 (2018).
38. M. A. Matilla, F. J. Leeper, G. P. C. Salmond, Biosynthesis of the antifungal haterumalide, oocydin A, in *Serratia*, and its regulation by quorum sensing, RpoS and Hfq. *Environ. Microbiol.* **17**, 2993-3008 (2015).
39. K. M. Fisch *et al.*, Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting. *Nat. Chem. Biol.* **5**, 494-501 (2009).
40. X. Jiang, N. Williams, J. K. De Brabander, Synthesis of psymberin analogues: probing a functional correlation with the pederin/mycalamide family of natural products. *Org. Lett.* **9**, 227-230 (2007).
41. N. Fusetani, T. Sugawara, S. Matsunaga, H. Hirota, Cytotoxic metabolites of the marine sponge *Mycale adhaerens* Lambe. *J. Org. Chem.* **56**, 4971-4974 (1991).
42. M. Rust *et al.*, A multiproducer microbiome generates chemical diversity in the marine sponge *Mycale hentscheli*. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9508-9518 (2020).
43. M. J. Ortega, E. Zubia, J. Luis Carballo, J. Salvá, New cytotoxic metabolites from the sponge *Mycale micracanthoxea*. *Tetrahedron* **53**, 331-340 (1997).
44. S.-C. Mao *et al.*, Lipophilic 2,5-disubstituted pyrroles from the marine sponge *Mycale* sp. inhibit mitochondrial respiration and HIF-1 activation. *J. Nat. Prod.* **72**, 1927-1936 (2009).
45. A. Kanakkanthara, P. T. Northcote, J. H. Miller, Peloruside A: a lead non-taxoid-site microtubule-stabilizing agent with potential activity against cancer, neurodegeneration, and autoimmune disease. *Nat. Prod. Rep.* **33**, 549-561 (2016).
46. E. J. N. Helfrich *et al.*, Evolution of combinatorial diversity in *trans*-acyltransferase polyketide synthase assembly lines across bacteria. *Nat. Commun.* **12**, 1422 (2021).
47. J. Piel, Metabolites from symbiotic bacteria. *Nat. Prod. Rep.* **26**, 338-362 (2009).
48. H. Zhang, Y. Wang, J. Wu, K. Skalina, B. A. Pfeifer, Complete biosynthesis of erythromycin A and designed analogs using *E. coli* as a heterologous host. *Chem. Biol.* **17**, 1232-1240 (2010).
49. H. Wang *et al.*, RecET direct cloning and Redαβ recombineering of biosynthetic gene clusters, large operons or single genes for heterologous expression. *Nat. Protoc.* **11**, 1175-1190 (2016).
50. C. Song *et al.*, Enhanced heterologous spinosad production from a 79-kb synthetic multioperon assembly. *ACS Synth. Biol.* **8**, 137-147 (2019).
51. C. Gurgui, J. Piel, Metagenomic approaches to identify and isolate bioactive natural products from microbiota of marine sponges. *Methods Mol. Biol.* **668**, 247-264 (2010).
52. D. Domik *et al.*, A terpene synthase is involved in the synthesis of the volatile organic compound sodorifen of *Serratia plymuthica* 4Rx13. *Front. Microbiol.* **7**, 737 (2016).
53. E. Martínez-García, V. de Lorenzo, "Transposon-Based and Plasmid-Based Genetic Tools for Editing Genomes of Gram-Negative Bacteria" in Synthetic Gene Networks: Methods in Molecular Biology (Methods and Protocols), W. Weber, M. Fussenegger, Eds. (Humana Press, Totowa, NJ, 2012), vol. 813, pp. 267-283.

Supplementary Information

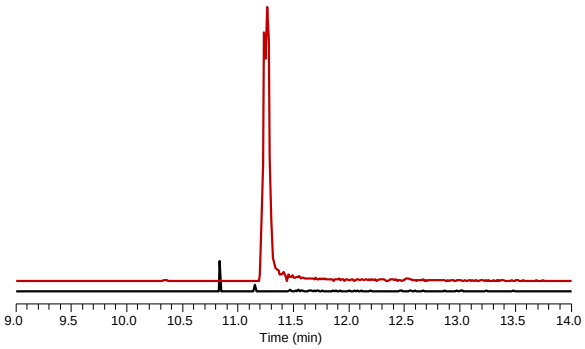


Figure S1. Extracted ion chromatogram of cinnamic acid *m/z* 149.0597 $[M+H]^+$. Extracts of uninduced (black) and induced (red) *E. coli* BAP1 harboring the plasmids pET28b-*lglD*, pCDFDuet-*lglE*, and pACYC-ATs-*lglK*.

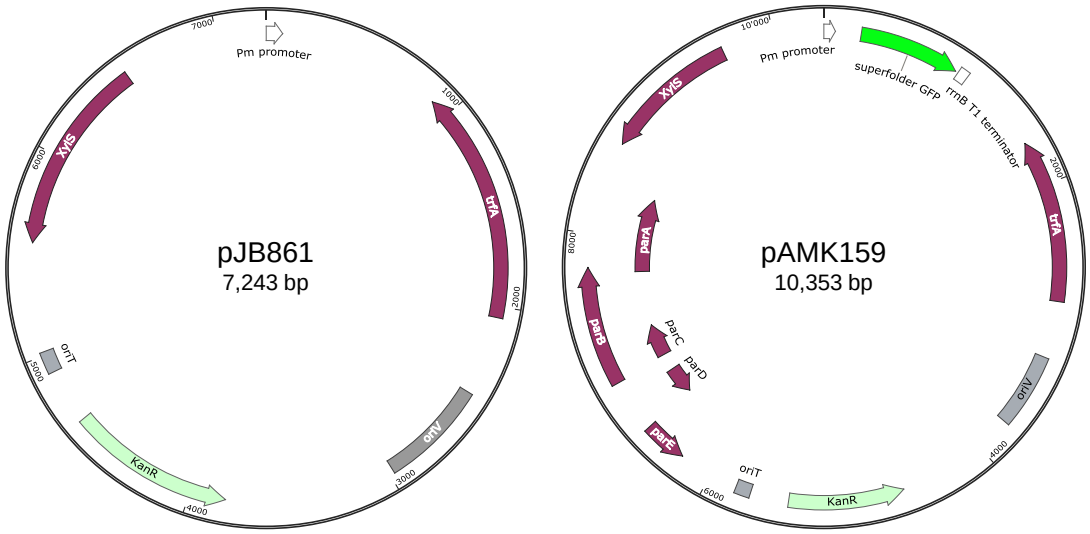


Figure S2. Vector maps of pJB861 and the modified version pAMK159.

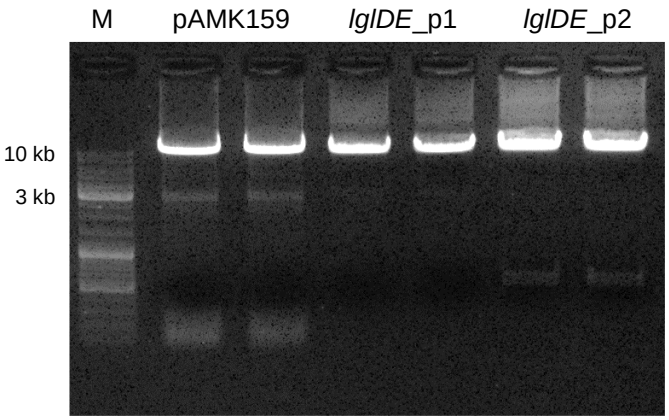


Figure S3. Amplification of pAMK159 and *lglDE* for ExoCET cloning. Preparative agarose gel with expected fragment sizes: pAMK159 (9.6 kb), *lglDE*_p1 (9.5 kb), *lglDE*_p2 (9.1 kb).

Table S1. Primers used to screen for the presence of the *tob* cluster.

Name	**Sequence (5'-3')**
Tob_start_F	CATGGAGATTAAGGACTGTATTGCTTTCGT
Tob_start_R	AACTTCGTTTGGATCTGCTTTTGGGATCTC
Tob_end_F	GGTGATCAACGAGGCTTTGC
Tob_end_R	TCAAGCCGTGTGCGGATCACGCT
TobC_F	AGACGCACACGAAACAAAACTTCT
TobC_R	CAGTCGAAGATAAAAGACCGACGC
TobD_F	AATGCCGGATGATTACCCACACTA
TobD_R	AGCAGACGATCAAAGTGCCATAGA
TobE_F	GGAAGCATACGTCGACAAGAATGG
TobE_R	GAAGGTGCCGTTCTTGATGAAGTC
TobF_F	CAGAAGTAATCAGGGATGCCTGGA
TobF_R	GGCTCGAAATGTAGATCGGAAGGA
TobG_F	GATGGGGTGCGATTTCATCCATTC
TobG_R	ATCTGATTTCCTGAGATGTCGCCC
TobJ_F	TACATGACCATGATGCGTCCGATC
TobJ_R	TCGATAAGGGACAGTGAAACCTCG
TobL_F	GCCTATTCTTCCCGAAGATGCAAC
TobL_R	GGTCCTGTATGACGTCACCTTCTT

Table S2. Primers used to screen for the presence of the *lgl* cluster.

Name	**Sequence (5'-3')**
LglD_F	TTGCTTATCGGCGATGACAATGAC
LglD_R	CGCAATTGTTCCTTTAATTCGGCG
LglE_F	CCAAAACAGATCCTGCCCATGTTT
LglE_R	AATCACACTGATAATGCCTGCAGC
LglK_F	TTTCTTGCGGTCTAGGCACAATTG
LglK_R	ACCACTGGCACCAATAGAACCATA

Table S3. Primers used to amplify *lglD* and *lglE* for restriction cloning.

Name	**Sequence (5'-3')**
NheI_lglD_F	TTAGCTAGCATGCAGAATATTGTTGATGTAG
XhoI_lglD_R	AAGTCTCGAGTCATTCATAAGTCAGCTCT
NotI_lglE_F	TTAGCGGCCGCTATGAATGATACAACGAATTTAA
XhoI_lglE_R	AAGTCTCGAGTTAGATCTCCGATTCTGC

Table S4. Primers used to amplify the acyltransferase and the PPTase genes and *lglK* for Gibson assembly.

Name	**Sequence (5'-3')**
NcoI_AT1_F	TATCCATGGCAATGACTATCTTTATGTTCCC
SD_AT1_R	ATGTATATCTCCTTTTAGATTTGCGTTACTAGG
SD_AT2_F	ATCTAAAAGGAGATATACATATGGTAAAAACAGCATTTG
SacI_AT2_R	TATGAGCTCCTATGCAAATTGCTCTTCAAC
BglII_lglK_F	TTAAGATCTGTTGGAATATCTAAAGTCTCTCAA
XhoI_lglK_R	ATATCTCGAGGAACTGAGCACTAACCTT

pAC_ATs_lglK_F	CCGCATAATGCTTAAGTC
pAC_ATs_lglK_R	TAATGATCATATGTATATCTCCTTCTATGCAAATTGCTCTTCAAC
PPTase_F	AAGGAGATATACATATGATCATTACCGAATTTAACC
PPTase_R	TCGACTTAAGCATTATGCGGCTAAAAAATCAATTGAGTATTTTGC

Table S5. Primers used to amplify the *tob* homology arms and the pJB861 vector backbone for Gibson assembly.

Name	**Sequence (5'-3')**
TobB_F	CATGAACATGGAGATTAAGGACTGTATTGCTTTCG
TobB_R	ATGAGACGGTCACGTGGGCGCAAGTTCCGCTCGA
TobL_F	AACTTGCGCCCACGTGACCGTCTCATCTACCATGACG
TobL_R	CCTGCATCGCTCACGCTCCCAAACGCAC
pJB861_F	GGGAGCGTGAGCGATGCAGGTGGCTGCT
pJB861_R	CCTTAATCTCCATGTTCATGACTCCATTATTATTGTTTC

Table S6. Primers used to amplify *lglDE* and pAMK159 for ExoCET cloning.

Name	**Sequence (5'-3')**
pAMK_Leg_F	GAAACTTCTGCATTTCAGTACTGCAGAATCGGAGATCTAATAATACTAGAGCCAGG CATC
pAMK_Leg_R	TCTGCGCCCTACTTTGTACTACATCAACAATATTCTGCATGTTCATGACTCCATTAT TATTG
Leg1_F	ATGCAGAATATTGTTGATGTAG
Leg1_R	TAAAGCAGTACCTGTACC
Leg2_F	TGATTCGTGCCTATAACC
Leg2_R	TTAGATCTCCGATTCTGC

Table S7. Primers used to generate *S. plymuthica* Δ*oocL*.

Name	**Sequence (5'-3')**
4Rx13_oocL_KO_F	CCTTAAGTTACTGGCGCAGTACCGTGAAGTACTCAATAGAAGGTAG
4Rx13_oocL_KO_R	CCTGTTCACGCAATTTTTTCATCAATAAGTTGGGGGTCGGATGATCG

Table S8. Proteins encoded in the proposed hiburipyranone BGC (*hib*) and their putative functions.

Protein	**Size, aa**	**Proposed function**	**Closest homolog, protein (origin)**	**S/I, %/%**	**GenBank accession number**
HibA	173	DUF697	DUF697 domain-containing protein (*Methylococcaceae* bacterium)	78/52	TSA39754.1
HibB	485	Acyl-CoA/ACP ligase	Acyl-CoA ligase (*Catenovulum* sp. CCB-QB4)	69/50	WP_108601778.1
HibC	236	KR	SDR family oxidoreductase (*Corallococcus* sp. CA053C)	74/56	WP_120609800.1
HibD	88	ACP	acyl carrier protein (*Lysobacter* sp.)	66/44	TXI48895.1
HibE	3,522	PKS	SDR family NAD(P)-dependent oxidoreductase (*Methylomusa anaerophila*)	57/42	WP_158618829.1
HibF	1,531	PKS	hybrid *trans*-AT PKS -NRPS (uncultured bacterium psy1)	64/48	ADA82585.1

Table S9. Primers used to amplify the *hib* cluster and pAMK159 for ExoCET cloning, and primers used to screen for correct constructs.

Name	**Sequence (5'-3')**
pAMK_Hib_F	CGCCGATAAACTCATGATTGATCCTTTGCCAAAAGGTTGATAATACTAGAGCCAGGCA
pAMK_Hib_R	CGCTTTTTTCCACTTCTTCTACTGTTGGCTTTTTAGGCATGTTCATGACTCCATT
Hib1_F	GGAATAAGCAGGATAAGAAGG
Hib1_R	TCACGAATTGAAGGCTGTT
Hib2_F	AATCCCTCTATCCGCTAA
Hib2_R	TCCAAACGAAGAGACTCC
Hib3_F	CGCCGTTTATTGTCCAACA
Hib3_R	CTATTTCTTCGGTTGGCT
Hib4_F	GGCGCCTCGATTAACAGA
Hib4_R	GGATGCTCGCTATACGGA
pAMK_Pm_F	AAGAAGCGGATACAGGAG
pAMK_sc_R	GATATATCATGAAAGGCTGGCT
Hib_start_R	AAAATCGGCATGGTCACT
Hib_end_F	AAGAACTATCAACAGGCAGG

Table S10. Primers used to amplify *pamM* and pSEVA328 and pSEVA438 for Gibson assembly.

Name	**Sequence (5'-3')**
PamM_F	ATAATGGAGTCATGACCATGAAGGAAGAACCCGTTATATG
PamM_F	AGGGTTTTCCCAGTCACGACTTAACAACTCAACTTATCCAG
pSEVA_F	GTCGTGACTGGGAAAACC
pSEVA_R	CATGGTCATGACTCCATTATTATTG

Table S11. Proteins encoded in the proposed mycalazol BGC (*mcz*) and their putative functions.

Protein	Size, aa	Proposed function	Closest homolog, protein (origin)	S/I, %/%	GenBank accession number
MczA	605	unknown	ubiquinone biosynthesis monooxygenase UbiB (*Anabaena* sp. UHCC 0451)	62/42	ATX68118.1
MczB	78	unknown	hypothetical protein (*Chondromyces apiculatus*)	61/42	WP_044250997.1
MczC	4789	PKS	SDR family NAD(P)-dependent oxidoreductase (*Fulvivirga* sp. 29W222)	57/39	WP_202859121.1
MczD	1289	PKS	SDR family NAD(P)-dependent oxidoreductase (*Enterovibrio* sp. CAIM 600)	64/52	WP_129123768.1
MczE	436	Baeyer-Villiger monooxygenase	NAD(P)-binding domain-containing protein (*Enterovibrio* sp. CAIM 600)	79/66	WP_129123767.1
MczF	439	PKS	alpha/beta fold hydrolase (*Enterovibrio* sp. CAIM 600)	62/46	WP_129123766.1
MczG	690	oxidoreductase	FAD-dependent oxidoreductase (*Hymenobacter actinosclerus*)	58/42	WP_092767548.1
MczH	201	unknown	DUF1295 domain-containing protein (Proteobacteria bacterium)	61/42	NIQ37354.1
MczI	595	AT	ACP S-malonyltransferase (Gammaproteobacteria bacterium)	54/36	NOX92126.1
MczJ	477	dehydrogenase	aldehyde dehydrogenase family protein (*Acaryochloris marina*)	90/77	WP_202947647.1

Table S12. Primers used for the subcloning of the *mcz* cluster.

Name	Sequence (5'-3')
Mcz1_F	GAACGACACAAACAACAAGA
Mcz1_R	AAGACGTTGTGAGAAATGG

Mcz2_F	CGGACGCAAAGAACTGAA
Mcz2_R	GGAAAGGGAGAGGCTAGT
Mcz3_F	AGAAGGAGGGATGGACAG
Mcz3_R	AAAGGTTGCGAGCGGATA
Mcz4_F	CATATCCGCTCGCAACCT
Mcz4_R	AGAGGACGCGATCAGGAA
Mcz5_F	GGTTTGGAGCTAGTCGTT
Mcz5_R	ACTGGGTATTGGTTTTGG
Mcz6_F	AAACCAGATCTCAACCAC
Mcz6_R	CCTTCAGCATTGACAAAC
Mcz7_F	CAGTACATCCCGTGCGAA
Mcz7_R	GTTGCGTGAATTGGGTCT
Mcz8_F	TCTCTCGCTGTTCACTCT
Mcz8_R	CAATGCCTACCTGACGAT

Table S13. Primers used to amplify the *pel* cluster and pAMK159 for ExoCET cloning.

Name	**Sequence (5'-3')**
pAMK_pelC_p1_F	AGTTGAAGCTGGACTATGCGGGGACAGGTCGGTTGCTGATTAATACTAGAGCCAGGCA
pAMK_pelC_p1_R	TAGCGATTTCGTTCTCCCCCCGACAACAATGATCCTCACTCATGTTCATGACTCCATT
Pel_pelC1.1_F	AGTGAGGATCATTGTTGTCG
Pel_pelC_p1.1_R	TCAAATCGATTGCTCACC
Pel_pelC_p1.2_F	AAAGAATCGTACCCGACG
Pel_pelC_p1.2_R	ATCAGCAACCGACCTGTCCC
pAMK_pelC_p2_F	ATGCTGAACTGGCTGTTCAGGAGGCTTGCTCACAATGATAATACTAGAGCCAGGCATC
pAMK_pelC_p2_R	TATGAAACCCGATTCGCCACGCTTGCGGGCGTCATCATGTTCATGACTCCATTATTATT

Pel_pelC2.1_F	ATGACGCCCGCAAGCGTG
Pel_pelC_p2.1_R	GCCGGCGCCAAACGAAGAT
Pel_pelC_p2.2_F	GGAAAAGACCTACGCTTG
Pel_pelC_p2.2_R	TCATTGTGAGCAAGCCTC
pAMK_pelD_p1_F	CGATGAGAGGCCCGAGAAATCCAGAATAGAAGGCAGCTAATACTAGAGCCAGGCATCA
pAMK_pelD_p1_R	AGATCGACGACACTCTGCTTTATACCGGTCGCATTCATGTTCATGACTCCATTATTAT
Pel_pelD1.1_F	ATGAATGCGACCGGTATA
Pel_pelD_p1.1_R	TCCGATATGCGATTTGGC
Pel_pelD1.2_F	ATGCATTAGTGCGGTCGT
Pel_pelD1.2_R	GCTGCCTTCTATTCTGGATT
pAMK_pelD_p2_F	TGCAGAAAAAACATGGTTGCATGGAGTTGAAAACATGATAATACTAGAGCCAGGCATCAA
pAMK_pelD_p2_R	TTTTTTCCTGCAATGCCTTTCGCAACCGCTCATCATTCATGTTCATGACTCCATTATTATTG
Pel_pelD2.1_F	AATGATGAGCGGTTGCGA
Pel_pelD_p2.1_R	CACCGTGTAGCCATTCGT
Pel_pelD2.2_F	TAGAGGATGGAGATCACA
Pel_pelD2.2_R	TCATGTTTTCAACTCCATGC

Table S14. Primers used to exchange the resistance gene of pAMK159.

Name	**Sequence (5'-3')**
cam_F	AACCAATTAACCAATTCTGATTACGCCCCGCCCTGCCACTCATCGCAGTA
cam_R	ACAGTAATACAAGGGGTGTTCAATTCGCCCTATAGTGAGTCGTATTACGCGCGCG
pAMK_F	ACTCACTATAGGGCGAATTGAACACCCCTTGTATTACTGTTTATGTAAGC
pAMK_R	AGTGGCAGGGCGGGGCGTAATCAGAATTGGTTAATTGGTTGTAACACTGGC

Chapter VI

Conclusions and Outlook

Specialized metabolites are substances produced by living organisms that are not inherently involved in growth, development, and reproduction but fulfil a variety of biological functions.[1, 2] Although these metabolites might not be essential for the short-term survival of the producing organism, they play key roles in the adaptation to changing environments and the colonization of ecological niches.[3] The various biological functions are reflected in the structural diversity of natural products. Based on their building blocks or chemical features, they are grouped into classes such as alkaloids, terpenes, polyketides, and peptides. The bioactivities of natural products outside their ecological context have rendered them important players in the drug discovery process.[4] While traditional medicines were mainly based on plant extracts,[5] bioactive compounds from microorganisms ushered in the "golden age" of antibiotic discovery.[6] The discovery of the first antibiotic penicillin[7] spurred a global search of microbe-derived molecules that resulted in the isolation of numerous lead structures for the development of clinically relevant therapeutics.[8] Natural products and molecules inspired by them constitute a considerable portion of all drugs approved for use in humans, and have proven effective as e.g. antibiotics, anticancer drugs, immunosuppressants, and cholesterol-lowering agents.[8] The improved bioactivity of natural products compared to compounds from synthetic libraries likely originates from the evolutionary pressure that shaped their interaction with macromolecular targets within biological communities.[9]

Initially, many compounds were isolated from soil bacteria,[3] among which actinomycetes have proven to be the richest source of specialized metabolites.[10] However, traditional bioactivity-guided isolation of natural products has been hampered by the rediscovery of known structures from common sources.[11] The advent of genome sequencing revealed that genes responsible for the biosynthesis of natural products cluster in microbial genomes,[12] and that many strains harbor biosynthetic gene clusters (BGCs) far exceeding the number of isolated products.[13] The "One Strain–Many Compounds" principle was employed as a useful strategy to isolate new metabolites by changing media components and cultivation conditions, or performing co-cultivation with other strains.[14] An alternative approach are high-throughput elicitor screens (HiTES), in which small molecules are used to activate silent BGCs.[15] The increasing amount of sequencing data and more sophisticated methods to mine genomes for BGCs have shifted the focus from soil bacteria to the global diversity of microorganisms including the numerous uncultivated strains.

The marine environment harbors a tremendous diversity of largely unexplored flora and fauna that offers seemingly infinite opportunities to discover chemical novelty.[16] With more than 5,000 isolated compounds, marine sponges are a particularly rich source of natural products contributing to about 30% of all marine natural products.[17] Among the oldest multicellular invertebrates,[18] the sessile filter-feeding animals have developed strategies to defend themselves against overgrowth, invasion, and predators. Many sponges live in symbiosis with astonishing numbers and diversities of microorganisms.[19] Numerous sponge-derived metabolites harbor chemical features commonly observed in bacterial natural products, leading to the hypothesis that these molecules might be produced by symbiotic microorganisms.[20] Striking examples are the *Haliclona* sp. that intracellularly harbor the renieramycin-producing symbiont "*Ca.* Endohaliclona renieramycinifaciens"[21] and *Theonella swinhoei* containing the "superproducer" symbionts "*Ca.* Entotheonella".[22] "Entotheonella" symbionts belong to a new candidate phylum and produce almost all bioactive peptides and polyketides known from the chemically rich sponge host.[22-25]

In Chapter II, we described a contrasting scenario for the sponge *Mycale hentscheli*, in which multiple phylogenetically diverse members of the microbiome contribute to the rich chemistry of their host.[26] The data provided a possible explanation for the different sponge chemotypes and raised important questions regarding the acquisition and regulation of symbionts. The assignment of biosynthetic genes to the three known polyketides mycalamide, pateamine, and peloruside was achieved by retrobiosynthetic dissection and *in silico* analysis of individual enzyme domains. The development of bioinformatic tools for the automated analysis of sequencing data greatly facilitated the identification of BGCs. antiSMASH is the most widely used detection tool[27] and identified numerous additional BGCs from diverse biosynthetic families in the metagenome of *M. hentscheli*. The architecture of modular pathways such as nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs) allows the prediction of core structures from protein sequences and *vice versa*. The *M. hentscheli* metagenome is particularly rich in a distinct family of PKSs, termed *trans*-acyltransferase (*trans*-AT) PKSs. The plethora of noncanonical features in these systems impedes structural predictions based on the colinearity principle[28] and requires specialized tools for BGC analysis. The recently developed TransATor uses the correlation between phylogenetically similar ketosynthase (KS) domains and chemical moieties introduced by the upstream module to predict polyketide structures.[29] Using this software in combination with manual analysis of BGCs enabled the construction of biosynthetic models for the various orphan pathways without known products in *M. hentscheli*. Despite their value in prioritizing BGCs and guiding the targeted isolation of novel metabolites, *in silico* predictions are usually complemented with the functional characterization of biosynthetic pathways.

Identifying and linking BGCs to natural products in complex communities can be challenging.[30] An alternative approach applied in Chapter III is the search of similar pathways in cultivated strains. Analysis of the small NRPS/PKS cluster in the *M. hentscheli* symbiont "*Ca.* Caria hoplita" revealed homologous clusters in bacteria of the *Rhodobacteraceae* family. The characterization of the BGC and the product in cultivated strains will guide the isolation of similar compounds from sponge extracts and provide a basis for studying the pathway evolution in different bacteria. The presence in free-living as well as symbiotic strains raises intriguing questions about the ecological function of these molecules. Another striking example is the large group of pederin-type compounds, which are distributed across a broad phylogenetic range of free-living and symbiotic bacteria (discussed in Chapter II).[26] However, homologous pathways in cultivated strains are rare and many BGCs in complex metagenomes are either incomplete or not assigned to individual genomes,[31] rendering their characterization more ambitious.

The assembly of metagenome-derived contigs from short-read data is complicated by the complexity of many datasets.[32] Specific assemblers developed for the reconstruction of metagenome-assembled genomes such as metaSPAdes[33] or MEGAHIT[34] have significantly improved the quality of metagenomic assemblies. However, the repetitive nature of NRPSs and PKSs is reflected in their nucleotide sequence and impedes the reliable assembly of biosynthetic pathways from short-read data.[35] The combination of short-read sequencing with single-molecule long-read technologies have greatly improved the overall quality of contigs and facilitated the assembly of complete biosynthetic pathways.[35] To reconstruct genomes of individual community members, contigs are grouped in a process called binning. Algorithms such as MetaBAT[36] group sequences according to genomic features such as

oligonucleotide frequency or contig coverage. More recent developments use the information of the assembly graph to refine existing binning results.[37] Despite the rapidly developing field of bioinformatic tools for assembly and binning, purely biosynthetic contigs often remain unbinned. The evolution of *trans*-AT PKSs by horizontal gene transfer and their presence in low-abundance organisms render the BGC assignment to producer genomes particularly challenging. In Chapter IV, we employed a novel indirect sequence capture method to enrich fragments harboring the terminal parts of unbinned BGCs prior to long-read sequencing. Once established, this workflow will resolve a broad range of metagenome-derived limitations in natural product research. The short detection sequence allows for the enrichment based on individual enzymatic domains and enables the rapid identification of complete biosynthetic pathways harboring a specific enzyme functionality. Furthermore, the enrichment from minute DNA amounts grants the analysis of scarce resources. The assignment of unbinned BGCs to individual genomes provides valuable information about the producing organism, which can guide targeted isolation and cultivation approaches. However, only a minor fraction of microorganisms have been cultivated in the laboratory,[35] restricting the analysis of many BGCs to heterologous hosts.

Heterologous expression of BGCs from uncultivated strains has generated a plethora of products from diverse biosynthetic families. Direct cloning methods based on homologous recombination are now routinely used to capture intact BGCs without the need to construct and screen libraries.[38] For example, the ExoCET method based on *in vitro* exonuclease activity and *in vivo* recombination was used to clone the 106 kb salinomycin BGC in one step.[39] In addition to the distinct assignment of a metabolite to a pathway, heterologous production offers various tools for engineering approaches. Modular pathways like NRPSs and PKSs seem predestined for the production of new molecules in a combinatorial fashion. Re-engineering approaches can target different stages of the biosynthesis including building block supply, extension unit selection, online modification of the backbone, chain release, and post-assembly line tailoring steps.[40] To exploit the combinatorial potential of assembly line pathways, the RedEx method was applied to rationally re-design the spinosad pathway, leading to the production of analogs with altered polyketide backbones and modified sugar methylation patterns.[41]

The unprecedented diversity of module variants and the combinatorial reorganization of gene fragments in *trans*-AT PKSs make them interesting candidates for engineering strategies.[42] However, the large size of many BGCs and the origin in strains taxonomically distant from commonly used hosts have significantly limited our understanding of *trans*-AT PKSs and complicated heterologous production efforts. In Chapter V, we applied the ExoCET method to efficiently clone complete *trans*-AT PKS clusters from cultivated and uncultivated sources, proving this method superior to traditional cloning strategies, particularly when assembling multiple fragments. Our experiments provide the basis towards the development of a heterologous production platform for *trans*-AT PKS pathways and the characterization of orphan BGCs from uncultivated sources, potentially leading to the identification of novel bioactive compounds. A broadly applicable system will advance our understanding on how these machineries function inside the cell and provide crucial insights for the rational pathway engineering on the way to generate "unnatural" natural products.

In summary, the goal of this thesis was to identify and characterize natural product pathways encoded in the microbiome of the sponge *M. hentscheli* and apply different strategies to exploit the biosynthetic potential encoded in these uncultivated strains. The discovery of numerous orphan BGCs distributed across a broad phylogenetic range of bacteria has revealed a contrasting scenario to the "superproducer" symbionts, suggesting that multiple mechanisms lead to the acquisition of chemical arsenals in sponges. The results reinforce uncultured microbes as promising source of chemical novelty and open avenues to characterize producing organisms and their biosynthetic pathways. The challenging task of genome-guided natural product identification from uncultivated or unknown strains was addressed in several ways. An indirect sequence capture technology in combination with long-read sequencing extended BGCs and putatively linked them to producer genomes. Detailed *in silico* analysis of gene clusters and the use of metabolic prediction methods allowed the construction of biosynthetic models that guide the isolation and structure elucidation process. Studies towards the development of a broad-host-range platform for the expression of *trans*-AT PKS gene clusters could pave the way for the biotechnological production of pharmaceutically relevant polyketides for which the producer is either unknown or uncultivated and serve as starting point for rational engineering approaches. Looking to the future, the increasing amount of (meta)genomic data, further development of predictive tools for mining sequence space, a better understanding of enzyme structure and function, and improved heterologous production systems will promote innovative strategies to generate chemical novelty.

References

1. A. L. Demain, A. Fang, "The Natural Functions of Secondary Metabolites" in History of Modern Biotechnology I. Advances in Biochemical Engineering/Biotechnology, A. Fiechter, Ed. (Springer, Berlin, Heidelberg, 2000), vol. 69, pp. 1-39.
2. J. Davies, Specialized microbial metabolites: functions and origins. *J. Antibiot.* **66**, 361-364 (2013).
3. M. F. Traxler, R. Kolter, Natural products in soil microbe interactions and evolution. *Nat. Prod. Rep.* **32**, 956-970 (2015).
4. H. Lachance, S. Wetzel, K. Kumar, H. Waldmann, Charting, navigating, and populating natural product chemical space for drug discovery. *J. Med. Chem.* **55**, 5989-6001 (2012).
5. A. G. Atanasov *et al.*, Discovery and resupply of pharmacologically active plant-derived natural products: a review. *Biotechnol. Adv.* **33**, 1582-1614 (2015).
6. G. D. Wright, Unlocking the potential of natural products in drug discovery. *Microb. Biotechnol.* **12**, 55-57 (2019).
7. A. Fleming, On the antibacterial action of cultures of a Penicillium, with special reference to their use in the isolation of *B. influenzæ*. *Br. J. Exp. Pathol.* **10**, 226-236 (1929).
8. D. J. Newman, G. M. Cragg, Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770-803 (2020).
9. H.-F. Ji, X.-J. Li, H.-Y. Zhang, Natural products and drug discovery. *EMBO Rep.* **10**, 194-200 (2009).
10. J. Bérdy, Bioactive microbial metabolites. *J. Antibiot.* **58**, 1-26 (2005).
11. C. R. Pye, M. J. Bertin, R. S. Lokey, W. H. Gerwick, R. G. Linington, Retrospective analysis of natural products provides insights for future discovery trends. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 5601-5606 (2017).
12. J. J. J. van der Hooft *et al.*, Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.* **49**, 3297-3314 (2020).

13. C. T. Nguyen, D. Dhakal, V. T. Pham, H. T. Nguyen, J.-K. Sohng, Recent advances in strategies for activation and discovery/characterization of cryptic biosynthetic gene clusters in *Streptomyces*. *Microorganisms* **8**, 616 (2020).
14. H. B. Bode, B. Bethe, R. Höfs, A. Zeeck, Big effects from small changes: possible ways to explore nature's chemical diversity. *ChemBioChem* **3**, 619-627 (2002).
15. B. K. Okada, M. R. Seyedsayamdost, Antibiotic dialogues: induction of silent biosynthetic gene clusters by exogenous small molecules. *FEMS Microbiol. Rev.* **41**, 19-33 (2017).
16. A. R. Carroll, B. R. Copp, R. A. Davis, R. A. Keyzers, M. R. Prinsep, Marine natural products. *Nat. Prod. Rep.* **37**, 175-223 (2020).
17. B.-N. Han *et al.*, "Natural Products from Sponges" in Symbiotic Microbiomes of Coral Reefs Sponges and Corals, Z. Li, Ed. (Springer, Dordrecht, 2019), pp. 329-463.
18. J. A. Zumberge *et al.*, Demosponge steroid biomarker 26-methylstigmastane provides evidence for Neoproterozoic animals. *Nat. Ecol. Evol.* **2**, 1709-1714 (2018).
19. N. S. Webster, M. W. Taylor, Marine sponges and their microbial symbionts: love and other relationships. *Environ. Microbiol.* **14**, 335-346 (2012).
20. J. Piel *et al.*, Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge *Theonella swinhoei*. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16222-16227 (2004).
21. M. D. Tianero, J. N. Balaich, M. S. Donia, Localized production of defence chemicals by intracellular symbionts of *Haliclona* sponges. *Nat. Microbiol.* **4**, 1149-1159 (2019).
22. M. C. Wilson *et al.*, An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58-62 (2014).
23. T. Mori *et al.*, Single-bacterial genomics validates rich and varied specialized metabolism of uncultivated *Entotheonella* sponge symbionts. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 1718-1723 (2018).
24. R. Ueoka *et al.*, Metabolic and evolutionary origin of actin-binding polyketides from diverse organisms. *Nat. Chem. Biol.* **11**, 705-712 (2015).
25. T. Wakimoto *et al.*, Calyculin biogenesis from a pyrophosphate protoxin produced by a sponge symbiont. *Nat. Chem. Biol.* **10**, 648-655 (2014).
26. M. Rust *et al.*, A multiproducer microbiome generates chemical diversity in the marine sponge *Mycale hentscheli*. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9508-9518 (2020).
27. K. Blin *et al.*, antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81-W87 (2019).
28. E. J. N. Helfrich, J. Piel, Biosynthesis of polyketides by *trans*-AT polyketide synthases. *Nat. Prod. Rep.* **33**, 231-316 (2016).
29. E. J. N. Helfrich *et al.*, Automated structure prediction of *trans*-acyltransferase polyketide synthase products. *Nat. Chem. Biol.* **15**, 813-821 (2019).
30. P. R. Jensen, Natural products and the gene cluster revolution. *Trends Microbiol.* **24**, 968-977 (2016).
31. D. Meleshko *et al.*, BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Res.* **29**, 1352-1362 (2019).
32. M. Ayling, M. D. Clark, R. M. Leggett, New approaches for metagenome assembly with short reads. *Briefings Bioinf.* **21**, 584-594 (2020).
33. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824-834 (2017).
34. D. Li, C. M. Liu, R. Luo, K. Sadakane, T. W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674-1676 (2015).
35. I. J. Miller, M. G. Chevrette, J. C. Kwan, Interpreting microbial biosynthesis in the genomic age: biological and practical considerations. *Mar. Drugs* **15**, 165 (2017).
36. D. D. Kang *et al.*, MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
37. V. Mallawaarachchi, A. Wickramarachchi, Y. Lin, GraphBin: refined binning of metagenomic contigs using assembly graphs. *Bioinformatics* **36**, 3307-3313 (2020).
38. Z. Lin, J. Nielsen, Z. Liu, Bioprospecting through cloning of whole natural product biosynthetic gene clusters. *Front. Bioeng. Biotechnol.* **8**, 526 (2020).
39. H. Wang *et al.*, ExoCET: exonuclease *in vitro* assembly combined with RecET recombination for highly efficient direct DNA cloning from complex genomes. *Nucleic Acids Res.* **46**, e28 (2018).

40. C. Beck, J. F. G. Garzón, T. Weber, Recent advances in re-engineering modular PKS and NRPS assembly lines. *Biotechnol. Bioprocess Eng.* **25**, 886-894 (2020).
41. C. Song *et al.*, RedEx: a method for seamless DNA insertion and deletion in large multimodular polyketide synthase gene clusters. *Nucleic Acids Res.* **48**, e130 (2020).
42. E. J. N. Helfrich *et al.*, Evolution of combinatorial diversity in *trans*-acyltransferase polyketide synthase assembly lines across bacteria. *Nat. Commun.* **12**, 1422 (2021).

Michael Rust

Address	Bergacker 66 8046 Zurich
Email	rustmi@ethz.ch
Date of birth	11.11.1991
Nationality	Swiss

Education

2017 - 2021 **Dr. sc. ETH Zurich Candidate** (Switzerland)
Member of the Microbiology and Immunology PhD program

2014 - 2016 **MSc in Interdisciplinary Sciences**, ETH Zurich (Switzerland)
Major in Chemistry and Biology

2011 - 2014 **BSc in Interdisciplinary Sciences**, ETH Zurich (Switzerland)
Biochemical-physical direction

Work Experience

2017 - 2021 **Graduate research with Prof. Jörn Piel, Institute of Microbiology, ETH Zurich**
Identification and Characterization of Natural Product Biosynthetic Pathways Encoded in the Microbiome of the Sponge Mycale hentscheli

2016 **Civilian Service, Institute for Sustainability, Agroscope, Zurich**

2016 **MSc Thesis with Prof. Jörn Piel, Institute of Microbiology, ETH Zurich**
Cloning and Heterologous Expression of Trans-AT PKS Gene Clusters

2015 - 2016 **Tutor at ABACUS Nachhilfeinstitut, Zurich**

2015 **Research Project with Prof. Ulrike Kutay, Institute of Biochemistry, ETH Zurich**
Stress Responses upon Impaired Ribosome Biogenesis

2014 - 2015 **Civilian Service, Teaching Assistant, Primary School Hedingen**

2014 **BSc Thesis with Prof. Jörn Piel, Institute of Microbiology, ETH Zurich**
Identification of Trans-AT PKS Gene Clusters by Structure-based Gene Targeting

2013 **Research Project with Prof. Sabine Werner, Institute of Molecular Health Sciences, ETH Zurich**
Identification and Characterization of FGF7-regulated Proteins in the Mouse Liver

2011 **Export Assistant, Essemtec AG, Aesch LU**

2010 **Military Service, Medic, Airolo**

Skills and Leadership Experience

2017 - 2021 **Student mentor:** Supervision of an exchange PhD student project (Pakjira Nanudorn), master projects (Tomas Kündig, Mariella Greutmann), and a high school student project (Nemo Milos), Institute of Microbiology, ETH Zurich

2017 - 2021 **Academic supervision:** Block courses (Bioactive Natural Products from Bacteria) and microbiology practical courses, Institute of Microbiology, ETH Zurich

2017-2020 **Workshops:**

- Learning to Teach, Lehrentwicklung und -technologie, ETH Zurich
- Basic Scientific Presentation Skills, MIM PhD program, ETH Zurich
- General Principles of Scientific Writing, MIM PhD program, ETH Zurich
- Nanopore Best Practice Workshop, German Network for Bioinformatics Infrastructure, Bielefeld University, Germany

Publications and Presentations

Peer-reviewed publications:

Schorn M.A., Verhoeven S., Ridder L., Huber F., ... , **Rust M.**, ... , Duncan K.R., Crüsemann M., Rogers S., Dorrestein P.C., Medema M.H., van der Hooft J.J.J. A community resource for paried genomic and metabolomic data mining. *Nat. Chem. Biol.* **17**, 363-368 (2021).

Piel J. and **Rust M.** Metagenome Mining. *Comprehensive Natural Products III*, **6**, 50-89 (2020).

Rust M., Helfrich E.J.N., Freeman M.F., Nanudorn P., Field C.M., Rückert C., Kündig T., Page M.J., Webb V.L., Kalinowski J., Sunagawa S., Piel J. A multiproducer microbiome generates chemical diversity in the marine sponge *Mycale hentscheli*. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9508-9518 (2020).

Helfrich E.J.N., Ueoka R., Dolev A., **Rust M.**, Meoded R.A., Bhushan A., Califano G., Costa R., Gugger M., Steinbeck C., Moreno P., Piel J. Automated structure prediction of *trans*-acyltransferase polyketide synthase products. *Nat. Chem. Biol.* **15**, 813-821 (2019).

Selected Posters and Talks

Rust M. A Multiproducer Microbiome Creates Chemical Diversity in the Marine Sponge *Mycale hentscheli*. Gordon Research Conference on Marine Natural Products 2020, Ventura, USA (Poster).

Rust M. A Multiproducer Microbiome Generates Natural Product Diversity in the Marine Sponge *Mycale hentscheli*. Swiss Society for Microbiology Section Prokaryotic Biology Meeting 2019, Coppet, Switzerland (Talk).

Language Skills

German	Mother tongue	French	Intermediate
English	Fluent	Spanish	Beginners' level