# Evolutionary investigation of bacterial biosynthetic gene clusters at multiple scales

**Dissertation**

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Athina Gavriilidou

aus Panorama Thessalonikis, Griechenland

Tübingen
2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

| | |
|---|---|
| Tag der mündlichen Qualifikation: | 04.06.2024 |
| Dekan: | Prof. Dr. Thilo Stehle |
| 1. Berichterstatterin: | Prof. Dr. Nadine Ziemert |
| 2. Berichterstatterin: | Prof. Dr. Kay Nieselt |
| 3. Berichterstatterin: | Prof. Dr. Avena C. Ross |

# Abstract

Specialised metabolites are chemical compounds that have a big impact on human life, being involved in our pharmaceuticals, our food and even in industrial production processes. Due to their relevance, the discovery of new molecules with desired bioactivities is imperative. From a bioinformatic point of view, these efforts are supported through genome mining tools, which can detect the genes relevant for the biosynthesis of specialised metabolites in the producers' genomes. These genes are often found in close proximity to one another, forming biosynthetic gene clusters (BGCs), especially in bacterial organisms. The advancements of sequencing methods has led to a rapid increase in available bacterial genomes, which can be explored for their biosynthetic capacity. Though a large number of bacterial BGCs, and the compounds whose biosynthesis they encode, have already been detected, much is still unclear about them. Their evolutionary history can be especially complex, with horizontal gene transfer (HGT) events more common in BGCs compared to the rest of the genome. However, evolutionary studies of biosynthetic genes have led to the development of certain genome mining as well as bioengineering methods and are necessary for the advancement of the field.

In the present dissertation, I am describing the efforts to promote the discovery of specialised metabolites by increasing our understanding of their distribution and evolution. Exploiting the available volume of sequencing information, a global analysis of bacterial BGCs revealed that there is great difference among the biosynthetic capacity of different taxa. Next, as I attempted to understand the observed distribution, the focus shifted on one specific system, glycopeptide antibiotics (GPAs), whose biosynthetic pathway is explained in detail. Subsequently, the phylogenetic reconstruction of the related BGCs' evolutionary history was possible and it revealed an important inaccuracy in the current classification system. Finally, an attempt to identify any significant associations between the presence of these and other kinds of BGCs was made. Even in a preliminary stage, the latter analysis revealed a promising lead that may constitute an adaptation mechanism to HGT of BGCs, though this hypothesis requires further investigation. Apart from the insights already gained, the methodologies and datasets presented here are expected to be the focus of various future studies.

# Kurzfassung

Spezialisierte Metaboliten sind chemische Verbindungen, die einen großen Einfluss auf das menschliche Leben haben, da sie in unseren Arzneimitteln, unseren Lebensmitteln und sogar in industriellen Produktionsprozessen enthalten sind. Aufgrund ihrer Bedeutung ist die Entdeckung neuer Moleküle mit gewünschten Bioaktivitäten zwingend erforderlich. Aus bioinformatischer Sicht werden diese Bemühungen durch Genome Mining Programme unterstützt, die die für die Biosynthese spezieller Metaboliten relevanten Gene im Genom der Produzenten aufspüren können. Diese Gene befinden sich oft in unmittelbarer Nähe zueinander und bilden Biosynthese Gencluster (BGCs), insbesondere in bakteriellen Organismen. Die Fortschritte bei den Sequenzierungsmethoden haben zu einer raschen Zunahme der verfügbaren Bakteriengenome geführt, die auf ihre biosynthetischen Fähigkeiten hin untersucht werden können. Obwohl bereits eine große Anzahl von bakteriellen BGCs und die Verbindungen, für deren Biosynthese sie kodieren, entdeckt wurden, ist noch vieles unklar über sie. Ihre Evolutionsgeschichte kann besonders komplex sein, wobei horizontale Gentransfers bei BGCs häufiger vorkommen als im übrigen Genom. Dennoch haben evolutionäre Studien zu biosynthetischen Genen zur Entwicklung bestimmter Genome Mining- und Bioengineering-Methoden geführt und sind für die Weiterentwicklung des Fachgebiets notwendig.

In der vorliegenden Dissertation beschreibe ich unsere Bemühungen, die Entdeckung spezialisierter Metaboliten zu fördern, indem wir unser Verständnis für ihre Verbreitung und Evolution verbessern. Eine globale Analyse der bakteriellen BGCs unter Ausnutzung der verfügbaren Sequenzierungsdaten ergab, dass es große Unterschiede zwischen den Biosynthesekapazitäten der verschiedenen Taxa gibt. Bei dem Versuch, die beobachtete Verteilung zu verstehen, konzentrierte ich mich dann auf ein spezifisches System, die Glykopeptid-Antibiotika (GPAs), deren Biosyntheseweg im Detail erläutert wird. Anschließend war eine phylogenetische Rekonstruktion der Entwicklungsgeschichte der verwandten BGCs möglich, die eine wichtige Ungenauigkeit im derzeitigen Klassifizierungssystem aufdeckte. Schließlich wurde versucht, signifikante Assoziationen zwischen dem Vorhandensein dieser und anderer Arten von BGCs zu identifizieren. Selbst in einem vorläufigen Stadium ergab

die letztgenannte Analyse eine vielversprechende Spur, die einen Anpassungsmechanismus an den horizontalen Gentransfer von BGCs darstellen könnte, auch wenn diese Hypothese weitere Untersuchungen erfordert. Abgesehen von den bereits gewonnenen Erkenntnissen dürften die hier vorgestellten Methoden und Datensätze im Mittelpunkt verschiedener künftiger Studien stehen.

# Acknowledgements

**T**he weight of this dissertation is carried by the support of many people that I had the pleasure to know. I will do my best to express my gratitude to them.

**H**ow else can I start, except with my family, who have always stood behind me. My mother Nikolina, who has always believed in me and instilled the values of hard work and responsibility. My father Vasileios, whose love of science I believe I inherited. Rozeta and Anestis, who always made me feel safe, no matter where I was in the world. Nikol and little Yannis, who were my home away from home during my time in Germany.

**A** most sincere appreciation for my first supervisor, Prof. Dr. Nadine Ziemert. I came across her group during my first months in Tübingen, when I was looking for a HiWi job. I can not believe my luck that the shaky photo of a job opening, sent by a random acquaintance of mine from my previous university in Greece, led me to this path. I have been part of the group for about six years now, first as a Master then as a PhD student and I have enjoyed every minute of it. Working in academia has its hardships, but the support that Nadine has displayed during not only the expected hurdles, but also during a pandemic, is a major factor that has led to the completion of my PhD. Know that you have set a high bar for me for the kind of atmosphere that a research group should have.

**N**urtured by the academic environment in the University of Tübingen, I am thankful for the guidance of my second supervisor, Prof. Dr. Daniel Huson, who I have learnt a lot from already during my Master's studies here. Also, Prof. Dr. Kay Nieselt, who was the Dean of Studies at the time and whose advice I have asked for my study and career plan.

**K**indred spirits among my colleagues, past and present, with whom I shared daily interactions. Martina, the well-rounded expert, who set the foundations of my investigations on glycopeptides and who we could all count on, be it bioinformaticians or microbiologists. Direnc, who took me in as a padawan in my early research endeavours and has fed me on multiple occasions in his house. Shrikant, who always had valuable insights on any scientific topic and who was the one to let me know about a PhD position opening up in the group. Caner, who is incredibly quick to solve any bioinformatics-related problem and quicker still to offer his help. Aileen, our beetle enthusiast, who opened my mind to different fields and always had time for a coffee and advice. Bita, my hobbit counterpart and pomodoro buddy, who I conversed with so often in the office, as much for research as not. To you and to all members of the group: I will miss our scientific exchange and our lunchtime discussions.

**Y**et more people were part of my daily life, on the 10th floor of the E-Bau auf der Morgenstelle. Among them, PD. Dr. Evi Stegmann, who I often ran to with questions about glycopeptides and was always happy to discuss them, often in Greek, as well as her student Jens, my microbiologist counterpart. Also, Aysun and Melanie, without whom a lot of things would not be possible, and Libera, whose input on scientific texts is so helpful to all of us.

**O**wing to the seamless cooperation with numerous collaborators, I extend my gratitude for their contributions that led to the realisation of the studies in this dissertation. Additionally, I express appreciation to the funding bodies and the resources, computational and otherwise, provided by the university.

**U**nderlining the significance of the people that I met here in Tübingen could not be missing from my acknowledgments. First of all, Manos, my partner and roommate, who has been with me for this entire journey. His loving support made everything so much easier. Also, my little bundle of joy Titika, who can not read this but could not be omitted. Then, my dear Monica, who was my first friend here and who became one of the closest people in my life and my travelling companion in our adventure in Ecuador. Which brings me to Mathias, who provided this opportunity, and, being also a PhD student, we have encouraged each other during this time. Finally, the Tübingen gang, for your friendship and support, especially the two academics. Melania, who understands my stressful nature and whose kindness I appreciate. And last but not least, Katerina, who has given me invaluable advice on so many occasions and, being our groups' unofficial therapist, always preaches that everything is going to be fine. You were right.

***so long and thanks for all the fish***

# Table of contents

# Abbreviations

Note: the abbreviations listed in this chapter were used in the *Introduction*, the *Research objectives* and the *Discussion and conclusion* chapters of this dissertation. For each of the main chapters (1-5), any abbreviations used are explained within the chapter.

BGC - Biosynthetic gene cluster

e.g. - exempli gratia (for the sake of example)

GCF - Gene Cluster Family

GPA - Glycopeptide Antibiotic

GPA BGC - BGC encoding the biosynthesis of a GPA

GRP - Glycopeptide Related Peptide

GRP BGC - BGC encoding the biosynthesis of a GRP

HGT - Horizontal Gene Transfer

i.e. - id est (that is)

MAG - Metagenome-Assembled Genome

NP - Natural Product

NRPS - Non-ribosomal Peptide Synthetase

OG - Orthologous Group

PKS - Polyketide Synthase

WHO - World Health Organisation

# Introduction

## Specialised metabolites and their significance

Every living organism employs certain primary functions to ensure their continuous survival. This includes processes such as nutrient acquisition, growth, and reproduction, all mediated by metabolic pathways and hardcoded into their genomes[1]. No matter how different two organisms may be, these operations are always present. However, there are other functions which do not play a direct role in survival, yet offer various evolutionary advantages to the hosts. These fall into what is called “specialised metabolism” and the specific roles of these natural products (NPs) that derive from it vary greatly, even among closely related organisms. Some plants produce substances to deter herbivores or attract carnivores for direct and indirect defence, accordingly[2]. There are symbiotic fungi that generate compounds that offer their insect hosts protection against predators, in exchange for improved chances of propagation[3]. Some bacteria excrete molecules to “capture” and bind the free iron in the environment and others to support their communities for the creation of biofilm structures[4].

Humans have been aware of the existence of specialised metabolites for quite some time, though under different names. They are involved in the making of beer, yoghurt, coffee, and tea[5], among others. Parts or extracts of plants and fungi have been used in folk medicine for aeons now, before modern researchers started to examine them methodologically and to discover the bioactive compounds that are responsible for the observed desired traits[6,7]. Today, we make use of specialised metabolites not only in the pharmaceutical industry, for example as drugs that can be used for cancer treatments or against infections from pathogens[6,7], but also in a range of other applications. Some compounds have insecticide activity and are valuable in agricultural settings[8], or are relevant for the food industry because of their traits that can affect the texture, taste or even shelf life of edible products[9]. Others have an ecological impact as they are being explored for possible bioremediation approaches[10]. These are only some examples that highlight their importance in our daily lives and underline the need to learn more about them and to discover more molecules with such applications. Thankfully, there are constant improvements in

technologies and methods that can be applied to the search of new bioactive compounds.

With that in mind, the overarching objective of the present thesis is to increase our knowledge and understanding of the distribution and evolutionary history of the genetic traits that are connected to the biosynthesis of these NPs, achieved through application of bioinformatic methods.

## Discovery through genome mining methods

The rapid growth of the ease and accuracy of DNA sequencing influenced the methodologies applied to the study of secondary metabolism and the derived bioactive compounds. Analysis of the genomes of known producers from multiple domains of life (e.g., bacteria, fungi and plants) revealed that the genes that encode the enzymatic machinery involved in secondary metabolism tend to appear in close proximity to one another, forming so-called biosynthetic gene clusters (BGCs)[4,11].

The connection of the observed biochemistry to the genetics behind it allowed various studies that steadily increased the knowledge about the biosynthetic pathways and their individual components, especially in regards to the phenotype. This opened the way to genetic manipulation, heterologous expression and the generation of semi-synthetic compounds[7]. Furthermore, thanks to the knowledge about the genetic elements behind specialised metabolism, genome mining approaches can be applied to detect BGCs in the genomes of potential producers, speeding up the discovery process[7,12].

Several approaches have been employed so far for the creation of genome mining tools. Several of them are described in chapter 1, along with the importance of their scalability. For instance, some genome mining tools are established upon knowledge-based rules, such as antiSMASH[13,14] (**Figure 1**), which is basing its detection algorithm on the genomic characteristics of known BGCs and is periodically being updated to incorporate the newest findings. Other approaches include machine learning methods[15], such as DeepBGC[16], that are trained on known datasets, but aspire to detect completely new BGCs, independently of established insights on the underlying pathways. However, apart from the discovery of the BGCs, it is vital to apply methods for their dereplication, as BGCs of various degrees of similarity may encode the biosynthesis of very similar or identical metabolites. To this

end, algorithms that can cluster BGCs based on similarity and can quantify their diversity have been established. One such example is the BiG-SLICE[17] tool, which is also efficient enough to be applicable on big data. The development of such programs and the ability to apply them on a grand scale gave us the opportunity to survey the biosynthetic diversity and potential of the bacterial domain and identify promising taxa for bioprospecting, as is demonstrated in chapter 2.

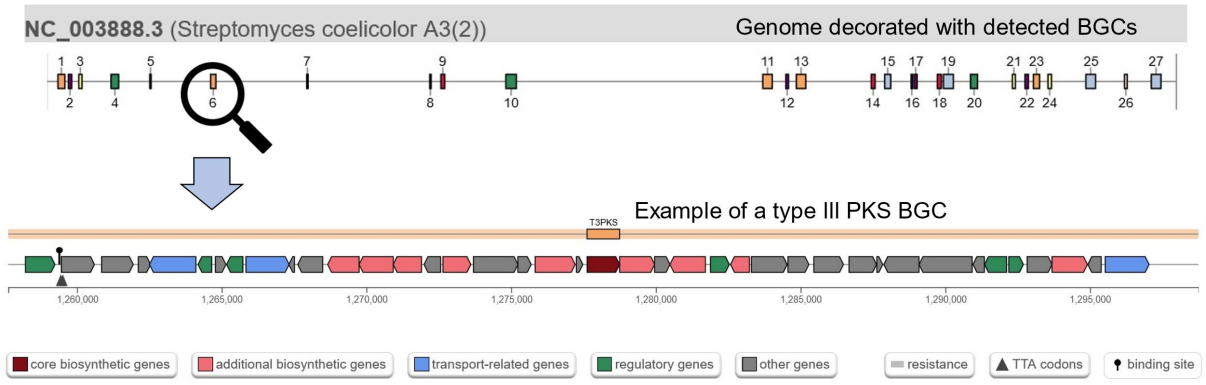



**Figure 1:** Example output of the antiSMASH[14] tool. Top: The example target genome (locus id and strain are as labelled) is represented by a line, decorated by numbered rectangles, which represent BGC locations. Bottom: example of a type III polyketide synthase (PKS) BGC. The coloured arrows represent genes in the BGC and their direction shows their encoding on the leading (direction to the right) or complementary strand (direction to the left). A scale under the arrows shows their location in the locus. The genes are coloured according to their general role and additional symbols indicate interesting regions (legend at the bottom). Image adapted from the antiSMASH example output.

## Glycopeptide antibiotics - a fitting model system to study specialised metabolism

Glycopeptide antibiotics (GPAs) have been the focus of a large part of this dissertation for two reasons. The first is their medical importance - they are used as last resort antibiotics against Gram-positive bacterial pathogens[18]. They are not the only product of specialised metabolism that has been applied against bacterial pathogens. The most well-known example is perhaps the finding of penicillin from a fungus, which gave rise to the golden age of antibiotic discovery, revolutionising the field of medicine[19]. In the decades after this discovery, a myriad of bioactive compounds that act as antibiotics were found, not only in fungal producers but also in bacteria, especially in actinomycetes [20].

Unfortunately, the rate of discovery of new antibiotics has sunk in the last few decades. The easiest to find producer candidates have been exhausted and it is no longer trivial to detect new ones. Discovery efforts from natural producers are associated with unattractive economic aspects which hinder the issue even more. The costs on the one side, and high chance or rediscovery of known compounds on the other side, have expedited a general loss of interest in pursuing such an endeavour[6,21]. However, the need for drugs with new modes of action against pathogenic microorganisms is higher than ever, due to the emergence of multidrug resistant pathogens[22]. In order to ameliorate future discovery efforts, our knowledge of the processes taking place in the bacterial producer cells of antibiotics needs to be expanded. Understanding in detail how the biosynthesis works and how the rest of the cellular mechanisms are affected by it, can open the way for new approaches, new targets to look for in candidate producers, be they specific enzymes or specific metabolic pathways that give away their potential for specialised metabolism. The GPAs are a good model system for such efforts, because their biosynthetic pathway, though highly complex, has been studied for decades[23]. A culmination of this knowledge has been summarised in chapter 3, which makes use of animated graphics as an effort to improve effective communication of the biosynthetic pathway of vancomycin, a type I GPA (**Figure 2**).

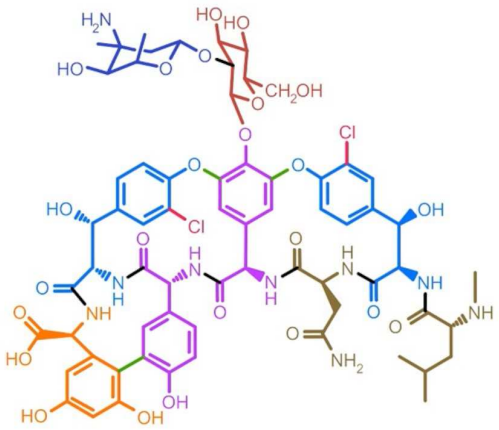


**Figure 2:** Structure of the GPA vancomycin. Vancomycin is a heptapeptide, whose backbone amino acids are crosslinked, chlorinated, glycosylated and methylated in various positions (all components coloured differently). Image adapted from the animation in chapter 3 (Supplementary Data 1).

The second reason for focusing on GPAs is their interesting evolutionary history. The BGCs encoding them have been found in multiple genera, especially in

*Streptomyces* and *Amycolatopsis*, and it has been proven that genes involved have been horizontally transmitted[24]. Studying the evolution of sequences related to specialised metabolism or of their producer organisms is vital for the development of new genome mining methods. As is explained in chapter 1, some genome mining tools are based on evolutionary theorems (e.g., ARTS[25,26]), which are concluded through phylogenetic studies. Inference of evolutionary history is heavily based on the construction of phylogenetic trees[27]. In the case of BGCs, phylogenetics has proven useful to determine possible connections of genes to the primary metabolism, for example duplication events and further diversification of the duplicate genes. Based on this, the detection of paralogues is employed in the ARTS genome mining tool to suggest putative BGCs[26].

Moreover, it has been established that both neutral and positive evolutionary pressure is applied on BGCs, making their history complex. Especially so, since after the incorporation of a gene into a working BGC, it becomes part of a larger biosynthetic unit and its evolutionary trajectory becomes intertwined with that of the entire cluster. At the same time, there are known cases where the evolutionary pressure does not act on a gene but on a functional domain within a gene. All this complexity leads to multiple signals being encoded in the sequences of the genes involved. The different evolutionary pressures applied on BGCs, as well as the opportunities for genome mining that this knowledge provides, are presented in the review of chapter 1. Aiming to increase such knowledge for the biosynthetic pathway of GPAs, an extensive evolutionary study of BGCs encoding for GPAs (from here on referred to as 'GPA BGCs') was conducted, demonstrated in chapter 4. These findings, which combined phylogenetic patterns and structural characteristics, led to a reassessment of their current classification system.

The complexity of GPA biosynthesis (explained in chapter 3) and the different types of enzymes necessary for it are translated to a complexity in their evolutionary history[24]. Knowing the main functions of most genes involved, it is possible to study the origins of these BGCs as a whole. This, coupled with the fact that there are strains very closely related to the known GPA producers which do not encode such BGCs in their genome, makes them an ideal system with which to study horizontal gene transfer (HGT) in BGCs. In fact, this statement is true not only in the confines of the gene clusters themselves, but also with any repercussions that such events may have on the rest of the genome. This is the main topic presented in chapter 5,

where the bacterial genomes of the *Amycolatopsis* genus are inspected for significant association relationships between genes involved and not involved in GPA biosynthesis. Though the analysis has not been completed, a candidate relationship between a group of genes and the emergence of GPA BGCs was detected and awaits further investigation.

# List of publications included in the thesis

Note: the projects here are presented in a different order as their corresponding chapters.

## Research articles:

1. Athina Gavriilidou*, Satria A. Kautsar*, Nestor Zaburannyi, Daniel Krug, Rolf Müller, Marnix H. Medema, and Nadine Ziemert. *Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes*. **Nature Microbiology**. 2022; 7: 726–735.

## Reviews and relevant articles:

1. Marc G. Chevrette*, Athina Gavrilidou*, Shrikant Mantri*, Nelly Selem-Mojica, Nadine Ziemert, and Francisco Barona-Gómez. *The confluence of big data and evolutionary genome mining for the discovery of natural products*. **Natural Products Reports**. 2021; 38:2024-2040.

## Manuscripts in progress:

1. Athina Gavriilidou, Martina Adamek, Jens-Peter Rodler, Noel Kubach, Anna Voigtländer, Leon Kokkoliadis, Chambers Hughes, Max J. Cryle, Evi Stegmann, and Nadine Ziemert. *Animating insights into the: Biosynthesis of glycopeptide antibiotics*. Advanced manuscript. Awaiting submission.
2. Athina Gavriilidou, Martina Adamek, Jens-Peter Rodler, Noel Kubach, Susanna Kramer, Daniel H. Huson, Max J. Cryle, Evi Stegmann, and Nadine Ziemert. *Phylogenetic distance and structural diversity directing a reclassification of glycopeptide antibiotics*. Advanced manuscript. Preprint available on bioRxiv.
3. Athina Gavriilidou, Franz Baumdicker and Nadine Ziemert. *BGC-aware gene coincidence analysis of the Amycolatopsis pangenome using the Goldfinder tool*. Initial manuscript.

*equal contributions

# List of publications not included in the thesis

## Research articles:

- Barbara R. Terlouw*, Kai Blin*, Jorge C. Navarro-Muñoz, Nicole E. Avalon, Marc G. Chevrette, Susan Egbert, Sanghoon Lee, David Meijer, Michael J.J. Recchia, Zachary L. Reitz, Jeffrey A. van Santen, Nelly Selem-Mojica, Thomas Tørring, Liana Zaroubi, Mohammad Alanjary, Gajender Aleti, C ´esar Aguilar, Suhad A.A. Al-Salihi, Hannah E. Augustijn, J. Abraham Avelar-Rivas, Luis A. Avitia-Dom´ınguez, Francisco Barona-Gómez, Jordan Bernaldo-Agüero, Vincent A. Bielinski, Friederike Biermann, Thomas J. Booth, Victor J. Carrion Bravo, Raquel Castelo-Branco, Fernanda O. Chagas, Pablo Cruz-Morales, Chao Du, Katherine R. Duncan, Athina Gavriilidou, Damien Gayrard, Karina Gutiiérrez-García, Kristina Haslinger, Eric J.N. Helfrich, Justin J.J. van der Hooft, Afif P. Jati, Edward Kalkreuter, Nikolaos Kalyvas, Kyo Bin Kang, Satria Kautsar, Wonyong Kim, Aditya M. Kunjapur, Yong-Xin Li, Geng-Min Lin, Catarina Loureiro, Joris J.R. Louwen, Nico L.L. Louwen, George Lund, Jonathan Parra, Benjamin Philmus, Bita Pourmohsenin, Lotte J.U. Pronk, Adriana Rego, Devasahayam Arokia Balaya Rex, Serina Robinson, L. Rodrigo Rosas-Becerra, Eve T. Roxborough, Michelle A. Schorn, Darren J. Scobie, Kumar Saurabh Singh, Nika Sokolova, Xiaoyu Tang, Daniel Udwary, Aruna Vigneshwari, Kristiina Vind, Sophie P.J.M. Vromans, Valentin Waschulin, Sam E. Williams, Jaclyn M. Winter, Thomas E. Witte, Huali Xie, Dong Yang, Jingwei Yu, Mitja Zdouc, Zheng Zhong, Jérôme Collemare, Roger G. Linington, Tilmann Weber, and Marnix H. Medema. *MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters*. **Nucleic Acids Research**. 2022; 51(D1): D603-610.
- Thierry Izoré*, Y. T. Candace Ho*, Joe A. Kaczmarski, Athina Gavriilidou, Ka Ho Chow, David L. Steer, Robert J. A. Goode, Ralf B. Schittenhelm, Julien Tailhades, Manuela Tosin, Gregory L. Challis, Elizabeth H. Krenske, Nadine Ziemert , Colin J. Jackson, and Max J. Cryle. *Structures of a non-ribosomal peptide synthetase condensation domain suggest the basis of substrate selectivity*. **Nature Communications**. 2021; 12:2511.

- Ira Handayani*, Hamada Saad*, Shanti Ratnakomala, Puspita Lisdiyanti, Wien Kusharyoto, Janina Krause, Andreas Kulik, Wolfgang Wohlleben, Saefuddin Aziz, Harald Gross, Athina Gavriilidou , Nadine Ziemert, and Yvonne Mast. *Mining Indonesian Microbial Biodiversity for Novel Natural Compounds by a Combined Genome Mining and Molecular Networking Approach*. **Marine Drugs**. 2021; 19(6):316.

## Manuscripts in progress:

- Christian Resl, Emilian Paulitz, Athina Gavriilidou, Nadine Ziemert, Anne Kupczok, and Franz Baumdicker. *Goldfinder: Unraveling Gene Co-occurrence and Gene-Avoidance in Bacterial Pangenomes while Considering Phylogenetic Relationships*. Advanced manuscript.

*equal contributions

# Contributions

Note: the projects here are presented in the same order as their corresponding chapters.

**Review 1 - chapter 1**: All authors wrote the manuscript together. All authors gathered and summarised the related literature. Main contributions per topic are as follows. Introduction: MGC and NSM. Evolution and phylogenetics in relation to Big Data: FBG and NSM. Genome and Metabolome Evolution of Natural Product Producers: NZ, SM and AG. BGC evolution: NZ and FBG. Evolutionary genome mining for NPs: FBG and NSM.

Personal contributions: Assessment of increase in sequenced genomes compared to taxonomic diversity and known producers. Collection and summarisation of literature related to the evolution of NP producer genomes.

**Research article 1 - chapter 2**: The tools used for the analysis (modified BiG-SLiCE and clust-o-matic) were developed by SAK and NZ. SAK also performed the detection of BGCs and both authors performed similarity clustering. DK conducted the statistical tests related to the variance of biosynthetic diversity in different taxonomic ranks. The manuscript was written by AG, DK, RM, MHM and NZ. All authors contributed to the conception and design of the analysis, read and approved the final manuscript. Author contributions are part of the published article as well.

Personal contributions: Design and application of the pipeline that handles the similarity clustering result from the two tools used. That includes: combination of the GCF assignments to the phylogenetic tree of bacteria from GTDB, conceptualization and creation of REDgroups from that tree, annotation of REDgroups with their unique GCFs, conduction of rarefaction analyses to extrapolate the potential GCFs of each group and of the total, uniqueness analyses of GCFs among groups, comparison of most promising groups to a database of known producers, connection of biogeography to MAG-derived GCFs. Writing of the first version of the manuscript.

**Manuscript 1 - chapter 3**: AG and NZ conceptualised the new format of the manuscript. AG, MA, JPR, NK, MJC, ES and EZ contributed to the literature review included in the manuscript. AG, NZ, NK, LK and ES were directly involved in the

animation content and design, which was mostly carried out by AV. All authors read and approved the latest manuscript.

Personal contributions: Conceptualisation of the new communication format. Contribution in the literature search and summarization. Writing of the first version of the manuscript. Careful curation of the animation content in each stage of the production.

**Manuscript 2 - chapter 4**: Dataset generation was conceptualised by NZ and implemented by AG. MA and AG conducted the manual trimming and curation of the BGC dataset, with contribution from ES and NK. Sample preparation and DNA extraction of the *S. varsoviensis* producer strain was accomplished by JPR. Stachelhaus code analysis was conducted by AG, MA and MJC. Development of the method for the concatenated phylogeny was done by AG and NZ with heavy contribution from SK. Creation and interpretation of the super network was conducted by AG, with heavy contribution from DHH. All authors contributed to the conception and design of the analysis, read and approved the latest manuscript.

Personal contributions: Design, implementation and execution of the semi-automatic pipeline for conducting the dataset creation and the full phylogenetic analysis of GPA and GRP encoding BGCs. This includes database searching via HMM, BGC detection, BGC similarity clustering, domain analysis, Stachelhaus code analysis, orthology inference, phylogenetic analyses of genes and domains, congruence check and concatenated phylogeny building. Manual trimming and curation of the BGC dataset. Genome assembly of the *S. varsoviensis* producer strain. Building and interpretation of the super network. Writing of the first version of the manuscript.

**Manuscript 3 - chapter 5**: All authors contributed to the conception and design of the analysis, read and approved the latest manuscript.

Personal contributions: Dataset generation. Design, implementation and execution of the pipeline for performing BGC-aware gene coincidence analysis using the Goldfinder tool. Writing of the first version of the manuscript.

# Research objectives

The principal ambition of this dissertation is to promote the discovery efforts for new bioactive compounds from natural sources by advancing our understanding of their distribution and evolutionary history. This is achieved by investigating the genetic elements of specialised metabolism in different scales.

The first chapter introduces genome mining approaches for the discovery of BGCs. It comprises a bibliographic review, describing the main evolutionary concepts behind many such methods, as well as the tools applying them. Several of the tools introduced in this chapter are applied in the following analyses. Additionally, because of the ever-increasing volume of sequencing data being produced, the notion of big data is discussed, in regards to its importance for the field, as well as the algorithms that are able to tackle such volumes of information.

The second chapter is focused on the comprehensive survey of biosynthetic potential encoded in bacterial genomes. The complete volume of publicly available sequenced genomes has been mined for potential BGCs and their distribution have been assessed across the bacterial domain. A quantification of the diversity of the encoded BGCs enabled the estimation of our current experimental coverage of the bacteria's specialised metabolites, which leaves much more to be discovered. Furthermore, a conjunction of quantified biosynthetic diversity with the phylogeny-based taxonomic placement of the producer genomes enabled an enumeration of promising taxa for future discovery efforts. What was evident from an analysis of such a broad scale was that only some bacterial taxa were very prolific producers of diverse specialised metabolites, whereas most were largely lacking in this regard. To understand why the distribution is as observed, it is necessary to delve into the evolutionary forces shaping these BGCs.

Shifting from the global overview to a specific model system (i.e., one category of BGCs) demands the comprehension of its encoded biosynthetic pathway. In this dissertation, the focus is redirected to the GPAs, a group of last-resort antibiotics. Their biosynthesis is complex and includes many steps which are difficult to visualise. To this end, an animation was created to ameliorate scientific communication of the known biosynthetic pathway of vancomycin, a GPA. This innovative approach is presented in chapter three, accompanied by a brief summary of the processes involved in GPA biosynthesis.

The fourth chapter combines the knowledge about GPA biosynthesis and the concepts discussed in the review of chapter 1 to investigate the complex evolutionary history of GPA BGCs. A carefully and manually curated dataset of GPA BGCs was used as input for a pipeline that was designed for constructing a representative phylogeny of the whole cluster, applicable to other BGC systems as well. Conclusions reached from the phylogenetic placement of the BGCs, in combination with known structural characteristics of the related compounds and their mode of action, mediated the suggestion of a dichotomous reclassification of the GPAs. The new system is based on predicted structural features that are concordant with the calculated phylogeny of the BGCs.

In the fifth chapter there is a switch of focus, from studying exclusively the genes involved in specialised metabolism to considering their possible connections to other genetic elements in the producers' genomes. The *Amycolatopsis* genus includes both strains that encode GPA biosynthesis and others that do not. By considering the genus' genomes as a whole, it was possible to conduct a BGC-aware gene coincidence analysis with the new Goldfinder tool, which allows the detection of strong association relationships between genes. The analysis resulted in a promising candidate BGC, whose presence possibly coincides with that of GPA BGCs. Though more research is required to confirm this relationship, it is hypothesised that its presence could in some way improve the chances of the host cell becoming a GPA producer.

# Chapter 1: The confluence of big data and evolutionary genome mining for the discovery of natural products

*(Manuscript published in Natural Product Reports, Aug 18 2021)*

Marc G Chevrette[1,*], Athina Gavrilidou[2,3,*], Shrikant Mantri[2,3,4,*], Nelly Selem-Mojica[5,†,@] , Nadine Ziemert[2,3,†], Francisco Barona-Gomez[5,†]

1 Wisconsin Institute for Discovery, Department of Plant Pathology, University of Wisconsin-Madison, Madison, WI, USA

2 Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Interfaculty Institute for Biomedical Informatics (IBMI), University Tübingen, Germany

3 German Centre for Infection Research (DZIF), Partner Site Tübingen, Germany.

4 Computational Biology Laboratory, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

5 Unidad de Genómica Avanzada (Langebio), Cinvestav-IPN, Irapuato, México

* authors contributed equally (MGC, AG, SM)

† corresponding authors (NSM, NZ, FBG)

@ Current address: Centro de Ciencias Matemáticas (CCM), UNAM, Morelia, Mexico

Personal contributions:

Scientific ideas/data generation/analysis & interpretation/writing: 20/20/20/20%

This review covers literature between 2014-2020

**Abstract**

The development and application of genome mining tools has given rise to ever-growing genetic and chemical databases and propelled natural products research into the modern age of Big Data. Likewise, an explosion of evolutionary studies has unveiled genetic patterns of natural products biosynthesis and function that support Darwin's theory of natural selection and other theories of adaptation and diversification. In this review, we aim to highlight how Big Data and evolutionary thinking converge in the study of natural products, and how this has led to an emerging sub-discipline of evolutionary genome mining of natural products. First, we outline general principles to best utilize Big Data in natural products research, addressing key considerations needed to provide evolutionary context. We then highlight successful examples where Big Data and evolutionary analyses have been combined to provide bioinformatic resources and tools for the discovery of novel natural products and their biosynthetic enzymes. Rather than an exhaustive list of evolution-driven discoveries, we highlight examples where Big Data and evolutionary thinking have been embraced for the evolutionary genome mining of natural products. After reviewing the nascent history of this sub-discipline, we discuss the challenges and opportunities of genomic and metabolomic tools with evolutionary foundations and/or implications and provide a future outlook for this emerging and exciting field of natural product research.

## 1. Introduction

Evolution is a process; therefore, evolutionary theory seeks to describe the series of events that have allowed life to appear, develop, and diversify. Natural selection, postulated by Charles Darwin more than one hundred and fifty years ago, is perhaps the most recognized of these theories, linking the natural histories of all living forms to their reproductive fitness(Sugden et al. 2009). In the years since Darwin, we have come to appreciate that evolutionary processes display enormous complexity and act through both selective and neutral forces of varying physicochemical, ecological, temporal, and population-level constraints(Goldman and Liberles 2021). Neutral, non-adaptive evolution was once thought to be discordant with Darwinian evolution; now we appreciate that evolutionary histories provide evidence of both selective pressures and neutral events(Lynch et al. 2016; Wideman et al. 2019). Founder effects, genetic drift, gene flow, and many other neutral mechanisms shape the genetic variation within populations upon which natural selection operates(Matthew B. Hamilton 2021). The enzymes of natural product (NP) biosynthesis are encoded in genomic information, and as such do not escape these forces of evolution (Chevrette, Gutiérrez-García, et al. 2020; Jensen 2016). This distinction is as important to recognize, as it is easy to neglect: NPs with antagonistic functions, like antibiotics or other biocides, are typically assumed to be under positive selection to maintain the interactions with their molecular target(s) necessary to retain function.

Paradoxically, the historical use of the term ‘secondary metabolism’, synonymous with trivial or unimportant metabolism, at the same time, suggests neutral evolution, free to drift from one structure to the next. This conundrum highlights the importance of better defining evolutionary principles during chemical and biological investigation of natural products.

In this review, we aim at providing basic evolutionary principles as they have been embraced by genome miners interested in natural products-based drug discovery and the development of bioinformatics tools useful for this purpose. We discussed the origins of this sub-discipline (sub-section 1.1), as well as working definitions and core evolutionary and Big Data principles, both generally and specifically regarding evolution-driven genome mining approaches (sub-sections 2.1 and 2.2). We distinguish and highlight selected examples in which the confluence of Big Data and evolutionary genome mining for the discovery of natural products is more evident; and provide information to better understand and efficiently use these tools, but also to prompt newcomers and pave the way for the development of tools embracing the predictive power of the theory of evolution and the wealth of Big Data. Both databases and algorithms with relevant evolutionary features are presented in sub-sections 2.3 and 2.4. Selected examples of NPs research embracing evolutionary thinking - from enzymes to whole microbiomes - are provided in sub-sections 3.1 and 3.2. The selected cases highlight evolutionary thinking and include the few examples that involve tools of what we call evolutionary genome mining of natural products. The final sub-section 4 provides future directions for the development of this emerging sub-discipline as an important area of research to better understand NPs as whole and direct their biotechnological exploitation.

### 1.1 Origins of evolutionary genome mining of natural products

Advances in DNA sequencing have allowed for the study of allelic variation and how it relates to different phenotypes and evolutionary pressures(Wolfe and Li 2003). These genetic investigations have developed into entire fields of molecular and genome evolution research, most notably advancing the areas of population genetics and phylogenetics. Population genetics investigates the frequencies and dynamics of genetic differences in and across populations, aiming to understand how some gene variants are more or less frequent than others(Matthew B. Hamilton 2021). In contrast, phylogenetics seeks to relate gene variants to each other by inferring an evolutionary history that explains differences between both genes and species(Masatoshi Nei and Sudhir Kumar 2000). Indeed, one might argue that phylogenetics was the first molecular biology Big Data method used broadly in biology, and remains so, as it aims to unveil hidden patterns otherwise ambiguous using empirical knowledge alone(Woese, Kandler, and Wheelis 1990). These inferences can be used to predict evolutionary histories through building networks of relatedness (e.g. phylogenetic trees) and reconstructing ancestral states, and therefore, in order to adopt evolutionary theory properly, these frameworks should be

considered when approaching the evolution of NPs, especially when mining large datasets.

While evolutionary frameworks increasingly appear in the study of NPs, the extreme interdisciplinarity of NP research has led to adoption of evolutionary principles at different rates in different subdisciplines, depending on scientific goals and availability of data and the technologies used for their generation and analysis. For example, NP chemists often focus on empirical and mechanistic data to direct future investigations, and by doing so, they reinforce working models of biosynthetic logic in well-studied enzymes, for instance, nonribosomal peptide synthetases (NRPS)(Süssmuth and Mainz 2017) and polyketide synthases (PKS)(Nivina et al. 2019). In contrast, phylogenetics, whether at the species, gene, or genome level, aims to unveil broader patterns and place them into evolutionary context. This is increasingly done for bacterial(Adamek et al. 2018; Gutiérrez-García et al. 2017; Larsen, Pearson, and Neilan 2021), fungal(Bushley and Turgeon 2010; Lind et al. 2017) and plant(Piatkowski et al. 2020; Wilson and Tian 2019) NP biosynthetic enzymes, and even across different taxonomic lineages that produce similar NPs(Jenke-Kodama et al. 2005; Shimizu, Ogata, and Goto 2017). Phylogenetic insights may have limited mechanistic value, but they can assist in posing novel mechanistic hypotheses that can be experimentally tested. The combination of both approaches is embraced by Dean and Thornton's functional synthesis, which proposes that sequence analyses should be coupled with empirical, molecular experiments to retrace the evolutionary histories of biochemical processes and their phenotypes(Dean and Thornton 2007).

In recent years, these two apparently disparate schools of thoughts have converged, yielding new protein evolution theory(DePristo, Weinreich, and Hartl 2005; Pál, Papp, and Lercher 2006) and NP genome-mining applications (Alanjary et al. 2017; Cruz-Morales et al. 2016; Sélem-Mojica et al. 2019). Indeed, the marriage of phylogenies and mechanistic insights, implicit in early protein evolution-rate studies(Alvarez-Ponce 2021), is the essence of evolutionary genome mining of NPs. The genes involved in NP biosynthesis and function, a subset of which have been validated through mechanistic studies, can be used to reconstruct large-scale phylogenies of multiple genes and their proteins. The genetic patterns uncovered by this Big Data approach can then feed back into more mechanistic predictions, providing hypotheses to further validate via new empirical, mechanistic studies. As these patterns can be affected by both evolutionary forces and genetic mechanisms underlying them (in bacteria (Chevrette, Gutiérrez-García, et al. 2020; Jensen 2016), fungi(Drott et al. 2021; Rokas et al. 2020; Rokas, Wisecaver, and Lind 2018) and plants(Moghe and Last 2015; Weng 2014) alike, yet each with their own intricacies) it is of utmost importance that these are clearly defined and appreciated by the natural products community when describing NP evolution.

## 2. Big Data and evolutionary genome mining of natural products: from key concepts to databases and algorithms

Genomic assemblies from DNA sequencing data and a strain's associated phenotypic and/or meta information are the source of Big Data needed for the development of NP evolutionary genome mining databases and applications. This stems from the fact that the interactions between the chemical products of natural product biosynthesis and their molecular targets are shaped by evolutionary processes that control chemical structure, regulation, and/or availability(Chevrette, Gutiérrez-García, et al. 2020). Thus, the enzymes that assemble natural products are subject to these evolutionary pressures as well(Chevrette, Gutiérrez-García, et al. 2020; Chevrette, Hoskisson, and Barona-Gómez 2020). Biosynthesis of natural products is typically a series of incorporating building blocks into a larger structure and adding stepwise chemical modifications. Precursors may be sourced from other parts of metabolism, the environment, or synthesized within the biosynthetic gene cluster itself (Chevrette, Gutiérrez-García, et al. 2020; Sélem-Mojica et al. 2019). Some biosynthesis belong to large macromolecular machinery, like NRPS (Süssmuth and Mainz 2017) or PKSs(Nivina et al. 2019), while others are single domain enzymes(Chevrette, Hoskisson, et al. 2020). BGCs can be as simple as a few genes or as complex as many dozens of genes whose encoded enzymes work in concert to produce the final product(s). The enzymes at work within natural product biosynthesis are as diverse and varied as the chemical structures they biosynthesize, the molecular targets with which they engage, and the interactions within and between species that they mediate. Taking this context into account, we next define evolutionary and Big Data key concepts as the foundations of evolutionary genome mining of natural products databases and algorithms.

### 2.1. Key Big Data concepts in Natural Products research

Big Data refers to datasets that fit four major criteria: volume, velocity, variety, and validation. First, volume: Big Data must be big(Megahed and Jones-Farmer 2015). This typically refers to having many different entries or examples or replicates, depending on your data type. The distinction between "normal" datasets and Big Data is an ever-changing definition: what is considered Big Data today will likely not be Big Data in the future. This is mainly due to scientific breakthroughs leading to technological improvements and data generation. Second, velocity: Big Data grows quickly, which is mainly prompted by technological advances. A useful example of volume and velocity is shown in Figure 1, highlighting the growth (volume) of genomes in NCBI over time (velocity). Third, variety: Big Data typically has several layers of information, which will be discussed below specifically for NP research. Finally, validation: a Big Data approach is only as good as its training data, so ensuring that training information is verified in some way is necessary for confidence in making forward predictions and identifying patterns. While validation is not strictly required for a dataset to be considered "Big", applications will have limited value if they are based on unverified information. This may sound fairly obvious yet is

something that needs to be explicitly stated. Gene annotations are a common example where validation becomes very important: comparing your gene of interest to a validated dataset (e.g. UniProt, SwissProt) yields classifications that are much higher confidence than if you were to compare to unvalidated datasets (e.g. NCBI-NR) where the annotations of the dataset itself are unvalidated and errors can compound (Barona-Gómez 2015).

As datasets grow bigger (volume) at faster rates (velocity), an unvalidated dataset made up only of predictions may have misannotations. These errors can lead to many more subsequent misannotations, which themselves can further exacerbate these errors (Cahan et al. 2019). Thus, understanding the level of validation for your dataset is necessary to properly interpret your results. Together, these four Vs present analysis challenges, as Big Data is often too large or complex such that non-traditional or parallel computing tools are needed for analysis with ad hoc algorithms (Jin et al. 2015; Marx 2013). In general, for a natural products researcher in the early 2020s, data becomes ‘Big Data’ when it is too large or too complex to do simple statistics in spreadsheet-based software (e.g. Microsoft Excel). These data, moreover, are hard to process and visualize with available tools within tolerable computing times.

Standard genome mining approaches to uncover NP biosynthesis have been used to explore a wide range of taxa and environments, identifying “microbial dark matter” as a promising source of hidden chemical treasures. In evolutionary genome mining of NPs this becomes an essential consideration with potentially confounding factors. As shown in Figure 1, the first two ‘Vs’, volume and velocity, are currently covered by the sequence data in large databases. In NP research, however, data is not limited to genetics, but it has many other layers, including chemical, gene expression, ecological, and evolutionary data. For instance, the MIBiG(Medema et al. 2015) data repository is a good example of ‘variety’, in that it includes multifaceted chemical and genetic data. It also has a high standard of validation, as the level of validation is listed for each entry. These advantages come at the cost of volume and velocity: keeping the standards of variety and validation high mean that this repository grows at slower rates than for example the NCBI genome database. Important to evolutionary genome mining, MIBiG and other repositories tend to be biased towards a limited number of taxa that have been investigated in great detail, like species of the genus Aspergillus in fungi(Drott et al. 2021; Lind et al. 2017) or Streptomyces(AbuSara et al. 2019; Barka et al. 2016; Belknap et al. 2020; Doroghazi and Metcalf 2013; Navarro-Muñoz et al. 2020) within the Actinobacteria. While a bias towards this bacterial genus clearly exists, this issue is slowly decreasing with other genera such as Nocardia(Männle et al. 2020), Amycolatopsis(Adamek et al. 2018), Salinispora(Ziemert et al. 2014), Micromonospora(Hifnawy et al. 2020), Pseudonocardia(Goldstein and Klassen 2020), Rhodococcus,(Schorn et al. 2016),(Agustina Undabarrena et al. 2021) etc. emerging as promising NP producers. Yet, bias in sampling remains a critical

consideration in evolutionary studies as they can confound results and sometimes lead to erroneous conclusions, as argued recently in the case of Aspergillus(Drott et al. 2021).

In summary, Big Data available for evolutionary studies and genome mining of natural products come from several sources, including both broad and specialized chemical and genetic databases (see Tables 1 and 2). As an example, NCBI database contains over 1.4 million bacterial and over 38 thousand archaeal samples at the writing of this manuscript, with data existing as either genomes, transcriptomes, or metagenomes. These data however are far from being informative into NP research unless they are organized and/or translated into other forms or layers of information and analyzed with suitable tools. Based on our own experience, Big Data for natural products research today implies algorithms fast enough to conveniently analyse the genomes and/or metabolomes of over 30 thousand strains or samples. These numbers will rapidly multiply in the future, and thus it is critical to continually reassess “natural classifications” seen in evolutionary relationships, keeping in mind that sampling bias of training data remains a fundamental, yet often overlooked, issue. Scalability of tools is also a consideration. For example, multiple sequence alignments and phylogenies of hundreds or thousands of genes was once considered Big Data, and remains so, yet now we can perform phylogenomic comparisons across entire kingdoms of life on an inexpensive laptop computer or free public web server(Alanjary et al. 2017; Sélem-Mojica et al. 2019). This scalability of datasets and analysis tools can provide the genetic context necessary to perform evolutionary genome mining.

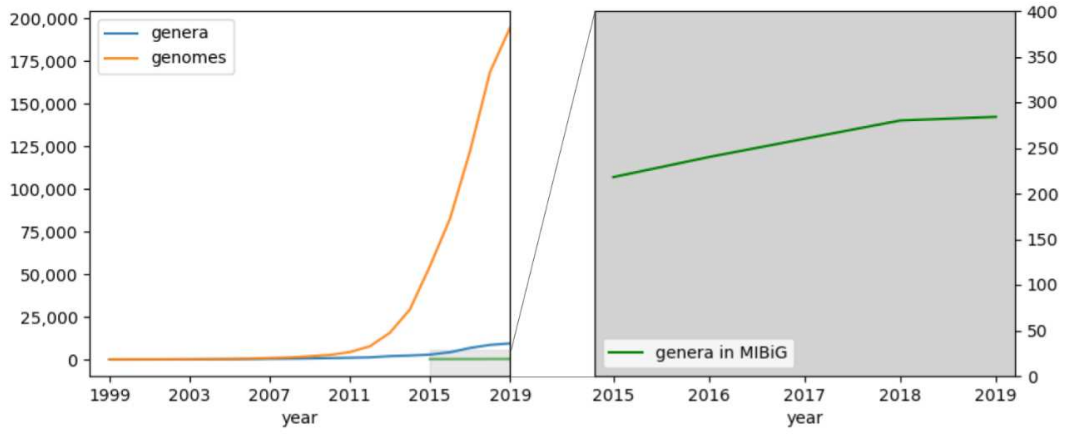



**Figure 1**. Growth of the number of NCBI genomes (bacteria and archaea) and genera per year from 1999 to 2019. Data from GTDB (release 95). Inset: number of genera represented by data in MIBiG.

### 2.2. Key evolutionary concepts in NP research

Genomic assemblies from DNA sequencing data and a strain’s associated phenotypic and/or meta information are the source of Big Data needed for the development of NP evolutionary genome mining databases and applications. This

stems from the fact that the interactions between the chemical products of natural product biosynthesis and their molecular targets are shaped by evolutionary processes that control chemical structure, regulation, and/or availability(Chevrette, Gutiérrez-García, et al. 2020). Thus, the enzymes that assemble natural products are subject to these evolutionary pressures as well(Chevrette, Gutiérrez-García, et al. 2020; Chevrette, Hoskisson, et al. 2020). Biosynthesis of natural products is typically a series of incorporating building blocks into a larger structure and adding stepwise chemical modifications. Precursors may be sourced from other parts of metabolism, the environment, or synthesized within the biosynthetic gene cluster itself (Chevrette, Gutiérrez-García, et al. 2020; Sélem-Mojica et al. 2019). Some biosynthesis belong to large macromolecular machinery, like NRPSs(Süssmuth and Mainz 2017) or PKSs(Nivina et al. 2019), while others are single domain enzymes(Chevrette, Hoskisson, et al. 2020). BGCs can be as simple as a few genes or as complex as many dozens of genes whose encoded enzymes work in concert to produce the final product(s). The enzymes at work within natural product biosynthesis are as diverse and varied as the chemical structures they biosynthesize, the molecular targets with which they engage, and the interactions within and between species that they mediate.

Evolutionary pressures that drive the appearance and that overall shape the physicochemical and biomolecular features of natural products biosynthesis, can be incredibly dynamic and complex. Nevertheless, overarching principles of evolution of NP enzymes and/or pathways emerge. Just as biochemical principles (e.g. adenylation (A) domain specificity of NRPSs or chain elongation during PKS-catalyzed synthesis) are mechanistically fundamental for the understanding of NP biosynthesis, the following broad evolutionary principles, with a mechanistic bearing, can be considered:

(i) Enzyme promiscuity drives pathway evolution through genetic expansion-and-recruitment events, providing the building blocks to assemble, shuffle, and combine NP biosynthetic pathways(Khersonsky and Tawfik 2010; Noda-Garcia, Liebermeister, and Tawfik 2018; Noda-Garcia and Tawfik 2020) .

(ii) Once enzymes (or domains) are recruited into NP biosynthesis, they tend to cluster together as multidomain megasynthases and/or biosynthetic gene clusters (BGC)(Chevrette, Gutiérrez-García, et al. 2020; Jensen 2016; Rokas et al. 2018).

These two corollaries are valid across bacteria(Chevrette, Gutiérrez-García, et al. 2020; Dittmann et al. 2015; Navarro-Muñoz et al. 2020; Sélem-Mojica et al. 2019), fungi ref, and plants(Fan et al. 2020; Liu, Cheema, et al. 2020; Liu, Duran, et al. 2020; Weng 2014) within their unique physiological, morphological, and chromosomal peculiarities. They also hold across different taxonomic lineages that share homologous NP biosynthetic enzymes(Montalbán-López et al. 2021; Tang et al. 2017). It is starting to be widely appreciated that the phenomena from which these corollaries derive can occur under strong positive selection, but growing

evidence and theory suggests a key role for negative selection and neutral forces on BGC dynamics(Chevrette, Gutiérrez-García, et al. 2020). Once recombination events cluster enzymes together, either as multidomain enzymes or BGCs, the resulting pathways can recruit other auxiliary elements, such as regulators, domain-domain interactors, transporters, and importantly, resistance genes(Chevrette, Hoskisson, et al. 2020). As these principles were comprehensively demonstrated n the last decade or so, they were exploited by researchers for the development of the four main evolutionary genome mining tools that the NP community has used to identify and investigate novel pathways: (i) EvoMining(Cruz-Morales et al. 2016; Sélem-Mojica et al. 2019), (ii) ARTS(Alanjary et al. 2017; Mungan et al. 2020) (iii) BiG-SCAPE(Navarro-Muñoz et al. 2020) and (iv) CORASON(Navarro-Muñoz et al. 2020). These tools are placed into the Big Data context and discussed in further detail in sub-section 2.4.

Using phylogenetics to unveil the evolutionary patterns of NPs follows two main approaches. On the one hand, gene trees can be used to infer a gene's evolutionary history and provide evidence for past events that have led to present-day data (i.e. branches or leaves of the tree). For evolutionary genome mining, gene trees can be useful in identifying expansions (e.g. duplications) and subsequent diversification of biosynthetic genes of interest. On the other hand, species trees describe the reconstructed evolutionary history of a set of species or individuals, and thus are useful for identifying larger-scale evolutionary events(Nakhleh 2013). Critically assessing how the topologies of genes and species agree and disagree can shed light on important evolutionary events, such as horizontal transfers(Avni and Snir 2020). While NP research is focused on BGCs (a collection of genes), much can be learned from studying single-gene and species trees. Understanding the distribution and evolution of NPs within taxa, for example, is a prerequisite for effective sampling and bioprospecting strategies.

For those interested in evolutionary genome mining of NPs, it is important to note that the above mentioned approaches are the result of properly embracing phylogenetics and evolutionary principles, often implementing concepts and principles not typically studied by NP chemists. Figure 2 shows the main concepts that those interested in the use and development of these tools should take into account. As mentioned, the main two evolutionary mechanisms driving the appearance of novel NP biosynthetic pathways are diversification (enzyme promiscuity and BGC dynamics) and selection (positive, negative and neutral). However, it is only when these forces combine and impact the fitness of the NP-producing organism that pathways are assembled and reassembled during the course of evolution(Chevrette, Hoskisson, et al. 2020). The main genetic mechanisms driving these evolutionary events have been identified and have been used in the development of NP evolutionary genome-mining tools (thicker arrows, Figure 2). Hocwever, much remains to be deciphered regarding the evolution of NPs, especially in terms of their expression and function in the real environmental settings

of their producing organisms, where fitness operates. Study cases are available (see sub-section 3), but their scarcity makes them anecdotal and thus more data is needed to develop mining tools based on Big Data principles to investigate this layer of complexity (thinner and/or dashed arrows, Figure 2).

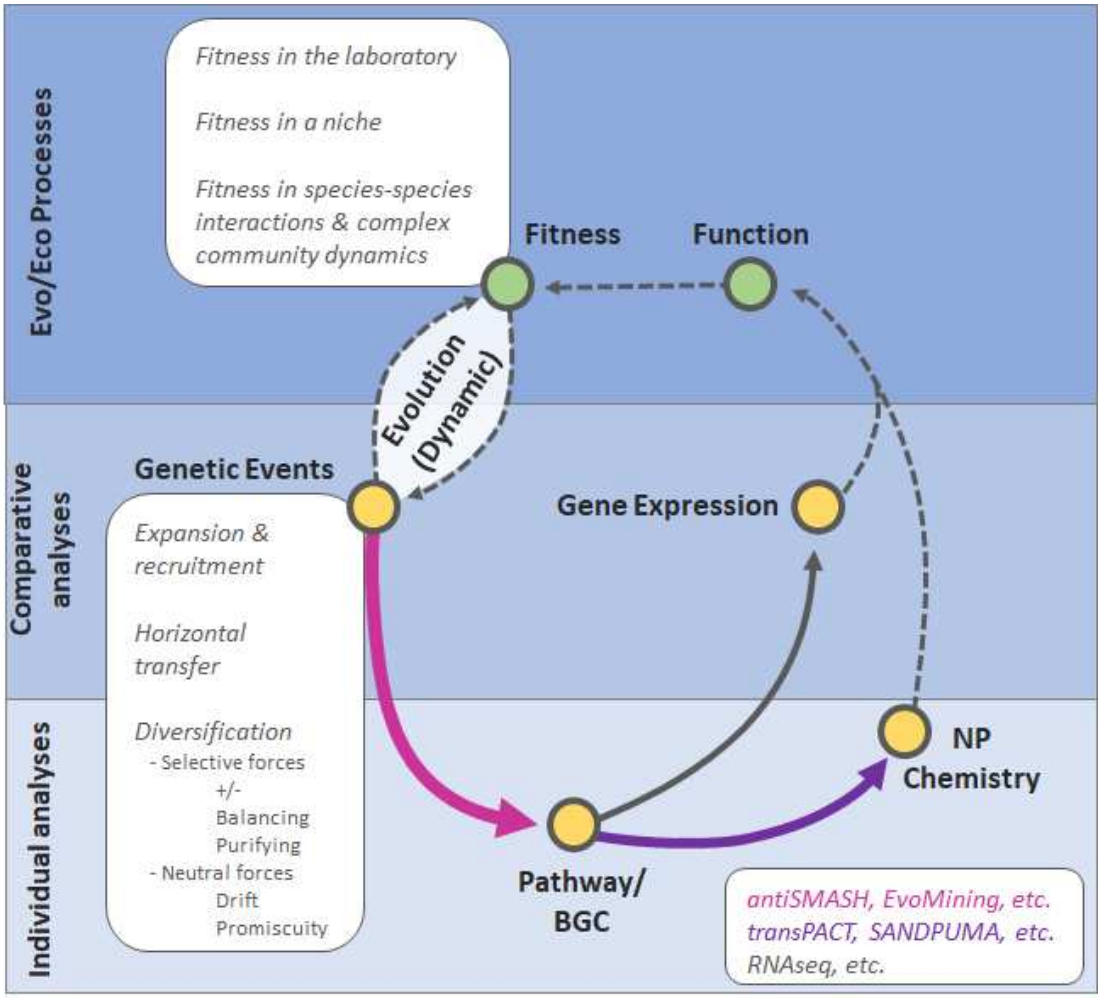



**Figure 2**. Evolutionary genome mining of natural products in a concept-driven framework. Studies on the evolutionary histories of NPs, their biosynthetic genes, and their producing organisms are driven by analyses at different levels of organization. Individual analyses (bottom) focus on a Pathway/BGC and their molecular product(s) or chemistry. Examples of tools that predict NP chemistry from BGCs are shown in purple. These individual data can then be contextualized with comparative analyses (middle) across many conditions or strains/species, with an emphasis in the genetic events underlying the evolution of NPs BGCs. One example is Gene Expression studies (gray, RNAseq) where comparisons of transcriptional patterns can place genes in a broader biological context. Analyses at the level of ecological and/or evolutionary processes (top) are the most challenging, and as a field we have only just begun to understand how Gene Expression, BGC, NP chemistry, and other "lower-level" data contribute to molecular function, and in turn how function contributes to an organism's fitness (linked by dotted lines to highlight that there are not yet standardized methods, but there is opportunity to develop them integrating Big Data). This remains a major challenge, as fitness is often a function of the environment. Evolution occurs as a dynamic process in which the fitness impact of a BGC's product influences the BGCs genetic components (e.g. diversification, selection, and other processes; see box). These in turn can feed back into fitness. Previously characterized

genes and/or patterns of genetic events can then be used to identify and characterize BGCs de novo from genomic data (pink), either through rules-based or evolutionary methods.

### 2.3 NP databases (training sets) available for NP evolutionary genome mining

As mentioned, data available for investigating natural products in the Big Data era comes from several sources. However, this information only becomes useful when organized on databases that can be coupled with metadata of the organisms themselves, but also with information about the technology and methods used to generate the data. Examples of well-executed databases include the GNPS mass spectra public database(Wang et al. 2016), the MIBiG repository with experimentally validated datasets(Kautsar et al. 2019; Medema et al. 2015), and the bioinformatically predicted BGCs of the antiSMASH DB (Blin et al. 2017, 2021) (Tables 1 and 2). Recently, the first evolutionary database, i.e. ActDES, which is specific for the Actinobacteria, has been reported(Schniete et al. 2021). All of these databases, despite complying with the four 'Vs' in one way or another, including variety, are useful in comparative or evolutionary studies, but not sufficient as none of them provide a comprehensive multi-layer database including or embracing evolution. In turn, at this stage, it is down to the evolutionary genome miner to select and integrate the most suitable and relevant DBs from those provided in Tables 1 and 2, within a phylogenomics framework. Selected DBs are highlighted throughout this review with the aim of emphasising their value in relation to the four 'Vs'.

### 2.4 Big Data and NP evolutionary genome mining algorithms

Communication between evolutionary biologists, computer scientists and mathematicians has historically led to biological insight, including the developments of population genetics theory and the transition matrices that are key to common genomic search algorithms like BLAST(Altschul et al. 1990). These disciplines have successfully converged again in recent years for the development of sophisticated NP genome-mining algorithms and platforms (Table 3). In this subsection, we list and explain major evolutionary genome mining of NPs approaches available to date with a focus on those that directly or indirectly rely on the use of the theory of evolution in any of its forms, either within the algorithms themselves or in their visualizations. The availability of genomic data (e.g. MIBiG, CARD, antiSMASH DB, Table 1) is fundamental, but probably more often will also be inputs from purely chemical DBs (Table 2), e.g. GNPS, Paired Omics Data Platform [PODP], which can also serve as training data in supervised algorithms. Notably, some of these genomic-based algorithms already include input from chemical databases (Kim et al. 2019; van Santen et al. 2019; Wang et al. 2016). Thus, the integration of data types, as in MIBiG or PODP, may provide training datasets with valuable links between genomic and chemical data, further embracing variety. This integration holds great promise and value to the field, but since it is only beginning to occur, it remains to be seen how regularly chemical data will be embraced by evolution-driven genome mining efforts.

**Table 1.** Genomic databases to explore natural products diversity and evolution.

| Database Name | Parameter Name | Parameter Value | Current Version (date) * |
|---|---|---|---|
| MIBiG (Kautsar et al. 2019) | BGCs | 1,923 | 2.0 (2019) |
| IMG-ABC (Palaniappan et al. 2019) | BGCs | 410,683 | 5.0 |
| antiSMASH-db (Blin et al. 2021) | BGCs | 147,517 | 3.0 |
| BiG-FAM (Kautsar, Blin, et al. 2021) | BGCs | 1,225,071 | 1.0 |
| NCBI Genome | Bacteria spp. | 278,820 | November 2020 |
| | Archaea spp. | 5,625 | November 2020 |
| | Eukaryote spp. | 14,486 | November 2020 |
| MGnify (Mitchell et al. 2020) | Metagenomes | 32,746 | November 2020 |
| IMG/M (Nayfach et al. 2020) | MAGs | 52,515 | November 2020 |
| | BGCs | 104,211 | November 2020 |
| CARD (Alcock et al. 2020) | Alleles | 213,809 | February 2021 |
| | Reference sequences | 3,146 | February 2021 |
| SRA (Bacteria) | Datasets | 1,466,494 | November 2020 |
| SRA (Archaea) | Datasets | 38,592 | November 2020 |
| NCBI WGS (Bacteria) | Projects | 941,266 | December 2020 |
| NCBI WGS (Archaea) | Projects | 6,225 | December 2020 |
| Resfinder 4.0 (Bortolaia et al. 2020) | Antibiotic resistance genes | 2,690 | December 2020 |
| MG-RAST 4.0.3 (Meyer et al. 2008) | Metagenome | 447,497 | January 2021 |

NB.- Most of these listed databases arguably satisfy the Big Data characteristics of volume and variety. Since there have been only few periodic releases for some of these databases, the velocity characteristics of Big Data can be appreciated for only a few of these. * The month and year (date) of each database, when last accessed, are provided. Exact dates for “current versions” are not provided as these were not available.

**Table 2.** Chemical databases to explore natural products diversity and evolution.

| Database Name | Parameter Name | Parameter Value | Current version (date) |
|---|---|---|---|
| MACADAM (Boulch et al, 2019) | Metabolites | 7,921 | 1 |
| PubChem (Kim et al, 2019) | Compounds | 111,456,896 | November 2020 |
| GNPS (Wang et al, 2016) | NP Compounds | 18,163 | 1 |
| | Spectra | 221,083 | 1 |
| NPAtlas (van Santen et al, 2019) | Compounds | 24,594 | v 2020_06 |
| COCONUT (Sorokina et al, 2021) | Compounds | 406,747 | March 2021 |
| StreptomeDB (Klementz et al, 2015) | Compounds | 4,000 | 2 |
| PoDP (Schorn et al, 2021) | Paired (meta)genomes and metabolomes | 4,853 | 2021 GitHub v0.9.2 |
| Siderophore DB () | Compounds | 262 | June 2021 |
| LOTUS (Rutz et al, 2022) | NP Compounds | 276,518 | February 2021 |

Currently, evolutionary genome-mining for the discovery of novel NPs(Hoskisson and Seipke 2020) aims to provide answers to two main questions, and by doing so, generate predictions: (i) which genes and/or BGCs produce metabolites not typically associated with central metabolism? and (ii) which genes or domains specific to a lineage represent innovation and diversification compared to ancestral states? As mentioned, several specialty databases (Table 1 & 2) are available and are used by the main evolutionary genome mining tools that the NP community has used to identify and investigate novel pathways: (i) EvoMining(Cruz-Morales et al. 2016; Sélem-Mojica et al. 2019), (ii) ARTS(Alanjary et al. 2017; Mungan et al. 2020) (iii) BiG-SCAPE(Navarro-Muñoz et al. 2020) and (iv) CORASON(Navarro-Muñoz et al. 2020). Following a similar rationale, a conceptual framework for mining siderophore BGCs based on their transporters has recently been reported(Crits-Christoph et al. 2020). Importantly, available tools can be used independently or in combination, and go in hand with species-level phylogenetic analyses which directly integrate NP biosynthesis (e.g. AutoMLST(Alanjary, Steinke, and Ziemert 2019)) or analyses that are part of more generalized phylogenetic pipelines(Adamek, Alanjary, and Ziemert 2019). The combination of the latter, i.e. a species tree, with large-scale BGC prediction and their taxonomic distribution, is BiG-SLiCE output 20

**Table 3.** Big Data algorithms for exploring natural products diversity and evolution.

| Algorithm Name | Validation dataset | Type of data | Method | Publication date |
|---|---|---|---|---|
| ARTS 2.0 (Mungan et al. 2020) | Bacterial kingdom genomes and metagenomes | Genomes | Duplication and BGC proximity, Phylogeny and resistance screen | May 2020 |
| BiG-SCAPE (Navarro-Muñoz et al. 2020) | Clusters from ~3,000 genomes | BGCs | Jaccard Index plus Maximum Likelihood FastTree | November 2019 |
| EvoMining 2.0 (Sélem-Mojica et al. 2019) | ~100 conserved families from ~1,000 genomes | Biosynthetic genes | Duplication and gene proximity to MIBiG, Phylogeny | December 2019 |
| BiG-SLICE (Kautsar, van der Hooft, et al. 2021) | BiG-FAM (1,225,071) | BGCs | Balanced Iterative Reducing and Clustering using Hierarchies | August 2020 |
| CORASON (Navarro-Muñoz et al. 2020) | ~3,000 | Genomes or BGCs (visualization) | Blast plus FastTree | November 2019 |
| clinker (Gilchrist and Chooi 2021) | NA | BGCs (visualization) | Hierarchical clustering | January 2021 |
| FlaGs (Saha et al. 2020) | 324 | BGCs (visualization) | BGC's Hidden Markov Model | September 2020 |
| TREND (Gumerov and Zhulin 2020) | NA | BGCs (visualization) | Hierarchical clustering | April 2020 |
| MicroReact (Argimón et al. 2016) | NA | Trees with metadata (visualization) | libraries:Chart.js, Leaflet, Phylocanvas, React, Sigma | November 2016 |
| Anvi'o (Eren et al. 2015) | NA | Pangenomes (visualization) | Hidden Markov Models | October 2015 |

Supervised algorithms make use of the DBs mentioned in the previous sub-section in the form of training sets with validated labels about what is an NP BGC and what is not36. Here, the “correct” classifications are known for training data and used to make predictions about new data. These methods typically require heavy (and often manual) curation of training sets, and thus the importance of the fourth V, validation. So far, most of NP research adopting genome mining approaches employs supervised algorithms, mainly used in classification problems that require prior knowledge(Bzdok, Krzywinski, and Altman 2018). Unsupervised algorithms, instead aim to extract patterns and trends from unlabeled data(Yang and Ersoy 2003), similar to phylogenies. These can be helpful to identify data features (e.g. genes and domains) that are important for categorization, but since no “true” answer is known false-positive errors may be more frequent. Clustering or other grouping methods used in unsupervised methods attempt to give some structure to a dataset. Typically, supervised and unsupervised strategies are complementary, as it is the case in NP evolutionary genome-mining (Figure 3).

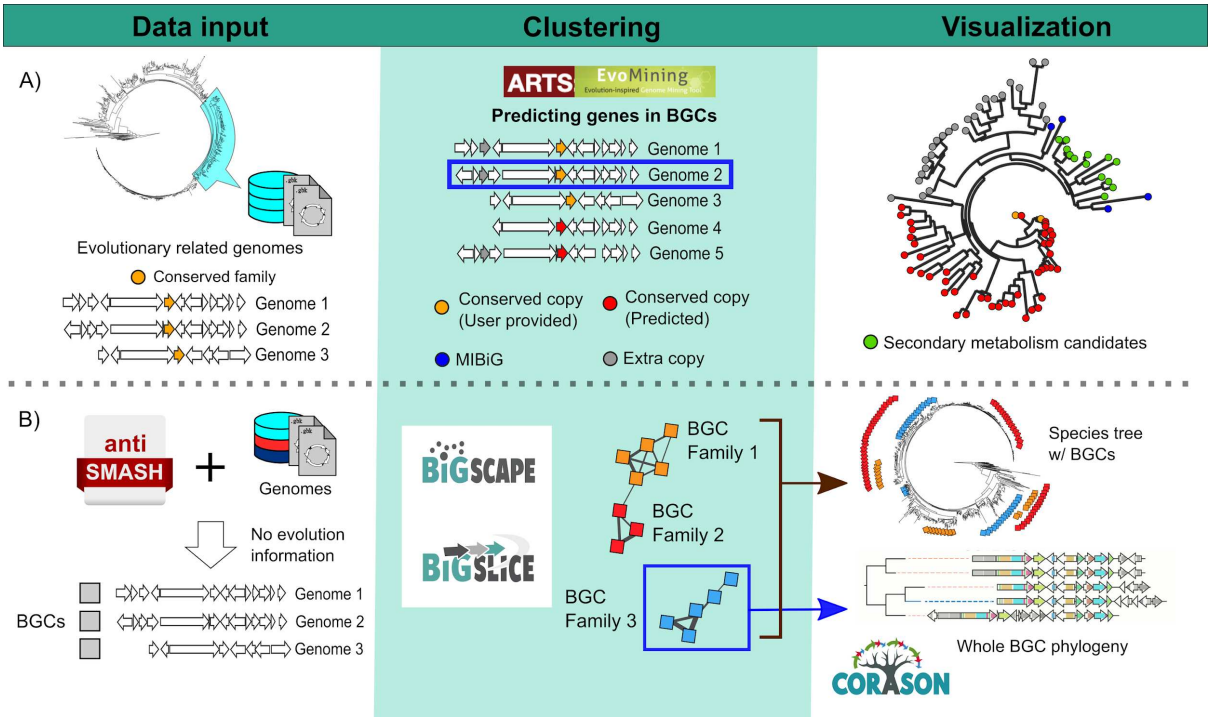



**Figure 3**. Evolution-driven genome mining tools. A. Evolutionary algorithms need as inputs genomes from taxonomically related lineages, where conserved protein families (orange) are selected for further exploration (ARTS/EvoMining). Conserved (orange and red) and extra (gray) copies of these families are identified and compared by a phylogenetic distance against proteins from NP databases (blue). Finally, the tree used in the phylogenetic distance is provided as a visualization, where predictions are included (green). B. Algorithms with an evolutionary visualizationbut without evolutionary driven distancesdoes not restrict their input genomes to be phylogenetically related. Gene clusters obtained from these algorithms are gathered in gene cluster families (GCF) by classification methods. Finally, evolutionary visualizations can be provided, either as a whole-BGC network of phylogenetic tree (BiG-SCAPE/CORASON) or as the occurrence of each GCF throughout a species tree (BiG-SLICE).

Within NP research, supervised problems are used to identify and classify domains, genes, and BGCs. ClusterFinder(Cimermancic et al. 2014) was one of the first algorithms that attempted to classify regions of the genome as NP BGC (or not) by calculating a moving average of a “biosynthetic score”, calculated based on domain- and gene-level agreement with profile Hidden Markov Models of biosynthetic enzymes. Although ClusterFinder does not directly leverage evolutionary theory in its algorithm, it is indirectly inferring the evolutionary processes that shaped BGC regions throughout the genome. Many of these algorithms have been trained primarily (or exclusively) on bacterial data, and thus accurate and reliable identification of fungal BGCs remains a challenge. Fortunately, recent work has begun to take fungal-specific genes and genetic structure into account to address this issue(Argimón et al. 2016; van der Lee and Medema 2016; Wolf et al. 2016). A similar scenario in plants(Kautsar et al. 2017) has now been encountered since the realization that BGCs actually exist in this large and prominent group of NP producing organisms.

Identifying shared and novel features within and between taxonomic lineages is attempted by unsupervised algorithms, such as BiG-SCAPE, BiG-SLICE and CORASON. For example, BiG-SCAPE, and more recently BiG-SLICE, clusters BGCs into gene cluster families (GCFs) without requiring prior knowledge of these families. This is done after calculating distance scores between BGCs on the basis of shared protein families and BGC organization. After clustering, it can be useful to sort and/or connect these GCFs with each other into bigger “clans”, that are related but more distantly so than members of the same GCF. This broader context can be used to track evolutionary events of related BGCs and investigate how these events are distributed across gene and/or strain phylogenies. An alternative-yet-complementary approach employed by CORASON involves phylogenetic trees of shared enzymatic features, including in some instances whole-BGCs phylogenies. Importantly, these processes use supervised classifications of genes and domains to perform unsupervised clustering into GCFs, so they too require high quality (i.e. validated, or at least carefully curated) genomic and chemical databases.

In contrast, EvoMining and ARTS, represent the first (and to our knowledge, thus far the only) heuristic algorithms that incorporate evolutionary thinking as part of the supervised approach itself, attempting to infer what is central metabolism and what may be secondary metabolism, with a certain degree of diversification hinting towards the appearance of an specialized pathway. Evolution is inferred as a distance metric, which can be seen as similar to a support vector machine algorithm(Kloosterman et al. 2020; Krause et al. 2007; Walker and Clardy 2021), but implemented using a tree to determine appropriate groupings (and thus classifications) for biosynthetic enzymes. Put in another way, it seeks to identify which query enzymes cluster more closely with central metabolism and which cluster more closely with secondary or specialized metabolism. Extra gene copies are

assessed by EvoMining as potential recruitments into NP biosynthesis, and these gene families may differ from one taxonomic lineage to another (Figure 3A).

After classification into BGC families (e.g. with BiG-SLICE and/or BiG-SCAPE), further evolutionary context can be added in the visualization stage with CORASON according to the phylogenetic history of genes within the BGC or the strain-level phylogeny of the producing organism itself. In turn, CORASON identifies gene clusters in a genomes database and sorts them according to their evolutionary relationships. Tools such as MicroReact(Argimón et al. 2016) can also allow for visual exploration of large phylogenetic trees annotated with metadata. EvoMining and ARTS both start with labeled sets (genes that are either the primary copy or specialized metabolism copies that belong to other databases, e.g. CARD/MIBiG) and employ supervised methods where evolutionary distance is used to classify putative BGCs. As a consequence their predictions are intuitively displayed phylogenetically. Other software suites that perform pangenomic visualization (e.g. Anvio (Eren et al. 2015)) are also useful in that they allow identification of families with potential gene expansion and/or recruitment events. Many recent tools aim to sort and visualize relations between BGCs: for example, MultiGeneBlast(Medema, Takano, and Breitling 2013) (implemented in antiSMASH), finds gene homologs in BGC comparisons. Given otherwise identified BGCs (e.g. by antiSMASH or other tools), BiG-SCAPE(Navarro-Muñoz et al. 2020) can classify them into BGC families and other visualization tools such as clinker(Gilchrist and Chooi 2021), FlaGs(Saha et al. 2020) and TREND(Gumerov and Zhulin 2020) allow for interactive visualizations (Figure 3B).

## 3. Genomic and enzymatic evolution of Natural Products

### 3.1 Evolution of the genome of NP-producing organisms

Multiple studies have been conducted on the evolution of NP producers, providing useful indications for targeted bioprospecting. Biosynthetic potential and diversity appear to be related to the ecological niche of the producers, as was confirmed in multiple instances (Caldera et al. 2019; Chevrette, Carlson, et al. 2019; Chevrette and Currie 2019; Gutiérrez-García et al. 2019; Iglesias et al. 2020; Miller, Chevrette, and Kwan 2017; Sharrar et al. 2020; Silva et al. 2019; Stubbendieck et al. 2019; Yang et al. 2019). In some cases though, phylogeny is more important, as observed in microbial taxa where secondary metabolism is most similar in closely related organisms rather than those isolated from the same source(Chevrette, Carlos-Shanley, et al. 2019; Silva et al. 2019). Such investigations showcase possible promising targets for NP research, be they specific known(Chevrette, Carlos-Shanley, et al. 2019; Gutiérrez-García et al. 2019) ref or understudied taxa(Gutiérrez-García et al. 2019; Schorn et al. 2016; Silva et al. 2019) or different environments/niches(Caldera et al. 2019; Chevrette and Currie 2019; Iglesias et al. 2020; Sharrar et al. 2020; Stubbendieck et al. 2019). As such, it is clear therefore

that evolution can be applied for the discovery of novel natural products, which can powerful if properly embraced.

Comparative genomic analyses have shown that most bacterial taxa harbor only a few BGCs while some dedicate a large proportion of their genomes to specialized or secondary metabolism (Adamek et al. 2019; Brito et al. 2020; Chevrette, Carlson, et al. 2019; Chevrette and Currie 2019; Doroghazi et al. 2014; Hoffmann et al. 2018; Männle et al. 2020; Miller et al. 2017; Sharrar et al. 2020; Silva et al. 2019; Yang et al. 2019).The quantity and diversity of BGC content differs among the taxa, with extreme cases reported(Caldera et al. 2019; Ziemert et al. 2014). How disperse the phylogenetic distribution of a BGC is, can allude to the various effects selection has had on its related pathways(Gluck-Thaler et al. 2020). Most notably, horizontal gene transfer (HGT) is a relatively frequent phenomenon in BGCs, which is one likely explanation for their extended distribution across distant taxa and their observed diversity (Adamek et al. 2018; Baldeweg, Hoffmeister, and Nett 2019; Brito et al. 2020; Chevrette, Carlson, et al. 2019; Chevrette, Gutiérrez-García, et al. 2020; Chevrette and Currie 2019; Koonin 2015; Medema et al. 2014; Miller et al. 2017; Vior et al. 2018; Wang et al. 2016). While HGT is observed frequently in BGCs compared to other genetic elements, it is important to note that the evolutionary timescales involved are still quite large(Chevrette, Gutiérrez-García, et al. 2020; Chevrette and Currie 2019; McDonald and Currie 2017) and depend on both population structure and genetic identity of donor and recipient(Chevrette, Gutiérrez-García, et al. 2020; Chevrette and Currie 2019; McDonald and Currie 2017). Vertical inheritance of BGCs within the same lineage is the dominant means through which biosynthetic information is transferred(Chase et al. 2021; Chevrette, Gutiérrez-García, et al. 2020). This is a key distinction that should be made when studying the evolution of BGCs, as the more subtle vertical evolutionary dynamics happen from generation to generation, while HGT events are typically observed at timescales closer to thousands, millions, or billions of years.

Thus far, all analyses mentioned in this subsection were not conducted on a Big Data scale. Indeed, the information discovered so far is being confirmed by multiple independent inquiries, yet still issues of small taxonomic coverage and sampling biases remain. In 2014, three articles were published that followed a more global approach to NP producer genomics. Cimermancic(Cimermancic et al. 2014) and co-authors analysed more than 1000 genomes from across the bacterial kingdom and created a "global map" of biosynthesis, encompassing ~33,000 predicted BGCs. Doroghazi(Doroghazi and Metcalf 2013) and co-authors focused on one phylum and, using different metrics and methods than Cimermancic, reached a similar conclusion by collecting information on the producers capacity and potential. At the same time, Medema(Medema et al. 2014) and co-authors examined a large number of known BGCs and proved that the rates of evolutionary events within such units are much higher than in clusters of primary metabolism. Since these studies were first published, the available data has multiplied and so too have the methods for

processing them; more universal-scope analyses will soon follow and give the answers to questions that remain open, including how and when biosynthetic diversity evolved(Hoffmann et al. 2018) or the capacity of nature to keep providing us with new compounds(Bérdy 2005).

The above mentioned studies have focused on microbes that have been cultured under laboratory conditions. However, the number of unculturable organisms is vast and metagenomic analyses have begun to unravel their hidden biosynthetic potential, indicating promising new sources for NP bioprospecting (see next paragraph). Furthermore, investigating evolutionary patterns based on environmental samples can shed light on the functions of the NPs found in nature as well as their raison d'etre within their microcosm(Traxler and Kolter 2015). This is important as NP evolution occurs at the population level, as highlighted by recent examples where population genomics frameworks have been adopted to mine NPs in genomic data, both in fungi and bacteria(Andam et al. 2016; Caldera et al. 2019; Drott et al. 2021; Li et al. 2019; McDonald et al. 2019; Tidjani et al. 2019). Such approaches have even proven valuable at the bacterial colony-level of a domesticated model laboratory strain, i.e. Streptomyces coelicolor(Zacharia et al. 2021; Zhang et al. 2020).

Soil metagenomic surveys in urban greenspaces, grassland meadows, and areas covering up to continent-wide scale have reported microbial diversity patterns(Bahram et al. 2018; Crits-Christoph et al. 2020; Delgado-Baquerizo et al. 2018; Thompson et al. 2017; Wang et al. 2018) . These patterns are drastically affected by the environment and massive sequencing efforts are required to comprehensively capture their diversity, even at kilometer scale. High throughput functional studies involving creation of large-insert metagenomic libraries provides a novel approach to examine the functional and phylogenetic diversity of sampled ecosystems(Handelsman et al. 1998; Nasrin et al. 2018; Santana-Pereira et al. 2020). Economically attractive approaches using amplicon sequencing have been used to prove the domain-level diversity of environmental NPs. Such approaches have provided clues to answer the long standing question of which sites should be surveyed to maximize the discovery of novel natural products (Crits-Christoph et al. 2020; Dror et al. 2020; Elfeki et al. 2018; Lemetre et al. 2017; Reddy et al. 2012; Sharrar et al. 2020; Wang et al. 2016). Massive amounts of shotgun metagenomic data are already easily available from public repositories. Analyzing these Big Data to infer significant NP patterns has now become the next bottleneck and faster algorithms and easy to use tools are badly required to mine the potential resource. Additionally, detailed documentation, standardized sampling procedures, and still more metadata are required to be incorporated into public databases in order to exploit patterns and extract useful information.

### 3.2. BGC and multidomain enzyme evolution

The evolutionary history of BGCs can be studied by building separate and/or concatenated trees of their genes and protein products. These can have very different topologies than the species trees of the NP producers themselves, suggesting unconventional sequence transmission events, such as Horizontal Gene Transfer (see previous section), gene conversion, intra-genomic recombination(Medema et al. 2014), and others. Together, these trees and functional information of NP genes can be used as a foundation to predict the activity of yet-unknown compounds and suggest potential links between fitness and the evolutionary forces at work.

Natural products exhibit extremely diverse chemistry. Their evolutionary complexity is no less complex. Domains evolve in the context of genes, genes in the context of BGCs, and BGCs in the context of their the producers' genomes(Chevrette, Gutiérrez-García, et al. 2020; Waglechner, McArthur, and Wright 2019). Further, how these metabolites contribute to the fitness of their producing organisms depends largely on their environmental niche, which is often completely unknown or has poorly-understood factors and boundaries(Firn and Jones 2003). Because of this interdependence between multiple levels of organization, evolution does not affect clusters uniformly(Medema et al. 2014). Indicatively, trans-acyltransferase (trans-AT) AT domains have evolved independently from cis-AT AT domains: the latter cluster into NP-specific clades and are known to be acquired vertically, while the prior are present in many different phyla and appear to be transferred horizontally(Nguyen et al. 2008). Based on the clades formed in trans-AT AT and KS trees, it appears their evolution is strongly linked to their elongation substrate specificities (Chevrette and Currie 2019; Masschelein, Jenner, and Challis 2017; Medema et al. 2014; Nguyen et al. 2008). Indeed, computational pipelines such as transPACT(Chevrette,Marc and Helfrich 2021) place KS sequence information into a phylogenetic framework to predict substrate specificity for unknown sequences. cis-AT and trans-AT PKS variants can produce similar metabolites even though they have distinct evolutionary histories. This case of evolution may be influenced by the modularity of Type I PKS clusters that can be more plastic due to intragenic recombinations and may allow for adaptability in a wide range of ecological niches (Nguyen et al. 2008)

Although much of NP evolution is thought of at the level of BGCs or genes, important evolutionary changes can also happen at even smaller scales. Substrate specificity of different NP enzymes is often dictated by the three dimensional organization of their active sites and/or protein-protein interaction surfaces, so subtle changes to the protein sequence of these areas can steer specificity (and promiscuity) in multiple evolutionary directions. In some cases these changes correlate with phylogeny, so knowledge of the evolutionary mechanisms behind BGCs can allow for collecting reliable information from domain phylogeny. NRPS domains also show evolutionary patterns linking phylogeny and chemistry(Nguyen et al. 2008). Similar to the trans-AT KS domains of the PKS clusters, A-domains of NRPSs cluster into clades according to substrate specificity, while C-domains are highly conserved and follow a

BGC-specific pattern (Chevrette and Currie 2019; Jenke-Kodama et al. 2005; Medema et al. 2014). Computational methods such as SANDPUMA(Chevrette et al. 2017) and others have used this phylogenetic signal to reliably predict the substrate specificity of A-domains. Recently, “Substrate level” evolutionary signals, like in trans-AT KS and NRPS A-domains, can be used to predict substrate specificity, while “pathway level” evolutionary signals, like in NRPS C-domains can be used to predict BGC-level patterns of similar molecules(Ziemert et al. 2014).

## 4. What lies ahead? Needs and opportunities for evolutionary genome mining of NPs.

Evolutionary genome mining of natural products in the Big Data era has inherited the tradition of phylogenetics, in the sense that natural history coupled with genetic and chemical observations can provide mechanistic insight. With this heritage comes the promise of discovering “The Known Unknowns, Unknown Knowns, and Unknown Unknowns of Secondary Metabolism”, which has important implications in gene expression and the distinctions between “cryptic” and “silent” BGCs.(Hoskisson and Seipke 2020) Although genomic and metabolomic speciality databases have made considerable progress, we envisage an ever-growing need for novel speciality datasets merging different layers of information. A promising current endeavor is the assemblage of metabologenomics databases, where genetic information and predictions are merged with chemical data (e.g. Paired Omics database(Schorn et al. 2021)). Nevertheless, the systematic inclusion of other data types, including evolutionary relationships, remains a challenge. One notable evolutionary database has been recently released for Actinobacteria 24, but those with larger scale and broader taxonomic coverage are much needed. These high-variety databases promise new insights in the NP field as a whole. Similarly, the accompanying algorithms needed to efficiently compute high volume datasets will allow us to perform these analyses at scale and keep pace with the technological advances that generate data at high velocity. In the near future we expect these data to go beyond only genomes, metabolomes, and metagenomes and begin to encompass ecological and functional metadata(Tracanna et al. 2021).

Biosynthetic enzyme domains are the focus of current, and likely future, algorithms. This presents unique challenges for enzyme families whose classifications are problematic and/or understudied in the community. For instance, chemists have provided insights into why sequence-based phylogenies are insufficient for certain enzymes: transition-state intermediaries can be highly reactive and plastic, and therefore sequence space is less constrained than in enzymes with well-defined active sites(Austin, O’Maille, and Noel 2008). Examples of this include the terpene cyclases, cytochrome P450s, hydrolases and type III polyketide synthases, amongst others. In these examples, analyses could benefit from alternative methods to establish relationships useful to provide classification and dataset structure. In turn, this may provide more informative training sets within well-structured databases, increasing the quality of predictions surrounding these important classes of natural

products biosynthetic enzymes. It should be noted that classification of some of these enzymes within abovementioned DBs, such as antiSMASH DB, does not necessarily mean that this problem has been sorted out (see validation; previous sections). Pangenomic analyses(Ding, Baumdicker, and Neher 2018; Eren et al. 2015) to identify expanded enzyme families within lineages may provide an interesting possibility to classify enzyme families on evolutionary grounds.

Here, by reviewing the nascent history of evolutionary genome mining of natural products as a sub-discipline, it has become apparent that a prerequisite for the development of successful algorithms is the realization and characterization of genetic events driving the evolution of biosynthetic enzymes in their genomic context (e.g. BGCs). As such, we highlight the following evolutionary concepts with the promise to link evolution to genetic and chemical mechanisms. It has become clearer that “natural” evolution of natural products can be governed by dynamic processes that result in functional replacements. For example, in convergent evolution of chemically related scaffolds with diverse biomolecular activities(Grenade, Howe, and Ross 2021), whose biosynthesis is directed by non-related BGCs that produce functionally similar molecules. It has also become clearer that biosynthetic pathways can be encoded by physically unrelated loci (in contrast to BGCs), which may consist of sub-clusters(Del Carratore et al. 2019), and that the same BGC can produce diverse natural products with different biological functions in response to the environmental conditions(Martinet et al. 2019). This intragenomic cross-talk might be seen as a simplified version of the metabolic exchange between different organisms within a microbiome, for which evolutionary experimental and conceptual frameworks have been developed (Cibrián-Jaramillo and Barona-Gómez 2016; Gutiérrez-García et al. 2019; Wiegand et al. 2020). Both levels of metabolic cross-talk represent an immanent Big Data challenge: to genomically mine large datasets to correlate physically unlinked loci and propose metabolic relationships(Nayfach et al. 2020; Sharrar et al. 2020) How to best embrace evolutionary processes, many of which we are only beginning to understand, in Big Data genome mining for natural products remains an exciting yet challenging endeavor; one that will surely provide many possibilities for the future of this emerging sub-discipline.

## References

AbuSara, Nader F., Brandon M. Piercey, Marcus A. Moore, Arshad Ali Shaikh, Louis-Félix Nothias, Santosh K. Srivastava, Pablo Cruz-Morales, Pieter C. Dorrestein, Francisco Barona-Gómez, and Kapil Tahlan. 2019. «Comparative Genomics and Metabolomics Analyses of Clavulanic Acid-Producing Streptomyces Species Provides Insight Into Specialized Metabolism». Frontiers in Microbiology 10:1-17. doi: 10.3389/fmicb.2019.02550.

Adamek, Martina, Mohammad Alanjary, Helena Sales-Ortells, Michael Goodfellow, Alan T. Bull, Anika Winkler, Daniel Wibberg, Jörn Kalinowski, and Nadine Ziemert. 2018. «Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species». BMC Genomics 19(1):426. doi: 10.1186/s12864-018-4809-4.

Adamek, Martina, Mohammad Alanjary, and Nadine Ziemert. 2019. «Applied Evolution: Phylogeny-Based Approaches in Natural Products Research». Natural Product Reports 36(9):1295-1312. doi: 10.1039/C9NP00027E.

Agustina Undabarrena, Ricardo Valencia, Andrés Cumsille, Leonardo Zamora-Leyva, Eduardo Castro- Nallar, Francisco Barona- Gomez, and Beatriz Cámara. 2021. «Rhodococcus comparative genomics reveals a phylogenomic- dependent non- ribosomal peptide synthetase distribution: insights into biosynthetic gene cluster connection to an orphan metabolite». doi: DOI 10.1099/mgen.0.000621.

Alanjary, Mohammad, Brent Kronmiller, Martina Adamek, Kai Blin, Tilmann Weber, Daniel Huson, Benjamin Philmus, and Nadine Ziemert. 2017. «The Antibiotic Resistant Target Seeker (ARTS), an Exploration Engine for Antibiotic Cluster Prioritization and Novel Drug Target Discovery». Nucleic Acids Research 45(W1):W42-48. doi: 10.1093/nar/gkx360.

Alanjary, Mohammad, Katharina Steinke, and Nadine Ziemert. 2019. «AutoMLST: An Automated Web Server for Generating Multi-Locus Species Trees Highlighting Natural Product Potential». Nucleic Acids Research 47(W1):W276-82. doi: 10.1093/nar/gkz282.

Alcock, Brian P., Amogelang R. Raphenya, Tammy T. Y. Lau, Kara K. Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V. Nguyen, Annie A. Cheng, Sihan Liu, Sally Y. Min, Anatoly Miroshnichenko, Hiu-Ki Tran, Rafik E. Werfalli, Jalees A. Nasir, Martins Oloni, David J. Speicher, Alexandra Florescu, Bhavya Singh, Mateusz Faltyn, Anastasia Hernandez-Koutoucheva, Arjun N. Sharma, Emily Bordeleau, Andrew C. Pawlowski, Haley L. Zubyk, Damion Dooley, Emma Griffiths, Finlay Maguire, Geoff L. Winsor, Robert G. Beiko, Fiona S. L. Brinkman, William W. L. Hsiao, Gary V. Domselaar, and Andrew G. McArthur. 2020. «CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database». Nucleic Acids Research 48(D1):D517-25. doi: 10.1093/nar/gkz935.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. «Basic Local Alignment Search Tool». Journal of Molecular Biology 215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2.

Alvarez-Ponce, David. 2021. «Richard Dickerson, Molecular Clocks, and Rates of Protein Evolution». Journal of Molecular Evolution 89(3):122-26. doi: 10.1007/s00239-020-09973-x.

Andam, Cheryl P., Mallory J. Choudoir, Anh Vinh Nguyen, Han Sol Park, and Daniel H. Buckley. 2016. «Contributions of Ancestral Inter-Species Recombination to the Genetic Diversity of Extant Streptomyces Lineages». The ISME Journal 10(7):1731-41. doi: 10.1038/ismej.2015.230.

Argimón, Silvia, Khalil Abudahab, Richard J. E. Goater, Artemij Fedosejev, Jyothish Bhai, Corinna Glasner, Edward J. Feil, Matthew T. G. Holden, Corin A. Yeats, Hajo Grundmann, Brian G. Spratt, and David M. Aanensen. 2016. «Microreact: visualizing and sharing data for genomic epidemiology and phylogeography». Microbial Genomics 2(11):e000093. doi: 10.1099/mgen.0.000093.

Austin, Michael B., Paul E. O'Maille, and Joseph P. Noel. 2008. «Evolving Biosynthetic Tangos Negotiate Mechanistic Landscapes». Nature Chemical Biology 4(4):217-22. doi: 10.1038/nchembio0408-217.

Avni, Eliran, and Sagi Snir. 2020. «A New Phylogenomic Approach For Quantifying Horizontal Gene Transfer Trends in Prokaryotes». Scientific Reports 10(1):12425. doi: 10.1038/s41598-020-62446-5.

Bahram, Mohammad, Falk Hildebrand, Sofia K. Forslund, Jennifer L. Anderson, Nadejda A. Soudzilovskaia, Peter M. Bodegom, Johan Bengtsson-Palme, Sten Anslan, Luis Pedro Coelho, Helery Harend, Jaime Huerta-Cepas, Marnix H. Medema, Mia R. Maltz, Sunil Mundra, Pål Axel Olsson, Mari Pent, Sergei Põlme, Shinichi Sunagawa, Martin Ryberg, Leho Tedersoo, and Peer Bork.

2018. «Structure and Function of the Global Topsoil Microbiome». Nature 560(7717):233-37. doi: 10.1038/s41586-018-0386-6.

Baldeweg, Florian, Dirk Hoffmeister, and Markus Nett. 2019. «A Genomics Perspective on Natural Product Biosynthesis in Plant Pathogenic Bacteria». Natural Product Reports 36(2):307-25. doi: 10.1039/C8NP00025E.

Barka, Essaid Ait, Parul Vatsa, Lisa Sanchez, Nathalie Gaveau-Vaillant, Cedric Jacquard, Jan P. Meier-Kolthoff, Hans-Peter Klenk, Christophe Clément, Yder Ouhdouch, and Gilles P. van Wezel. 2016. «Taxonomy, Physiology, and Natural Products of Actinobacteria». Microbiology and Molecular Biology Reviews: MMBR 80(1):1-43. doi: 10.1128/MMBR.00019-15.

Barona-Gómez, Francisco. 2015. «Re-Annotation of the Sequence > Annotation: Opportunities for the Functional Microbiologist». Microbial Biotechnology 8(1):2-4. doi: 10.1111/1751-7915.12242.

Belknap, Kaitlyn C., Cooper J. Park, Brian M. Barth, and Cheryl P. Andam. 2020. «Genome Mining of Biosynthetic and Chemotherapeutic Gene Clusters in Streptomyces Bacteria». Scientific Reports 10(1):2003. doi: 10.1038/s41598-020-58904-9.

Bérdy, János. 2005. «Bioactive Microbial Metabolites». The Journal of Antibiotics 58(1):1-26. doi: 10.1038/ja.2005.1.

Blin, Kai, Marnix H. Medema, Renzo Kottmann, Sang Yup Lee, and Tilmann Weber. 2017. «The AntiSMASH Database, a Comprehensive Database of Microbial Secondary Metabolite Biosynthetic Gene Clusters». Nucleic Acids Research 45(D1):D555-59. doi: 10.1093/nar/gkw960.

Blin, Kai, Simon Shaw, Satria A. Kautsar, Marnix H. Medema, and Tilmann Weber. 2021. «The AntiSMASH Database Version 3: Increased Taxonomic Coverage and New Query Features for Modular Enzymes». Nucleic Acids Research 49(D1):D639-43. doi: 10.1093/nar/gkaa978.

Bortolaia, Valeria, Rolf S. Kaas, Etienne Ruppe, Marilyn C. Roberts, Stefan Schwarz, Vincent Cattoir, Alain Philippon, Rosa L. Allesoe, Ana Rita Rebelo, Alfred Ferrer Florensa, Linda Fagelhauer, Trinad Chakraborty, Bernd Neumann, Guido Werner, Jennifer K. Bender, Kerstin Stingl, Minh Nguyen, Jasmine Coppens, Basil Britto Xavier, Surbhi Malhotra-Kumar, Henrik Westh, Mette Pinholt, Muna F. Anjum, Nicholas A. Duggett, Isabelle Kempf, Suvi Nykäsenoja, Satu Olkkola, Kinga Wieczorek, Ana Amaro, Lurdes Clemente, Joël Mossong, Serge Losch, Catherine Ragimbeau, Ole Lund, and Frank M. Aarestrup. 2020. «ResFinder 4.0 for predictions of phenotypes from genotypes». Journal of Antimicrobial Chemotherapy 75(12):3491-3500. doi: 10.1093/jac/dkaa345.

Brito,  ngela, Jorge Vieira, Cristina P. Vieira, Tao Zhu, Pedro N. Leão, Vitor Ramos, Xuefeng Lu, Vitor M. Vasconcelos, Muriel Gugger, and Paula Tamagnini. 2020. «Comparative Genomics Discloses the Uniqueness and the Biosynthetic Potential of the Marine Cyanobacterium Hyella Patelloides». Frontiers in Microbiology 11(1527):1-15. doi: 10.3389/fmicb.2020.01527.

Bushley, Kathryn E., and B. Gillian Turgeon. 2010. «Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships». BMC Evolutionary Biology 10(1):26. doi: 10.1186/1471-2148-10-26.

Bzdok, Danilo, Martin Krzywinski, and Naomi Altman. 2018. «Machine Learning: Supervised Methods». Nature Methods 15(1):5-6. doi: 10.1038/nmeth.4551.

Cahan, Eli M., Tina Hernandez-Boussard, Sonoo Thadaney-Israni, and Daniel L. Rubin. 2019. «Putting the Data before the Algorithm in Big Data Addressing Personalized Healthcare». Npj Digital Medicine 2(1):1-6. doi: 10.1038/s41746-019-0157-2.

Caldera, Eric J., Marc G. Chevrette, Bradon R. McDonald, and Cameron R. Currie. 2019. «Local Adaptation of Bacterial Symbionts within a Geographic Mosaic of Antibiotic Coevolution». Applied and Environmental Microbiology 85(24):e01580-19. doi: 10.1128/AEM.01580-19.

Chase, Alexander B., Douglas Sweeney, Mitchell N. Muskat, Dulce Guillén-Matus, and Paul R. Jensen. 2021. «Vertical Inheritance Governs Biosynthetic Gene Cluster Evolution and Chemical Diversification». BioRxiv 2020.12.19.423547. doi: 10.1101/2020.12.19.423547.

Chevrette, Marc G., Fabian Aicheler, Oliver Kohlbacher, Cameron R. Currie, and Marnix H. Medema. 2017. «SANDPUMA: Ensemble Predictions of Nonribosomal Peptide Chemistry Reveal Biosynthetic Diversity across Actinobacteria». Bioinformatics 33(20):3202-10. doi: 10.1093/bioinformatics/btx400.

Chevrette, Marc G., Camila Carlos-Shanley, Katherine B. Louie, Benjamin P. Bowen, Trent R. Northen, and Cameron R. Currie. 2019. «Taxonomic and Metabolic Incongruence in the Ancient Genus Streptomyces». Frontiers in Microbiology 10(2170):1-12. doi: 10.3389/fmicb.2019.02170.

Chevrette, Marc G., Caitlin M. Carlson, Humberto E. Ortega, Chris Thomas, Gene E. Ananiev, Kenneth J. Barns, Adam J. Book, Julian Cagnazzo, Camila Carlos, Will Flanigan, Kirk J. Grubbs, Heidi A. Horn, F. Michael Hoffmann, Jonathan L. Klassen, Jennifer J. Knack, Gina R. Lewin, Bradon R. McDonald, Laura Muller, Weilan G. P. Melo, Adrián A. Pinto-Tomás, Amber Schmitz, Evelyn Wendt-Pienkowski, Scott Wildman, Miao Zhao, Fan Zhang, Tim S. Bugni, David R. Andes, Monica T. Pupo, and Cameron R. Currie. 2019. «The Antimicrobial Potential of Streptomyces from Insect Microbiomes». Nature Communications 10(1):516. doi: 10.1038/s41467-019-08438-0.

Chevrette, Marc G., and Cameron R. Currie. 2019. «Emerging evolutionary paradigms in antibiotic discovery». Journal of Industrial Microbiology and Biotechnology 46(3-4):257-71. doi: 10.1007/s10295-018-2085-6.

Chevrette, Marc G., Karina Gutiérrez-García, Nelly Selem-Mojica, César Aguilar-Martínez, Alan Yañez-Olvera, Hilda E. Ramos-Aboites, Paul A. Hoskisson, and Francisco Barona-Gómez. 2020. «Evolutionary Dynamics of Natural Product Biosynthesis in Bacteria». Natural Product Reports 37(4):566-99. doi: 10.1039/C9NP00048H.

Chevrette, Marc G., Paul A. Hoskisson, and Francisco Barona-Gómez. 2020. «Enzyme Evolution in Secondary Metabolism». Pp. 90-112 en Comprehensive Natural Products III. Elsevier.

Chevrette,Marc, and Helfrich. 2021. «transPACT v1.0». bioRxiv. doi: 10.5281/zenodo.4148258.

Cibrián-Jaramillo, Angélica, and Francisco Barona-Gómez. 2016. «Increasing Metagenomic Resolution of Microbiome Interactions Through Functional Phylogenomics and Bacterial Sub-Communities». Frontiers in Genetics 7:1-8. doi: 10.3389/fgene.2016.00004.

Cimermancic, Peter, Marnix H. Medema, Jan Claesen, Kenji Kurita, Laura C. Wieland Brown, Konstantinos Mavrommatis, Amrita Pati, Paul A. Godfrey, Michael Koehrsen, Jon Clardy, Bruce W. Birren, Eriko Takano, Andrej Sali, Roger G. Linington, and Michael A. Fischbach. 2014. «Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters». Cell 158(2):412-21. doi: 10.1016/j.cell.2014.06.034.

Crits-Christoph, Alexander, Nicholas Bhattacharya, Matthew R. Olm, Yun S. Song, and Jillian F. Banfield. 2020. «Transporter Genes in Biosynthetic Gene Clusters Predict Metabolite Characteristics and Siderophore Activity». Genome Research 31(2):239-50. doi: 10.1101/gr.268169.120.

Cruz-Morales, Pablo, Johannes Florian Kopp, Christian Martínez-Guerrero, Luis Alfonso Yáñez-Guerra, Nelly Selem-Mojica, Hilda Ramos-Aboites, Jörg Feldmann, and Francisco Barona-Gómez. 2016. «Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters

Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes». Genome Biology and Evolution 8(6):1906-16. doi: 10.1093/gbe/evw125.

Dean, Antony M., and Joseph W. Thornton. 2007. «Mechanistic approaches to the study of evolution». Nature reviews. Genetics 8(9):675-88. doi: 10.1038/nrg2160.

Del Carratore, Francesco, Konrad Zych, Matthew Cummings, Eriko Takano, Marnix H. Medema, and Rainer Breitling. 2019. «Computational Identification of Co-Evolving Multi-Gene Modules in Microbial Biosynthetic Gene Clusters». Communications Biology 2(1):1-10. doi: 10.1038/s42003-019-0333-6.

Delgado-Baquerizo, Manuel, Angela M. Oliverio, Tess E. Brewer, Alberto Benavent-González, David J. Eldridge, Richard D. Bardgett, Fernando T. Maestre, Brajesh K. Singh, and Noah Fierer. 2018. «A Global Atlas of the Dominant Bacteria Found in Soil». Science 359(6373):320-25. doi: 10.1126/science.aap9516.

DePristo, Mark A., Daniel M. Weinreich, and Daniel L. Hartl. 2005. «Missense Meanderings in Sequence Space: A Biophysical View of Protein Evolution». Nature Reviews Genetics 6(9):678-87. doi: 10.1038/nrg1672.

Ding, Wei, Franz Baumdicker, and Richard A. Neher. 2018. «panX: pan-genome analysis and exploration». Nucleic Acids Research 46(1):e5-e5. doi: 10.1093/nar/gkx977.

Dittmann, Elke, Muriel Gugger, Kaarina Sivonen, and David P. Fewer. 2015. «Natural Product Biosynthetic Diversity and Comparative Genomics of the Cyanobacteria». Trends in Microbiology 23(10):642-52. doi: 10.1016/j.tim.2015.07.008.

Doroghazi, James R., Jessica C. Albright, Anthony W. Goering, Kou-San Ju, Robert R. Haines, Konstantin A. Tchalukov, David P. Labeda, Neil L. Kelleher, and William W. Metcalf. 2014. «A Roadmap for Natural Product Discovery Based on Large-Scale Genomics and Metabolomics». Nature Chemical Biology 10(11):963-68. doi: 10.1038/nchembio.1659.

Doroghazi, James R., and William W. Metcalf. 2013. «Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes». BMC Genomics 14(1):611. doi: 10.1186/1471-2164-14-611.

Dror, Barak, Zongqiang Wang, Sean F. Brady, Edouard Jurkevitch, and Eddie Cytryn. 2020. «Elucidating the Diversity and Potential Function of Nonribosomal Peptide and Polyketide Biosynthetic Gene Clusters in the Root Microbiome». MSystems 5(6):e00866-20. doi: 10.1128/mSystems.00866-20.

Drott, Milton T., Tomás A. Rush, Tatum R. Satterlee, Richard J. Giannone, Paul E. Abraham, Claudio Greco, Nandhitha Venkatesh, Jeffrey M. Skerker, N. Louise Glass, Jesse L. Labbé, Michael G. Milgroom, and Nancy P. Keller. 2021. «Microevolution in the Pansecondary Metabolome of Aspergillus Flavus and Its Potential Macroevolutionary Implications for Filamentous Fungi». Proceedings of the National Academy of Sciences 118(21). doi: 10.1073/pnas.2021683118.

Elfeki, Maryam, Mohammad Alanjary, Stefan J. Green, Nadine Ziemert, and Brian T. Murphy. 2018. «Assessing the Efficiency of Cultivation Techniques To Recover Natural Product Biosynthetic Gene Populations from Sediment». ACS Chemical Biology 13(8):2074-81. doi: 10.1021/acschembio.8b00254.

Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. «Anvi'o: An Advanced Analysis and Visualization Platform for 'omics Data». PeerJ 3:e1319. doi: 10.7717/peerj.1319.

Fan, Pengxiang, Peipei Wang, Yann-Ru Lou, Bryan J. Leong, Bethany M. Moore, Craig A. Schenck, Rachel Combs, Pengfei Cao, Federica Brandizzi, Shin-Han Shiu, and Robert L. Last. 2020. «Evolution of a plant gene cluster in Solanaceae and emergence of metabolic diversity» editado por D. J. Kliebenstein, C. S. Hardtke, and R. Peters. eLife 9:e56717. doi: 10.7554/eLife.56717.

Firn, Richard D., and Clive G. Jones. 2003. «Natural Products ? A Simple Model to Explain Chemical Diversity». Natural Product Reports 20(4):382. doi: 10.1039/b208815k.

Gilchrist, Cameron L. M., and Yit-Heng Chooi. 2021. «clinker & clustermap.js: automatic generation of gene cluster comparison figures». Bioinformatics (btab007):btab007. doi: 10.1093/bioinformatics/btab007.

Gluck-Thaler, Emile, Sajeet Haridas, Manfred Binder, Igor V. Grigoriev, Pedro W. Crous, Joseph W. Spatafora, Kathryn Bushley, and Jason C. Slot. 2020. «The Architecture of Metabolism Maximizes Biosynthetic Diversity in the Largest Class of Fungi». Molecular Biology and Evolution 37(10):2838-56. doi: 10.1093/molbev/msaa122.

Goldman, Aaron D., and David A. Liberles. 2021. «The Journal of Molecular Evolution Turns 50». Journal of Molecular Evolution 89(3):119-21. doi: 10.1007/s00239-021-10000-w.

Goldstein, Sarah L., and Jonathan L. Klassen. 2020. «Pseudonocardia Symbionts of Fungus-Growing Ants and the Evolution of Defensive Secondary Metabolism». Frontiers in Microbiology 11(621041):1-8. doi: 10.3389/fmicb.2020.621041.

Grenade, Neil L., Graeme W. Howe, and Avena C. Ross. 2021. «The Convergence of Bacterial Natural Products from Evolutionarily Distinct Pathways». Current Opinion in Biotechnology 69:17-25. doi: 10.1016/j.copbio.2020.10.009.

Gumerov, Vadim M., and Igor B. Zhulin. 2020. «TREND: A Platform for Exploring Protein Function in Prokaryotes Based on Phylogenetic, Domain Architecture and Gene Neighborhood Analyses». Nucleic Acids Research 48(W1):W72-76. doi: 10.1093/nar/gkaa243.

Gutiérrez-García, Karina, Edder D. Bustos-Díaz, José Antonio Corona-Gómez, Hilda E. Ramos-Aboites, Nelly Sélem-Mojica, Pablo Cruz-Morales, Miguel A. Pérez-Farrera, Francisco Barona-Gómez, and Angélica Cibrián-Jaramillo. 2019. «Cycad Coralloid Roots Contain Bacterial Communities Including Cyanobacteria and Caulobacter Spp. That Encode Niche-Specific Biosynthetic Gene Clusters». Genome Biology and Evolution 11(1):319-34. doi: 10.1093/gbe/evy266.

Gutiérrez-García, Karina, Adriana Neira-González, Rosa Martha Pérez-Gutiérrez, Giovana Granados-Ramírez, Ramon Zarraga, Kazimierz Wrobel, Francisco Barona-Gómez, and Luis B. Flores-Cotera. 2017. «Phylogenomics of 2,4-Diacetylphloroglucinol-Producing Pseudomonas and Novel Antiglycation Endophytes from Piper auritum». Journal of Natural Products 80(7):1955-63. doi: 10.1021/acs.jnatprod.6b00823.

Handelsman, Jo, Michelle R. Rondon, Sean F. Brady, Jon Clardy, and Robert M. Goodman. 1998. «Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products». Chemistry & Biology 5(10):R245-49. doi: 10.1016/S1074-5521(98)90108-9.

Hifnawy, Mohamed S., Mohamed M. Fouda, Ahmed M. Sayed, Rabab Mohammed, Hossam M. Hassan, Sameh F. AbouZid, Mostafa E. Rateb, Alexander Keller, Martina Adamek, Nadine Ziemert, and Usama Ramadan Abdelmohsen. 2020. «The Genus Micromonospora as a Model Microorganism for Bioactive Natural Product Discovery». RSC Advances 10(35):20939-59. doi: 10.1039/D0RA04025H.

Hoffmann, Thomas, Daniel Krug, Nisa Bozkurt, Srikanth Duddela, Rolf Jansen, Ronald Garcia, Klaus Gerth, Heinrich Steinmetz, and Rolf Müller. 2018. «Correlating Chemical Diversity with Taxonomic Distance for Discovery of Natural Products in Myxobacteria». Nature Communications 9(1):803. doi: 10.1038/s41467-018-03184-1.

Hoskisson, Paul A., and Ryan F. Seipke. 2020. «Cryptic or Silent? The Known Unknowns, Unknown Knowns, and Unknown Unknowns of Secondary Metabolism». mBio 11(5):e02642-20. doi: 10.1128/mBio.02642-20.

Iglesias, Alba, Adriel Latorre-Pérez, James E. M. Stach, Manuel Porcar, and Javier Pascual. 2020. «Out of the Abyss: Genome and Metagenome Mining Reveals Unexpected Environmental Distribution of Abyssomicins». Frontiers in Microbiology 11. doi: 10.3389/fmicb.2020.00645.

Jenke-Kodama, Holger, Axel Sandmann, Rolf Müller, and Elke Dittmann. 2005. «Evolutionary Implications of Bacterial Polyketide Synthases». Molecular Biology and Evolution 22(10):2027-39. doi: 10.1093/molbev/msi193.

Jensen, Paul R. 2016. «Natural Products and the Gene Cluster Revolution». Trends in Microbiology 24(12):968-77. doi: 10.1016/j.tim.2016.07.006.

Jin, Xiaolong, Benjamin W. Wah, Xueqi Cheng, and Yuanzhuo Wang. 2015. «Significance and Challenges of Big Data Research». Big Data Research 2(2):59-64. doi: 10.1016/j.bdr.2015.01.006.

Kautsar, Satria A., Kai Blin, Simon Shaw, Jorge C. Navarro-Muñoz, Barbara R. Terlouw, Justin J. J. van der Hooft, Jeffrey A. van Santen, Vittorio Tracanna, Hernando G. Suarez Duran, Victòria Pascal Andreu, Nelly Selem-Mojica, Mohammad Alanjary, Serina L. Robinson, George Lund, Samuel C. Epstein, Ashley C. Sisto, Louise K. Charkoudian, Jérôme Collemare, Roger G. Linington, Tilmann Weber, and Marnix H. Medema. 2019. «MIBiG 2.0: A Repository for Biosynthetic Gene Clusters of Known Function». Nucleic Acids Research gkz882. doi: 10.1093/nar/gkz882.

Kautsar, Satria A., Kai Blin, Simon Shaw, Tilmann Weber, and Marnix H. Medema. 2021. «BiG-FAM: The Biosynthetic Gene Cluster Families Database». Nucleic Acids Research 49(D1):D490-97. doi: 10.1093/nar/gkaa812.

Kautsar, Satria A., Hernando G. Suarez Duran, Kai Blin, Anne Osbourn, and Marnix H. Medema. 2017. «plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters». Nucleic Acids Research 45(W1):W55-63. doi: 10.1093/nar/gkx305.

Kautsar, Satria A., Justin J. J. van der Hooft, Dick de Ridder, and Marnix H. Medema. 2021. «BiG-SLiCE: A Highly Scalable Tool Maps the Diversity of 1.2 Million Biosynthetic Gene Clusters». GigaScience 10(1):giaa154. doi: 10.1093/gigascience/giaa154.

Khersonsky, Olga, and Dan S. Tawfik. 2010. «Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective». Annual Review of Biochemistry 79:471-505. doi: 10.1146/annurev-biochem-030409-143718.

Kim, Sunghwan, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. 2019. «PubChem 2019 Update: Improved Access to Chemical Data». Nucleic Acids Research 47(D1):D1102-9. doi: 10.1093/nar/gky1033.

Kloosterman, Alexander M., Peter Cimermancic, Somayah S. Elsayed, Chao Du, Michalis Hadjithomas, Mohamed S. Donia, Michael A. Fischbach, Gilles P. van Wezel, and Marnix H. Medema. 2020. «Expansion of RiPP Biosynthetic Space through Integration of Pan-Genomics and Machine

Learning Uncovers a Novel Class of Lanthipeptides». PLOS Biology 18(12):e3001026. doi: 10.1371/journal.pbio.3001026.

Koonin, Eugene V. 2015. «Archaeal ancestors of eukaryotes: not so elusive any more». BMC Biology 13(84):1-7. doi: 10.1186/s12915-015-0194-5.

Krause, Lutz, Alice C. McHardy, Tim W. Nattkemper, Alfred Pühler, Jens Stoye, and Folker Meyer. 2007. «GISMO—gene identification using a support vector machine for ORF classification». Nucleic Acids Research 35(2):540-49. doi: 10.1093/nar/gkl1083.

Larsen, Joachim Steen, Leanne Andrea Pearson, and Brett Anthony Neilan. 2021. «Genome Mining and Evolutionary Analysis Reveal Diverse Type III Polyketide Synthase Pathways in Cyanobacteria». Genome Biology and Evolution 13(4):1-15. doi: 10.1093/gbe/evab056.

van der Lee, Theo A. J., and Marnix H. Medema. 2016. «Computational Strategies for Genome-Based Natural Product Discovery and Engineering in Fungi». Fungal Genetics and Biology 89:29-36. doi: 10.1016/j.fgb.2016.01.006.

Lemetre, Christophe, Jeffrey Maniko, Zachary Charlop-Powers, Ben Sparrow, Andrew J. Lowe, and Sean F. Brady. 2017. «Bacterial Natural Product Biosynthetic Domain Composition in Soil Correlates with Changes in Latitude on a Continent-Wide Scale». Proceedings of the National Academy of Sciences 114(44):11615-20. doi: 10.1073/pnas.1710262114.

Li, Yisong, Adrián A. Pinto-Tomás, Xiaoying Rong, Kun Cheng, Minghao Liu, and Ying Huang. 2019. «Population Genomics Insights into Adaptive Evolution and Ecological Differentiation in Streptomycetes». Applied and Environmental Microbiology 85(7):e02555-18. doi: 10.1128/AEM.02555-18.

Lind, Abigail L., Jennifer H. Wisecaver, Catarina Lameiras, Philipp Wiemann, Jonathan M. Palmer, Nancy P. Keller, Fernando Rodrigues, Gustavo H. Goldman, and Antonis Rokas. 2017. «Drivers of Genetic Diversity in Secondary Metabolic Gene Clusters within a Fungal Species». PLOS Biology 15(11):e2003583. doi: 10.1371/journal.pbio.2003583.

Liu, Zhenhua, Jitender Cheema, Marielle Vigouroux, Lionel Hill, James Reed, Pirita Paajanen, Levi Yant, and Anne Osbourn. 2020. «Formation and Diversification of a Paradigm Biosynthetic Gene Cluster in Plants». Nature Communications 11(1):5354. doi: 10.1038/s41467-020-19153-6.

Liu, Zhenhua, Hernando G. Suarez Duran, Yosapol Harnvanichvech, Michael J. Stephenson, M. Eric Schranz, David Nelson, Marnix H. Medema, and Anne Osbourn. 2020. «Drivers of Metabolic Diversification: How Dynamic Genomic Neighbourhoods Generate New Biosynthetic Pathways in the Brassicaceae». New Phytologist 227(4):1109-23. doi: 10.1111/nph.16338.

Lynch, Michael, Matthew S. Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, W. Kelley Thomas, and Patricia L. Foster. 2016. «Genetic Drift, Selection and the Evolution of the Mutation Rate». Nature Reviews Genetics 17(11):704-14. doi: 10.1038/nrg.2016.104.

Männle, Daniel, Shaun M. K. McKinnie, Shrikant S. Mantri, Katharina Steinke, Zeyin Lu, Bradley S. Moore, Nadine Ziemert, and Leonard Kaysser. 2020. «Comparative Genomics and Metabolomics in the Genus Nocardia» editado por D. F. Savage. MSystems 5(3):e00125-20. doi: 10.1128/mSystems.00125-20.

Martinet, Loïc, Aymeric Naômé, Benoit Deflandre, Marta Maciejewska, Déborah Tellatin, Elodie Tenconi, Nicolas Smargiasso, Edwin de Pauw, Gilles P. van Wezel, and Sébastien Rigali. 2019. «A Single Biosynthetic Gene Cluster Is Responsible for the Production of Bagremycin Antibiotics and Ferroverdin Iron Chelators». mBio 10(4):e01230-19. doi: 10.1128/mBio.01230-19.

Marx, Vivien. 2013. «The Big Challenges of Big Data». Nature 498(7453):255-60. doi: 10.1038/498255a.

Masatoshi Nei, and Sudhir Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford, New York: Oxford University Press.

Masschelein, J., M. Jenner, and G. L. Challis. 2017. «Antibiotics from Gram-Negative Bacteria: A Comprehensive Overview and Selected Biosynthetic Highlights». Natural Product Reports 34(7):712-83. doi: 10.1039/c7np00010c.

Matthew B. Hamilton. 2021. Population Genetics, 2nd Edition | Wiley. 2.a ed.

McDonald, Bradon R., Marc G. Chevrette, Jonathan L. Klassen, Heidi A. Horn, Eric J. Caldera, Evelyn Wendt-Pienkowski, Matias J. Cafaro, Antonio C. Ruzzini, Ethan B. Van Arnam, George M. Weinstock, Nicole M. Gerardo, Michael Poulsen, Garret Suen, Jon Clardy, and Cameron R. Currie. 2019. «Biogeography and Microscale Diversity Shape the Biosynthetic Potential of Fungus-Growing Ant-Associated Pseudonocardia». BioRxiv 545640. doi: 10.1101/545640.

McDonald, Bradon R., and Cameron R. Currie. 2017. «Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus Streptomyces». MBio 8(3):e00644-17. doi: 10.1128/mBio.00644-17.

Medema, Marnix H., Peter Cimermancic, Andrej Sali, Eriko Takano, and Michael A. Fischbach. 2014. «A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis». PLOS Computational Biology 10(12):e1004016. doi: 10.1371/journal.pcbi.1004016.

Medema, Marnix H., Renzo Kottmann, Pelin Yilmaz, Matthew Cummings, John B. Biggins, Kai Blin, Irene de Bruijn, Yit Heng Chooi, Jan Claesen, R. Cameron Coates, Pablo Cruz-Morales, Srikanth Duddela, Stephanie Düsterhus, Daniel J. Edwards, David P. Fewer, Neha Garg, Christoph Geiger, Juan Pablo Gomez-Escribano, Anja Greule, Michalis Hadjithomas, Anthony S. Haines, Eric J. N. Helfrich, Matthew L. Hillwig, Keishi Ishida, Adam C. Jones, Carla S. Jones, Katrin Jungmann, Carsten Kegler, Hyun Uk Kim, Peter Kötter, Daniel Krug, Joleen Masschelein, Alexey V. Melnik, Simone M. Mantovani, Emily A. Monroe, Marcus Moore, Nathan Moss, Hans-Wilhelm Nützmann, Guohui Pan, Amrita Pati, Daniel Petras, F. Jerry Reen, Federico Rosconi, Zhe Rui, Zhenhua Tian, Nicholas J. Tobias, Yuta Tsunematsu, Philipp Wiemann, Elizabeth Wyckoff, Xiaohui Yan, Grace Yim, Fengan Yu, Yunchang Xie, Bertrand Aigle, Alexander K. Apel, Carl J. Balibar, Emily P. Balskus, Francisco Barona-Gómez, Andreas Bechthold, Helge B. Bode, Rainer Borriss, Sean F. Brady, Axel A. Brakhage, Patrick Caffrey, Yi-Qiang Cheng, Jon Clardy, Russell J. Cox, René De Mot, Stefano Donadio, Mohamed S. Donia, Wilfred A. van der Donk, Pieter C. Dorrestein, Sean Doyle, Arnold J. M. Driessen, Monika Ehling-Schulz, Karl-Dieter Entian, Michael A. Fischbach, Lena Gerwick, William H. Gerwick, Harald Gross, Bertolt Gust, Christian Hertweck, Monica Höfte, Susan E. Jensen, Jianhua Ju, Leonard Katz, Leonard Kaysser, Jonathan L. Klassen, Nancy P. Keller, Jan Kormanec, Oscar P. Kuipers, Tomohisa Kuzuyama, Nikos C. Kyrpides, Hyung-Jin Kwon, Sylvie Lautru, Rob Lavigne, Chia Y. Lee, Bai Linquan, Xinyu Liu, Wen Liu, Andriy Luzhetskyy, Taifo Mahmud, Yvonne Mast, Carmen Méndez, Mikko Metsä-Ketelä, Jason Micklefield, Douglas A. Mitchell, Bradley S. Moore, Leonilde M. Moreira, Rolf Müller, Brett A. Neilan, Markus Nett, Jens Nielsen, Fergal O'Gara, Hideaki Oikawa, Anne Osbourn, Marcia S. Osburne, Bohdan Ostash, Shelley M. Payne, Jean-Luc Pernodet, Miroslav Petricek, Jörn Piel, Olivier Ploux, Jos M. Raaijmakers, José A. Salas, Esther K. Schmitt, Barry Scott, Ryan F. Seipke, Ben Shen, David H. Sherman, Kaarina Sivonen, Michael J. Smanski, Margherita Sosio, Evi Stegmann, Roderich D. Süssmuth, Kapil Tahlan, Christopher M. Thomas, Yi Tang, Andrew W. Truman, Muriel Viaud, Jonathan D. Walton, Christopher T. Walsh, Tilmann Weber, Gilles P. van Wezel, Barrie Wilkinson, Joanne M. Willey, Wolfgang Wohlleben, Gerard D. Wright, Nadine Ziemert, Changsheng Zhang, Sergey B. Zotchev, Rainer Breitling, Eriko Takano, and Frank Oliver Glöckner.

2015. «Minimum Information about a Biosynthetic Gene Cluster». Nature Chemical Biology 11(9):625-31. doi: 10.1038/nchembio.1890.

Medema, Marnix H., Eriko Takano, and Rainer Breitling. 2013. «Detecting Sequence Homology at the Gene Cluster Level with MultiGeneBlast». Molecular Biology and Evolution 30(5):1218-23. doi: 10.1093/molbev/mst025.

Megahed, Fadel M., and L. Allison Jones-Farmer. 2015. «Statistical Perspectives on “Big Data”». Pp. 29-47 en Frontiers in Statistical Quality Control 11, Frontiers in Statistical Quality Control, editado por S. Knoth and W. Schmid. Cham: Springer International Publishing.

Meyer, F., D. Paarmann, M. D’Souza, R. Olson, EM Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and RA Edwards. 2008. «The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes». BMC Bioinformatics 9(1):386. doi: 10.1186/1471-2105-9-386.

Miller, Ian J., Marc G. Chevrette, and Jason C. Kwan. 2017. «Interpreting Microbial Biosynthesis in the Genomic Age: Biological and Practical Considerations». Marine Drugs 15(6):165. doi: 10.3390/md15060165.

Mitchell, Alex L., Alexandre Almeida, Martin Beracochea, Miguel Boland, Josephine Burgin, Guy Cochrane, Michael R. Crusoe, Varsha Kale, Simon C. Potter, Lorna J. Richardson, Ekaterina Sakharova, Maxim Scheremetjew, Anton Korobeynikov, Alex Shlemov, Olga Kunyavskaya, Alla Lapidus, and Robert D. Finn. 2020. «MGnify: the microbiome analysis resource in 2020». Nucleic Acids Research 48(D1):D570-78. doi: 10.1093/nar/gkz1035.

Moghe, Gaurav D., and Robert L. Last. 2015. «Something Old, Something New: Conserved Enzymes and the Evolution of Novelty in Plant Specialized Metabolism». Plant Physiology 169(3):1512-23. doi: 10.1104/pp.15.00994.

Montalbán-López, Manuel, Thomas A. Scott, Sangeetha Ramesh, Imran R. Rahman, Auke J. van Heel, Jakob H. Viel, Vahe Bandarian, Elke Dittmann, Olga Genilloud, Yuki Goto, María José Grande Burgos, Colin Hill, Seokhee Kim, Jesko Koehnke, John A. Latham, A. James Link, Beatriz Martínez, Satish K. Nair, Yvain Nicolet, Sylvie Rebuffat, Hans-Georg Sahl, Dipti Sareen, Eric W. Schmidt, Lutz Schmitt, Konstantin Severinov, Roderich D. Süssmuth, Andrew W. Truman, Huan Wang, Jing-Ke Weng, Gilles P. van Wezel, Qi Zhang, Jin Zhong, Jörn Piel, Douglas A. Mitchell, Oscar P. Kuipers, and Wilfred A. van der Donk. 2021. «New Developments in RiPP Discovery, Enzymology and Engineering». Natural Product Reports 38(1):130-239. doi: 10.1039/D0NP00027B.

Mungan, Mehmet Direnç, Mohammad Alanjary, Kai Blin, Tilmann Weber, Marnix H. Medema, and Nadine Ziemert. 2020. «ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining». Nucleic Acids Research 48(W1):W546-52. doi: 10.1093/nar/gkaa374.

Nakhleh, L. 2013. «Evolutionary Trees». Pp. 549-50 en Brenner’s Encyclopedia of Genetics. Elsevier.

Nasrin, Shamima, Suresh Ganji, Kavita S. Kakirde, Melissa R. Jacob, Mei Wang, Ranga Rao Ravu, Paul A. Cobine, Ikhlas A. Khan, Cheng-Cang Wu, David A. Mead, Xing-Cong Li, and Mark R. Liles. 2018. «Chloramphenicol Derivatives with Antibacterial Activity Identified by Functional Metagenomics». Journal of Natural Products 81(6):1321-32. doi: 10.1021/acs.jnatprod.7b00903.

Navarro-Muñoz, Jorge C., Nelly Selem-Mojica, Michael W. Mullowney, Satria A. Kautsar, James H. Tryon, Elizabeth I. Parkinson, Emmanuel L. C. De Los Santos, Marley Yeong, Pablo Cruz-Morales, Sahar Abubucker, Arne Roeters, Wouter Lokhorst, Antonio Fernandez-Guerra, Luciana Teresa Dias Cappelini, Anthony W. Goering, Regan J. Thomson, William W. Metcalf, Neil L. Kelleher, Francisco

Barona-Gomez, and Marnix H. Medema. 2020. «A Computational Framework to Explore Large-Scale Biosynthetic Diversity». Nature Chemical Biology 16(1):60-68. doi: 10.1038/s41589-019-0400-9.

Nayfach, Stephen, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, David Paez-Espino, I.-Min Chen, Marcel Huntemann, Krishna Palaniappan, Joshua Ladau, Supratim Mukherjee, T. B. K. Reddy, Torben Nielsen, Edward Kirton, José P. Faria, Janaka N. Edirisinghe, Christopher S. Henry, Sean P. Jungbluth, Dylan Chivian, Paramvir Dehal, Elisha M. Wood-Charlson, Adam P. Arkin, Susannah G. Tringe, Axel Visel, Tanja Woyke, Nigel J. Mouncey, Natalia N. Ivanova, Nikos C. Kyrpides, and Emiley A. Eloe-Fadrosh. 2020. «A Genomic Catalog of Earth's Microbiomes». Nature Biotechnology 1-11. doi: 10.1038/s41587-020-0718-6.

Nguyen, TuAnh, Keishi Ishida, Holger Jenke-Kodama, Elke Dittmann, Cristian Gurgui, Thomas Hochmuth, Stefan Taudien, Matthias Platzer, Christian Hertweck, and Jörn Piel. 2008. «Exploiting the Mosaic Structure of Trans-Acyltransferase Polyketide Synthases for Natural Product Discovery and Pathway Dissection». Nature Biotechnology 26(2):225-33. doi: 10.1038/nbt1379.

Nivina, Aleksandra, Kai P. Yuet, Jake Hsu, and Chaitan Khosla. 2019. «Evolution and Diversity of Assembly-Line Polyketide Synthases: Focus Review». Chemical Reviews 119(24):12524-47. doi: 10.1021/acs.chemrev.9b00525.

Noda-Garcia, Lianet, Wolfram Liebermeister, and Dan S. Tawfik. 2018. «Metabolite–Enzyme Coevolution: From Single Enzymes to Metabolic Pathways and Networks». Annual Review of Biochemistry 87(1):187-216. doi: 10.1146/annurev-biochem-062917-012023.

Noda-Garcia, Lianet, and Dan S. Tawfik. 2020. «Enzyme Evolution in Natural Products Biosynthesis: Target- or Diversity-Oriented?» Current Opinion in Chemical Biology 59:147-54. doi: 10.1016/j.cbpa.2020.05.011.

Pál, Csaba, Balázs Papp, and Martin J. Lercher. 2006. «An Integrated View of Protein Evolution». Nature Reviews Genetics 7(5):337-48. doi: 10.1038/nrg1838.

Palaniappan, Krishnaveni, I.-Min A. Chen, Ken Chu, Anna Ratner, Rekha Seshadri, Nikos C. Kyrpides, Natalia N. Ivanova, and Nigel J. Mouncey. 2019. «IMG-ABC v.5.0: An Update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase». Nucleic Acids Research gkz932. doi: 10.1093/nar/gkz932.

Piatkowski, Bryan T., Karn Imwattana, Erin A. Tripp, David J. Weston, Adam Healey, Jeremy Schmutz, and A. Jonathan Shaw. 2020. «Phylogenomics Reveals Convergent Evolution of Red-Violet Coloration in Land Plants and the Origins of the Anthocyanin Biosynthetic Pathway». Molecular Phylogenetics and Evolution 151:106904. doi: 10.1016/j.ympev.2020.106904.

Reddy, Boojala Vijay B., Dimitris Kallifidas, Jeffrey H. Kim, Zachary Charlop-Powers, Zhiyang Feng, and Sean F. Brady. 2012. «Natural Product Biosynthetic Gene Diversity in Geographically Distinct Soil Microbiomes». Applied and Environmental Microbiology 78(10):3744-52. doi: 10.1128/AEM.00102-12.

Rokas, Antonis, Matthew E. Mead, Jacob L. Steenwyk, Huzefa A. Raja, and Nicholas H. Oberlies. 2020. «Biosynthetic Gene Clusters and the Evolution of Fungal Chemodiversity». Natural Product Reports 37(7):868-78. doi: 10.1039/C9NP00045C.

Rokas, Antonis, Jennifer H. Wisecaver, and Abigail L. Lind. 2018. «The Birth, Evolution and Death of Metabolic Gene Clusters in Fungi». Nature Reviews Microbiology 16(12):731-44. doi: 10.1038/s41579-018-0075-3.

Saha, Chayan Kumar, Rodrigo Sanches Pires, Harald Brolin, Maxence Delannoy, and Gemma Catherine Atkinson. 2020. «FlaGs and webFlaGs: discovering novel biology through the analysis of

gene neighbourhood conservation». Bioinformatics 37(9):1312-14. doi: 10.1093/bioinformatics/btaa788.

Santana-Pereira, Alinne L. R., Megan Sandoval-Powers, Scott Monsma, Jinglie Zhou, Scott R. Santos, David A. Mead, and Mark R. Liles. 2020. «Discovery of Novel Biosynthetic Gene Cluster Diversity From a Soil Metagenomic Library». Frontiers in Microbiology 11(585398):1-17. doi: 10.3389/fmicb.2020.585398.

van Santen, Jeffrey A., Grégoire Jacob, Amrit Leen Singh, Victor Aniebok, Marcy J. Balunas, Derek Bunsko, Fausto Carnevale Neto, Laia Castaño-Espriu, Chen Chang, Trevor N. Clark, Jessica L. Cleary Little, David A. Delgadillo, Pieter C. Dorrestein, Katherine R. Duncan, Joseph M. Egan, Melissa M. Galey, F. P. Jake Haeckl, Alex Hua, Alison H. Hughes, Dasha Iskakova, Aswad Khadilkar, Jung-Ho Lee, Sanghoon Lee, Nicole LeGrow, Dennis Y. Liu, Jocelyn M. Macho, Catherine S. McCaughey, Marnix H. Medema, Ram P. Neupane, Timothy J. O'Donnell, Jasmine S. Paula, Laura M. Sanchez, Anam F. Shaikh, Sylvia Soldatou, Barbara R. Terlouw, Tuan Anh Tran, Mercia Valentine, Justin J. J. van der Hooft, Duy A. Vo, Mingxun Wang, Darryl Wilson, Katherine E. Zink, and Roger G. Linington. 2019. «The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery». ACS Central Science 5(11):1824-33. doi: 10.1021/acscentsci.9b00806.

Schniete, Jana K., Nelly Selem-Mojica, Anna S. Birke, Pablo Cruz-Morales, Iain S. Hunter, Francisco Barona-Gomez, and Paul A. Hoskisson. 2021. «ActDES – a curated Actinobacterial Database for Evolutionary Studies». Microbial Genomics. doi: 10.1099/mgen.0.000498.

Schorn, Michelle A., Mohammad M. Alanjary, Kristen Aguinaldo, Anton Korobeynikov, Sheila Podell, Nastassia Patin, Tommie Lincecum, Paul R. Jensen, Nadine Ziemert, and Bradley S. Moore. 2016. «Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters». Microbiology, 162(12):2075-86. doi: 10.1099/mic.0.000386.

Schorn, Michelle A., Stefan Verhoeven, Lars Ridder, Florian Huber, Deepa D. Acharya, Alexander A. Aksenov, Gajender Aleti, Jamshid Amiri Moghaddam, Allegra T. Aron, Saefuddin Aziz, Anelize Bauermeister, Katherine D. Bauman, Martin Baunach, Christine Beemelmanns, J. Michael Beman, María Victoria Berlanga-Clavero, Alex A. Blacutt, Helge B. Bode, Anne Boullie, Asker Brejnrod, Tim S. Bugni, Alexandra Calteau, Liu Cao, Víctor J. Carrión, Raquel Castelo-Branco, Shaurya Chanana, Alexander B. Chase, Marc G. Chevrette, Leticia V. Costa-Lotufo, Jason M. Crawford, Cameron R. Currie, Bart Cuypers, Tam Dang, Tristan de Rond, Alyssa M. Demko, Elke Dittmann, Chao Du, Christopher Drozd, Jean-Claude Dujardin, Rachel J. Dutton, Anna Edlund, David P. Fewer, Neha Garg, Julia M. Gauglitz, Emily C. Gentry, Lena Gerwick, Evgenia Glukhov, Harald Gross, Muriel Gugger, Dulce G. Guillén Matus, Eric J. N. Helfrich, Benjamin-Florian Hempel, Jae-Seoun Hur, Marianna Iorio, Paul R. Jensen, Kyo Bin Kang, Leonard Kaysser, Neil L. Kelleher, Chung Sub Kim, Ki Hyun Kim, Irina Koester, Gabriele M. König, Tiago Leao, Seoung Rak Lee, Yi-Yuan Lee, Xuanji Li, Jessica C. Little, Katherine N. Maloney, Daniel Männle, Christian Martin H, Andrew C. McAvoy, Willam W. Metcalf, Hosein Mohimani, Carlos Molina-Santiago, Bradley S. Moore, Michael W. Mullowney, Mitchell Muskat, Louis-Félix Nothias, Ellis C. O'Neill, Elizabeth I. Parkinson, Daniel Petras, Jörn Piel, Emily C. Pierce, Karine Pires, Raphael Reher, Diego Romero, M. Caroline Roper, Michael Rust, Hamada Saad, Carmen Saenz, Laura M. Sanchez, Søren Johannes Sørensen, Margherita Sosio, Roderich D. Süssmuth, Douglas Sweeney, Kapil Tahlan, Regan J. Thomson, Nicholas J. Tobias, Amaro E. Trindade-Silva, Gilles P. van Wezel, Mingxun Wang, Kelly C. Weldon, Fan Zhang, Nadine Ziemert, Katherine R. Duncan, Max Crüsemann, Simon Rogers, Pieter C. Dorrestein, Marnix H. Medema, and Justin J. J. van der Hooft. 2021. «A Community Resource for Paired Genomic and Metabolomic Data Mining». Nature Chemical Biology 1-6. doi: 10.1038/s41589-020-00724-z.

Sélem-Mojica, Nelly, César Aguilar, Karina Gutiérrez-García, Christian E. Martínez-Guerrero, and Fancisco Barona-Gómez. 2019. «EvoMining reveals the origin and fate of natural product biosynthetic enzymes». Microbial Genomics. doi: 10.1099/mgen.0.000260.

Sharrar, Allison M., Alexander Crits-Christoph, Raphaël Méheust, Spencer Diamond, Evan P. Starr, and Jillian F. Banfield. 2020. «Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type». MBio 11(3):e00416-20. doi: 10.1128/mBio.00416-20.

Shimizu, Yugo, Hiroyuki Ogata, and Susumu Goto. 2017. «Type III Polyketide Synthases: Functional Classification and Phylogenomics». ChemBioChem 18(1):50-65. doi: 10.1002/cbic.201600522.

Silva, Sandra G., Jochen Blom, Tina Keller-Costa, and Rodrigo Costa. 2019. «Comparative Genomics Reveals Complex Natural Product Biosynthesis Capacities and Carbon Metabolism across Host-associated and Free-living Aquimarina ( Bacteroidetes, Flavobacteriaceae ) Species». Environmental Microbiology 21(11):4002-19. doi: 10.1111/1462-2920.14747.

Stubbendieck, Reed M., Daniel S. May, Marc G. Chevrette, Mia I. Temkin, Evelyn Wendt-Pienkowski, Julian Cagnazzo, Caitlin M. Carlson, James E. Gern, and Cameron R. Currie. 2019. «Competition among Nasal Bacteria Suggests a Role for Siderophore-Mediated Interactions in Shaping the Human Nasal Microbiota». Applied and Environmental Microbiology 85(10):e02406-18. doi: 10.1128/AEM.02406-18.

Sugden, Andrew, Caroline Ash, Brooks Hanson, and Laura Zahn. 2009. «Happy Birthday, Mr. Darwin». Science 323(5915):727-727. doi: 10.1126/science.323.5915.727.

Süssmuth, Roderich D., and Andi Mainz. 2017. «Nonribosomal Peptide Synthesis—Principles and Prospects». Angewandte Chemie International Edition 56(14):3770-3821. doi: 10.1002/anie.201609079.

Tang, Man-Cheng, Yi Zou, Kenji Watanabe, Christopher T. Walsh, and Yi Tang. 2017. «Oxidative Cyclization in Natural Product Biosynthesis». Chemical Reviews 117(8):5226-5333. doi: 10.1021/acs.chemrev.6b00478.

Thompson, Luke R., Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J. Locey, Robert J. Prill, Anupriya Tripathi, Sean M. Gibbons, Gail Ackermann, Jose A. Navas-Molina, Stefan Janssen, Evguenia Kopylova, Yoshiki Vázquez-Baeza, Antonio González, James T. Morton, Siavash Mirarab, Zhenjiang Zech Xu, Lingjing Jiang, Mohamed F. Haroon, Jad Kanbar, Qiyun Zhu, Se Jin Song, Tomasz Kosciolek, Nicholas A. Bokulich, Joshua Lefler, Colin J. Brislawn, Gregory Humphrey, Sarah M. Owens, Jarrad Hampton-Marcell, Donna Berg-Lyons, Valerie McKenzie, Noah Fierer, Jed A. Fuhrman, Aaron Clauset, Rick L. Stevens, Ashley Shade, Katherine S. Pollard, Kelly D. Goodwin, Janet K. Jansson, Jack A. Gilbert, and Rob Knight. 2017. «A Communal Catalogue Reveals Earth's Multiscale Microbial Diversity». Nature 551(7681):457-63. doi: 10.1038/nature24621.

Tidjani, Abdoul-Razak, Jean-Noël Lorenzi, Maxime Toussaint, Erwin van Dijk, Delphine Naquin, Olivier Lespinet, Cyril Bontemps, and Pierre Leblond. 2019. «Massive Gene Flux Drives Genome Diversity between Sympatric Streptomyces Conspecifics». mBio 10(5):e01533-19. doi: 10.1128/mBio.01533-19.

Tracanna, Vittorio, Adam Ossowicki, Marloes L. C. Petrus, Sam Overduin, Barbara R. Terlouw, George Lund, Serina L. Robinson, Sven Warris, Elio G. W. M. Schijlen, Gilles P. van Wezel, Jos M. Raaijmakers, Paolina Garbeva, and Marnix H. Medema. 2021. «Dissecting Disease-Suppressive Rhizosphere Microbiomes by Functional Amplicon Sequencing and 10× Metagenomics». mSystems 6(3):e01116-20. doi: 10.1128/mSystems.01116-20.

Traxler, Matthew F., and Roberto Kolter. 2015. «Natural Products in Soil Microbe Interactions and Evolution». Natural Product Reports 32(7):956-70. doi: 10.1039/C5NP00013K.

Vior, Natalia M., Rodney Lacret, Govind Chandra, Siobhán Dorai-Raj, Martin Trick, and Andrew W. Truman. 2018. «Discovery and Biosynthesis of the Antibiotic Bicyclomycin in Distantly Related

Bacterial Classes». Applied and Environmental Microbiology 84(9):e02828-17. doi: 10.1128/AEM.02828-17.

Waglechner, Nicholas, Andrew G. McArthur, and Gerard D. Wright. 2019. «Phylogenetic Reconciliation Reveals the Natural History of Glycopeptide Antibiotic Biosynthesis and Resistance». Nature Microbiology 4(11):1862-71. doi: 10.1038/s41564-019-0531-5.

Walker, Allison S., and Jon Clardy. 2021. «A Machine Learning Bioinformatics Method to Predict Biological Activity from Biosynthetic Gene Clusters». Journal of Chemical Information and Modeling 61(6):2560-71. doi: 10.1021/acs.jcim.0c01304.

Wang, Haitao, Minying Cheng, Melissa Dsouza, Pamela Weisenhorn, Tianling Zheng, and Jack A. Gilbert. 2018. «Soil Bacterial Diversity Is Associated with Human Population Density in Urban Greenspaces». Environmental Science & Technology 52(9):5115-24. doi: 10.1021/acs.est.7b06417.

Wang, Mingxun, Jeremy J. Carver, Vanessa V. Phelan, Laura M. Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A. Kapono, Tal Luzzatto-Knaan, Carla Porto, Amina Bouslimani, Alexey V. Melnik, Michael J. Meehan, Wei-Ting Liu, Max Crüsemann, Paul D. Boudreau, Eduardo Esquenazi, Mario Sandoval-Calderón, Roland D. Kersten, Laura A. Pace, Robert A. Quinn, Katherine R. Duncan, Cheng-Chih Hsu, Dimitrios J. Floros, Ronnie G. Gavilan, Karin Kleigrewe, Trent Northen, Rachel J. Dutton, Delphine Parrot, Erin E. Carlson, Bertrand Aigle, Charlotte F. Michelsen, Lars Jelsbak, Christian Sohlenkamp, Pavel Pevzner, Anna Edlund, Jeffrey McLean, Jörn Piel, Brian T. Murphy, Lena Gerwick, Chih-Chuang Liaw, Yu-Liang Yang, Hans-Ulrich Humpf, Maria Maansson, Robert A. Keyzers, Amy C. Sims, Andrew R. Johnson, Ashley M. Sidebottom, Brian E. Sedio, Andreas Klitgaard, Charles B. Larson, Cristopher A. Boya P, Daniel Torres-Mendoza, David J. Gonzalez, Denise B. Silva, Lucas M. Marques, Daniel P. Demarque, Egle Pociute, Ellis C. O'Neill, Enora Briand, Eric J. N. Helfrich, Eve A. Granatosky, Evgenia Glukhov, Florian Ryffel, Hailey Houson, Hosein Mohimani, Jenan J. Kharbush, Yi Zeng, Julia A. Vorholt, Kenji L. Kurita, Pep Charusanti, Kerry L. McPhail, Kristian Fog Nielsen, Lisa Vuong, Maryam Elfeki, Matthew F. Traxler, Niclas Engene, Nobuhiro Koyama, Oliver B. Vining, Ralph Baric, Ricardo R. Silva, Samantha J. Mascuch, Sophie Tomasi, Stefan Jenkins, Venkat Macherla, Thomas Hoffman, Vinayak Agarwal, Philip G. Williams, Jingqui Dai, Ram Neupane, Joshua Gurr, Andrés M. C. Rodríguez, Anne Lamsa, Chen Zhang, Kathleen Dorrestein, Brendan M. Duggan, Jehad Almaliti, Pierre-Marie Allard, Prasad Phapale, Louis-Felix Nothias, Theodore Alexandrov, Marc Litaudon, Jean-Luc Wolfender, Jennifer E. Kyle, Thomas O. Metz, Tyler Peryea, Dac-Trung Nguyen, Danielle VanLeer, Paul Shinn, Ajit Jadhav, Rolf Müller, Katrina M. Waters, Wenyuan Shi, Xueting Liu, Lixin Zhang, Rob Knight, Paul R. Jensen, Bernhard Ø. Palsson, Kit Pogliano, Roger G. Linington, Marcelino Gutiérrez, Norberto P. Lopes, William H. Gerwick, Bradley S. Moore, Pieter C. Dorrestein, and Nuno Bandeira. 2016. «Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking». Nature Biotechnology 34(8):828-37. doi: 10.1038/nbt.3597.

Weng, Jing-Ke. 2014. «The Evolutionary Paths towards Complexity: A Metabolic Perspective». New Phytologist 201(4):1141-49. doi: 10.1111/nph.12416.

Wideman, Jeremy G., Aaron Novick, Sergio A. Muñoz-Gómez, and W. Ford Doolittle. 2019. «Neutral Evolution of Cellular Phenotypes». Current Opinion in Genetics & Development 58-59:87-94. doi: 10.1016/j.gde.2019.09.004.

Wiegand, Sandra, Mareike Jogler, Christian Boedeker, Daniela Pinto, John Vollmers, Elena Rivas-Marín, Timo Kohn, Stijn H. Peeters, Anja Heuer, Patrick Rast, Sonja Oberbeckmann, Boyke Bunk, Olga Jeske, Anke Meyerdierks, Julia E. Storesund, Nicolai Kallscheuer, Sebastian Lücker, Olga M. Lage, Thomas Pohl, Broder J. Merkel, Peter Hornburger, Ralph-Walter Müller, Franz Brümmer, Matthias Labrenz, Alfred M. Spormann, Huub J. M. Op den Camp, Jörg Overmann, Rudolf Amann, Mike S. M. Jetten, Thorsten Mascher, Marnix H. Medema, Damien P. Devos, Anne-Kristin Kaster, Lise

Øvreås, Manfred Rohde, Michael Y. Galperin, and Christian Jogler. 2020. «Cultivation and Functional Characterization of 79 Planctomycetes Uncovers Their Unique Biology». Nature Microbiology 5(1):126-40. doi: 10.1038/s41564-019-0588-1.

Wilson, Alexander E., and Li Tian. 2019. «Phylogenomic Analysis of UDP-Dependent Glycosyltransferases Provides Insights into the Evolutionary Landscape of Glycosylation in Plant Metabolism». The Plant Journal 100(6):1273-88. doi: 10.1111/tpj.14514.

Woese, C. R., O. Kandler, and M. L. Wheelis. 1990. «Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya». Proceedings of the National Academy of Sciences of the United States of America 87(12):4576-79.

Wolf, Thomas, Vladimir Shelest, Neetika Nath, and Ekaterina Shelest. 2016. «CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes». Bioinformatics 32(8):1138-43. doi: 10.1093/bioinformatics/btv713.

Wolfe, Kenneth H., and Wen-Hsiung Li. 2003. «Molecular Evolution Meets the Genomics Revolution». Nature Genetics 33(3):255-65. doi: 10.1038/ng1088.

Yang, Jack Y., and Okan K. Ersoy. 2003. «Combined Supervised and Unsupervised Learning in Genomic Data Mining». 143.

Yang, Yang, Xiaobao Liu, Jimiao Cai, Yipeng Chen, Boxun Li, Zhikai Guo, and Guixiu Huang. 2019. «Genomic characteristics and comparative genomics analysis of the endophytic fungus Sarocladium brachiariae». BMC Genomics 20(1):782. doi: 10.1186/s12864-019-6095-1.

Zacharia, Vineetha M., Yein Ra, Catherine Sue, Elizabeth Alcala, Jewel N. Reaso, Steven E. Ruzin, and Matthew F. Traxler. 2021. «Genetic Network Architecture and Environmental Cues Drive Spatial Organization of Phenotypic Division of Labor in Streptomyces coelicolor». mBio 12(3):e00794-21. doi: 10.1128/mBio.00794-21.

Zhang, Zheren, Chao Du, Frédérique de Barsy, Michael Liem, Apostolos Liakopoulos, Gilles P. van Wezel, Young H. Choi, Dennis Claessen, and Daniel E. Rozen. 2020. «Antibiotic Production in Streptomyces Is Organized by a Division of Labor through Terminal Genomic Differentiation». Science Advances 6(3):eaay5781. doi: 10.1126/sciadv.aay5781.

Ziemert, N., A. Lechner, M. Wietz, N. Millan-Aguinaga, K. L. Chavarria, and P. R. Jensen. 2014. «Diversity and Evolution of Secondary Metabolism in the Marine Actinomycete Genus Salinispora». Proceedings of the National Academy of Sciences 111(12):E1130-39. doi: 10.1073/pnas.1324161111.

## Acknowledgments

We are grateful to Jorge Navarro-Muñoz for useful discussions and Erika V. Cruz for help with figures. Support for M.G.C. provided by grant 2020-67012-31772 (accession 1022881) from the USDA National Institute of Food and Agriculture. F.B.G. and N.S.M. are supported by Conacyt, Mexico (grant No. 285746) and the Royal Society of the United Kingdom, Newton Advanced Fellowship (NAF\R2\180631) to F.B.G. A.G. is grateful for the support of the Deutsche Forschungsgemeinschaft (DFG; Project ID # 398967434-TRR 261). S.M. is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)

under Germany's Excellence Strategy – EXC 2124 – 390838134. N.Z. is funded by the German Center for Infection Research (TTU09.716).

# Chapter 2: Compendium of secondary metabolite biosynthetic diversity encoded in bacterial genomes

*(Manuscript published in Nature Microbiology, May 02 2022)*

Athina Gavriilidou*[1], Satria A. Kautsar*[2,3], Nestor Zaburannyi[4,5], Daniel Krug[4,5], Rolf Müller[4,5], Marnix H. Medema**[+2], Nadine Ziemert**[+1,6,7]

*These authors contributed equally: Athina Gavriilidou, Satria A. Kautsar.

**These authors jointly supervised this work: Marnix H. Medema, Nadine Ziemert.

[+]e-mail: marnix.medema@wur.nl; nadine.ziemert@uni-tuebingen.de

[1]Translational Genome Mining for Natural Products, Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany.

[2]Bioinformatics Group, Wageningen University, the Netherlands

[3]Chemistry Department, Scripps Research Florida, United States

[4]Helmholtz Institute for Pharmaceutical Research Saarland (HIPS)—Helmholtz Centre for Infection Research (HZI), Campus E8 1, 66123 Saarbrücken, Germany.

[5]German Center for Infection Research (DZIF), Partner site Hannover-Braunschweig, Inhoffenstr. 7, 38124 Braunschweig, Germany.

[6]Cluster of Excellence 'Controlling Microbes to Fight Infections' (CMFI), University of Tübingen, Tübingen, Germany.

[7]German Centre for Infection Research (DZIF), Partnersite Tübingen, Tübingen, Germany.

Personal contributions:

Scientific ideas/data generation/analysis & interpretation/writing: 50/40/90/80%

## Abstract

Bacterial specialized metabolites are a proven source of antibiotics and cancer therapies, but whether we have sampled all the secondary metabolite chemical diversity of cultivated bacteria is not known. We analysed ~ 170,000 bacterial genomes and ~ 47,000 metagenome assembled genomes (MAGs) using a modified BiG-SLiCE and the new clust-o-matic algorithm. We found that only 3% of the natural products potentially encoded in bacterial genomes have been experimentally characterized. We show that the variation of secondary metabolite biosynthetic diversity drops significantly at the genus level, identifying it as an appropriate taxonomic rank for comparison. Equal comparison of genera based on Relative Evolutionary Distance revealed that *Streptomyces* bacteria encode the largest biosynthetic diversity by far, with *Amycolatopsis*, *Kutzneria* and *Micromonospora* also encoding substantial diversity. Finally we find that several less-well-studied taxa, such as Weeksellaceae (Bacteroidota), Myxococcaceae (Myxococcota), *Pleurocapsa* and Nostocaceae (Cyanobacteria), have potential to produce highly diverse sets of secondary metabolites that warrant further investigation.

## Introduction

Specialized metabolites (also called secondary metabolites) are biomolecules that are not essential for life but rather offer specific ecological or physiological advantages to their producers allowing them to thrive in particular niches. These Natural Products (NPs) are more chemically diverse than the molecules of primary metabolism, varying in both structure and mode of action among different organisms[1]. Historically, microbial NPs and their derivatives have contributed and continue to contribute a substantial part of chemical entities brought to the clinic, especially as anticancer compounds and antibiotics[2–4]. Regrettably, the emergence of antibiotic-resistant pathogens[3] concomitant to a stagnation of antimicrobial discovery pipelines[2,4] is leading to a global public health crisis[3].

Nonetheless, genomics-based approaches to NP discovery[5,6] have revealed a largely untapped and much more diverse source of biosynthetic potential within genomes[3,7]. These findings were possible following the discovery that bacterial genes encoding the biosynthesis of secondary metabolites are usually located in close proximity to each other, forming recognizable Biosynthetic Gene Clusters (BGCs). However, while the numbers and kinds of BGCs clearly differ across microbial genomes[7,8] and metabolomic data indicate that some biosynthetic pathways are unique to specific taxa[9], a systematic analysis of the taxonomic distribution of BGCs has not yet been performed. Similarly, while useful estimates of the chemical diversity of specific taxa have been provided[8], methodical comparisons across taxa are lacking. Because of this, the scientific community appears undecided on the best strategy for natural product discovery: should the established known NP producers be studied further or should the community be investigating underexplored taxa[7,10]? A relatively recent question is how much chemical diversity is hidden in uncultured bacteria. Metagenomic assembled genomes from uncultured bacteria

have demonstrated a big potential of unknown BGCs[7]. It is unclear to what extent unexplored associated ecological niches and (micro)environments are also associated with unique and unexplored chemistry.

Here, we harnessed recent advances in computational genomic analysis of BGCs to survey the enormous amount of genome data accumulated by the scientific community so far. Using a global approach based on more than 170,000 publicly available genomes, we created a comprehensive overview of the biosynthetic diversity found across the entire bacterial kingdom. We clustered 1,185,995 BGCs into 62,449 Gene Cluster Families (GCFs), and calibrated the granularity of the clustering to make it directly comparable to chemical classes as defined in NP Atlas[11]. This facilitated an analysis of the variance of diversity across major taxonomic ranks, which showed the genus rank to be the most appropriate to compare biosynthetic diversity across homogeneous groups. This finding allowed us to conduct comparisons within the bacterial kingdom. Evident patterns emerged from our analysis, revealing popular taxa as prominent sources of both actual and potential biosynthetic diversity, and multiple yet uncommon taxa as promising producers.

## Main text

### *Biosynthetic diversity of the bacterial kingdom*

To assess the global number of Gene Cluster Families found in sequenced bacterial strains, we ran AntiSMASH[12] on ~170,000 genomes from the NCBI RefSeq database[13] (Supplementary Table 1), spanning 48 bacterial phyla containing 464 families (according to the Genome Taxonomy DataBase classification - GTDB[14]). We also included almost 50,000 bacterial Metagenome Assembled Genomes (MAGs) from 6 metagenomic projects of various origins[15–20] (Table 1 and Supplementary Table 1). To accurately group similar BGCs – which likely encode pathways towards the production of similar compounds – into Gene Cluster Families (GCFs) across such a large dataset, we used a slightly modified version of the BiG-SLiCE tool[21], which has been calibrated to output GCFs that match the grouping of known compounds in the NP Atlas database[11] (see Methods: Quantification of biosynthetic diversity with BiG-SLiCE). The resulting GCFs were then used to measure biosynthetic diversity across taxa.

The number of GCFs in RefSeq ranged from 19,152 to 51,052 depending on the cut-off used by BiG-SLiCE (Table 1). While, as expected, the pure numbers of the analysis changed based on the l2-normalized euclidean threshold, the overall tendencies observed remained the same (Figure 1a, Supplementary Figure 1). The effect that the chosen threshold has on these results presented a challenge to our investigation, as previous estimations have also shown great heterogeneity when different thresholds were used[7,8], precluding direct comparisons of their predictions. As each BGC can be considered a proxy for its encoded pathways and their products, differing thresholds will result in different degrees of granularity in the

grouping of compound structures (Extended Data Figure 1). Nevertheless, linear relationships are not always applicable, as shown previously[22], and a specific threshold will need to be set anyway to make comparisons possible. For this, we sought to directly relate the choice of our BGC clustering threshold to the clustering of their compound structures. NPAtlas, a database of known microbial small molecules, provides hierarchical clustering of the compound structures via Morgan fingerprinting and Dice similarity scoring[11]. As many as 947 compounds in NPAtlas are mapped to a known BGC in MIBiG repository[23], giving us the opportunity to use them as an anchor for choosing our clustering threshold. After mapping the BiG-SLiCE groupings of known BGCs from the MIBiG to the compound clusters in NPAtlas (Supplementary Figure 2), we chose a threshold of 0.4, as it provided the most congruent agreements between the two groupings, with v-score=0.94 (out of 1.00) and ΔGCF=-17.

**Table 1. Input datasets and biosynthetic diversity with different BiG-SLiCE cut-offs**. The "Complete Dataset" was used for the computation of the actual and potential biosynthetic diversity found in all cultured (and some uncultured) bacteria. The dataset "RefSeq bacteria with known species taxonomy" was used for pinpointing the emergence of biosynthetic diversity, for which accurate taxonomic information was needed, and for identifying groups of promising producers. The "T"s under Gene Cluster Families represent different BiG-SLiCE l2-normalized euclidean thresholds; the values under T=0.4 stand out due to it being considered the most suitable cut-off. BGC to GCF assignment for each threshold can be found in Supplementary Tables 2-5. *MAG sources: bovine rumen[15], chicken caecum[16], human gut[17], ocean[18], uncultivated bacteria[19], various sources[20].

| **Dataset** | | **Genomes** | **BGCs** | **Gene Cluster Families** | | | |
|---|---|---|---|---|---|---|---|
| | | | | **T = 0.4** | T = 0.5 | T = 0.6 | T = 0.7 |
| Complete Dataset | All RefSeq bacteria | 170,549 | 1,060,592 | **51,052** | 37,785 | 28,057 | 19,152 |
| | Bacterial MAGs* | 47,098 | 125,403 | **21,354** | - | - | - |
| | Total | 217,647 | 1,185,995 | **62,449** | - | - | - |
| RefSeq bacteria with known species taxonomy | Complete Genomes | 16,004 | 94,904 | **16,984** | 13,546 | 10,399 | 7,151 |
| | Draft Genomes | 147,265 | 913,642 | **37,123** | 27,748 | 20,638 | 14,016 |
| | Total | 163,269 | 1,008,546 | **41,870** | 31,237 | 23,227 | 15,766 |

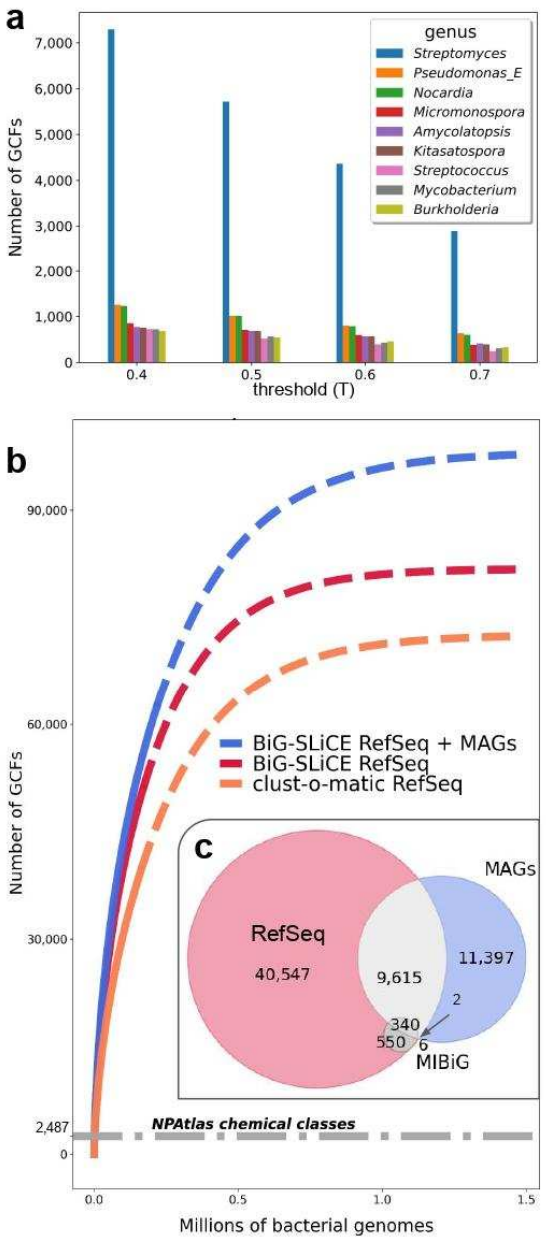



**Figure 1. Biosynthetic diversity of the sequenced bacterial kingdom**. Panel **a**: Bar plots of Gene Cluster Families (GCFs, as defined by BiG-SLiCE) of nine most biosynthetically diverse genera using different thresholds (T). The absolute number of GCFs changes from threshold to threshold, but the general tendencies (highest to lowest GCF count) are consistent between them. Panel **b**: Rarefaction curves of all RefSeq bacteria based on BiG-SLiCE (red) and based on clust-o-matic (orange), and rarefaction curve of the Complete Dataset, which includes bacterial MAGs (blue), based on BiG-SLiCE. BiG-SLiCE GCFs were calculated with T=0.4. Clust-o-matic GCFs were calculated with T=0.5. The solid lines represent interpolated and actual data, while the dotted lines represent extrapolated data. The number of chemical classes documented in NPAtlas[11], which come from bacterial producers (gray dotted line - 2,487), corresponds to 2.5% - 3.3% of the predicted potential of the bacterial kingdom (number of GCFs at 1.6 million genomes). The Y values (number of extrapolated GCFs) at the right end of the graph are 97,760.12 (blue), 81,748.32 (red) and 72,411.11 (orange). Panel **c**: Venn Diagram of GCFs (as defined by BiG-SLiCE, T=0.4) of the bacterial RefSeq, Minimum Information about a Biosynthetic Gene cluster (MIBiG[23]) and bacterial MAGs datasets. More information on the MiBIG dataset can be found in Supplementary Table 6. About 53,4% of the GCFs of MAGs are unique (blue shape) to this dataset.

This calibration of thresholds of GCFs to families of chemical structures allowed us to perform a rarefaction analysis to assess how genomically encoded biochemical

diversity (expressed as the number of distinct GCFs) increases with the number of sequenced and screened genomes (Figure 1b). The curve appears far from saturated, while the slope is steeper still if the bacterial MAGs are included in the analysis. When compared to the number of chemical classes documented in the NPAtlas[11] database (Figure 1b), it appears that, to date, only about 3% of the kingdom's biosynthetic diversity has been experimentally accessed.

In an attempt to evaluate the potential contribution of metagenomic data to Natural Product (NP) discovery, we studied how many of the GCFs found in the MAGs datasets were unique to this dataset (Figure 1c). Around 53,4% of GCFs in the MAGs were not found in the RefSeq strains or in the Minimum Information about a Biosynthetic Gene cluster database (MIBiG[23]). Paradoxically, in Figure 1b, the contribution of MAGs does not reflect this finding, but this is most likely because the metagenomic dataset is of limited size and does not cover the full microbial diversity of the biosphere. An analysis of the uniqueness of GCFs found in different environments, although only limited to one[20] of the MAGs datasets, suggests that a connection exists between the biogeography of microbiomes and the uniqueness of their biosynthetic diversity, as the majority of GCFs (74.43 %) are biome-specific (Extended Data Figure 2, Supplementary Table 7).The latter finding is concordant with recent proof that most genes have a strong biogeography signal[24].

### *Variation in biosynthetic diversity drops at genus level*

To identify the most promising bacterial producers, it is important to compare them at a specific taxonomic level. Several studies indicate that there is significant discontinuity in how BGCs are distributed across taxonomy: 'lower' taxonomic ranks like species within a genus carry more similar biosynthetic diversity, than 'higher' taxonomic ranks like phyla within a kingdom. To assess which taxonomic rank is the most appropriate to evaluate biosynthetic potential, we aimed to determine up to which taxonomic level the biosynthetic diversity remains homogeneous within that taxon. For this analysis, from our initial dataset, we left out the MAGs and only used the RefSeq bacterial strains as taxonomic assignment up to species rank (based on GTDB[14]) was available only for the latter dataset(Table 1).

We first decorated the GTDB[14] bacterial tree with GCF values from the BiG-SLiCE analysis (Figure 2a), revealing the biosynthetic diversity found within currently sequenced genomes at the phylum rank. It immediately stood out that biosynthetic diversity was differently dispersed among the bacterial phyla, in accordance with published data[7,25]. As expected for known NP producers, the phyla Proteobacteria and Actinobacteria appeared particularly diverse[8,26,27]. However, these phyla are amongst the most studied and therefore the most sequenced[8,26,27], a bias that was addressed later in the study.

Next, we examined whether the diversity of each phylum contributed to the domain's total diversity, or if there was overlap among them. For this reason, we depicted the

number of unique GCFs within each phylum, as well as the pairwise overlaps (Figure 2b). In most phyla, the vast majority (on average 73.81 ± 20.35%) of their GCFs appeared to be unique to them and not found anywhere else. This is coherent with the fact that HGT events, although relatively frequent for BGCs[28], are much more common among closely related taxa[29].

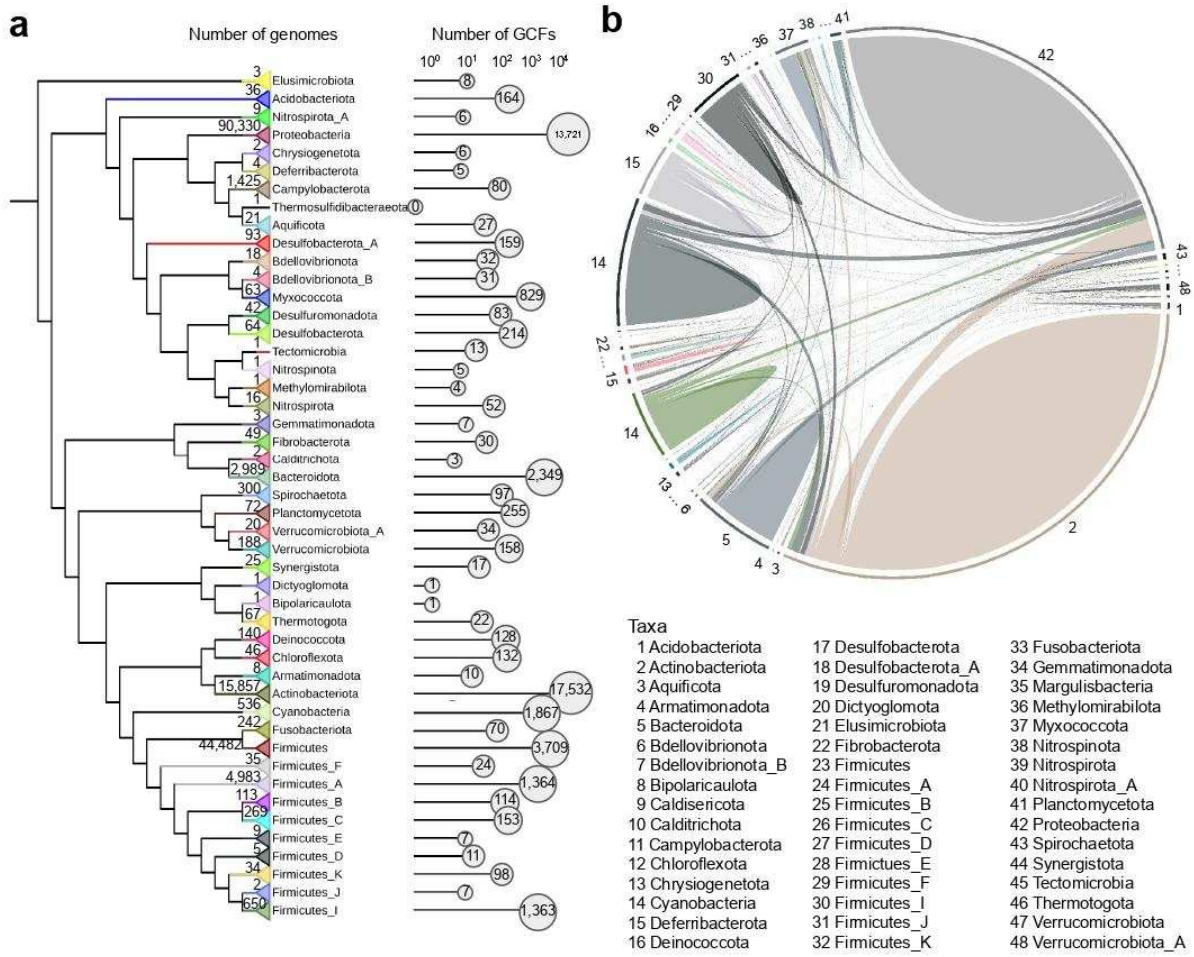



**Figure 2. Comparison of biosynthetic diversity among phyla**. Panel **a**: The Genome Taxonomy DataBase (GTDB[14]) bacterial tree was visualized with iTOL[30] v6.5.2, decorated with Gene Cluster Families (GCFs) values (as defined by BiG-SLiCE at T=0.4), collapsed at the phylum rank and accompanied by bar plot of GCFs in logarithmic scale ($10^0$ to $10^4$). The number of genomes belonging to each phylum is displayed next to the tree's leaf nodes. Panel **b**: GCFs, as defined by BiG-SLiCE (T=0.4), unique to phyla (solid shapes) and with pairwise overlaps between phyla (ribbons), visualized with circlize[31]. Each phylum has a distinct color. Actinobacteriota (2) and Proteobacteria (40) seem particularly rich in unique GCFs.

Once we obtained information on the diversity of different phyla, as well as the rest of the major taxonomic ranks (classes, orders, families, genera, species), we proceeded to determine from which taxonomic rank biosynthetic diversity levels no longer show high variability. Therefore, we conducted a variance analysis that included each taxonomic rank, from phylum to species. For each rank, the variance value was computed based on the #GCFs values of immediately lower-ranked taxa

(see Methods: Variance Analysis). The distribution of these variance values for each rank is visualized in Figure 3a.

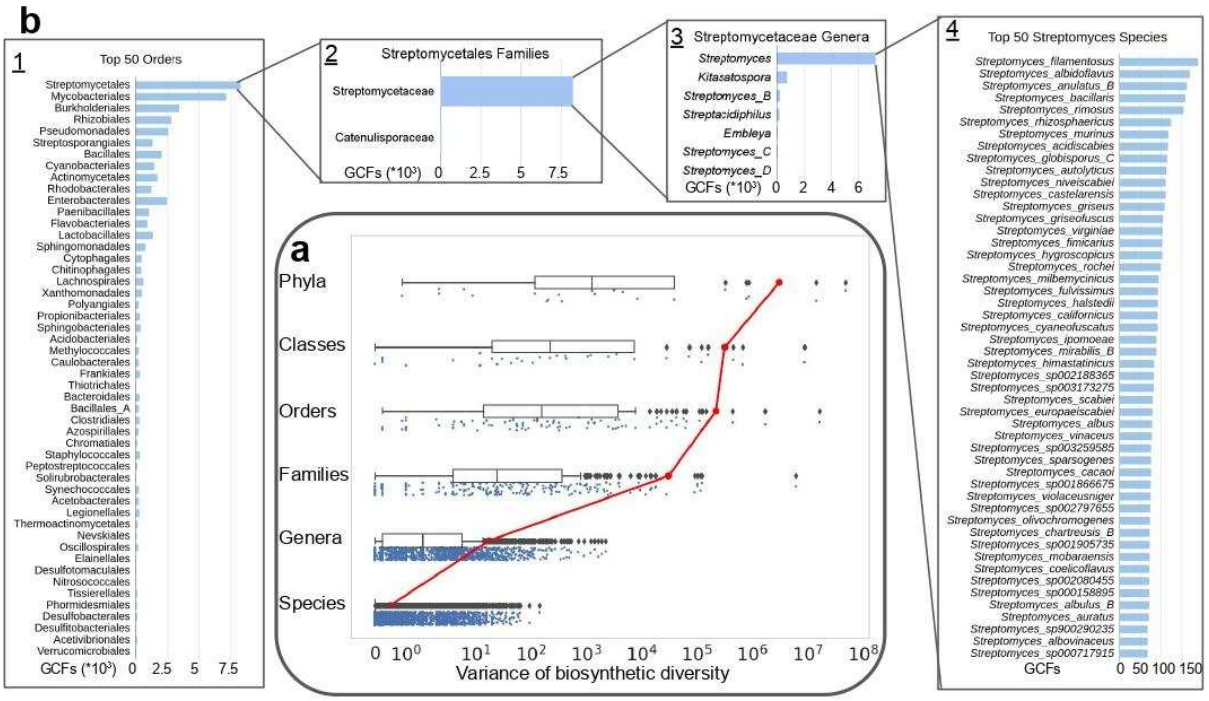



**Figure 3. Relations of taxonomic levels to variability in biosynthetic diversity**. Panel **a**: Modified "raincloud plots"[32] of major taxonomic ranks (X axis in logarithmic scale). Each boxplot represents the dispersion of variance values of a certain taxonomic rank, computed from the number of Gene Cluster Families (GCFs as defined by BiG-SLiCE at T=0.4) of the immediately lower rank. The boxplots' center line represents the median value; the box limits represent the upper and lower quartiles. Whiskers represent a 1.5x interquartile range. Points outside of the whiskers are outliers. Sample sizes are: Phyla n=21, Classes n=33, Orders n=89, Families n=224, Genera n=1,607, Species n=13,065. Jittered raw data points are plotted under the boxplots for better visualization of the values' distribution. The red line connects the mean variance values of each rank. There is a noticeable drop in dispersion of variance values from the family rank to the genus rank (see also Supplementary Figure 3), indicating that the genera are suitable taxonomic groups to be characterised as diverse and be compared to each other. Panel **b**: Biosynthetic diversity of various taxa, measured in absolute numbers of distinct GCFs as defined by BiG-SLiCE (T=0.4) from currently sequenced genomes. Top 50 most diverse orders (1), Streptomycetales families (2), Streptomycetaceae genera (3), top 50 most diverse *Streptomyces* species (4). The difference in variance is visible in the graphs 1,2,3, but becomes homogeneous at the species level as is shown in graph 4.

There is a noticeable drop in the range of variance values for each rank, while diversity becomes highly homogeneous at the species level (Figures 3a,b). The plunge is most striking from the family to the genus level (Figure 3a), with even the outliers all falling under the $10^3$- line in the genus rank. Different species within a genus are likely to display uniform biosynthetic diversity, while much dissimilarity is observed between different genera belonging to the same family (Figure 3b). Additional statistical analysis confirmed the significance of this observation (Supplementary Figure 3) thus pinpointing, for the first time, the genus rank as the most appropriate for comparative analyses.

### *Taxa that are sources of substantial biosynthetic diversity*

The identification of the genus level as the most informative rank to measure biosynthetic diversity across taxonomy paved the way for a comprehensive comparative analysis of biosynthetic potential across the bacterial tree of life. However, to be able to systematically compare diversity values among groups, said groups need to be uniform. In this case, a common phylogenetic metric was necessary. We chose Relative Evolutionary Divergence (RED) and a specific threshold that was based on the GTDB's range of RED values for the genus rank[14] to define REDgroups: groups of bacteria analogous to genera but characterized by equal evolutionary distance (see Methods: Definition of REDgroups). Our classification revealed the inequalities in within-taxon phylogenetic similarities among the genera, with some being divided into multiple REDgroups (for example the *Streptomyces* genus was split into 21 REDgroups: Streptomyces_RG1, Streptomyces_RG2 etc.) and some being joined together with other genera to form mixed REDgroups (for example Burkholderiaceae_mixed_RG1 includes the genera *Paraburkholderia*, *Paraburkholderia_A*, *Paraburkholderia_B*, *Burkholderia*, *Paraburkholderia_E* and *Caballeronia*). This disparity among the genera reaffirmed the importance of defining the REDgroups as a technique that allowed for fair comparisons among bacterial producers.

The resulting 3,779 REDgroups showed huge differences in biosynthetic diversity as measured by the numbers of GCFs found in genomes sequenced from these groups so far, with the maximum diversity at 3,339 GCFs, average at 17 GCFs and minimum at 1 GCF. Nevertheless, the variance of diversity within the REDgroups was even more uniform than in the genera (Supplementary Figure 4). Some of the top groups (Supplementary Table 8) included known rich NP producers, such as *Streptomyces*, *Pseudomonas_E* and *Nocardia*[23,26,27,30].

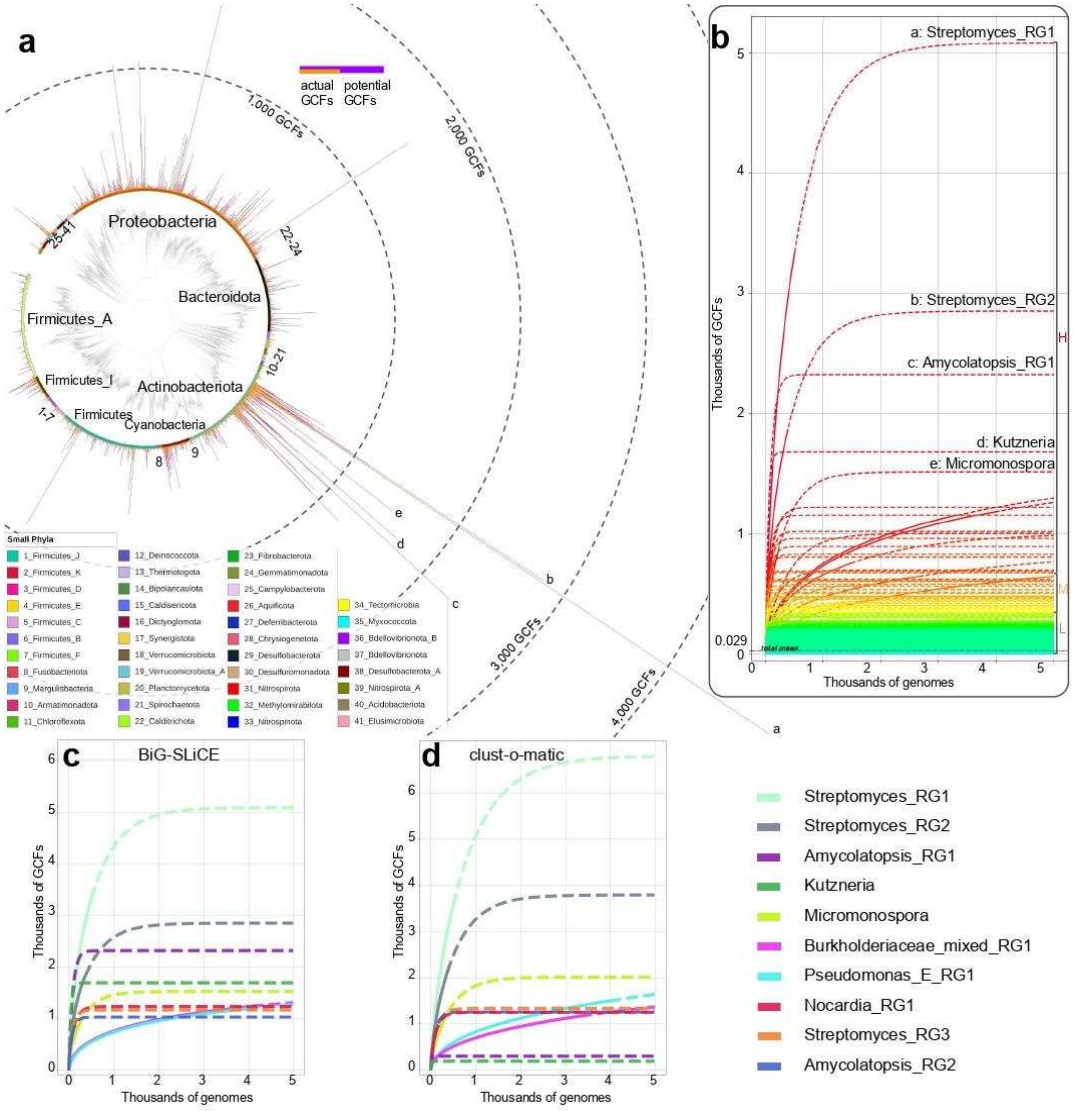



**Figure 4. Overview of actual and potential biosynthetic diversity of bacterial kingdom, compared at REDgroup level.** Panel **a:** GTDB[14] bacterial tree up to REDgroup level, visualized with iTOL[30] v6.5.2, colour coded by phylum, decorated with barplots of actual (orange) and potential (purple) Gene Cluster Families (GCFs), as defined by BiG-SLiCE (T=0.4). Top REDgroups with most potential GCFs include the following: A: Streptomyces_RG1, B: Streptomyces_RG2, C: Amycolatopsis_RG1, D: Kutzneria, E: Pseudomonas_E. Phyla known to be enriched in NP producers are immediately visible (Actinobacteriota, Protobacteriota), with the most promising groups coming from the Actinobacteriota phylum (the highest peak belongs to a REDgroup containing *Streptomyces* strains). Simultaneously, within the underexplored phyla, there seems to be significant biosynthetic diversity and potential. An interactive version of Figure 4a can be accessed online (Extended Data Figure 3). Panel **b:** Rarefaction curves of REDgroups (BiG-SLiCE T=0.4). In panels b, c and d the solid lines represent interpolated and actual data, while the dotted lines represent extrapolated data. The letters “L”, “M” and “H” correspond to Low- (0-389 pGCFs), Medium- (390-649 pGCFs) and High-diversity (more than 650 pGCFs) producers. The “L” range includes 3,737 REDgroups (shades of green), the “M” range includes 22 (shades of yellow/orange), while the “H” range includes 20 REDgroups (shades of red). The vast majority of REDgroups belong to the low-diversity producers (the mean of all REDgroups’ pGCFs is 29). The labels of most promising REDgroups are indicated (the letters a-e correspond to the peaks in panel a). *Streptomyces* strains are included in several of them. Panel **c:** Rarefaction curves of the most promising REDgroups (BiG-SLiCE T=0.4). Panel **d:** Rarefaction curves of the most promising REDgroups (clust-o-matic T=0.5). Though the exact numbers differ, the similarities between the two methods are apparent.

Although very informative, this analysis is biased because of large differences in the number of sequenced strains among the groups, with the economically or medically important strains having been sequenced more systematically than others. To overcome this bias, rarefaction analyses were conducted for each REDgroup (Figure 4b, Supplementary Table 8), as performed in previous studies[31,32]. Additionally, to examine how effectively this method overcomes the sequencing bias, a random sampling approach was taken (see Methods: Random sampling), which showed comparable results to the original analysis (Supplementary Table 9). With all the information on REDgroups, and in order to provide a global overview of the actual biosynthetic diversity and the potential number of GCFs, we modified and complemented the bacterial tree from Parks *et. al.*[14], as shown in Figure 4a (Extended Data Figure 3). The dispersion of these values across the various phyla can also be seen, with the exceptional outliers standing out: Streptomyces_RG1, Streptomyces_RG2, Amycolatopsis_RG1, Kutzneria, and Micromonospora. All these are groups known for their NP producers[8,26,27,33] and they remain in the top (Supplementary Table 8), seemingly having much unexplored biosynthetic potential.

To ensure that our conclusions are not the product of algorithmic artifacts, we reran the analysis using an alternative method of quantifying biosynthetic diversity, which was developed independently, yet for the same purpose. This alternative approach, called clust-o-matic, is based on a sequence similarity all-versus-all distance matrix of BGCs and subsequent agglomerative hierarchical clustering in order to form GCFs (see Methods: Quantification of biosynthetic diversity with clust-o-matic). Like for BiG-SLiCE, we calibrated the threshold for clust-o-matic based on NP Atlas clusters. When comparing the results (Figure 4c,d, Supplementary Table 8), despite slight differences in absolute numbers, the two algorithms appeared to identify very similar trends.

*Streptomyces*, even when split into multiple REDgroups, is in the top groups both based on the known biosynthetic diversity and based on the estimated potential values. 5,908 (+103 Streptomyces_B, +39 Streptomyces_C, +16 Streptomyces_D) GCFs appear to be unique to the group, even among other phyla (Figure 5a). This is in agreement with previous studies investigating how much overlap there is among the main groups of producers[34]. What is more, streptomycetes appear to be the source of a good percentage of the biosynthetic diversity attributed to the Actinobacteria phylum, as seen in Figure 5b.

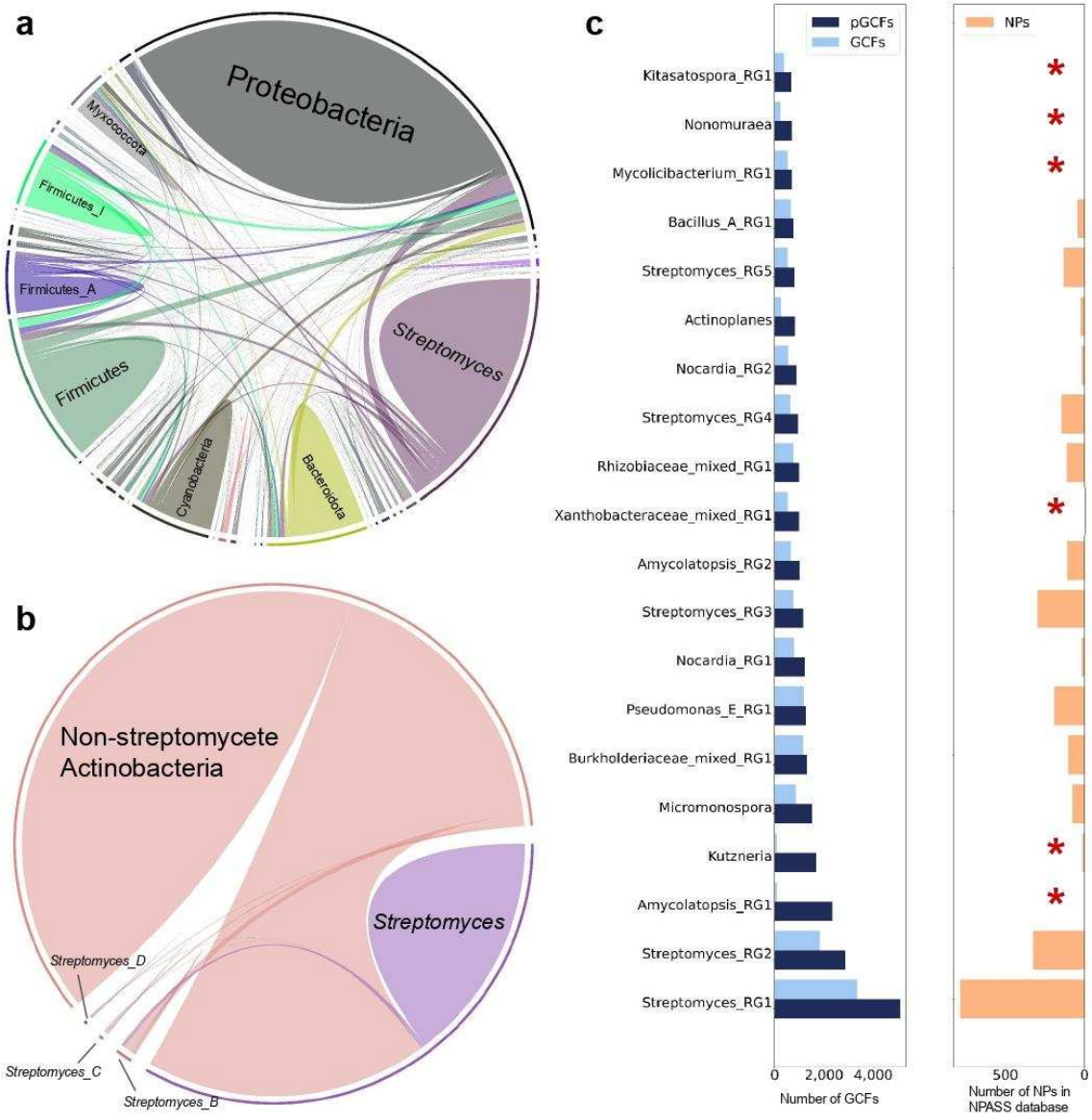



**Figure 5. Unique diversity in the known producer *Streptomyces* and promising potential of less popular taxa**. Panel **a**: Unique Gene Cluster Families (GCFs) as defined by BiG-SLiCE (T=0.4), of phyla and *Streptomyces* (solid shapes) and pairwise overlaps of phyla - phyla and phyla - *Streptomyces* (ribbons), visualized with circlize[31]. Each taxon has a distinct color. The smaller shapes and ribbons represent smaller phyla that can be seen in Extended Data Figure 4. The genus *Streptomyces* appears to have a very high amount of unique GCFs comparable to entire phyla, such as Proteobacteria. Panel **b**: Unique GCFs as defined by BiG-SLiCE (T=0.4), of non-streptomycete Actinobacteriota and all *Streptomyces* genera (solid shapes) and pairwise overlaps between Actinobacteriota and *Streptomyces* (ribbons), visualized with circlize[31]. The *Streptomyces* genus, only one of many belonging to the Actinobacteriota phylum, appears to be responsible for a big percentage of the phylum's unique diversity (see big pink ribbon). Panel **c**: Left: Potential (pGCFs) and actual (GCFs) number of Gene Cluster Families as defined by BiG-SLiCE (T=0.4), of top 20 most promising REDgroups. Right: number of Natural Products (NPs) found in the NPASS database[38], that originate from species included in each REDgroup. The REDgroups with few (< 15) to no known NPs associated with them are marked with red stars on the right side of the graph. Several of the displayed groups are in the latter category: Amycolatopsis_RG1, Kutzneria, Xanthobacteraceae_mixed_RG1 (containing the genera *Bradyrhizobium, Rhodopseudomonas, Tardiphaga* and *Nitrobacter*), Mycolicibacterium_RG1, Nonomuraea, Kitasatospora_RG1.

However, taxa less popular for NP discovery also show promise, as was evidenced by a comparison of our results with data from the NPASS database of Natural

Products[35] (Figure 5c). Among the 20 overall most promising REDgroups we found at least 6 groups that show promise but whose members are either not catalogued in the database as NP sources or are connected to few (<15) known compounds: Amycolatopsis_RG1, Kutzneria, Xanthobacteriaceae_mixed_RG1, Mycolicibacterium_RG1, Nonomuraea, Kitasatospora_RG1. The Amycolatopsis_RG1 group only includes three rare species: *Amycolatopsis antarctica, marina* and *nigrescens*. Other promising REDgroups with very few known producers include Cupriavidus (from Proteobacteria phylum), Weeksellaceae_mixed_RG1 (from Bacteroidota phylum) and Pleurocapsa (from Cyanobacteria phylum). More information about the promising underexplored taxa can be found in Supplementary Table 8.

## Discussion

Using two different algorithms, we mined deposited bacterial sequencing data to identify Biosynthetic Gene Clusters (BGCs) and grouped them into gene cluster families (GCFs) according to chemical families of encoded compounds. We identified maximal emergence of the highest biosynthetic diversity close to the genus rank and chose to further investigate analogous taxonomic groups (REDgroups). Rarefaction analysis identified the highest biosynthetic potential and the most promising bacterial taxa among many known diverse groups as well as multiple promising understudied producers. To the best of our knowledge, this is the largest survey of secondary metabolite production to date, and our study provides a reproducible pipeline to underpin drug discovery efforts.

The biosynthetic capacity of the bacterial kingdom was previously assessed by Cimermancic *et. al.*[7], but the dataset analysed was 33,000 BGCs compared with the 1,185,995 BGCs we analysed. Additionally they used ClusterFinder, which is known as a more exploratory identification tool[7,36]. Projects that exploit publicly available genomic data are reliant on the quality of genomes sequenced as well as the efficiency of available genome mining methods, which have some limitations[37]. For instance, the study of GCF uniqueness among taxa may be affected by antiSMASH's imperfect BGC boundary prediction[12]. Even though BiG-SLiCE converts BGCs into features based only on domains related to biosynthesis[21], genomic context unrelated to the biosynthetic pathway of a BGC could still have a role in the GCF assignment; this issue cannot be fully addressed with currently available tools. However, antiSMASH's ability to discern cluster limits and detect BGCs from cultured strains and MAGs is comparable to alternative tools, while its ability to predict different BGC types is unparalleled[38], as is apparent from its common use in Natural Product (NP) research[7,9,25,30,32,39]. What is more, the fact that it is rule-based[12] implies the possibility of undetected types of clusters and increases the likelihood that our calculations have underestimated the true biosynthetic potential of bacterial organisms.

Furthermore, our pipeline was the first to use the GTDB[14] taxonomy for studying global bacterial biosynthetic diversity. This enabled us to avoid misclassifications of NCBI taxonomic placement[40–43]. The use of rarefaction curves allowed us to infer the

biosynthetic potential of bacterial groups, as done in some smaller-scaled projects[7,8,31,32]. This method aims to enable fair comparisons among incomplete samples[44]. However, while overestimation is not expected to happen, for those groups that contain very few genomes, there is a tendency to underestimate their potential capacity[44]. Hence, sequencing bias of popular taxa still affects our results. We tried to minimize the bias within the pipeline as much as possible while retaining high diversity of bacterial taxa; therefore, we decided not to exclude REDgroups with very few members from the dataset. We also ran an additional random sampling analysis using the most populated REDgroups and confirmed the reproducibility of our results. Nonetheless, the remaining bias will only be eliminated with the inclusion of increased biodiversity in sequencing projects[17,20].

Our analysis identified a plethora of unexplored taxonomic groups with substantial biosynthetic potential[9,10,45–47]. At the same time, it revealed that undiscovered biosynthetic diversity present in well-characterized NP producers. For example, multiple Proteobacteria taxa were identified among the top producers: *Pseudomonas*, *Pseudoalteromonas*, *Paracoccus*, *Serratia* among others. This is in accordance with the known biosynthetic potential of the Proteobacteria phylum[33]. Furthermore, we identified taxa that are less well represented in sequence databases as being potentially useful sources of secondary metabolites, including myxobacterial genera *Cystobacter*, *Melittangium*, *Archangium*, *Vitiosangium*, *Sorangium* and *Myxococcus*[9,30,48], and *Chryseobacterium* and *Chryseobacterium_A*[49] from the Bacteroidota phylum. However, the most diverse groups of metabolites are predicted to be produced by actinobacterial strains of well-known and well-studied NP producers such as *Actinoplanes*, *Amycolatopsis*, *Micromonospora*, *Mycobacterium*, *Nocardia* and *Streptomyces*[8,26,27,34]. These bacteria produce most of the natural product antibiotics[26] and our analysis confirms that recent analyses of biosynthetic novelty in the genomes of rare actinobacteria suggest that there is still much more natural product diversity to be discovered in this group as more diversified strains get sequenced[8,26,27,50].

*Streptomyces* is a genus of the Actinobacteria phylum that contains some of the most complex bacteria that we know of, though by far not the most sequenced in our dataset (Supplementary Figure 5). These bacteria have been known as NP producers for a long time[34], as single strains containing a high number of BGCs have been discovered, taking up to 10% of their genome[51]. However, members of other genera contain comparable absolute numbers of BGCs. This is the first time that a systematic comparison of the diversity of the encoded compounds within bacterial genera has been conducted, revealing how diverse *Streptomyces* are compared to all others[34]. The factors that cause this taxonomic group to stand out are not completely clear but probably related to their sophisticated lifestyle. Many observations suggest that NP biosynthesis drives speciation within the *Streptomyces* genus[8]. The exploration of factors that led to the rise of biosynthetic diversity in

*Streptomyces* to such an impressive degree will be the subject of further investigations in the future.

Having the genomic capacity for the biosynthesis of secondary metabolites does not always herald the discovery of a novel chemistry[52,53]. Sometimes, the bacterium in question cannot be grown or BGCs are not expressed in laboratory conditions[26,45,47,52,53]. This issue is related to the complexity of BGCs; we have only just scratched the surface of their intricate regulation and connection to primary metabolism[5,45,52,54]. However, efforts to decode biosynthetic mechanisms for the activation of silent clusters need to be tailored to specific producer groups[26,27,53], such as groups phylogenetically related to promising producers, e.g. members of the Pseudonocardiaceae family (REDgroups Amycolatopsis_RG1 & Kutzneria in Figure 4, these and more REDgroups in Supplementary Table 8), partly on the grounds that each phylum has unique diversity (Figure 2b).

Original approaches to the prioritization issue of NP research continue to emerge, fuelled by the advances in metagenomics and computational tools that enable the use of the biosynthetic potential of unculturable bacteria from environmental samples[55]. Furthermore, apart from the few metagenomic projects whose MAGs we incorporated in the first part of our analysis, there are multiple such projects publicly available, some of which have been the focus of NP studies[56]. Although the reconstruction of genomes from metagenomes remains a challenge[57] and the assembly will often miss BGCs[58], which has indirectly prevented their comparison to the cultured bacteria in the current project, metagenomics is proving a promising source of information on NPs and their producers[7,34,45,55,56], as made apparent in the present investigation. We expect the effect of this field on NP research to become more evident in the following years.

The collection of microbial data from a large variety of habitats points to another interesting aspect, namely the relation between the biome of origin of the producers and the uniqueness of their biosynthetic diversity. Although this connection has been investigated to some extent[24,25,32,33,47] drawing more definitive conclusions will require the use of a wider-scale dataset and the availability of more detailed and standardized metadata of producers' genomes.

Our analysis provides a global overview of diverse known and promising understudied NP-producing taxa. We expect this to greatly help overcome one of the main bottlenecks of Natural Product discovery: the prioritization of producers for research[55].

## Online Methods

### *BGC data set*

We obtained 170,585 complete and draft bacterial genomes (Table 1) from RefSeq[13] on 27 March 2020. Furthermore, a dataset of 47,098 MAGs was included in the first part of the analysis (see Results: Biosynthetic diversity of the bacterial kingdom). For the rest of the study, we used only 161,290 RefSeq bacterial genomes whose taxonomic classification up to the species level was known (Table 1). All genomes were analyzed with antiSMASH (version 5)[12], which identified their BGCs (Supplementary Table 1). The entirety of the MIBiG[23] database (accessed on 27 March 2020) was included in parts of our analysis (their IDs can be found in Supplementary Table 6).

### *Taxonomic classification*

Due to multiple indications regarding a lack of accuracy of NCBI's taxonomic classification of bacterial genomes[40–43], we chose to use the Genome Taxonomy Database (GTDB[14]) instead. The bacterial tree of 120 concatenated proteins (GTDB release 89), as well as the classifications of organisms up to the species level, were included in the analysis.

### *Quantification of biosynthetic diversity with BiG-SLiCE*

For a bacterium to be regarded as biosynthetically diverse, we considered not the number of BGCs important, but rather how different these BGCs are to each other. In order to quantify this diversity, we analyzed all BGCs with the new BiG-SLiCE tool[21], which groups similar clusters into Gene Cluster Families (GCFs). However, the first version of this tool has an inherent bias towards multi-protein families BGCs, producing uneven coverage between BGCs of different classes (i.e., due to their lack of biosynthetic domain diversity, all lanthipeptide BGCs may be grouped together using the Euclidean threshold of T=900, which in contrast is ideal for clustering Type-I Polyketide BGCs). To alleviate this issue and provide a fair measurement of biosynthetic diversity between the taxa, we modified the original distance measurement by normalizing the BGC features under L^2-norm, which will produce a cosine-like distance when processed by the Euclidean-based BIRCH algorithm. This usage of cosine-like distance will virtually balance the measured distance between BGCs with “high” and “low” feature counts (Supplementary Figure 6a), in the end providing an improved clustering performance when measured using the reference data of manually-curated MIBiG GCFs (Supplementary Figure 6b).

The GTDB[14] (release 89) bacterial tree was pruned so that it included only the organisms that are part of our dataset. Then, having both the taxonomic classification of all bacteria, as well as how many GCFs their BGCs group into, the pruned GTDB tree was decorated with #GCFs values at each node. This allowed for the evaluation of the biosynthetic diversity of any clade, including the main taxonomic ranks. To pick a single threshold for subsequent taxonomy richness analysis, we compared BiG-SLiCE results on 947 MIBiG BGCs versus the compound-based clustering provided by the NPAtlas database[11] (Supplementary

Figure 2). A final threshold of T=0.4 was chosen based on its similarity to NPAtlas's compound clusters (V-score=0.9X, GCF counts difference=+XX).

### *Quantification of biosynthetic diversity with clust-o-matic*

We aimed to repeat and evaluate the reproducibility of the BGC-to-GCF quantification step of BiG-SLiCE with an alternative, independently derived algorithm. For that instead of grouping BGCs into GCFs based on biosynthetic domain diversity, we developed an algorithm that considers full core biosynthetic genes. Biosynthetic gene clusters that were detected in the input data by antiSMASH 5.1 were parsed to deliver core biosynthetic protein sequences. Those protein sequences were subjected to all-against-all multi-gene sequence similarity search with DIAMOND[59] 2.0 using default settings. Only one best hit per query core gene per BGC was allowed divided by a total core protein length, resulting in the final pairwise BGC score always being within range of 0 to 1. Pairwise BGC similarity scores were used to build a distance matrix that was later subjected to agglomerative hierarchical clustering in python programming language (package scipy.cluster.hierarchy). The same process as described in the paragraph above (for BiG-SLiCE in that case) was performed for identification of the most suitable threshold for the clust-o-matic algorithm. The determined optimal threshold of 0.5 was then used to generate GCFs, which were then fed into the next steps in parallel to the original set of GCFs obtained from BiG-SLiCE.

### *Biogeography Analysis*

One[20] of the MAGs datasets was accompanied by sufficient metadata that allowed for a study of a potential connection between biosynthetic diversity patterns and the biomes of origin of the corresponding MAGs. The GCFs for each ecosystem type were collected by combining information from Supplementary Tables 1, 2 of this project and from the Nayfach paper[20] Supplementary Information. This led to the creation of Supplementary Table 7. Then, the largest occurring intersections were computed and visualised in Extended Data Figure 2 using the UpSet[60] visualisation technique.

### *Variance Analysis*

In order to pinpoint the emergence of biosynthetic diversity, the within-taxon homogeneity was compared among the main taxonomic ranks. For each rank, the variance value was computed (with NumPy[61]) based on the #GCFs values of immediately lower-ranked taxa, as long as there were at least two such taxa. For example, a phylum that includes only one class in our dataset was omitted from this computation. But a phylum with two or more classes would be assigned a variance value computed from its classes' #GCFs values. The distribution of these variance values was plotted for each rank in Figure 3a. We noticed a significant reduction in variance from the family to the genus rank, which was confirmed with an additional statistical test (Supplementary Figure 3, Supplementary Methods). A similar variance analysis was performed to compare genera and REDgroups (Supplementary Figure

4) but in this case variance was calculated based on the strains' biosynthetic diversity.

### *Definition of REDgroups*

To study the biosynthetic diversity of genera, we attempted to achieve uniform taxa. The creators of GTDB used Relative Evolutionary Divergence (RED) for taxonomic rank normalization[14]; it is a metric that relies heavily on the branch length of a phylogenetic tree and is consequently dependent on the rooting. The GTDB developers provided us with a bacterial tree decorated with the average RED values of all plausible rootings at each node. Since GTDB accepts a range of RED values for each taxonomic rank placement[14], we chose the median of GTDB genus RED values, namely 0.934, as a cutoff threshold. Any clade in the GTDB bacterial tree with an assigned RED value higher than the threshold was considered one group (Supplementary Figure 7) that we named "REDgroup". For REDgroup naming conventions, see Supplementary Figure 7.

### *Rarefaction analysis*

The extrapolation of potential #GCFs values was achieved by conducting rarefaction analyses, by use of the iNEXT R package[62]. A GCF presence/absence table (GCF-by-strain matrix) was constructed for each group considered and was then used as "incidence-raw" data in the iNEXT main function, where 500 points were inter- or extrapolated with an endpoint of 5000 for the REDgroups, and of 1.6 million (about 8 times the number of strains in the Complete Dataset) in each group for the RefSeq analyses (where 2000 points were inter- or extrapolated). By default, the number of bootstrap replications is 50.

### *Random sampling*

In order to test whether the above methods (creation of REDgroups and the subsequent rarefaction analyses) overcome the inherent sequencing bias in our dataset, a random sampling technique was used. A reduced dataset was tested that included only those REDgroups containing at least 20 members. For each REDgroup, a sample of 20 genomes was randomly chosen (using the Python "random" module), while preserving the species diversity of the group. The latter was achieved by ensuring that genomes belonging to as many species as possible are included in each sample; if all species of a REDgroup were included but the genomes were fewer than 20, the remaining "spots" were distributed evenly among a random sample of the REDgroup's species. One hundred iterations of this process were calculated for all REDgroups in this reduced dataset and rarefaction analyses were conducted for the random samples in each iteration. Finally, the average potential GCFs (pGCFs) value for each REDgroup from all iterations was calculated and reported in Supplementary Table 9.

### *Identification of unknown producers*

We investigated the genera included in the most promising REDgroups, to find out whether they include species that are producers of known compounds. Hence, the species names were cross-referenced with the species named as producers in the NPASS depository[35] (accessed on 15 October 2020), taking care to match the GTDB-given names to the NCBI-given names that the database uses.

## Data availability

The datasets generated and analyzed during the current study are available in the following zenodo repository: https://doi.org/10.5281/zenodo.6365726.

## Code availability

The clust-o-matic code is available here: https://github.com/Helmholtz-HIPS

The modified BiG-SLiCE script (that accepts as input a regular BiG-SLiCE output folder, then outputs the GCF membership in a tsv file) is available both in our zenodo repository (file name: perform_l2norm_clustering.py) and under the following link: https://github.com/medema-group/bigslice/blob/master/misc/useful_scripts/perform_l2norm_clustering.py

## References


1. O'Connor, S. E. Engineering of Secondary Metabolism. *Annu. Rev. Genet.* **49**, 71–94 (2015).
2. Newman, D. J. & Cragg, G. M. 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).
3. Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature* **529**, 336–343 (2016).
4. Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., International Natural Product Sciences Taskforce & Supuran, C. T. Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* **20**, 200–216 (2021).
5. Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes - a review. *Nat. Prod. Rep.* **33**, 988–1005 (2016).
6. Medema, M. H., de Rond, T. & Moore, B. S. Mining genomes to illuminate the specialized chemistry of life. *Nat. Rev. Genet.* (2021) doi:10.1038/s41576-021-00363-7.
7. Cimermancic, P. *et al.* Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
8. Doroghazi, J. R. *et al.* Aroadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **10**, 963–968 (2014).
9. Hoffmann, T. *et al.* Correlating chemical diversity with taxonomic distance for discovery of natural products in myxobacteria. *Nat. Commun.* **9**, 803 (2018).
10. Lewis, K. The Science of Antibiotic Discovery. *Cell* vol. 181 29–45 (2020).
11. van Santen, J. A. *et al.* The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Cent Sci* **5**, 1824–1833 (2019).
12. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
13. Haft, D. H. *et al.* RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* **46**, D851–D860 (2018).
14. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
15. Stewart, R. D. *et al.* Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961 (2019).
16. Glendinning, L., Stewart, R. D., Pallen, M. J., Watson, K. A. & Watson, M. Assembly of hundreds of novel bacterial genomes from the chicken caecum. *Genome Biol.* **21**, 34 (2020).
17. Almeida, A. *et al.* A unified sequence catalogue of over 280,000 genomes obtained from the human gut microbiome. doi:10.1101/762682.
18. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft

metagenome-assembled genomes from the global oceans. *Sci Data* **5**, 170203 (2018).
19. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).
20. Nayfach, S. *et al.* Author Correction: A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 521 (2021).
21. Kautsar, S. A., van der Hooft, J. J. J., de Ridder, D. & Medema, M. H. BiG-SLiCE: A Highly Scalable Tool Maps the Diversity of 1.2 Million Biosynthetic Gene Clusters. *Bioinformatics* (2020).
22. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
23. Kautsar, S. A. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).
24. Coelho, L. P. *et al.* Towards the biogeography of prokaryotic genes. *Nature* (2021) doi:10.1038/s41586-021-04233-4.
25. Sharrar, A. M. *et al.* Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type. *MBio* **11**, (2020).
26. Barka, E. A. *et al.* Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiol. Mol. Biol. Rev.* **80**, 1–43 (2016).
27. Genilloud, O. Actinomycetes: still a source of novel antibiotics. *Nat. Prod. Rep.* **34**, 1203–1232 (2017).
28. Chevrette, M. G. *et al.* The confluence of big data and evolutionary genome mining for the discovery of natural products. *Nat. Prod. Rep.* (2021) doi:10.1039/d1np00013f.
29. Chase, A. B., Sweeney, D., Muskat, M. N., Guillén-Matus, D. & Jensen, P. R. Vertical inheritance governs biosynthetic gene cluster evolution and chemical diversification. doi:10.1101/2020.12.19.423547.
30. Männle, D. *et al.* Comparative Genomics and Metabolomics in the Genus Nocardia. *mSystems* **5**, (2020).
31. Ziemert, N. *et al.* Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E1130–9 (2014).
32. Adamek, M. *et al.* Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species. *BMC Genomics* **19**, 426 (2018).
33. Buijs, Y. *et al.* Marine Proteobacteria as a source of natural products: advances in molecular tools and strategies. *Nat. Prod. Rep.* **36**, 1333–1350 (2019).
34. Berdi, J. Bioactive Microbial Metabolites: A Personal View. *Journal of Antibiotics. Antibiotics* **58**, 1–26 (2005).
35. Zeng, X. *et al.* NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.* **46**, D1217–D1222 (2018).
36. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11**, 639–648 (2015).
37. Miller, M. E. *et al.* Increased virulence of Puccinia coronata f. sp.avenae populations through allele frequency changes at multiple putative Avr loci. *PLoS Genet.* **16**, e1009291 (2020).
38. Chavali, A. K. & Rhee, S. Y. Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. *Brief. Bioinform.* **19**, 1022–1034 (2018).
39. Adamek, M., Alanjary, M. & Ziemert, N. Applied evolution: Phylogeny-based approaches in natural products research. *Natural Product Reports* vol. 36 1295–1312 (2019).
40. Ciufo, S. *et al.* Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int. J. Syst. Evol. Microbiol.* **68**, 2386–2392 (2018).
41. Martínez-Romero, E. *et al.* Genome misclassification of Klebsiella variicola and Klebsiella quasipneumoniae isolated from plants, animals and humans. *Salud Publica Mex.* **60**, 56–62 (2018).
42. Mateo-Estrada, V., Graña-Miraglia, L., López-Leal, G. & Castillo-Ramírez, S. Phylogenomics Reveals Clear Cases of Misclassification and Genus-Wide Phylogenetic Markers for Acinetobacter. *Genome Biol. Evol.* **11**, 2531–2541 (2019).
43. Rekadwad, B. N. & Gonzalez, J. M. Correcting names of bacteria deposited in National Microbial Repositories: an analysed sequence data necessary for taxonomic re-categorization of misclassified bacteria-ONE example, genus Lysinibacillus. *Data Brief* **13**, 761–778 (2017).
44. Chao, A. *et al. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies*. vol. 84 45–67 http://purl.oclc.org/estimates (2014).
45. Hug, J. J., Bader, C. D., Remškar, M., Cirnski, K. & Müller, R. Concepts and Methods to Access Novel Antibiotics from Actinomycetes. *Antibiotics (Basel)* **7**, (2018).
46. Ling, L. L. *et al.* A new antibiotic kills pathogens without detectable resistance. *Nature* **517**,

455–459 (2015).
47. Subramani, R. & Sipkema, D. Marine rare actinomycetes: A promising source of structurally diverse and unique novel natural products. *Marine Drugs* vol. 17 (2019).
48. Weissman, K. J. & Müller, R. Myxobacterial secondary metabolites: bioactivities and modes-of-action. *Nat. Prod. Rep.* **27**, 1276–1295 (2010).
49. Dahal, R. H., Chaudhary, D. K., Kim, D.-U., Pandey, R. P. & Kim, J. Chryseobacterium antibioticum sp. nov. with antimicrobial activity against Gram-negative bacteria, isolated from Arctic soil. *J. Antibiot.* **74**, 115–123 (2021).
50. Schorn, M. A. *et al.* Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. *Microbiology* **162**, 2075–2086 (2016).
51. Baltz, R. H. Gifted microbes for genome mining and natural product discovery. *J. Ind. Microbiol. Biotechnol.* **44**, 573–588 (2017).
52. Seyedsayamdost, M. R. Toward a global picture of bacterial secondary metabolism. *Journal of Industrial Microbiology and Biotechnology* vol. 46 301–311 (2019).
53. Wohlleben, W., Mast, Y., Stegmann, E. & Ziemert, N. Antibiotic drug discovery. *Microb. Biotechnol.* **9**, 541–548 (2016).
54. van Bergeijk, D. A., Terlouw, B. R., Medema, M. H. & van Wezel, G. P. Ecology and genomics of Actinobacteria: new concepts for natural product discovery. *Nat. Rev. Microbiol.* **18**, 546–558 (2020).
55. Tracanna, V., de Jong, A., Medema, M. H. & Kuipers, O. P. Mining prokaryotes for antimicrobial compounds: From diversity to function. *FEMS Microbiology Reviews* vol. 41 417–429 (2017).
56. Chen, R. *et al.* Discovery of an Abundance of Biosynthetic Gene Clusters in Shark Bay Microbial Mats. *Front. Microbiol.* **11**, 1950 (2020).
57. Ghurye, J. S., Cepeda-Espinoza, V. & Pop, M. Metagenomic Assembly: Overview, Challenges and Applications. *Yale J. Biol. Med.* **89**, 353–362 (2016).
58. Mantri, S. S. *et al.* Metagenomic Sequencing of Multiple Soil Horizons and Sites in Close Vicinity Revealed Novel Secondary Metabolite Diversity. *mSystems* **6**, e0101821 (2021).
59. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
60. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).
61. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
62. Hsieh, T. C., Ma, K. H. & Chao, A. iNEXT: an R package for rarefaction and extrapolation of species diversity ( H ill numbers). *Methods Ecol. Evol.* **7**, 1451–1456 (2016).
63. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
64. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* vol. 30 2811–2812 (2014).
65. Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R. & Kievit, R. A. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res* **4**, 63 (2019).

## Acknowledgements

A.G. is grateful for the support of the Deutsche Forschungsgemeinschaft (DFG; Project ID # 398967434-TRR 261). N.Zi. is supported by the German Center for Infection Research (DZIF) (TTU 09.716). M.H.M. is supported by an European Research Council Starting Grant 948770-DECIPHER. S.K. was supported by the Graduate School for Experimental Plant Sciences (EPS) of Wageningen University. Work in the lab of R.M. is supported by BMBF (16GW0243), DFG and DZIF (807-5-8-0982600). A.G. and N.Zi. thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2124 – 390838134 for the infrastructural support. A.G. thanks M. Direnc Mungan for valued discussions on optimizing the analysis, as well as Caner Bagci for his imaginative suggestion on dealing with large data. We also thank Dr. Libera do Presti for invaluable comments on the manuscript.

## Author contributions

A.G., S.A.K., N.Za. and D.K. have performed the analysis. S.A.K. and N.Za. have contributed analysis tools. A.G., D.K., R.M., M.H.M. and N.Zi. have written the paper. All authors have contributed to the conception and design of the analysis. All authors have read and agreed to the published version of the manuscript.

## Competing interests

M.H.M. is a co-founder of Design Pharmaceuticals and a member of the scientific advisory board of Hexagon Bio. The other authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at https://doi.org/10.1038/s41564-022-01110-2.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41564-022-01110-2.

**Correspondence and requests for materials** should be addressed to Marnix H. Medema or Nadine Ziemert.

**Peer review information** Nature Microbiology thanks Nigel Mouncey and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

# Chapter 3: Animating insights into the biosynthesis of glycopeptide antibiotics

*(Advanced manuscript awaiting submission)*

Athina Gavriilidou[1], Martina Adamek[1,2,3], Jens-Peter Rodler[2,4], Noel Kubach[1,2], Anna Voigtländer[5], Leon Kokkoliadis[2], Chambers Hughes[2,3,4], Max J. Cryle[6,7], Evi Stegmann*[2,3,4], Nadine Ziemert*[1,2,3]

1: Translational Genome Mining for Natural Products, Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Tübingen, Germany

2: Cluster of Excellence 'Controlling Microbes to Fight Infections' (CMFI), University of Tübingen, Tübingen, Germany

3: German Centre for Infection Research (DZIF), Partnersite Tübingen, Tübingen, Germany

4: Microbial Bioactive Compounds, Interfaculty Institute of Microbiology and Infection Medicine Tübingen, University of Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany

5: Center for Media Competence (ZFM), University of Tübingen, Tübingen, Germany

6: Department of Biochemistry and Molecular Biology, The Monash Biomedicine Discovery Institute, Monash University, Clayton, VIC, 3800, Australia

7: EMBL Australia, Monash University, Clayton, Victoria 3800, Australia

* co-corresponding author

Personal contributions:

Scientific ideas/data generation/writing: 90/10/60%

## Abstract

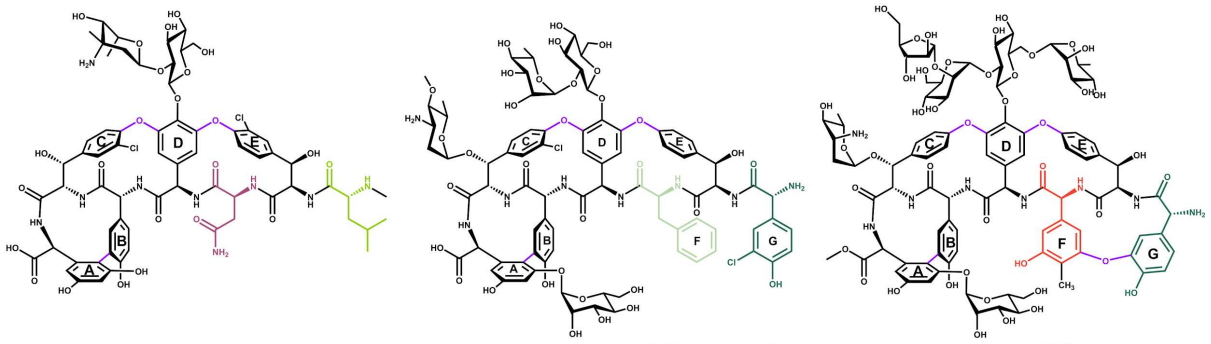
The realm of natural product (NP) research is continually expanding, with diverse applications in both medicine and industry. In this interdisciplinary field, scientists come together to investigate various aspects, ranging from comprehending the mode of action of compounds to unravelling biosynthetic pathways, studying evolutionary aspects, and attempting to heterologously express the enzymes involved. Collaboration is pivotal to grasp the intricacies of biosynthesis, but it is challenging considering that all parties involved come from very different backgrounds (such as microbiology, synthetic chemistry, biochemistry, and bioinformatics) and may not employ the same terminology. Thankfully, contemporary technologies, like videos, provide novel avenues for effective engagement. Recognizing that visual stimuli can be much more effective in explaining a complex process than written words, we envision a future where animations become a common tool of communication in interdisciplinary realms, accompanying perspectives or reviews. As a demonstration, here we employed animation to elucidate the biosynthesis of a known glycopeptide antibiotic, vancomycin, providing an example of how such approaches can enhance understanding.

## Introduction

Glycopeptide antibiotics (GPA) (Type I-IV?) are an important group of antibiotics, with the first member - vancomycin - discovered in 1953. Since then, 27 natural GPAs have been identified, complemented by the synthesis of numerous semi-synthetic derivatives, some of which are currently used in the clinic for the treatment of infections with multi-resistant Gram-positive bacterial pathogens [1,2]. Over the years, extensive studies have been carried out on the biosynthesis (*in vivo* and *in vitro*) and mechanism of action of GPAs [1,3–10]. The nomenclature “glycopeptide antibiotic” succinctly summarises their structural characteristics: a peptidic backbone typically consisting of seven amino acids, decorated by one or more sugar moieties. Beyond this, GPAs are crosslinked through their aromatic amino acid residues, a feature that confers rigidity to the core structure and represents a distinctive hallmark of this class of natural products that is essential for their bioactivity. The peptide backbone is further modified by the addition of halogen atoms, sulphate moieties, sugar residues and methyl groups [1,11,12].

### GPA classification

To date, GPAs have been classified into five types (I-V) primarily based on their structural attributes (**Figure 1**). Type I GPAs are characterised by a backbone comprising two aliphatic amino acids at positions 1 and 3 (Leu and Asn) of the peptide, and five non-proteinogenic aromatic amino acids ($\beta$-hydroxytyrosine (Bht); 4-hydroxyphenylglycine (Hpg) and 3,5-dihydroxyphenylglycine (Dpg)), which are linked through three phenolic/biaryl crosslinks [11]. Type II GPAs diverge by substituting aliphatic residues at positions 1 and 3 with aromatic amino acids, encompassing non-proteinogenic (Hpg) and proteinogenic (phenylalanine) constituents. However, these aromatic amino acids are not linked together [11].

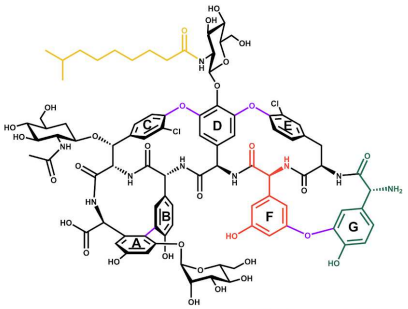


7-Dpg 6-Bht 5-Hpg 4-Hpg 3-Asn 2-Bht 1-Leu
**Type I - vancomycin**

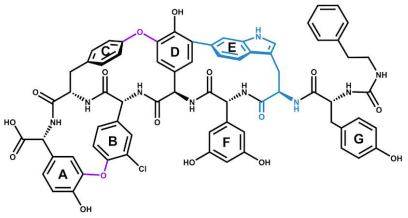
7-Dpg 6-Bht 5-Hpg 4-Hpg 3-Phe 2-Bht 1-Hpg
**Type II - keratinimicin A**

7-Dpg 6-Bht 5-Hpg 4-Hpg 3-Dpg 2-Bht 1-Hpg
**Type III - ristocetin A**

| Type | Side chains of AAs in position 1 and 3 of the backbone | Number of crosslinks | Additional unique characteristics |
|---|---|---|---|
| I | aliphatic (leu/ala & asn/glu) | 3 | - |
| II | aromatic (hpg & phe) | 3 | - |
| III | aromatic (hpg & dpg) | 4 | - |
| IV | aromatic (hpg & dpg) | 4 | Acylated sugar moiety |
| V | variable | variable | Trp in the backbone |



7-Dpg 6-Bht 5-Hpg 4-Hpg 3-Dpg 2-Tyr 1-Hpg
**Type IV - teicoplanin A2-2**



7-Hpg 6-Tyr 5-Hpg 4-Hpg 3-Dpg 2-Trp 1-Tyr
**Type V - kistamicin B**

**Figure 1**: Structural characteristics of the GPA types. The most important features for the classification of GPAs are listed and accompanied by examples of structures. The most significant characteristics for each type are coloured. The aromatic rings of the AAs are labelled A-G based on prior publications [13]. Abbreviations: Tyr, tyrosine; Leu, leucine; Asn, asparagine; Bht, $\beta$-hydroxytyrosine; Hpg, 4-hydroxyphenylglycine; Dpg, 3,5-dihydroxyphenylglycine; Phe, phenylalanine, Trp, tryptophane. The SMILES of the structures can be found in **Supplementary Table 1**.

In contrast to those of type II GPAs, the aromatic amino acids (Hpg$^1$-Bht$^2$-Dpg$^3$-Hpg$^4$-Hpg$^5$-Bht$^6$-Dpg$^7$) of type III GPAs are all crosslinked. Consequently, type III GPAs possess an additional crosslink compared to type I/II GPAs (4 vs 3) [11]. Type IV GPAs maintain an identical core peptide sequence to type III GPAs, distinguishing themselves by the inclusion of an acyl group attached to one of the pendant sugar residues. Notwithstanding these structural disparities, type I-IV GPAs share a common mechanism of antibiotic action involving the sequestration of bacterial cell wall precursors (lipid II). Thereby, type I-IV GPAs impede correct formation of the cell wall. In contrast, type V GPAs constitute an outlier group with respect to both the structure and mode of action [12]. Structurally, they exhibit substantial variations, particularly the absence of glycosylation[12], and diverse peptide

sequences. Nonetheless, type V GPAs do exhibit certain similarities to other GPAs, particularly in the crosslinking patterns between the aromatic residues.

## GPA biosynthesis

Due to their clinical importance and the need to produce glycopeptide antibiotics (GPAs) by in vivo fermentation, their biosynthesis has been extensively studied [1,3–10]. Understanding GPA biosynthesis is especially important if new derivatives of these compounds are to be produced on a large scale, given the limitations of current chemical syntheses for these complex molecules [10,14,15]. To this end and to ameliorate the communication between scientists of different backgrounds who are interested in GPAs, the authors have created an animated video which demonstrates each step in the biosynthesis of vancomycin, a type I GPA (**Supplementary Data 1**), which is also further explained in this manuscript.

Most of the biosynthetic genes required for the production of GPAs are localised in the bacterial genome within so-called biosynthetic gene clusters (BGCs), which can also include genes for export, regulation, and self-resistance. In principle, the biosynthesis of GPAs can be divided into three main steps: the synthesis of non-proteinogenic amino acids (1), the formation of the peptide backbone by the action of non-ribosomal peptide synthetases (NRPS), which are large multi-modular enzymes that assemble peptides without the involvement of the ribosome by the stepwise incorporation of individual amino acids (2), and the modification of the peptide backbone (3).

### (1) Biosynthesis of the non-proteinogenic amino acids

All established examples of GPAs contain non-proteinogenic amino acids, predominantly derivatives of phenylglycine [13] (**Figure 2**). These non-proteinogenic amino acids are primarily generated by dedicated biosynthetic pathways that are typically encoded as subclusters within the GPA BGCs. The precursors essential for the biosynthesis of the non-proteinogenic amino acids, notably hydroxyphenylpyruvate (4-HPP) and tyrosine, derive from the shikimate pathway. Given the tightly regulated nature of the shikimate metabolism, bacteria have evolved alternative mechanisms to circumvent this regulatory control and ensure the supply of precursors for secondary metabolite production. The producers of GPAs, for example, have acquired a second copy of the key enzymes of the shikimate pathway, Dahp and Pdh, which are also located within the GPA BGCs[16,17].

Hpg is synthesised from 4-HPP by the action of a 4-hydroxymandelate oxidase (HmO) and a 4-hydroxymandelate synthase (HmaS). The amino group of Hpg is derived from tyrosine, which is transferred by the aminotransferase HptT/Pgat [18–20].

Dpg is synthesised through a multistep biosynthetic process from acetyl-coA. This pathway involves orchestrated enzymatic activities of DpgA, DpgB, DpgC, and DpgD, which represent a type III polyketide synthase as well as modifying enzymes responsible for the generation 3,5-dihydroxyphenylglyoxylate. The ultimate

transformation entails a transamination reaction, which utilizes tyrosine as the substrate and is catalyzed by PgT/Pgat [18,21–23].

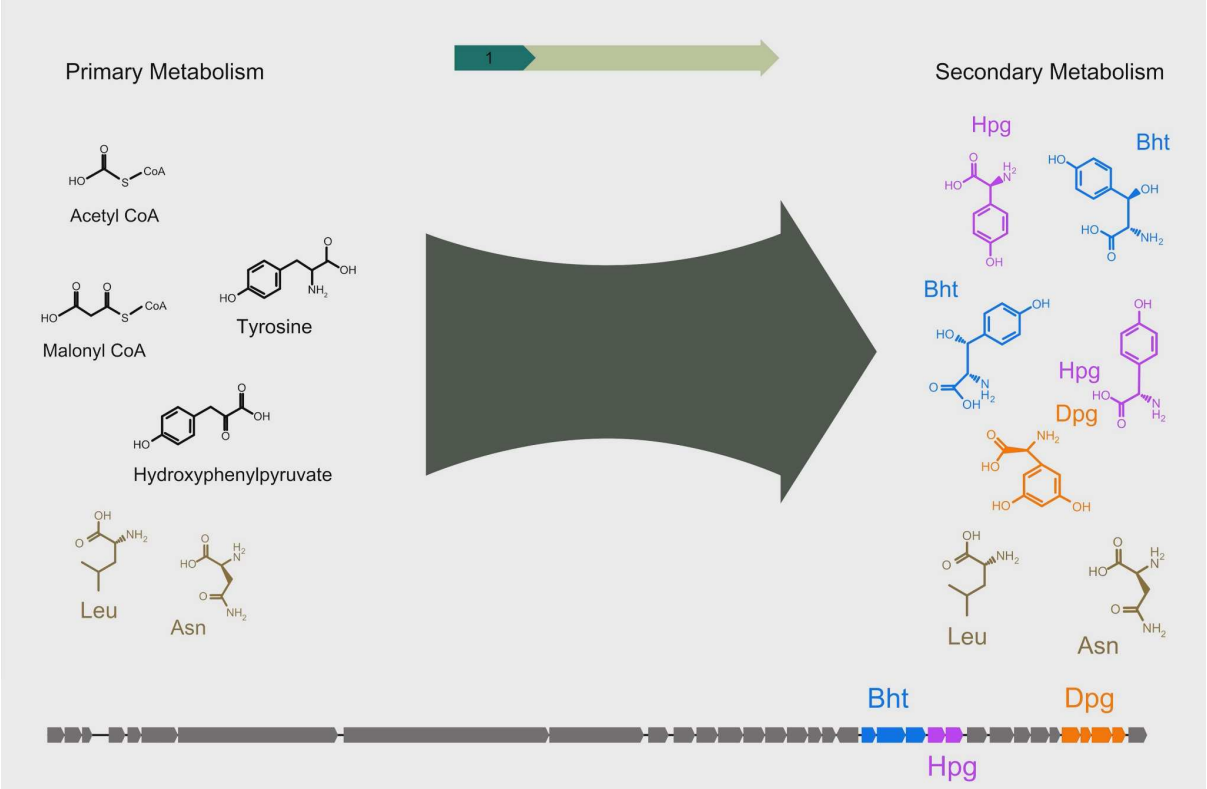



**Figure 2**: Precursor supply of non-proteinogenic amino acids. On the left, molecules present as part of the primary metabolism, including leucine (Leu) and asparagine (Asn), which are part of the vancomycin backbone. The molecules in black participate in the chemical reactions (symbolised by the big arrow) which lead to the synthesis of the non-proteinogenic amino acids hydroxyphenylglycine (Hpg), dihydroxyphenylglycine (Dpg) and $\beta$-hydroxytyrosine (Bht), all parts of the vancomycin backbone, which are shown on the right. On the bottom, the BGC encoding the biosynthesis of vancomycin is shown, with the genes involved in the synthesis of each non-proteinogenic amino acid coloured and labelled accordingly. This figure was created from screenshots of the animation (**Supplementary Data 1**).

The biosynthesis of phenylglycine derivatives is common to all GPAs. However, the production of Bht follows different pathways depending on the type of GPA. The first mechanism involves hydroxylation of Tyr by a non-heme iron oxygenase subsequent to its selection/activation by the nonribosomal peptide synthetase (NRPS)[19]. The correct modification of Tyr is controlled by the atypical catalytic activities of a peptide bond forming domain within the main enzyme complex[20]. Conversely, the second mechanism entails the biosynthesis of Bht prior to its incorporation into the main assembly line[24,25]. In this scenario, an NRPS module activates Tyr [26], which undergoes subsequent hydroxylation catalysed by a cytochrome P450 monooxygenase [27]. Bht is then cleaved from the minimal NRPS module by a specific thioesterase [28], resulting in free Bht for subsequent activation by the main NRPS.

### (2) The formation of the peptide backbone

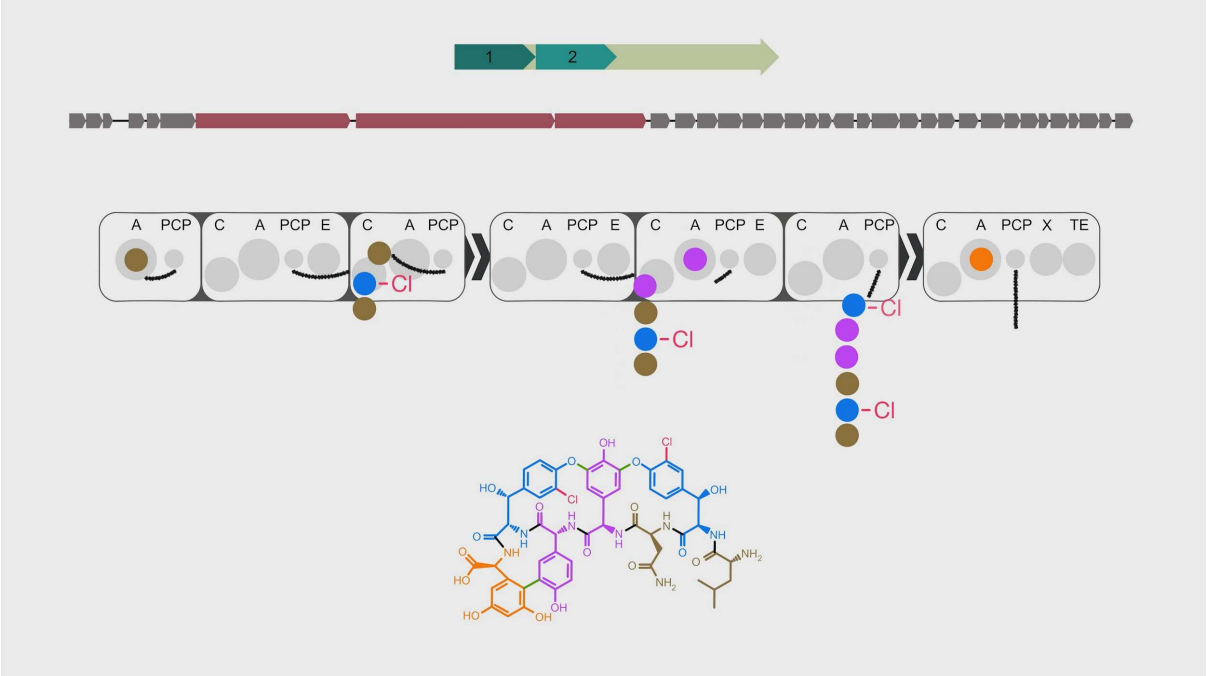



**Figure 3**: Backbone assembly of vancomycin via non-ribosomal peptide synthetases (NRPS). On the top, the vancomycin BGC is shown, with the genes encoding the NRPS enzymes coloured red. Under it, a schematic represents the three NRPS joined with their docking domains (black arrows). Each module is drawn separately and each functional domain within is labelled accordingly. The coloured circles each represent an amino acid either waiting to be incorporated or already part of the growing peptide. The structure of the heptapeptide as it is when released from the final module is shown below, with the amino acids coloured the same as their representative circles. This figure was adapted from the animation (**Supplementary Data 1**).

The synthesis of the peptide backbone of GPAs is catalysed by NRPSs [21,27] (**Figure 3**). Each NRPS module contains distinct catalytic domains that perform specific functions. The adenylation (A) domains recognize and activate specific amino acids priming them for subsequent incorporation into the growing peptide chain [21,27]. Upon activation, the amino acid is covalently linked as a thioester to the thiol group of the phosphopantetheine arm within an adjacent peptidyl carrier protein (PCP) domain. This enables *trans* acting enzymes to further modify the PCP-bound amino acid, for example, via halogenation [8] and hydroxylation [20]. The PCP domain then facilitates the translocation of the amino acid to the acceptor pocket of the condensation (C) domain [29]. C-domains, usually located at the N-terminus of a module, catalyse the formation of amide bonds between two PCP-bound substrates [29], by adding the amine group of the downstream acceptor amino acid to the thioester linkage of the upstream donor substrate [21,27]. In addition, certain modules contain an epimerization domain (E domain), which converts the L-configured amino acid residue of the C-terminal fragment of the peptide to its D-form [30]. In this way, the peptide chain is progressively elongated until the final - for GPAs typically the seventh - module [21].

## (2.1) Oxidative crosslinks and aglycone release

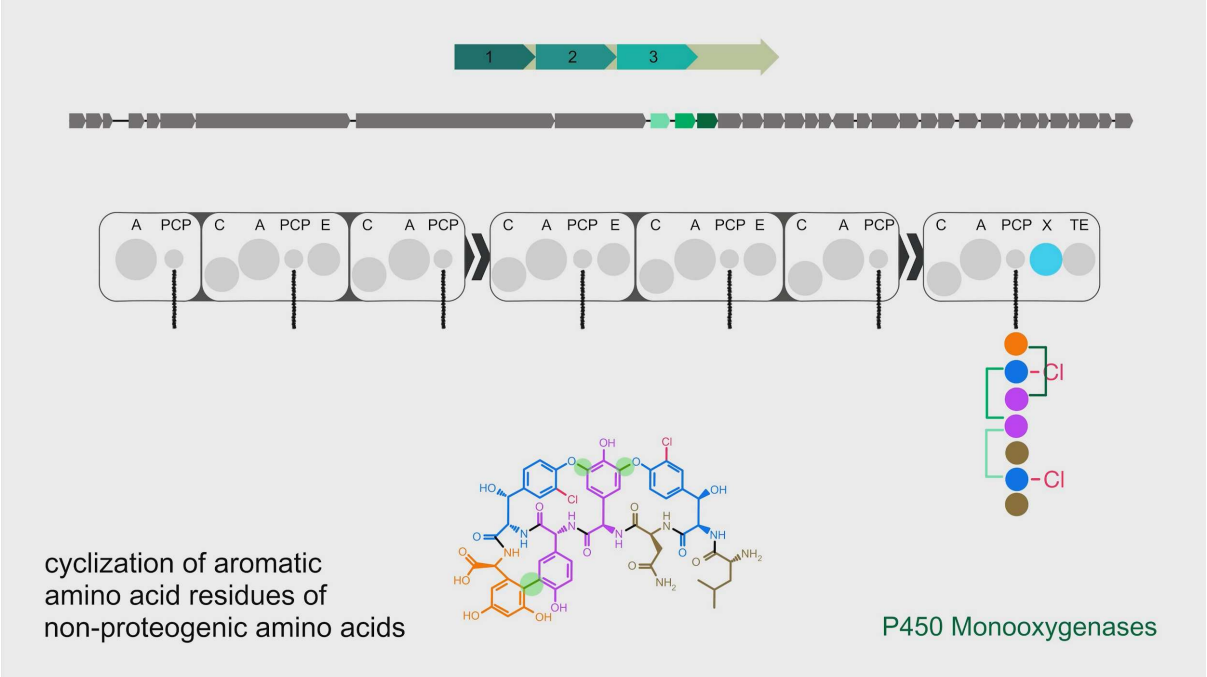



**Figure 4**: Cyclization reactions during vancomycin biosynthesis. On the top, the vancomycin BGC is shown, with the genes encoding the p450 monooxygenase enzymes coloured green. Under it, a schematic represents the three NRPSs joined with their docking domains (black arrows). Each module is drawn separately and each functional domain within is labelled accordingly. In the final module, the X-domain is coloured blue and the final PCP-domain is connected to a heptapeptide. The coloured circles each represent an amino acid with the same colour in the structure shown in the middle. The crosslinked amino acids are connected by brackets in the PCP-connected visualisation and highlighted in green in the structure visualisation. This figure was adapted from the animation (**Supplementary Data 1**).

A key step in the assembly of GPAs by NRPSs is the recruitment of up to four cross-linking enzymes to the final NRPS module. These enzymes, known as Oxy enzymes, are members of the cytochrome P450 family and are responsible for crosslink formation between the aromatic side chains within the peptide structure. The recruitment of the Oxy enzymes to the peptide is facilitated by a unique domain exclusively found in GPA biosynthesis, known as the X-domain. This domain, which is structurally related to C-domains, uses a conserved interface to sequentially recruit P450 enzymes to the NRPS-bound peptide via a shuffling mechanism [31]. The terminal thioesterase (TE) domain catalyses the release of the fully cyclised peptide products from the enzyme complex.

### (3) Modification of the peptide backbone

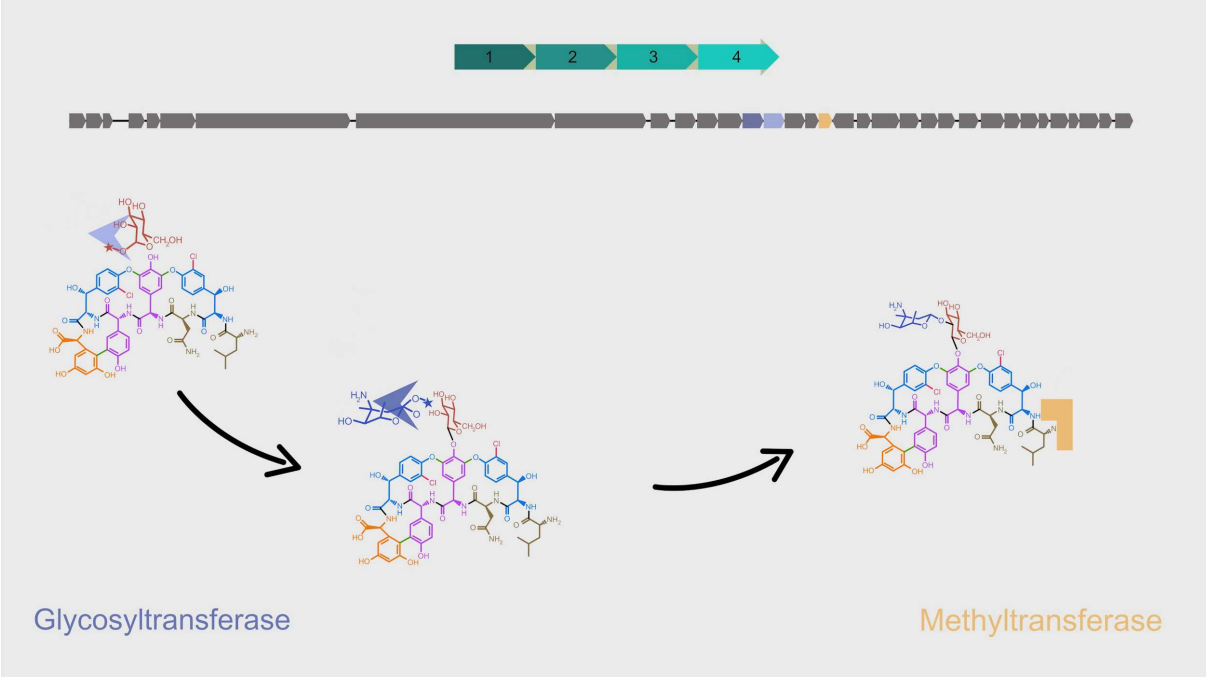



**Figure 5**: Post-assembly modifications of vancomycin. On the top, the vancomycin BGC is shown, with the genes encoding the glycosyltransferases and the methyltransferase coloured blue and yellow accordingly. Under it, the three tailoring reactions are shown in order. The first glycosyltransferase (light blue shape) attaches the first sugar moiety to the fourth amino acid in the backbone. Then, the second glycosyltransferase (dark blue shape) attaches the second sugar moiety to the first sugar. Additionally, a methyltransferase (yellow shape) adds a methyl group on the first amino acid in the backbone. The colouring scheme of the backbone amino acids follows that of Figure 2. This figure was adapted from the animation (**Supplementary Data 1**).

In addition to the cross-linked peptide backbone, type I-IV GPAs are characterised by the incorporation of sugar moieties. The synthesis of these sugar components requires various enzymes, such as epimerases, transaminases, dehydratases, and methyltransferases, most of which are typically encoded in the GPA BGC [32]. Once synthesized, the sugar moieties are linked to the cyclic peptide (aglycone) by specific glycosyltransferases [21]. The abundance and diversity of these sugar moieties significantly contribute to the structural heterogeneity found in GPAs [31].

Moreover, many GPAs exhibit optional modifications, including methylation [21,33], sulfation of amino acid side chains [21,34] and the acylation of sugar moieties [21].

### (4) Export and mode of action

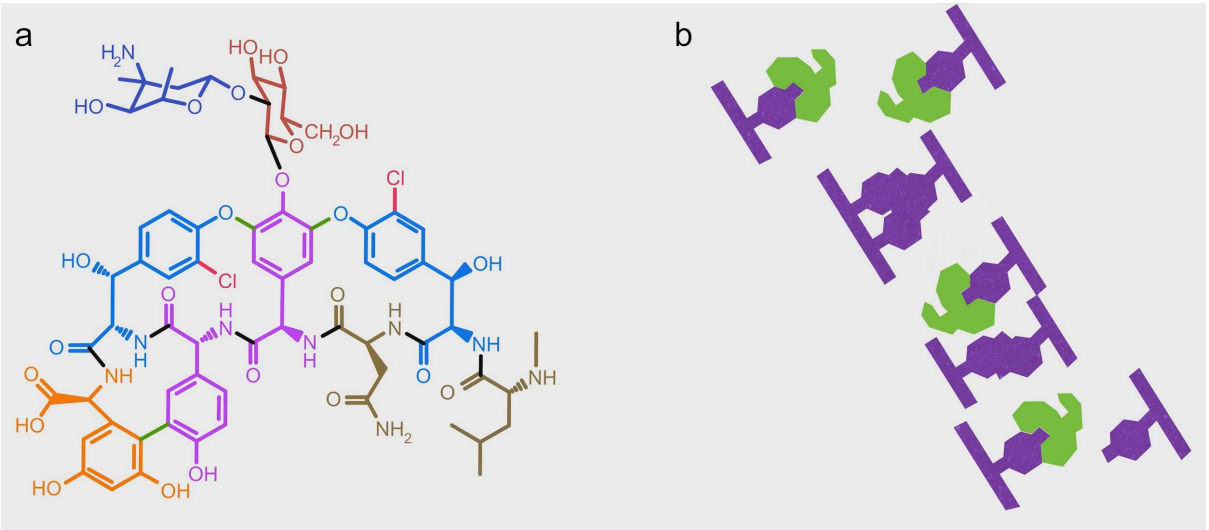



**Figure 6**: Final structure and mode of action of vancomycin. Panel a: The final structure of vancomycin after its biosynthesis has been completed. The colouring scheme follows that of Figure 2. Panel b: visualisation of the mode of action of vancomycin. The green shapes represent the compound molecules while the purple shapes represent the lipid II component of the bacterial cell wall in the target cell. Wherever vancomycin has bound on the lipid II, it blocks other vital enzymes from catalysing cell wall reactions, leading to cell death. This figure was adapted from the animation (**Supplementary Data 1**).

Upon completion of GPA biosynthesis (**Figure 6a**), the active compound is transported out of the intracellular environment by a specific ATP-binding cassette (ABC) transporter [35]. Type I-IV GPAs have lethal consequences on bacterial (Gram positive) competitors of the producer strain by selectively binding to the D-alanyl-D-alanine (D-Ala-D-Ala) terminus of lipid II. Lipid II is a fundamental component of bacterial cell walls and plays a key role in the peptidoglycan biosynthetic pathway [36]. By binding on lipid II, vancomycin blocks vital enzymatic reactions of the cell wall biosynthesis (**Figure 6b**), leading to bacterial cell death.

## References

"Arrow" symbol in Figure 5 is by 4B Icons from thenounproject.com.

1. Stegmann, E., Frasch, H. J. & Wohlleben, W. Glycopeptide biosynthesis in the context of basic cellular functions. *Current Opinion in Microbiology* **13**, 595–602 (2010).
2. van Groesen, E., Innocenti, P. & Martin, N. I. Recent Advances in the Development of Semisynthetic Glycopeptide Antibiotics: 2014–2022. *ACS Infect Dis* **8**, 1381–1407 (2022).
3. Butler, M. S., Hansford, K. A., Blaskovich, M. A. T., Halai, R. & Cooper, M. A. Glycopeptide antibiotics: Back to the future. *J Antibiot* **67**, 631–644 (2014).
4. Mccormick, M. H., Mcguire, J. M., Pittenger, G. E., Pittenger, R. C. & Stark, W. M. Vancomycin, a new antibiotic. I. Chemical and biologic properties. *Antibiot Annu* **3**, 606–611 (1955).
5. Hansen, M. H., Stegmann, E. & Cryle, M. J. Beyond vancomycin: recent advances in the modification, reengineering, production and discovery of improved glycopeptide antibiotics to tackle multidrug-resistant bacteria. *Curr Opin Biotechnol* **77**, 102767 (2022).

6. Greule, A. *et al.* Kistamicin biosynthesis reveals the biosynthetic requirements for production of highly crosslinked glycopeptide antibiotics. *Nature Communications 2019 10:1* **10**, 1–15 (2019).
7. Schoppet, M. *et al.* The biosynthetic implications of late-stage condensation domain selectivity during glycopeptide antibiotic biosynthesis. *Chemical Science* **10**, 118–133 (2019).
8. Kittilä, T. *et al.* Halogenation of glycopeptide antibiotics occurs at the amino acid level during non-ribosomal peptide synthesis. *Chemical Science* **8**, 5992–6004 (2017).
9. Wohlleben, W., Stegmann, E. & Süssmuth, R. D. Chapter 18. Molecular genetic approaches to analyze glycopeptide biosynthesis. *Methods Enzymol* **458**, 459–486 (2009).
10. Truman, A. W. *et al.* The Role of Cep15 in the Biosynthesis of Chloroeremomycin: Reactivation of an Ancestral Catalytic Function. *Chemistry & Biology* **15**, 476–484 (2008).
11. Chen, S., Wu, Q., Shen, Q. & Wang, H. Progress in Understanding the Genetic Information and Biosynthetic Pathways behind Amycolatopsis Antibiotics, with Implications for the Continued Discovery of Novel Drugs. *ChemBioChem* **17**, 119–128 (2016).
12. Xu, M. *et al.* GPAHex-A synthetic biology platform for Type IV–V glycopeptide antibiotic production and discovery. *Nature Communications 2020 11:1* **11**, 1–12 (2020).
13. Toma, R. S. A., Brieke, C., Cryle, M. J. & Süssmuth, R. D. Structural aspects of phenylglycines, their biosynthesis and occurrence in peptide natural products. *Nat. Prod. Rep.* **32**, 1207–1235 (2015).
14. Kegler, C. & Helge B, B. Artificial Splitting of a Non-Ribosomal Peptide Synthetase by Inserting Natural Docking Domains. *Angew Chem Int Ed Engl.* **10**, (2020).
15. Bozhüyük, K. A. J. *et al.* De novo design and engineering of non-ribosomal peptide synthetases. *Nature Chemistry* **10**, 275–281 (2018).
16. Thykaer, J. *et al.* Increased glycopeptide production after overexpression of shikimate pathway genes being part of the balhimycin biosynthetic gene cluster. *Metabolic Engineering* **12**, 455–461 (2010).
17. Goldfinger, V. *et al.* Metabolic engineering of the shikimate pathway in *Amycolatopsis* strains for optimized glycopeptide antibiotic production. *Metabolic Engineering* **78**, 84–92 (2023).
18. Tseng, C. C., McLoughlin, S. M., Kelleher, N. L. & Walsh, C. T. Role of the Active Site Cysteine of DpgA, a Bacterial Type III Polyketide Synthase. *Biochemistry* **43**, 970–980 (2004).
19. Stinchi, S. *et al.* A derivative of the glycopeptide A40926 produced by inactivation of the β-hydroxylase gene in Nonomuraea sp. ATCC39727. *FEMS Microbiology Letters* **256**, 229–235 (2006).
20. Kaniusaite, M. *et al.* A proof-reading mechanism for non-proteinogenic amino acid incorporation into glycopeptide antibiotics. *Chemical Science* **10**, 9466–9482 (2019).
21. Yim, G., Thaker, M. N., Koteva, K. & Wright, G. Glycopeptide antibiotic biosynthesis. *Journal of Antibiotics* **67**, 31–41 (2014).
22. Chen, H., Tseng, C. C., Hubbard, B. K. & Walsh, C. T. Glycopeptide antibiotic biosynthesis: Enzymatic assembly of the dedicated amino acid monomer (S)-3,5-dihydroxyphenylglycine. *Proceedings of the National Academy of Sciences* **98**, 14901–14906 (2001).
23. Pfeifer, V. *et al.* A Polyketide Synthase in Glycopeptide Biosynthesis: THE BIOSYNTHESIS OF THE NON-PROTEINOGENIC AMINO ACID (S)-3,5-DIHYDROXYPHENYLGLYCINE*. *Journal of Biological Chemistry* **276**, 38370–38377 (2001).
24. Chen, H. & Walsh, C. T. Coumarin formation in novobiocin biosynthesis: beta-hydroxylation of the aminoacyl enzyme tyrosyl-S-NovH by a cytochrome P450 NovI. *Chem Biol* **8**, 301–312 (2001).
25. Chen, H., Hubbard, B. K., O'Connor, S. E. & Walsh, C. T. Formation of beta-hydroxy histidine in the biosynthesis of nikkomycin antibiotics. *Chem Biol* **9**, 103–112 (2002).
26. Mulyani, S. *et al.* The thioesterase Bhp is involved in the formation of beta-hydroxytyrosine during balhimycin biosynthesis in Amycolatopsis balhimycina. *Chembiochem* **11**, 266–271 (2010).
27. Miller, B. R. & Gulick, A. M. Structural Biology of Non-Ribosomal Peptide Synthetases. *Methods Mol Biol* **1401**, 3–29 (2016).
28. Haslinger, K., Peschke, M., Brieke, C., Maximowitsch, E. & Cryle, M. J. X-domain of peptide synthetases recruits oxygenases crucial for glycopeptide biosynthesis. *Nature* **521**, 105–109

(2015).
29. Izoré, T. *et al.* Structures of a non-ribosomal peptide synthetase condensation domain suggest the basis of substrate selectivity. *Nature communications* **12**, 2511 (2021).
30. Luo, L. *et al.* Timing of Epimerization and Condensation Reactions in Nonribosomal Peptide Assembly Lines: Kinetic Analysis of Phenylalanine Activating Elongation Modules of Tyrocidine Synthetase B. *Biochemistry* **41**, 9184–9196 (2002).
31. Nicolaou, K. C., Boddy, C. N. C., Bräse, S. & Winssinger, N. Chemistry, Biology, and Medicine of the Glycopeptide Antibiotics. *Angewandte Chemie International Edition* **38**, 2096–2152 (1999).
32. Donadio, S., Sosio, M., Stegmann, E., Weber, T. & Wohlleben, W. Comparative analysis and insights into the evolution of gene clusters for glycopeptide antibiotic biosynthesis. *Molecular Genetics and Genomics* **274**, 40–50 (2005).
33. Brieke, C., Yim, G., Peschke, M., Wright, G. D. & Cryle, M. J. Catalytic promiscuity of glycopeptide N-methyltransferases enables bio-orthogonal labelling of biosynthetic intermediates †. *Chem. Commun* **52**, 13679 (2016).
34. Kalan, L., Perry, J., Koteva, K., Thaker, M. & Wright, G. Glycopeptide sulfation evades resistance. *Journal of Bacteriology* **195**, 167–171 (2013).
35. Menges, R., Muth, G., Wohlleben, W. & Stegmann, E. The ABC transporter Tba of Amycolatopsis balhimycina is required for efficient export of the glycopeptide antibiotic balhimycin. *Appl Microbiol Biotechnol* **77**, 125–134 (2007).
36. Müller, A., Klöckner, A. & Schneider, T. Targeting a cell wall biosynthesis hot spot. *Natural Product Reports* **34**, 909–932 (2017).

## Supplementary Material

All supplementary material of unpublished projects are available for download (upon request) from a zenodo repository: https://doi.org/10.5281/zenodo.10879735

Supplementary Table 1: SMILES of selected known GPA structures.
Supplementary Data 1: animation of the biosynthesis of vancomycin.

# Chapter 4: Phylogenetic distance and structural diversity directing a reclassification of glycopeptide antibiotics

*(Advanced manuscript; awaiting submission to journal)*

Athina Gavriilidou[1], Martina Adamek[1,2], Jens-Peter Rodler[3], Noel Kubach[1], Susanna Kramer[1], Daniel H. Huson[4], Max J. Cryle[5,6], Evi Stegmann* [3], Nadine Ziemert* [1,2]

1: Translational Genome Mining for Natural Products, Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Tübingen, Germany / Cluster of Excellence 'Controlling Microbes to Fight Infections' (CMFI), University of Tübingen, Tübingen, Germany

2: German Centre for Infection Research (DZIF), Partnersite Tübingen, Tübingen, Germany

3: Microbial Bioactive Compounds, Interfaculty Institute of Microbiology and Infection Medicine Tübingen, University of Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany / Cluster of Excellence 'Controlling Microbes to Fight Infections' (CMFI), University of Tübingen, Tübingen, Germany

4: Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, 72076, Germany / International Max Planck Research School "From Molecules to Organisms", Max Planck Institute for Biology Tübingen, Max-Planck-Ring 5, Tübingen, 72076, Germany / Cluster of Excellence 'Controlling Microbes to Fight Infections' (CMFI), University of Tübingen, Tübingen / Germany, Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Tübingen, Germany

5: Department of Biochemistry and Molecular Biology, The Monash Biomedicine Discovery Institute, Monash University, Clayton, VIC, 3800, Australia

6: EMBL Australia, Monash University, Clayton, Victoria 3800, Australia

* co-corresponding author

Personal contributions:

Scientific ideas/data generation/analysis & interpretation/writing: 70/95/90/70%

## Abstract

Antibiotics have been an essential part of modern medicine since their initial discovery. The continuous search for new antibiotic candidates remains a necessity given the increasing emergence of resistance to antimicrobial compounds among pathogens. The glycopeptide antibiotics (GPAs) represent an important group of last resort antibiotics which inhibit bacterial growth through non-covalent binding to the cell wall precursor lipid II. The so far reported GPAs exhibit an enormous diversity in the biosynthetic gene clusters that encode their production, which is in turn reflected in the variety of their structures. GPAs are typically composed of seven amino acids, which are highly crosslinked and decorated with a variable collection of sugar moieties as well as other modifications. Based on their structural characteristics, they have been classified into four main types. More recently, atypical GPAs have been identified that differ from type I-IV GPAs in both their structure and function, and have consequently been classified as type V GPAs. Given these differences, we studied the phylogeny of all gene sequences related to the biosynthesis of the GPAs, and observed a clear evolutionary diversification between the lipid II binding GPA classes and the so-called type V GPAs. Here we suggest the adoption of a phylogeny-driven reclassification and a separation of classical lipid II binding GPAs from type V GPAs, which we propose to identify instead as glycopeptide- related peptides (GRPs).

## Introduction

Glycopeptide antibiotics (GPAs) constitute a crucial class of clinical antibiotics, exemplified by the discovery of vancomycin in 1953. Since then, the identification of 27 natural GPAs and the synthesis of numerous semi-synthetic derivatives, some of which are actively utilised in clinical settings, underscore their significance in combating infections caused by multi-resistant Gram-positive bacterial pathogens [1] [2]. Both the biosynthesis (*in vivo* and *in vitro*) and the mode of action of GPAs have now been studied for decades [1], [2], [3], [4], [5], [6], [7], [8], [9].

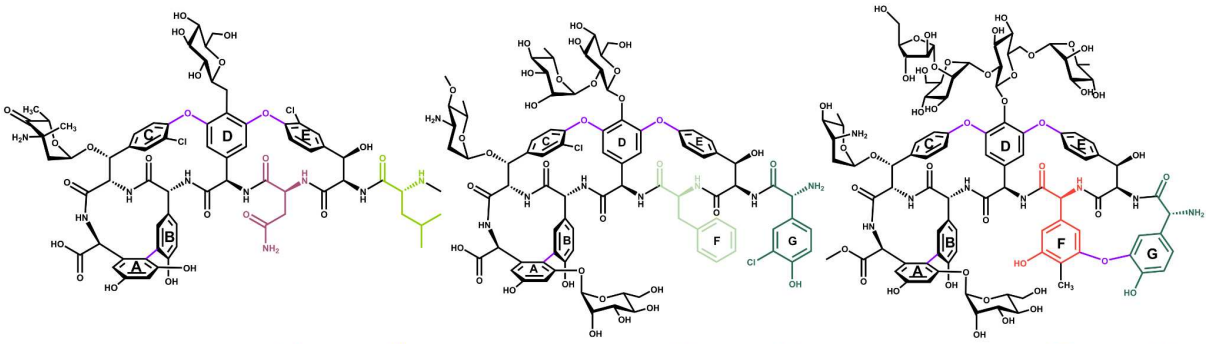
The term GPA reflects their characteristic structure: a peptidic backbone typically comprising seven amino acids, often glycosylated. Additionally, GPAs undergo substantial crosslinking via the side chains of aromatic amino acid residues, imparting structural rigidity to the core- an inherent characteristic crucial for their bioactivity. Additional modifications include halogenation, sulfation, glycosylation, and methylation of the peptide backbone (**Table 1**) [1], [10], [11].

#### Existing GPA classification

To date, GPAs have been categorised into five types (I-V) according to their structural features (**Figure 1**) [1], [10], [11]. Type I GPAs feature a backbone comprising two aliphatic amino acids (leucine and asparagine or alanine and glutamic acid) at positions 1 and 3, respectively, along with five non-proteinogenic aromatic amino acids (β-hydroxytyrosine (Bht), 4-hydroxyphenylglycine (Hpg), and 3,5-dihydroxyphenylglycine (Dpg)). These aromatic amino acids are linked via three phenolic/biaryl crosslinks [10]. Type II GPAs share a similar crosslinking pattern to

the type I GPAs, although this class possesses aromatic amino acids, non proteinogenic (Hpg) and proteinogenic (Phe) at positions 1 and 3 of the backbone instead of aliphatic residues [10]. The backbone of type III GPAs consists exclusively of aromatic amino acids (e.g. $Hpg^{1}$–$Bht^{2}$–$Dpg^{3}$–$Hpg^{4}$–$Hpg^{5}$-$Bht^{6}$–$Dpg^{7}$), all of which are crosslinked, presenting an additional crosslink compared to type I/II GPAs (4 vs 3) [10]. The backbone of the type IV GPAs is also made up entirely of aromatic amino acids, but in addition they contain an acyl group attached to one of the pendant sugar residues. Despite these differences in structure, type I-IV GPAs all share a common mechanism of action, which involves the sequestration of bacterial cell wall precursors (lipid II), thus preventing correct cell wall formation. Still, their structural characteristics do cause minor differences in target affinity, as is the case of teicoplanin (type IV), whose acyl chain on the sugar moiety hinders the dimer formation observed in vancomycin, but instead interacts with the bacterial membrane, bringing about a alternatively mediated binding of the compound to the same target (D-alanyl-D-alanine, D-Ala-D-Ala, tail of lipid II)[12].

Type V GPAs represent an outlier group compared to type I-IV GPAs, as they have distinct characteristics. Unlike their counterparts, type V GPAs have variable peptide backbone lengths of up to 9 amino acids and lack glycosylation [11]. In addition, their peptide sequences vary, although a common feature of all known structures is the presence of a Trp amino acid, cross-linked to the central Hpg [13] (**Figure 1**). Notably, type V GPAs have a unique mode of action, binding to autolysin molecules and inhibiting their hydrolytic activity on peptidoglycan during cell division [14]. Despite these differences, type V GPAs share similarities with other GPAs, particularly in the cross-linking patterns between aromatic residues.

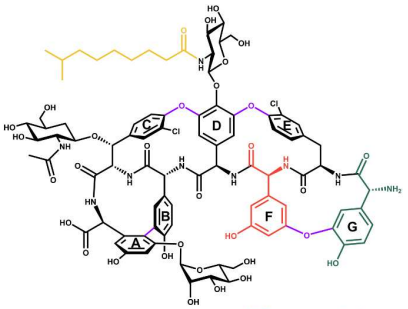


7-Dpg 6-Bht 5-Hpg 4-Hpg 3-Asn 2-Bht 1-Leu
**Type I - balhimycin**

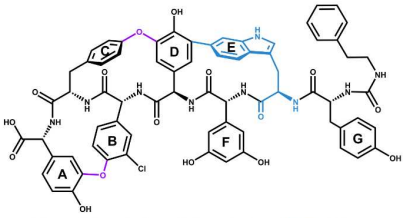
7-Dpg 6-Bht 5-Hpg 4-Hpg 3-Phe 2-Bht 1-Hpg
**Type II - keratinimicin A**

7-Dpg 6-Bht 5-Hpg 4-Hpg 3-Dpg 2-Bht 1-Hpg
**Type III - ristocetin A**

| Type | Side chains of AAs in position 1 and 3 of the backbone | Number of crosslinks | Additional unique characteristics |
|---|---|---|---|
| I | aliphatic (leu/ala & asn/glu) | 3 | - |
| II | aromatic (hpg & phe) | 3 | - |
| III | aromatic (hpg & dpg) | 4 | - |
| IV | aromatic (hpg & dpg) | 4 | Acylated sugar moiety |
| V | variable | variable | Trp in the backbone |



7-Dpg 6-Bht 5-Hpg 4-Hpg 3-Dpg 2-Tyr 1-Hpg
**Type IV - teicoplanin A2-2**



7-Hpg 6-Tyr 5-Hpg 4-Hpg 3-Dpg 2-Trp 1-Tyr
**Type V - kistamicin B**

**Figure 1**: Structural characteristics of the GPA types. The most important features for the classification of GPAs are listed and accompanied by examples of structures. The most significant characteristics for each type are coloured. The aromatic rings of the AAs are labelled A-G based on prior publications [15]. Abbreviations: Tyr, tyrosine; Leu, leucine; Asn, asparagine; Bht, $\beta$-hydroxytyrosine; Hpg, 4-hydroxyphenylglycine; Dpg, 3,5-dihydroxyphenylglycine; Phe, phenylalanine, Trp, tryptophane. The SMILES of the structures can be found in **Supplementary Table 1**. Figure adapted from chapter 3 (Figure 1).

## Biosynthesis of GPA in types I-IV

The biosynthesis of GPAs, which has been extensively studied for types I-IV [1], [2], [3], [4], [5], [6], [7], [8], involves several key processes: precursor supply of the non-proteinogenic amino acids, stepwise backbone formation via non-ribosomal peptide synthetases (NRPS), and subsequent tailoring reactions. Genes responsible for these processes are clustered within biosynthetic gene clusters (BGCs), which also contain genes encoding transporters, regulators, and enzymes associated with resistance (**Table 1**).

The NRPSs are organised into modules, each of which is responsible for the incorporation of an AA into the backbone in a stepwise, assembly line-like manner

[16]. This is achieved by the sequential activities of functional domains within each module [17].

Several domain types play critical roles in the biosynthesis of GPAs, with some being optional while others are indispensable. Adenylation domains (A-domains) are the initial players, each specifically recruiting an amino acid [16], [17]. Then the AA or the growing peptide is loaded onto a peptidyl carrier domain (PCP-domain), which transfers it to the donor pocket of the downstream condensation domain (C-domain). Simultaneously, the downstream A-domain recruits the next AA to be incorporated, which is loaded onto the downstream PCP-domain and brought into the acceptor pocket of the C-domain. The C-domain catalyses the peptide bond formation and releases the peptide from the upstream PCP-domain [18]. This process elongates the backbone, with the PCP-domain subsequently transporting it to the next C-domain. Finally, once the polypeptide reaches the final module, a thioesterase domain (TE-domain) catalyses the breaking of the final PCP-peptidyl bond, releasing the assembled peptide into the cytosol. This final step is preceded by another domain, the X-domain, which is responsible for recruiting the P450 monooxygenase enzymes (OxyABCE) that create the distinguishing cross-links of GPAs. These domains are known to be a unique characteristic of GPAs and are not found in NRPS enzymes from other systems. Additional, optional domains fulfil various functions within the GPA biosynthetic pathway. For instance, epimerization (E) domains can alter the stereochemistry of specific amino acids [19]. During the backbone assembly, other (optional) enzymes may perform halogenations [7] and hydroxylations [20]. The latter are referred to as “online hydroxylation” of Tyr AAs into Bht AAs to underline the contrast to the “offline” hydroxylation of Tyr into Bht via a three enzyme system before its incorporation into the backbone. The former hydroxylation reaction occurs exclusively at the Tyr in position 2, resulting in the presence of both Bht and Tyr in the backbone, as opposed to the “offline” Bht production, which typically results in no Tyr AAspresent in the heptapeptide. After the release of the assembled backbone from the NRPS enzymatic complex, further modifications may take place by tailoring enzymes, encompassing methylations, sulfations, acylations and glycosylations [15], [17], [21] (**Table 1**).

**Table 1**: Most common gene categories found in GPA-synthesis (types I-IV) encoding BGCs. The genes and their known functions are summarised in this table, organised into categories and subcategories based on their role. An example gene or domain is provided in the corresponding columns (References in **Supplementary Table 2**).

| Category | Subcategory | example Gene/domain | Function |
|---|---|---|---|
| precursor supply | shikimate pathway - related | *pdh* | prephenate dehydrogenase. Related to Tyr biosynthesis |
| | | *dahp* | 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase. Related to Tyr biosynthesis |
| | Hpg/Dhpg synthesis | *pgat* | p-hydroxy- and 3,5-dihydroxyphenylglycine aminotransferase |
| | Dpg synthesis | *dpgA* | 3,5-dihydroxyphenylacetyl-CoA synthase (type III PKS) |
| | | *dpgB* | enoyl-CoA Hydratase |
| | | *dpgC* | 3,5-Dihydroxyphenylacetyl-CoA oxidase |
| | | *dpgD* | enoyl-CoA Hydratase |
| | Hpg synthesis | *hmaS* | p-hydroxymandelate synthase |
| | | *hmo* | p-hydroxymandelate oxidase |
| | Bht synthesis | *bhp* | alpha/beta hydrolase |
| | | *bpsD* | additional non-ribosomal peptite synthetase. Its A-domain is selecting a Tyr |
| | | *oxyD* | P450 monooxygenase. Hydroxylation of Tyr |
| core biosynthesis | NRPS vital domains | A-domain | adenylation domain: recognition and activation of backbone AAs |
| | | PCP-domain | peptidyl carrier protein domain: loading and shuttling of AAs via its phosphopantetheine arm |
| | | C-domain | condensation domain: amide bond formation between two typically PCP-bound AAs |
| | | X-domain | recruitment of p450 mono-oxygenases (Oxys) |
| | | TE-domain | thioesterase domain: release of polypeptide from the enzyme complex |
| | NRPS optional domains | E-domain | epimerization domain: conversion of L-configured AAs to D-configured AAs |
| | | TIGR01720 domain | unknown - possibly post-condensation modifications |
| | A-domain related | *mbtH* | folding, stability, and activity of A-domains |
| AA modifications during backbone biosynthesis | crosslinking | *oxyB* | 1st-acting oxidative crosslinking of C and D aromatic rings |
| | | *oxyE* | oxidative crosslinking of F and G aromatic rings. Acts after OxyB |
| | | *oxyA* | oxidative crosslinking of D and E aromatic rings. Acts after OxyB (and OxyE). |
| | | *oxyC* | oxidative crosslinking of A and B aromatic rings. Acts after OxyA |
| | halogenation | *bhaA* | halogenation of PCP-bound AAs |
| | hydroxylation | *tei12* | hydroxylation of Tyr into Bht |
| AA modifications after backbone biosynthesis | methylation | *mtfA* | methylation of AAs |
| | sulfation | *staL* | sulfation of AAs |
| sugar-related | sugar synthesis | *evaA* | C2 deoxygenation |
| | | *evaB* | C3 amination |

| Category | Subcategory | example Gene/domain | Function |
|---|---|---|---|
| | | *evaC* | methylation |
| | | *evaD* | C5 epimerization |
| | | *evaE* | C4 ketoreduction |
| | sugar transport | *bgtfA* | glycosyltransferase |
| | | *bgtfB* | UDP-N-acetylglucosamine transferase |
| | | *bgtfC* | UDP-N-acetylglucosamine transferase |
| | | *tei3* | mannosyltransferase |
| | sugar modification | *tei11* | acylation of sugar moiety |
| regulation & resistance | resistance genes | *vanH* | D-lactate dehydrogenase. Part of VanHAX resistance cassette |
| | | *vanA* | D-Ala-D-Lac ligase. Part of vanHAX resistance cassette |
| | | *vanX* | D-Ala-D-Ala dipeptidase. Part of vanHAX resistance cassette |
| | | *vanY* | carboxypeptidase. Alternative resistance mechanism to vanHAX. |
| | resistance regulators | *vanR* | response regulator, part of of two component system vanRS |
| | | *vanS* | histidine kinase, part of two component system vanRS |
| | biosynthesis regulators | *bbr* | Str-like pathway specific regulator |
| | | *luxR* | transcriptional regulator, key player in quorum sensing |
| transport | ABC transporter related | *abc* | ABC transporter. Contains an ATP-binding domain and a permease domain. Sometimes the domains are split into two genes |
| | other transporters | *dbv35* | putative Na+/H- antiporter |

Abbreviations: Tyr: tyrosine; Hpg: hyrdoxyphenylglycine; Dpg: dihydroxyphenylglycine; Bht: $\beta$-hydroxytyrosine; PKS: polyketide synthase; NRPS: non-ribosomal peptide synthetase; AA: amino acid.

## How do the newly discovered GPAs fit into the classification?

The main steps in the biosynthesis of type V GPAs are analogous to the ones described above, with some differences. The precursor supply of non-proteinogenic AAs, which does not include Bht, is established first and the relevant genes are found in the biosynthetic gene clusters associated with the biosynthesis of type V GPAs (from now on referred to as type V GPA BGCs). The stepwise assembly of the backbone is conducted via NRPS complexes, whose organisation into modules and domains is the same as in the types I-IV GPAs, though the total number of modules and hence, the number of AAs in the backbone, can be 7 or 9 [11], [14]. There are several differences in the functional domain composition of the NRPS involved in the biosynthesis of type V GPAs, compared to the rest (**Supplementary Table 2**). As mentioned above, a characteristic of all known type V GPA compounds is the presence of a Trp AA, which corresponds to an adenylation domain (A-domain) selecting for this AA in their NRPS genes [22]. A condensation starter domain (C-starter domain) is sometimes observed in the first module, whose role is either the acylation of the N-terminal AA or the initiation of the NRPS assembly line [22]. Additionally, an N-methyltransferase domain (nMT-domain) has been detected in the 6th module of NRPS in a few type V GPA BGCs, which translates to a methylated

Tyr in the structure of the compounds [22]. Finally, in some BGCs there is an inactive A-domain observed, which is located between the X-domain and the terminal thioesterase domain (TE-domain).

Apart from the backbone assembly, the peptides of type V GPAs are also crosslinked and modified by tailoring enzymes, though the specific reactions have diverged from the ones observed in GPAs of types I-IV. One notable difference is related to the number and function of the genes encoding for p450 monooxygenases. Type V GPA BGCs often include a lower number of such genes compared to the number of observed crosslinks in the compound structure, which for the types I-IV GPAs always had a ratio of 1:1. In some cases, the reason for this is that a single enzyme catalyses more than one crosslinking reaction [5] and in others that fewer crosslinks are formed in the compound [11], [14]. Furthermore, the presence of a gene encoding ferredoxin, which supplies electrons to the p450 monooxygenases, is common only in type V GPA BGCs, whereas this gene has been found in locations distant to the GPA BGCs in the genomes of type I-IV GPA producers [23].

Further modifications do take place in the biosynthesis of type V GPAs, either during the backbone assembly, or after the release of the assembled peptide. The first category includes crosslinking reactions and halogenations, like in types I-IV, though the specificity of the enzymes differ to compensate for the substrate of type V (different backbone length and composition) [7], [24]. The second category is rather poor compared to types I-IV, as no genes related to methylation or sulfation have been found in type V GPA BGCs, though methyltransferase action of NRPS-domains has been observed [22]. However, there is a four-gene cassette found in the misaugamycin (type V GPA) BGC, which has not been observed in any other [22]. These enzymes are hypothesised to mediate the production of the N-terminal acyl chain moiety of the compound.

One of the most striking differences concerns the glycosylation of the compounds. No genes encoding enzymes related to sugar synthesis or transport or modification have been detected in any of the type V GPA BGCs [22]. Though this has been reported on only one occasion in a type III GPA [25], it is an outstanding contrast to the rest of the GPAs, whose name originates from the presence of sugar moieties on their compounds' structure.

The other conspicuous dissimilarity is the lack of the *vanHAXY* resistance genes from all type V GPA BGCs [14]. These genes are commonly found in type I-IV GPA BGCs, though they are sometimes present elsewhere in the genome [26]. This characteristic could very well be connected to the recently discovered alternative mode of action of type V GPAs, which requires an adapted way of self-resistance.

Based on our understanding of GPAs biosynthesis and their mode of action, it is evident that the recently identified type V GPAs exhibit significant differences from type I-IV GPAs. These differences include variable length, lack of glycosylations and

a distinct mode of action [11], [14] and raise the critical question regarding the classification of these compounds: do they truly belong to the GPAs class, or do they represent a separate class? To address this question, we analysed an extensive dataset of GPA BGCs. Our comprehensive phylogenetic investigations spanning multiple levels such as whole BGCs, specific genes and domains, revealed a substantial 'distance' between type V GPAs and all other known types, encompassing both structural and evolutionary differences. Based on these results, we re-evaluated the classification of the types of GPAs and herein propose a revised classification system that integrates phylogeny and chemical structure to more appropriately define these peptide natural products.

## Results

### Analysis of the differences in type V GPAs structure and BGCs

The investigations started by exploring the evolutionary history of GPA-synthesis encoding biosynthetic gene clusters (from here on referred to as “GPA BGCs”). To this end, we generated a comprehensive dataset of GPA BGCs from previously published BGCs [9], [11], [14], [22], [25], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45] supplemented by an extensive search of public sequence databases targeting the characteristic GPA signature X-domain. The candidate loci were used as input for an antiSMASH [46] analysis, which revealed the location and likely limits of the GPA BGCs. Due to the large number of the BGCs, BiG-SCAPE [47] was used to group them into gene cluster families (GCFs) based on similarity, and all BGCs belonging to the same GCF were processed at the same time in the next step, the clinker [48] visualisation of related BGCs. The clinker figures of related BGC, in combination with the antiSMASH results, were necessary for the identification of incomplete BGCs and for the uniform manual trimming of similar BGCs (**Supplementary Figure 1**), whose associated compounds likely belonged to the same GPA type. The latter step was necessary both to overcome the chance of including unrelated genes in antiSMASH-detected BGCs and to ensure uniformity of the choice of genes that were considered related to GPA biosynthesis.

After this meticulous manual review of the candidate BGCs for completeness and the correction of their limits, the final dataset comprised 182 GPA BGCs that were found in the genomes of at least 9 different bacterial genera (**Supplementary Table 3**). Within our dataset, numerous BGCs lacked connections to known natural products or exhibited low similarities to known BGCs. To gain insight into their potential structures and their resemblance to known compounds, the AA sequences of the A-domains were extracted and an analysis of their Stachelhaus selection code was conducted (**Supplementary Table 4**).

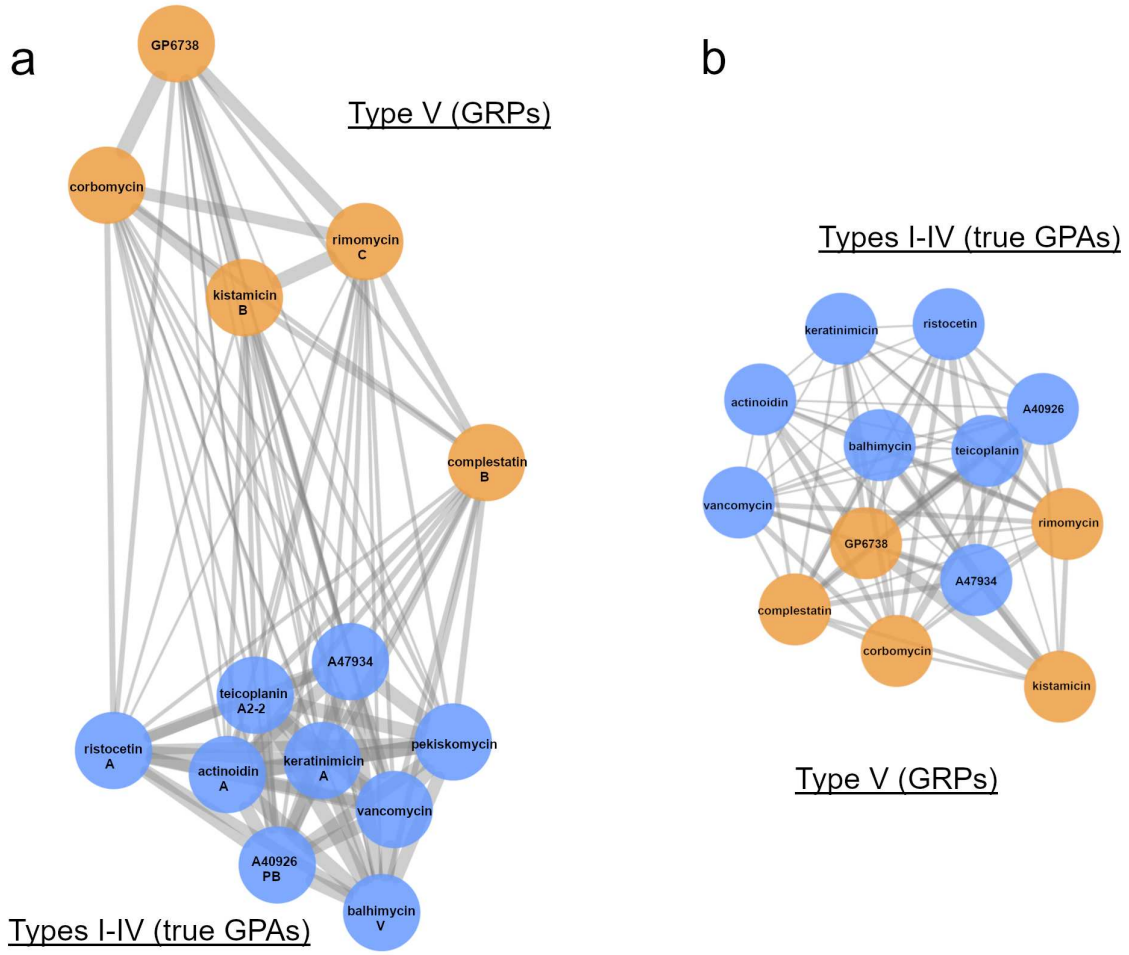



**Figure 2: Structural and biosynthetic distance of type V GPAs**. Panel a: Tanimoto similarity network of selected GPA structures (see Methods, **Supplementary Table 1**). Panel b: domain sequence similarity (DSS) network of selected GPA BGCs (corresponding to the structures of panel a), as calculated by BiG-SCAPE. Both networks were visualised with Cytoscape [49] using the Prefuse Force-directed layout based on their similarity metric. The type V GPAs are coloured orange, while the type I-IV GPAs are shown in blue. The width of the edges is analogous to the value of the similarity metric (legends are found in **Supplementary Figure 2**).

The BGCs in our final dataset were classified into types I-V based on their similarity to a prototypical BGC of each GPA type and on their genes and predicted backbone composition. Our analysis revealed a significant diversity and distinctiveness in the BGCs of type V GPAs compared to those of types I-IV. Significant differences were observed between BGCs encoding type V GPAs and BGCs encoding type I-IV GPAs. The former lacked several genes commonly found in the latter, while also containing additional genes encoding supplementary enzymatic functions (**Supplementary Table 5**).

Given these differences, we conducted further investigations using known structures of GPAs, which were available for some of the BGCs in our dataset (see Methods). Their PubChem fingerprints were compared using the Tanimoto similarity metric [50], to investigate how alike the structures of the different types are. The result was visualised in a network based on their pairwise similarity values (**Figure 2a**).

Although there are notable structural similarities among type I-IV GPAs, there is a considerable divergence with type V GPAs. Type V GPAs appear to occupy a distinct (and expansive) chemical space. This is in accordance with the structural differences detected in type V, including, but not limited to, the lack of sugar moieties (Introduction: Existing GPA classification). Given this, the term “glycopeptide” becomes inaccurate when applied to molecules of type V. Therefore, we propose a dichotomous reclassification of the current GPAs into two categories: true GPAs, a term introduced by Culp *et al* in 2020 [14], which include the types I-IV GPAs, and glycopeptide-related peptides (GRPs), which include the current members of type V.

The differences in structure between true GPAs and GRPs were detectable, though less clear, when the corresponding BGCs are compared using their domain sequence similarity (DSS) as estimated by BiG-SCAPE [47] (**Figure 2b**). However, the NRPS domains involved in this calculation are most closely correlated to the backbone composition of the resulting compounds, while there are many more structural differences to take into account, which can be attributed to other genes in the BGCs (**Table 1**).

**Phylogenetic analysis of full BGCs supports new naming convention**

To take the full genetic content of the BGCs into account and examine if the structural differences of the GRPs (type V GPAs) compounds are mirrored in the evolutionary history of the BGCs, phylogenetic analyses of each gene had to be carried out. Due to the size of the dataset and the lack of conserved gene synteny/order in true GPA/GRP-encoding BGCs (from here on referred to as GPA BGCs and GRP BGCs), an orthology inference analysis was conducted before selecting sequences for phylogenetic analysis. The zol tool [51] was used to identify homologous groups of all genes (OGs), except the core biosynthetic NRPS genes, and phylogenetic trees were calculated based on their translated protein sequences.

Due to the established unusual events taking place in the evolution of NRPS genes [52], such as gene fusion and separation, shuffling and recombination of domains and modules, they were instead analysed on a domain level. Phylogenetic trees were built for each type of functional domain and, when applicable, separated by order of module. The latter distinction is important for two reasons. The first one is, that to reconstruct the phylogenetic history of all true GPA and GRP BGCs based on the trees of their genes and domains, only one copy can be included in each tree per BGC. The NRPS are modular enzymes, and each module includes one of certain types of functional domains, which means that there can be multiple domains of the same type within one BGC (e.g. 7 A-domains for a BGC that produces a heptapeptide). Therefore, by considering each module separately, all the domains can be included in the analysis, retaining the information they carry, but at the same time conclusions for the BGC as a whole can be reached. The second reason is that by considering each module there is emphasis on the connection of the NRPS genes with the backbone of the resulting compound, especially its length, which directly correlates to the number of modules present in the NRPS genes.

The diversity of genes and domains present was very high in our dataset, but the sequences present in at least half of the BGCs in the dataset were deemed informative. A graphical summary of their corresponding trees is visualised in a super network (**Figure 3**). The difference in phylogenetic diversity between true GPAs and GRPs is evident once again. The network illustrates the extensive gene flow between the BGCs of the two classes. The level of conservation between different genes and domains is not equal, increasing the complexity of their evolutionary history. However, there is a clear separation between GPAs and GRPs. The true GPAs are occupying a smaller space but can be further divided into I, II and III/IV types in the network. Types III and IV form a mixed clade due to their definition - the two types can be differentiated only by the presence of one gene: an acyltransferase. Moreover, the super network displayed in **Figure 3** indicates that the set gene trees, despite an otherwise high level of incompatibility among the trees, largely agree on the partitioning of the GRP BGCs into clear clades, which we propose to use for their subsequent categorisation into A-E types.

To further support the previous analyses, an additional phylogenetic analysis was performed by concatenation of (congruent) core gene and domain protein sequences from all true GPA and GRP BGCs. The genes and domains considered as 'core' were those present in at least 90% of the clusters in the dataset (**Supplementary Data 2**), while their congruence was established by an additional analysis (see Methods). As expected, the concatenated phylogeny turned out to be in good agreement with the previous evolutionary picture in **Figure 3**, with the same clades and subclades forming. A combined visualisation of the representative phylogeny together with the gene content of each BGC (**Figure 4, Figure 5**) revealed several patterns (gene presence/absence) that are characteristic for each suggested new type, which are presented in the corresponding sections.

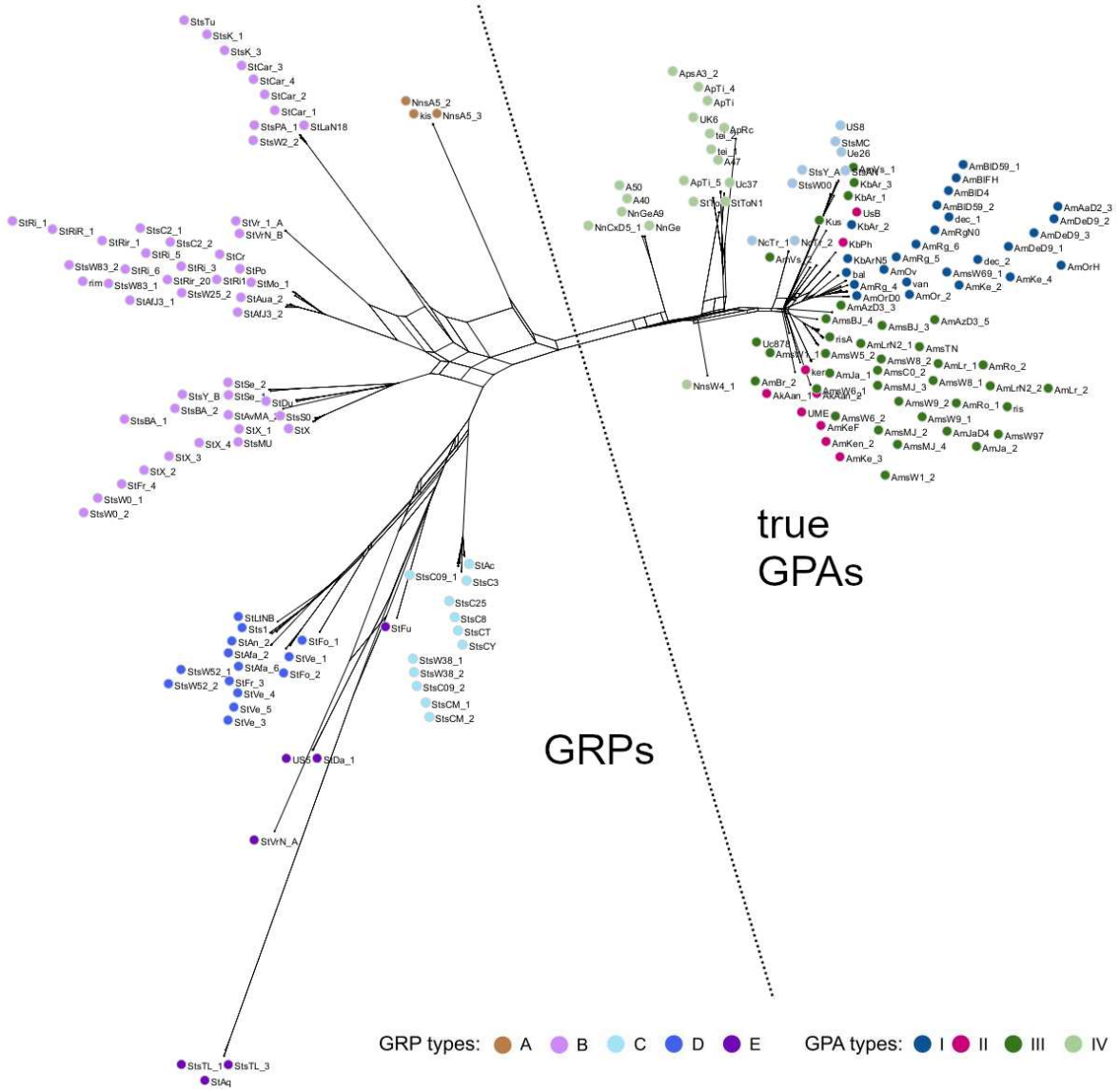



**Figure 3: Phylogenetic network of GPA BGCs**. Super network constructed from ML trees of all genes and domains from true GPA and GRP encoding BGCs (**Supplementary Table 3**) (seed=0), computed with the SplitsTree program, using “greedy weak compatibility” filtering to reduce visual complexity (**Supplementary Data 1**). The true GPAs and GRPs are separated with a dotted line, to make the separation clearer. Each suggested new type is coloured differently next to the BGC ID. The network is built in three dimensions, which is why some branches are not completely clear.


**Figure 4: Phylogeny and gene patterns in GRP BGCs**. Left: concatenated phylogeny of true GPA (collapsed) and GRP encoding BGCs (**Supplementary Data 2**, **Supplementary Table 3**). The types are coloured differently next to the tree leaves. Right: presence/absence heatmap of genes and domains, organised per general function (labels). For more information see **Supplementary Table 5**, **Supplementary Figure 3**.


**Figure 5: Phylogeny and gene patterns in GPA BGCs**. Left: concatenated phylogeny of true GPA and GRP (collapsed) encoding BGCs (**Supplementary Data 2**, **Supplementary Table 3**). The types are coloured differently next to the tree leaves. Right: presence/absence heatmap of genes and domains, organised per general function (labels). For more information see **Supplementary Table 5**, **Supplementary Figure 3**.

## Examination of the true glycopeptide antibiotics (GPAs) BGCs and predicted structures

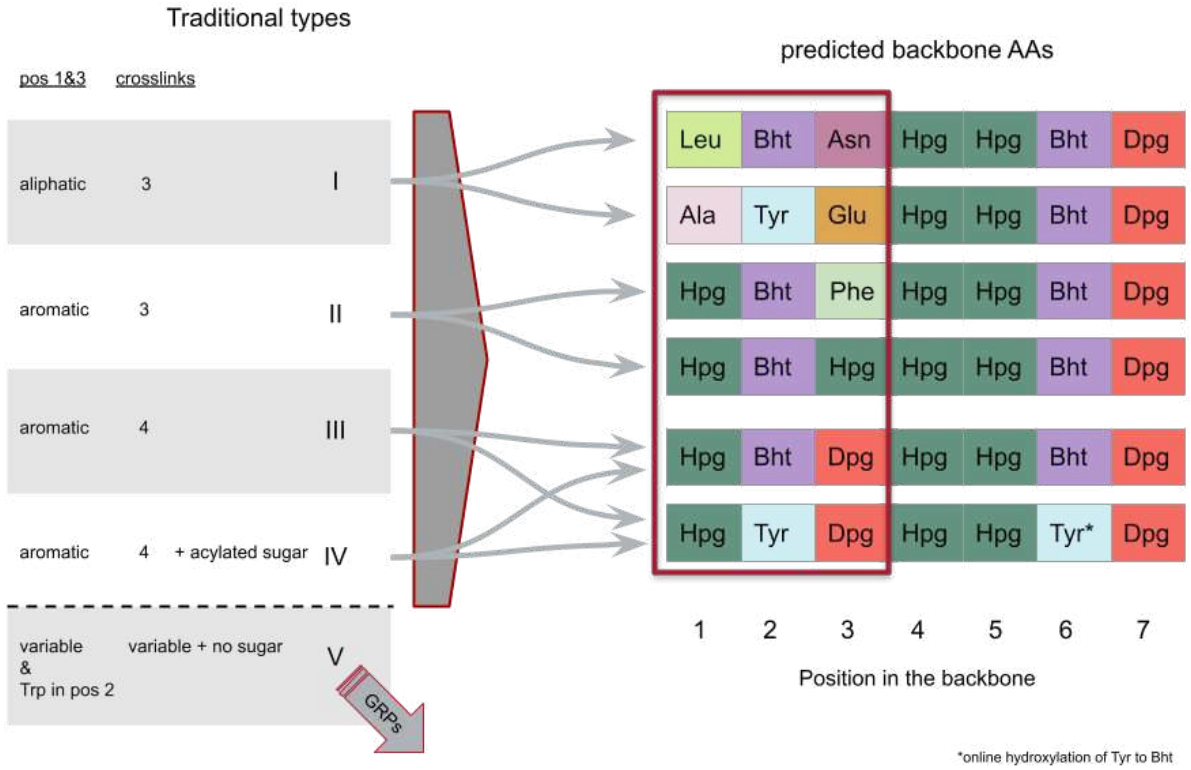



**Figure 6: GPA types I-IV traditional and observed predicted backbone compositions.** On the left, the current structural characteristics that define the GPA types I-V are listed. Arrows connect the types with the backbone AAs of their compounds (as inferred in this analysis) on the right. The type V type description is shown but an arrow leads it away from the rest and towards the GRPs, to highlight the fact that it is no longer considered among true GPAs and is discussed with the GRPs in Figure 7. The region in the red frame shows the variance among the types and their distinctive "motifs" (AAs in positions 1-3). The backbone composition is not discernable between types III and IV, since two variations, which depend on the method of Bht synthesis (online or offline), were observed in both types. Types III and IV differ only on the presence of an acyl group on their sugar moiety. Amino acids are determined based on the Stachelhaus code [53] analysis (**Supplementary Table 4**). Ala, alanine; Asn, asparagine; Bht, b-hydroxytyrosine; Dpg, dihydroxyphenylglycine; Glu, glutamate; Hpg, hydroxyphenylglycine; Leu, leucine; Phe, phenylalanine; Tyr, tyrosine; Tyr*, A-domain specificity suggests Tyr but gene pattern suggests online hydroxylation to Bht.

Our phylogenetic analyses (**Figure 5**) reaffirmed the established classification of GPAs, while the predicted backbone composition of the compounds associated with these BGCs is highly conserved within the GPA types (**Figure 6**), especially in positions 4-7, in agreement with other findings [54]. Additionally, our analysis agrees with the observed types of NRPS genes and domains described in the introduction and in **Table 1** (**Supplementary Table 5**). Enzymes related to sugar synthesis, transport and modification were encoded almost exclusively in GPA BGCs and resistance genes (**Table 1**) followed a similar pattern. Furthermore, genes related to the precursor supply of Dpg and BhT were observed in all GPA BGCs. Tailoring enzymes were present more often and in greater variety in the GPA BGCs compared to the GRP BGCs. In general, there were some gene presence/absence patterns

discernable among the GPA types, each of which is described in detail in the corresponding paragraph of the type.

Type I GPAs are recognisable from the first three AAs in their predicted backbone (from now on referred to as their motif), featuring aliphatic amino acids (AAs) at positions 1 and 3, which are not observed in any other type, and they form a monophyletic clade (**Figure 5**). Their members form two subtypes, which are supported both by their phylogeny and gene patterns, as well as their backbone composition (**Figure 6**). The first subtype comprises most known type I GPAs, such as vancomycin and balhimycin, and can be distinguished by a starting tripeptide of Leu-Bht-Asn. The second subtype, which includes pekiskomycin, differs mainly by the different (non-aromatic) amino acids in positions 1 to 3 of the core peptide: Ala-Tyr-Glu. The remaining positions, which are highly conserved among all true GPAs, are occupied by aromatic amino acids, forming three cyclizations encoded by the corresponding *oxy* genes. Type I GPAs have a small difference in gene patterns among the two subtypes: genes involved in sugar transport are common for the vancomycin subtype but not for the pekiskomycin subtype.

Type II is similar to type I in its phylogeny: it forms a monophyletic clade (**Figure 5**) and has a characteristic motif (**Figure 6**) with subtypes (and subclades) associated with the 3rd AA in the backbone, which is either an Hpg, as in VEG, or a Phe, as in pekiskomycin. Like type I GPAs, these aromatic AAs are not interconnected. Type II GPAs also exhibit three cyclizations as witnessed by their *oxy* genes. Their gene presence/absence pattern is very similar to that of the type I GPAs, making the domains related to their backbone composition (A-domains) their most distinct features.

The exception in the agreement between phylogeny and classification lies within types III and IV, which form mixed clades. Both types feature four crosslinked amino acids in their backbone, which is mirrored by the presence of four *oxy* genes in their BGCs (**Figure 5**). The current structural distinction between Type III and Type IV GPAs is based on the presence of an acyl chain attached to the sugar, a characteristic feature of Type IV GPAs. Due to this fact, their BGCs differ on the presence (type IV) or absence (type III) of a single acyltransferase gene. Therefore, their mixed phylogeny can be expected and it is in agreement with prior analyses [54]. Another notable difference is the way in which these GPAs incorporate Bht into their backbones. In type IV GPAs, only online hydroxylation of Tyr into Bht had been observed, whereas in type III GPAs both online and offline Bht hydroxylation is possible. Our recent analyses have revealed an unexpected finding: type IV GPAs, traditionally associated with online Bht synthesis, which harbour the three-gene system for offline Bht synthesis within their BGCs. These BGCs are not associated with any known compounds and were classified into type IV in our dataset due to the presence of the characteristic acyltransferase gene. Furthermore, in the case of the types III and IV, it is the method of producing Bht that seems the most congruent with the phylogeny (**Supplementary Table 5**), since the BGCs that include the

three-gene system responsible for offline Bht production (**Table1**) are forming a monophyletic clade, which includes all of type I and II GPA BGCs, and some type III and IV GPA BGCs. On the other hand, the type III and IV GPAs that encode the gene for online hydroxylation of Tyr form mixed clades. For both types, Hpg is incorporated at position 1 and Dpg at position 3 of the backbone (**Figure 6**). These aromatic amino acids undergo cyclization, resulting in a fourth cyclization in both type III and type IV GPAs, and an additional *oxy* gene in their BGCs. Having both on- and offline Tyr hydroxylation present in types III and IV and otherwise lacking any differences in their observed (in known structures) and predicted backbone composition, they are the only GPA or GRP types which can not be discriminated based on their BGCs' A-domains' specificities. Given these complexities, a clear demarcation between type III and type IV GPAs has become challenging. Therefore, we propose to maintain the previous classification scheme: type III GPAs lacking the acyl chain and type IV GPAs possessing the acyl chain, irrespective of the backbone structure, which is indistinguishable between these two types. It is worth noting that online hydroxylation of Bht has not been detected in any other type of GPA or GRP. As far as the rest of the observed genes are concerned, there is no obvious type-specific pattern.

It is important to note that there were some true GPA BGCs which posed an exception to the well-defined characteristic of sugar-related genes included in the BGC limits (**Figure 5**). The BGCs in question belonged to type I, and were most similar to pekiskomycin (BGC IDs: StsY_A, StsAN), and to type III, which includes the known case of the BGC encoding the biosynthesis of A47934 [40] and closely related BGCs (BGC IDs: StTo, StToN1).

## Classification of glycopeptide-related peptides (GRPs)

The glycopeptide related peptides (GRPs, former type V GPAs), have shown significant divergence from the true GPAs, both in their mode of action (Introduction), their structure (Introduction, **Figure 2**), their phylogenetic distribution (**Figure 3**, **Figure 4**) and their gene presence/absence patterns (**Figure 4**). This phenomenon is also observed in their predicted backbone length and composition (**Figure 7**), which are clearly discernible from the ones belonging to the true GPAs (**Figure 6**). These observations support our proposal to distinguish the GRPs from the true GPAs. In general, the GRP BGCs lack genes associated with sugars, as well as resistance genes (**Table 1**), but seem to employ more regulators and transport enzymes. They are quite poor in their tailoring enzymes that act after the assembly of the backbone, compared to the GPAs, but there are some genes found only in GRP BGCs whose role has not been determined. No GRP compound that includes a Bht in the backbone has been discovered so far, and they mostly lack the associated genes. Their NRPS genes can include more domains than GPAs: they display additional E-domains and TIGR01720 domains, as well as domains not observed in GPAs at all. The latter include C-starter domains, nMT-domains as well as an additional final inactive A-domain. Furthermore, the phylogeny of the GRPs reveals

clear subclades, which can be associated with a characteristic backbone motif (**Figure 7**), and which we propose to associate with the subcategorization of the GRPs into types A-E. These types, which all show a characteristic NRPS domain pattern, are described in detail with regards to their predicted backbone and common genes.

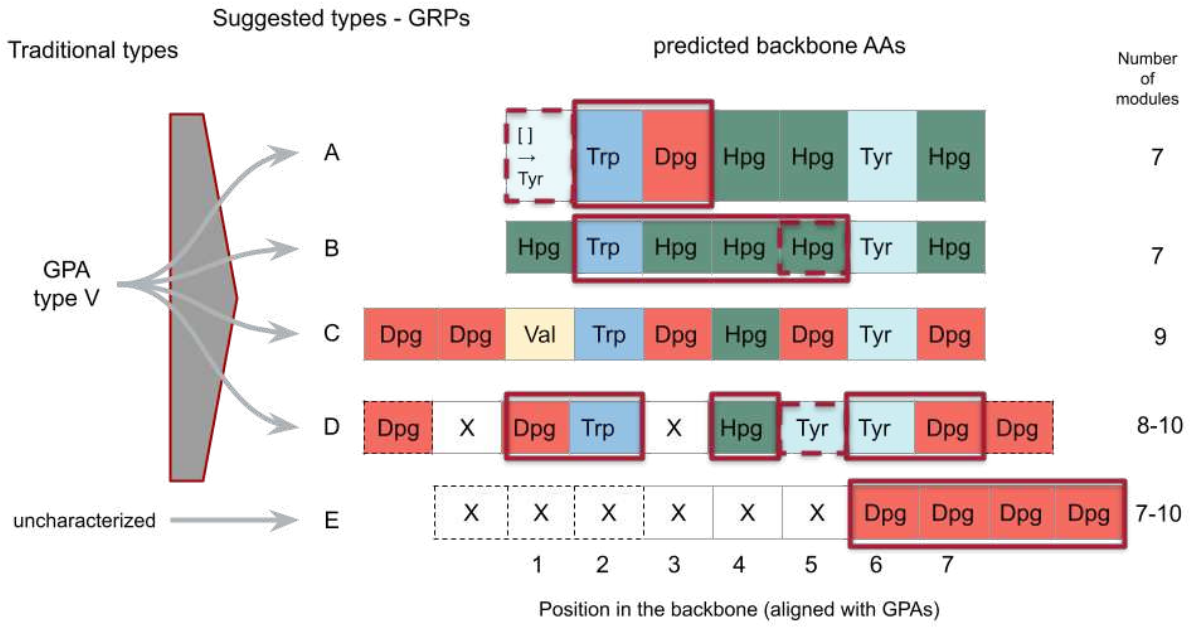



**Figure 7: GRP suggested new types.** On the left, the current classification of type V GPAs is depicted. Arrows show the assignment of its members to the new types of GRPs, since the type V is no longer considered among true GPAs. There are also a few BGCs with uncharacterized structure that were detected in the present analysis, which are also assigned to the GRPs (type E). On the right, the names of the new suggested types are displayed (A, B, C, D, E), along with the pattern of their predicted backbone AAs. The region in the red box shows the variance among the types and their distinctive pattern that allows quick recognition (their "motifs") and the dotted boxes highlight positions with exceptions. On the right, the number of modules is displayed, since this is variable for GRPs. For easy reference, the backbone AAs of the GRPs are shown as they would align with the GPAs (**Figure 6**). The modules marked X and coloured white are variable within the type. The dotted lines around a module indicate that its presence is optional (variable module number within a type). More details on the types themselves: The first module of type A is actually not present in the NRPS genes, but the AA is present in the structure, which is why it is included here. Type B has an exception clade (**Figure 4**) where the 5th AA is a Dpg instead of an Hpg. Type C is very conserved and the backbone AAs always appear as displayed, which is why there is no box indicating the characteristic region. Type D has a fragmented pattern, which is completely discernable from type C BGCs with the same amount of modules, and again an exception clade where the Tyr in position 5 is a Dpg. Type E shows very high variety, but can be distinguished by the presence of four Dpgs at the end of the backbone. Amino acids are determined based on either confirmed or predicted A-domain specificity (Stachelhaus code [53] analysis, **Supplementary Table 4**).

## Type A GRPs

The proposed type A GRPs include only the BGC encoding kistamicin biosynthesis. It is characterised by a Trp-Dpg motif in positions 2 and 3 and the presence of 7 NRPS modules (**Figure 7**). Notably, the first module is missing the A-domain, though

the compound contains a Tyr in the first position, a phenomenon still unresolved. Type A is the GRP type most closely resembling the gene presence/absence pattern of the true GPAs, and it does form an outgroup to the rest of the GRPs. It includes all genes associated with the precursor supply of dpg (**Table 1**), which is not the case for any other GRP but is common in GPAs. Also, the same gene encoding a halogenase that is common in GPAs is present in this type too. Finally, type A BGCs encode a unique *oxyC*, which performs two crosslinking reactions [5], but lack *oxyB*.

**Type B GRPs**

Type B GRPs also include 7 amino acids but they differ from type A in the absence of Dpg-specific A domains within the NRPS. Instead their unique motif is Trp-Hpg-Hpg-[Hpg] in positions 2-5 (**Figure 7**). Position 5 is in brackets because though it is extremely conserved within the type, there is one exception where instead of an Hpg there is a Dpg, in a subclade comprising the BGCs StVr_1_A and StVrN_B. Surprisingly, genes related to the precursor supply of Dpg are absent from the whole type, including these two exceptions that are predicted to include Dpg in their backbone. These two BGCs are unusually small in length and are lacking a lot of other characteristic genes as well. No structure has been associated with any of them yet so the effect these absences have to the resulting compound is uncertain. There are a lot of genes that are present in the majority of the type B BGCs, but few that are absolutely conserved in the type, such as a gene encoding a halogenase, which is also found in type A and in a lot of true GPAs. Type B GRPs include the following known structures, which are phylogenetically distinct (**Figure 4**): complestatin, rimomycin and misaugamycin. There are some differences in gene patterns among these subclades, though not enough to place them in subtypes. For example in the regulation-related genes, the BGCs in the complestatin subclade include only the vanRS two-component system and a StrR-like regulator, the BGCs in the rimomycin subclade contain only a StrR-like and a LuxR-like regulator, and the BGCs in the misaugamycin subclade carry a copy of all of these genes. Some BGCs from the rimomycin and completatin subclade include nMT-domains in their sixth NRPS module. Only in the misaugamycin subclade, there is sometimes also a four-gene cassette which is hypothesised to be involved with the N-acylation of the compound [22].

**Type C GRPs**

The GRP encoding BGCs of type C are very uniform, as evidenced by their clearly defined monophyletic clade in all phylogenetic representations (**Figures 3, 4**) and their gene patterns (**Figure 4**). These GRPs, which include the known GP6738 structure, always contain nine amino acids in their backbone, with the third amino acid of the peptide being a valine (Val), followed by the - highly conserved in most GRPs - Trp (**Figure 7**). They include genes related to the precursor supply of Dpg and Hpg, as well as an alpha/beta hydrolase known to be involved in the offline hydroxylation of Bht, though no A-domain specificity for Bht was detected an the known structure also does not include this AA. BGCs of this type only include two

*oxy* genes, and the known compound from this type includes two crosslinked AAs. Finally, type C GRP BGCs are the only ones encoding two copies of the vanR/vanS two-component system.

**Type D GRPs**

The number of amino acids in the backbone of type D GRPs, which include corbomycin, ranges between 8 and 10 (**Figure 7**). Though more variant in their backbone composition than other types discussed so far, there is a discernible pattern (**Supplementary Figure 5**). They have the conserved Dpg-Trp motif in (aligned) positions 1-2, which is followed by a variable amino acid and then an Hpg in (aligned) position 4, which is highly conserved among GRP types A-D, likely forming a crosslink with the Trp. In addition, we see a motif of Tyr-Dpg in (aligned) positions 6 and 7 accordingly. Their gene presence/absence pattern follows that of type C, when genes related to precursor supply are considered. There is a subclade featuring a gene encoding a halogenase not present in other BGCs. Type D BGCs commonly include a LysR-like regulator and two types of ABC transporters. There are also a few other genes routinely observed within the limits of type D BGCs, whose role is not known.

**Type E GRPs**

Type E comprises GRPs that are either heptapeptides or decapeptides and display high within-type variance in their predicted backbone composition (**Supplementary Figure 6**). At the same time, there is no known structure from any BGC coming from this clade. Their backbone contains neither Trp, nor Hpg, but they are characterised by a Dpg-rich tetrapeptide tail (**Figure 7**). Type E BGCs always lack the genes whose enzymes are involved in the shikimate pathway (**Table 1**), which is not a unique phenomenon, but no other type consistently lacks these genes. Their within-type differences in the backbone composition are mirrored by differences in their NRPS domains. They are the only type to display two different patterns in their *oxy* genes. Within type E there is a subclade of two BGCs, which are the only GRP BGCs to encode a mannosyltransferase. Finally, there is another exception BGC (like in type B), which is minimal and lacking a lot of genes characteristically found in GPA/GRP BGCs, such as transporters and regulators. Due to a lack of information on GRP E BGCs, the structures whose biosynthesis they encode, and also their mode of action, it is likely that they will be further reclassified in the future, when more information is discovered.

## Discussion

Following up on the blatant differences of type V glycopeptide antibiotics (GPAs) with all other types both in structure and mode of action, an extensive phylogenetic analysis of the encoding BGCs was conducted. We were able to confirm those discrepancies on the evolutionary level and to underline the contrast in variety within this type compared to the rest. These observations were the basis for the suggested reclassification of GPAs, dichotomizing them first into true GPAs and glycopeptide-related peptides (GRPs) and then further categorising them into types.

The original types I-IV were retained. Type V GPAs were rebranded as GRPs, and their divergent characteristics compelled their division into the new types A-E.

The present dataset of BGCs is the most comprehensive set used in a study about GPAs, to the knowledge of the authors. This was achieved thanks to the extensive sequencing databases accessible to the scientific community. Naturally, our analysis was, to a degree, constrained by the availability and quality of this public genomic information. In fact, some of the BGCs in our dataset came from lower quality or incomplete assemblies of the producer genomes. Whenever possible, a better quality sequence containing the BGC was extracted from more recently sequenced and assembled genomes (and in one case the producer strain was resequenced). Though without experimental evidence it can not be absolutely concluded if the limits of an open reading frame or a gene sequence is accurate, a conscious effort was made to ameliorate this issue. If there was any indication that the BGC was not complete or had fragmented genes, it was not included in our analysis. This process eliminated about half of the BGCs originally detected, but ensured the high quality of the remaining BGCs.

A milestone of our analysis was the separation of the genes involved in GPA BGCs into homologous groups, suitable for alignment and for generating a biologically meaningful phylogenetic tree. The use of suitable bioinformatics tools for solving this problem, in our case the new zol pipeline [51], developed specifically for gene clusters, helped greatly to achieve this, despite the difficulties. Different kinds of genes have different degrees of sequence similarity, which can not be completely handled by this tool, though it performed better than orthology inference tools designed for whole proteomes. For example, there are five types of P450 monooxygenase genes found in GPA BGCs: *oxyA-E* [5] The algorithm was able to discern the evolutionary history of most of them, but *oxyA* and *oxyE* were put together into one group (**Supplementary Figure 4**). The separation of *oxyA* and *oxyE* was conducted manually in this case, with the help of the genes with known functions. Furthermore, there were some occurrences of BGCs having multiple copies of a gene, e.f. *vanR* and *vanS* in the case of GRP type C, but for the phylogenetic analysis it was necessary to keep a maximum of only one gene copy per BGC. To resolve this, whenever multiple options were present, the average phylogenetic distance to the rest of the tree was calculated and the closest clade was kept (see also methods), ensuring that the most closely related sequences would be considered to belong to a group (OG).

The phylogenetic analysis presented in this manuscript was not trivial to design, as a number of issues had to be overcome. Firstly, such an analysis is dependent on the inclusion of genes related to the biosynthesis, regulation and transport of GPAs (or GRPs) and the exclusion of unrelated genes. In our dataset, this distinction was made based on the presence of genes homologous to those identified so far in GPA BGCs [27]. Consequently, one temporal constraint of this analysis was the fact that the BGCs needed to be manually and thoroughly checked, in order to both confirm

that they are, in fact, GPA/GRP encoding BGCs and to determine their precise borders. The borders were defined in a first step with the antiSMASH algorithm [55] which, though very efficient, is known to be charitable with the proposed length of a BGC, often including neighbouring genes that do not really belong [56]. Even though this is preferable to excluding genes that should be part of the BGC, it still demands a manual inspection of the BGCs, if the exact limits are important for further analyses, like in our case. Therefore, each antiSMASH-predicted BGC in our dataset had to be extensively investigated gene by gene (and sometimes domain by domain) to locate the most likely borders of the BGC based on the predicted function of the genes/domains within the BGC. To ensure uniform trimming as well as speed up the process without sacrificing accuracy, the clinker tool [48] was used (**Supplementary Figure 1**) for visualisation of similar BGCs (similar based on BiG-SCAPE analysis with T=0.2 [47]). This means that the determination of the exact BGC limits relied heavily on the existing knowledge of the biosynthesis of known compounds. However, several new and unusual BGCs were brought to light in the present study and their study is expected to unearth new information on the biosynthesis of GPAs and GRPs and most importantly, the genes involved in it. For example, it would be vital to learn more on the (as yet unknown) role of certain enzymes in the biosynthesis, especially among the GRP BGCs.

Another challenging aspect of GPA BGC phylogeny is the fact that evolution is acting on different levels: on the separate genes, but also on the whole BGCs and on distinct domains [52], [57], [58], [59], [60]. As explained earlier, the backbone of GPAs and GRPs is assembled by NRPS enzymes, which can contain from 7 to 10 modules. In each module there are always A and T domains, very often C domains and sometimes also E domains. This means that for each position of the backbone, the corresponding domain sequences needed to be collected to form a homologous group that could be used for constructing phylogeny [61]. The variation in the number of modules introduced an additional difficulty, since a “centre” position had to be chosen around which the rest of the domains would be aligned. This was achieved through a concatenated multiple sequence alignment (MSA) built from the extended (34AA) Stachelhaus codes [53] of the A-domains, as detected by antiSMASH [46]. This MSA and the well-known balhimycin backbone was used as the base for choosing the positions of the domains of all other BGCs (see Methods). This process allowed us to align the modules of all BGCs, regardless of length, and enabled the incorporation of very different BGCs in the same analysis, which so far had not been conducted for GPAs on a BGC level.

Inference of the evolutionary history of a full BGC was complicated by the known fact that many GPA encoding BGCs have been, in part or in their entirety, horizontally transferred [21], [59], [62]. The use of the super network (**Figure 3**) as a method of visualising the complicated evolutionary history of the GPA BGCs was not burdened by this. As opposed to binary trees, which is the most common graph used for phylogenetic analyses, a super network can incorporate data from discongruent

phylogenies. In fact, it enables the visualisation of the harmonious information from the underlying set of trees, such that the evolutionary relationships supported by the majority of the trees are observable [63]. However, the potential presence of horizontal gene transfer (HGT) is always an obstacle to the construction of a biologically sound, concatenated phylogeny. To exclude the use of sequences of suspicious origin, we conducted a congruence analysis prior to the choice of the sequences to include in the concatenated phylogeny of the GPA and GRP BGCs (see Methods). While reducing the sensitivity of the final phylogenetic tree, since fewer genes and domains could be involved (22 out of 37 candidates), this process guaranteed a reduction of the "noise" that HGTed sequences would have introduced and therefore increased the accuracy of the result.

Having overcome these obstacles, we presented an extensive and reliable phylogenetic analysis of the GPA and GRP BGCs, accompanied by information on their predicted backbone and gene presence/absence patterns. The separation between GPAs and GRPs was clear in all our results. However, considering the complexity of the data generated, we struggled in some instances to choose a level of dissimilarity as a criterion for declaring the new types, especially for the GRPs. And there were debatable cases, especially when the phylogeny of the BGCs with an uneven number of modules placed them closely together. Such was the case for type E GRPs (**Figure 7**), where no compound has been isolated and thereby no structure could be used as a guide for those classifications. We expect that further analysis and experimental work on type E will elucidate its biosynthesis and shed light on the process involved, which may lead to changes in the definition of the type. Additionally, we detected a few "minimal" BGCs in GRP types B and E, which were lacking a lot of characteristic genes of GPA and GRP BGCs. It is possible that these genes are present in unrelated locations in the genomes, which was not investigated in the present study. An alternative interpretation is that they encode a "minimal" biosynthetic pathway resulting in a peptide without many additional modifications, unlike most GPA compounds. Furthermore, a quite surprising exception to the expected gene patterns was the existence of GPA encoding BGCs without sugar-related genes in some type IV and type II clusters. Based on the phylogenetic relation to their respective types, it can be assumed that they originally did include sugar-related genes, which were lost later in their evolutionary history. A more focused study on these clusters is needed to elucidate such events.

Though the path to calculating a reliable phylogeny of GPA and GRP BGCs was fraught with challenges, the current analysis does provide the - to our knowledge - most comprehensive study of the evolution of these BGCs. The evidence supporting our suggested reclassification system is present and compelling. The distinct structural characteristics of GPAs prompted their original (traditional) classification into different types, [10], [11]. It was a necessary act and it promoted exchange between scientists from many different disciplines (e.g., biochemistry, synthetic chemistry, microbiology, bioinformatics), each studying GPAs from a different

perspective. Since then though, new knowledge has come to light that needs to be considered. Putting together in the same group compounds and BGCs that are in reality so different would have only hindered progress. The inaccuracies of the current classification system were a necessity that was born from the fact that type V GPAs were the last ones discovered [65], and every new and differently looking GPA BGC has been sorted there since [11], [14], [22], [66], [67]. However, the new classification system we suggest, backed by an exhaustive phylogenetic analysis of a large dataset of BGCs, is expected to fuel new, type-specific research of GPAs and GRPs. The use of both phylogeny and structure-based criteria for the declaration of the new types will ensure the reconciliation of scientific communities focused on evolution and chemical structure, respectively. We hope that this new classification system will get adopted by the community and will improve with new insights.

## Methods

### BGC dataset creation

In order for the evolutionary analysis of GPAs BGCs to be as complete as possible, we aimed to create a dataset of all sequenced clusters. The initial dataset contained all known clusters found in the literature [9], [11], [14], [22], [25], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45] and was extended by searching in publicly available sequence databases (MIBiG [68], most of NCBI RefSeq and Genbank [69], [70], [71], JGI IMG DB [72], MGnify [73]) and some published projects [73]–[[74]–[76]. Information on the origin of the BGCs, the date of the search and more metadata are available in **Supplementary Table 3**. **Supplementary Data 4** includes the accession numbers of all NCBI (RefSeq and Genbank) entries checked and whether they were a hit (True), not a hit (False) and the ones where the search failed due to lack of protein sequences associated with the accession number (Failed). The target of the search, accommodated via HMMER (v. 3.3.1, accessed from http://hmmer.org/, RRID:SCR_005305), was the X-domain in the last NRPS module (its HMM was extracted from antiSMASH v6 [55], RRID:SCR_022060), which is found only in GPA encoding BGCs so far [76]. However, the X-domain has evolved from a condensation domain (C domain) and their sequences remain quite similar [6]. To avoid false positives, the known clusters were searched with the X-domain HMM and the lowest local score of the confirmed X-domains (300) was used as a minimal threshold. Lower local scores (up to 250) were manually checked for the MIBiG dataset and indeed no X-domains were found below the chosen threshold. The custom scripts used for this and all other parts of the analysis can be found in **Supplementary Data 5**.

The sequences that were a hit in the search were used as input for an antiSMASH v7 [46] (RRID:SCR_022060) analysis (default parameters + MIBiG cluster comparison) for the detection of the full BGC. At this stage, after manual inspection, some candidates were dropped due to low quality assemblies which led to obviously incomplete clusters, and due to a few false positives (where a C domain was

mistakenly detected as an X-domain). The cluster regions that passed this inspection were assigned an ID reflecting their taxonomic placement (**Supplementary Table 3**) and their coding domains were re-annotated with bakta (v1.8.1) [78] to ensure homogeneity. Bakta was run with default parameters + skipping trna and tmrna detection (--skip-trna --skip-tmrna) and on metagenome mode when appropriate. The original contig headers were kept (--keep-contig-headers) but the assigned ID was used as a locus and locus tag prefix. The bakta-annotated regions were re-analyzed with antiSMASH v7 [46] with all features on (--fullhmmer –clusterhmmer –tigrfam –asf –cc-mibig –cb-general –cb-subclusters –cb-knownclusters –pfam2go –rre –smcog-trees –tfbs).

One exception to this process was the case of *Streptomyces varsoviensis*, which included a very unusual BGC, which could not be conclusively labelled complete. Due to its interesting characteristics though, the strain was ordered for resequencing. *S. varsoviensis* was generally grown in tryptic soy broth (TSB) (BD Bacto™ Tryptic Soy Broth; Becton, Dickinson and Company, Franklin Lakes, NJ, USA). For isolation of high molecular weight genomic DNA *S. varsoviensis* was cultured in R5 medium [79] for 2 days on a rotary shaker at 28 °C. DNA isolation was performed by using DNA isolation kit (NucleoBond® HMW DNA Kit, Machery-Nagel, Düren, Germany) following the manufacturer's protocol. Genome sequencing was performed by the NGS competence center (NCC in Tübignen, Germany) on a PromethION (Nanopore) with 9.4.1 chemistry (details in **Supplementary Data 6**). Assembly was done with Unicycler [80] (for long reads) and corrected with medaka (accessed from https://github.com/nanoporetech/medaka) based on the sequencing parameters (**Supplementary Data 6**). The resulting genome sequence was searched for the X-domain as described above and BGCs were included in the analysis (**Supplementary Table 3**).

To overcome antiSMASH's generous selection of BGC borders, a manual inspection was necessary, to ensure the quality of the evolutionary reconstruction, which would be affected by the accidental inclusion of unrelated sequences. The process was sped up by annotating the BGCs in groups of GCFs as defined by a BiG-SCAPE analysis [47] (v1.0.1 2020-01-27, RRID:SCR_022561). The chosen threshold of 0.2 defined the largest possible groups that never have more than one type GPA (based on the MIBiG dataset). BiG-SCAPE was run on glocal mode, mixing all classes (as some BGCs are marked NRPS and others are marked other), but using the score weights of the NRPS class. The GCFs were then visualized with clinker [48] (v0.0.28). We meticulously manually curated each cluster by investigating every single gene for its function and possible role in the biosynthesis and the clinker visualization ensured uniform trimming of the most closely related BGCs. Thus, the final dataset of 182 trimmed clusters from 9 different genera was created (**Supplementary Table 3**).

**Chemical structure similarity network**

The structures of the glycopeptides (**Figure 1, Figure 2a**) were collected from the original publications [11], [14], [22], [32], [33], [35], [39], [67], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90] (entries under database: MIBiG in **Supplementary Table 3**), formatted and converted to SMILES with ChemDraw (https://revvitysignals.com/products/research/chemdraw, RRID:SCR_016768) and RDKit (v. 2020.09.1, http://www.rdkit.org/, RRID:SCR_014274).

The examination of the structural variation within and among GPA types was done by use of the Tanimoto similarity metric [50]. The SMILES of known GPA compounds were inserted into the Cytoscape program (v. 3.9.1, RRID:SCR_003032) [49], which accommodates the calculation and visualisation of a structure similarity network based on Tanimoto with *chemViz2* (fingerprint: Pubchem), a cheminformatics app for Cytoscape. The nodes of the resulting network were annotated by GPA type (I-IV or V) and the edge width was adjusted to represent the degree of similarity among the connecting parts.

**Stachelhaus code analysis**

A-domain specificity was predicted (**Supplementary Table 4**) using the 34 AA long Stachelhaus-code prediction from NRPSpredictor2, implemented in antiSMASH (v. 5.1.2, RRID:SCR_022060) [53], [91], [92]. Due to the fact that certain A-domain specificities can not be predicted *in silico* [39], all the known amino acid specificities (for glycopeptides whose structure is known) were used in a blastp (v2.14.0+) search [93] and best scoring alignments were used to determine the annotations used for the rest of the analysis, which sometimes differed from the antiSMASH assigned ones. A summary of A domain specificity was visualised in **Figure 6** and **Figure 7**, while the predicted backbone of all BGCs can be seen in **Supplementary Figure 3.**

**Detection of homologous genes**

The identification of homologous groups of genes from the full dataset (**Supplementary Table 5**) was mostly carried out by the zol tool (v 1.3.9) [51], which infers phylogenetic orthology for comparative genomics of gene clusters. The platform (run with default parameters) generated 309 so-called orthogroups (OGs). There was one case of related genes with distinct functions being placed in the same Orthogroup, namely the oxyAs and oxyEs, which are known to be closely related [5], [22]. Those proteins were used for a multiple sequence alignment (MSA) and their separation into groups was guided by their phylogenetic placement compared to proteins of known function. Cases of OGs with only a few clusters containing multiple copies of a gene were dealt with via a custom script, discarding some copies based on similarity criteria [94]. Finally, the NRPS domains were extracted from the genes and split into groups based on the position of their module compared to the rest.

The latter was achieved by building an MSA from the concatenated 34 AA Stachelhaus codes [53] (as provided in the antiSMASH results [46] with the help of the NRPSpredictor2 tool [91]) of the underlying A-domains (**Supplementary Data 3**). The longest conserved region (positions 4-7 of balhimycin) was taken as a center

and all internal gaps were rejected, since the positions directly correspond to the backbone of the compounds. The alignment was in agreement with the fact that positions 1-3 of the types I-IV glycopeptides are known to be the most variable among known compounds and their corresponding BGCs [54]. All module positions (00, 0, 1-9) were annotated based on this fixed MSA, with the balhimycin encoding BGC as a template for positions 1-7.

**Visualising complete evolutionary history**

All multiple sequence alignments (MSAs) were performed with the mafft tool (v7.490, 2021/Oct/30, RRID:SCR_011811) with default parameters [95] (**Supplementary Data 7**). Phylogenetic trees for every occasion in this study (153 gene and domain trees, as well as concatenated phylogenies described below) were built with iqtree (multicore v2.2.0.3 COVID-edition for Linux 64-bit built Aug 2 2022) [96], [97] (**Supplementary Data 7**). For the gene and domain trees, iqtree was first run only in model testing mode, checking for all bacteria-suitable evolutionary models (suitable models are listed in **Supplementary table 7**) and then in tree-building mode based on the best fitting model. The resulting trees can be seen in **Supplementary Data 1**.

A graphical summary of the adequately populated (n>50%) 47 separate (partial) gene and domain trees was calculated and visualised by the super network algorithm [98] (default options) implemented in the SplitsTree CE tool (version 6.0.10-beta) [99], [100] (**Figure 3**) after greedily selecting a weakly compatible set of splits of maximum support (GreedyWeaklyCompatible splits filter). The super network (**Supplementary Data 1**) summarises the set of input trees, taking into account that many of the trees are incomplete. It is a splits network in which each band of parallel edges represents one of the splits or branches found in the set of input trees. Incompatibilities among the input trees give rise to parallelograms in the network. Edges in the network are scaled to represent the average relative length of the corresponding edges in the input trees. The set of splits computed by the super network method was greedily filtered by decreasing support (number of trees that contain a given split), so as to obtain a subset of “weakly compatible splits” that maintains major incompatibilities, while avoiding higher-dimensional edge configurations in the network, thus avoiding visual clutter.

**Concatenated phylogeny**

A species phylogeny can be constructed from concatenated sequences of core genes, as long as their separate trees are congruent [101]. Following this concept on the BGCs, 22 genes or domains found in at least 90% of the clusters in the dataset were identified. Cases where there were duplications in max 5% of the cases were acceptable but otherwise these groups were single-copy genes. These limits come from the methodology of a wide-scale phylogenetic study [101]. However, it was necessary to ensure the congruence of the participating genes before concatenating them. The underlying MSAs were first trimmed with trimAl (v1.4.rev15 build[2013-12-17], RRID:SCR_017334) [102] with default parameters and then the trees were recalculated (**Supplementary Data 7**). The latter were used in a

congruence analysis as performed by Parks and colleagues [101]: The well supported splits were calculated by checking their existence in random subsampled concatenated phylogenies and then for each gene tree, their presence was used to calculate ‘normalised compatible split length’ ,a metric that reflects congruence of this tree to the rest in the group (**Supplementary Table 6**). This value was computed with a python script, implemented with Biopython [103] (RRID:SCR_007173), FastTree v2.1.11 [104] (RRID:SCR_015501) and a function from the (still in development) GeneTreeTk toolbox (accessed from https://github.com/dparks1134/GeneTreeTk). 11 genes or domains passed a specific threshold (0.67) and were used for the concatenated phylogeny representing the evolutionary history of the GPA and GRP BGCs (**Supplementary Data 2**). The tree was then rooted with the MAD algorithm [105]. This tree, together with an absence/presence heatmap of the various genes and domains present in the clusters was visualised with iTOL [106] (RRID:SCR_018174) (**Figure 4, Figure 5, Supplementary Figure 3**).

## References


[1] E. Stegmann, H. J. Frasch, and W. Wohlleben, “Glycopeptide biosynthesis in the context of basic cellular functions,” *Current Opinion in Microbiology*, vol. 13, no. 5, pp. 595–602, 2010, doi: 10.1016/j.mib.2010.08.011.

[2] M. S. Butler, K. A. Hansford, M. A. T. Blaskovich, R. Halai, and M. A. Cooper, “Glycopeptide antibiotics: Back to the future,” *J Antibiot*, vol. 67, no. 9, Art. no. 9, Sep. 2014, doi: 10.1038/ja.2014.111.

[3] M. H. Mccormick, J. M. Mcguire, G. E. Pittenger, R. C. Pittenger, and W. M. Stark, “Vancomycin, a new antibiotic. I. Chemical and biologic properties,” *Antibiot Annu*, vol. 3, pp. 606–611, 1956 1955.

[4] M. H. Hansen, E. Stegmann, and M. J. Cryle, “Beyond vancomycin: recent advances in the modification, reengineering, production and discovery of improved glycopeptide antibiotics to tackle multidrug-resistant bacteria,” *Curr Opin Biotechnol*, vol. 77, p. 102767, Oct. 2022, doi: 10.1016/j.copbio.2022.102767.

[5] A. Greule *et al.*, “Kistamicin biosynthesis reveals the biosynthetic requirements for production of highly crosslinked glycopeptide antibiotics,” *Nature Communications*, vol. 10, no. 1, Dec. 2019, doi: 10.1038/s41467-019-10384-w.

[6] M. Schoppet *et al.*, “The biosynthetic implications of late-stage condensation domain selectivity during glycopeptide antibiotic biosynthesis,” *Chemical Science*, vol. 10, no. 1, pp. 118–133, 2019, doi: 10.1039/C8SC03530J.

[7] T. Kittilä *et al.*, “Halogenation of glycopeptide antibiotics occurs at the amino acid level during non-ribosomal peptide synthesis,” *Chemical Science*, vol. 8, no. 9, pp. 5992–6004, 2017, doi: 10.1039/c7sc00460e.

[8] W. Wohlleben, E. Stegmann, and R. D. Süssmuth, “Chapter 18. Molecular genetic approaches to analyze glycopeptide biosynthesis,” *Methods Enzymol*, vol. 458, pp. 459–486, 2009, doi: 10.1016/S0076-6879(09)04818-6.

[9] A. W. Truman, M. J. Kwun, J. Cheng, S. H. Yang, J.-W. Suh, and H.-J. Hong, “Antibiotic resistance mechanisms inform discovery: identification and characterization of a novel amycolatopsis strain producing ristocetin,” *Antimicrob Agents Chemother*, vol. 58, no. 10, pp. 5687–5695, Oct. 2014, doi: 10.1128/AAC.03349-14.

[10] S. Chen, Q. Wu, Q. Shen, and H. Wang, “Progress in Understanding the Genetic Information and Biosynthetic Pathways behind Amycolatopsis Antibiotics, with Implications for the Continued Discovery of Novel Drugs,” *ChemBioChem*, vol. 17, no. 2, pp. 119–128, Jan. 2016, doi: 10.1002/cbic.201500542.

[11] M. Xu, W. Wang, N. Waglechner, E. J. Culp, A. K. Guitor, and G. D. Wright, “GPAHex-A synthetic biology platform for Type IV–V glycopeptide antibiotic production and discovery,” *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1–12, Oct. 2020, doi: 10.1038/s41467-020-19138-5.

[12] D. Zeng *et al.*, “Approved Glycopeptide Antibacterial Drugs: Mechanism of Action and Resistance,” *Cold Spring Harb Perspect Med*, vol. 6, no. 12, p. a026989, Dec. 2016, doi: 10.1101/cshperspect.a026989.
[13] L. Tian, S. Shi, X. Zhang, F. Han, and H. Dong, “Newest perspectives of glycopeptide antibiotics: biosynthetic cascades, novel derivatives, and new appealing antimicrobial applications,” *World J Microbiol Biotechnol*, vol. 39, no. 2, p. 67, 2023, doi: 10.1007/s11274-022-03512-0.
[14] E. J. Culp *et al.*, “Evolution-guided discovery of antibiotics that inhibit peptidoglycan remodelling.,” *Nature*, no. May, 2020, doi: 10.1038/s41586-020-1990-9.
[15] K. C. Nicolaou, C. N. C. Boddy, S. Bräse, and N. Winssinger, “Chemistry, Biology, and Medicine of the Glycopeptide Antibiotics,” *Angewandte Chemie International Edition*, vol. 38, no. 15, pp. 2096–2152, 1999, doi: 10.1002/(SICI)1521-3773(19990802)38:15<2096::AID-ANIE2096>3.0.CO;2-F.
[16] B. R. Miller and A. M. Gulick, “Structural Biology of Non-Ribosomal Peptide Synthetases,” *Methods Mol Biol*, vol. 1401, pp. 3–29, 2016, doi: 10.1007/978-1-4939-3375-4_1.
[17] G. Yim, M. N. Thaker, K. Koteva, and G. Wright, “Glycopeptide antibiotic biosynthesis,” *Journal of Antibiotics*, vol. 67, no. 1, pp. 31–41, Jan. 2014, doi: 10.1038/ja.2013.117.
[18] T. Izoré *et al.*, “Structures of a non-ribosomal peptide synthetase condensation domain suggest the basis of substrate selectivity.,” *Nature communications*, vol. 12, no. 1, p. 2511, May 2021, doi: 10.1038/s41467-021-22623-0.
[19] L. Luo, R. M. Kohli, M. Onishi, U. Linne, M. A. Marahiel, and C. T. Walsh, “Timing of Epimerization and Condensation Reactions in Nonribosomal Peptide Assembly Lines: Kinetic Analysis of Phenylalanine Activating Elongation Modules of Tyrocidine Synthetase B,” *Biochemistry*, vol. 41, no. 29, pp. 9184–9196, Jul. 2002, doi: 10.1021/bi026047+.
[20] M. Kaniusaite, J. Tailhades, E. A. Marschall, R. J. A. Goode, R. B. Schittenhelm, and M. J. Cryle, “A proof-reading mechanism for non-proteinogenic amino acid incorporation into glycopeptide antibiotics,” *Chemical Science*, vol. 10, no. 41, pp. 9466–9482, 2019, doi: 10.1039/C9SC03678D.
[21] S. Donadio, M. Sosio, E. Stegmann, T. Weber, and W. Wohlleben, “Comparative analysis and insights into the evolution of gene clusters for glycopeptide antibiotic biosynthesis,” *Molecular Genetics and Genomics*, vol. 274, no. 1, pp. 40–50, Aug. 2005, doi: 10.1007/s00438-005-1156-3.
[22] M. Xu, W. Wang, N. Waglechner, E. J. Culp, A. K. Guitor, and G. D. Wright, “Phylogeny-Informed Synthetic Biology Reveals Unprecedented Structural Novelty in Type V Glycopeptide Antibiotics,” vol. 46, p. 18, doi: 10.1021/acscentsci.1c01389.
[23] N. Geib, T. Weber, T. Wörtz, K. Zerbe, W. Wohlleben, and J. A. Robinson, “Genome mining in Amycolatopsis balhimycina for ferredoxins capable of supporting cytochrome P450 enzymes involved in glycopeptide antibiotic biosynthesis,” *FEMS Microbiol Lett*, vol. 306, no. 1, pp. 45–53, May 2010, doi: 10.1111/j.1574-6968.2010.01933.x.
[24] Y. T. C. Ho, R. B. Schittenhelm, D. Iftime, E. Stegmann, J. Tailhades, and M. J. Cryle, “Exploring the Flexibility of the Glycopeptide Antibiotic Crosslinking Cascade for Extended Peptide Backbones,” *ChemBioChem*, vol. 24, no. 6, p. e202200686, 2023, doi: 10.1002/cbic.202200686.
[25] J. Pootoolal *et al.*, “Assembling the glycopeptide antibiotic scaffold: The biosynthesis of A47934 from Streptomyces toyocaensis NRRL15009.” [Online]. Available: www.pnas.orgcgidoi10.1073pnas.102285099
[26] R. Kilian, H.-J. Frasch, A. Kulik, W. Wohlleben, and E. Stegmann, “The VanRS Homologous Two-Component System VnlRSAb of the Glycopeptide Producer Amycolatopsis balhimycina Activates Transcription of the vanHAXSc Genes in Streptomyces coelicolor, but not in A. balhimycina,” *Microb Drug Resist*, vol. 22, no. 6, pp. 499–509, Sep. 2016, doi: 10.1089/mdr.2016.0128.
[27] S. Pelzer *et al.*, “Identification and analysis of the balhimycin biosynthetic gene cluster and its use for manipulating glycopeptide biosynthesis in Amycolatopsis mediterranei DSM5908,” *Antimicrob Agents Chemother*, vol. 43, no. 7, pp. 1565–1573, Jul. 1999, doi: 10.1128/AAC.43.7.1565.
[28] J. M. Wink *et al.*, “Three new antibiotic producing species of the genus Amycolatopsis, Amycolatopsis balhimycina sp. nov., A. tolypomycina sp. nov., A. vancoresmycina sp. nov., and description of Amycolatopsis keratiniphila subsp. keratiniphila subsp. nov. and A. keratiniphila subsp. nogabecina subsp. nov,” *Syst Appl Microbiol*, vol. 26, no. 1, pp. 38–46, Mar. 2003, doi: 10.1078/072320203322337290.
[29] L. Xu *et al.*, “Complete genome sequence and comparative genomic analyses of the

vancomycin-producing Amycolatopsis orientalis," *BMC Genomics*, vol. 15, no. 1, May 2014, doi: 10.1186/1471-2164-15-363.
[30] D. P. Labeda, "Amycolatopsis coloradensis sp. nov., the avoparcin (LL-AV290)-producing strain," *International Journal of Systematic Bacteriology*, vol. 45, no. 1, pp. 124–127, Jan. 1995, doi: 10.1099/00207713-45-1-124/CITE/REFWORKS.
[31] J. Wink *et al.*, "Amycolatopsis decaplanina sp. nov., a novel member of the genus with unusual morphology," *International Journal of Systematic and Evolutionary Microbiology*, vol. 54, no. 1, pp. 235–239, doi: 10.1099/ijs.0.02586-0.
[32] F. Xu *et al.*, "A genetics-free method for high-throughput discovery of cryptic microbial metabolites," *Nat Chem Biol*, vol. 15, no. 2, pp. 161–168, Feb. 2019, doi: 10.1038/s41589-018-0193-2.
[33] G. Yim *et al.*, "Harnessing the synthetic capabilities of glycopeptide antibiotic tailoring enzymes: characterization of the UK-68,597 biosynthetic cluster," *Chembiochem*, vol. 15, no. 17, pp. 2613–2623, Nov. 2014, doi: 10.1002/cbic.201402179.
[34] M. Sosio, H. Kloosterman, A. Bianchi, P. de Vreugd, L. Dijkhuizen, and S. Donadio, "Organization of the teicoplanin gene cluster in Actinoplanes teichomyceticus," *Microbiology (Reading)*, vol. 150, no. Pt 1, pp. 95–102, Jan. 2004, doi: 10.1099/mic.0.26507-0.
[35] O. Yushchuk *et al.*, "Genomic-Led Discovery of a Novel Glycopeptide Antibiotic by Nonomuraea coxensis DSM 45129," *ACS Chemical Biology*, vol. 16, no. 5, pp. 915–928, May 2021, doi: 10.1021/acschembio.1c00170.
[36] M. Sosio, S. Stinchi, F. Beltrametti, A. Lazzarini, and S. Donadio, "The gene cluster for the biosynthesis of the glycopeptide antibiotic A40926 by nonomuraea species," *Chem Biol*, vol. 10, no. 6, pp. 541–549, Jun. 2003, doi: 10.1016/s1074-5521(03)00120-0.
[37] M. R. Bardone, M. Paternoster, and C. Coronelli, "Teichomycins, new antibiotics from Actinoplanes teichomyceticus nov. sp. II. Extraction and chemical characterization," *J Antibiot (Tokyo)*, vol. 31, no. 3, pp. 170–177, Mar. 1978, doi: 10.7164/antibiotics.31.170.
[38] H. T. Chiu *et al.*, "Molecular cloning and sequence analysis of the complestatin biosynthetic gene cluster," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 15, pp. 8548–8553, Jul. 2001, doi: 10.1073/PNAS.151246498.
[39] M. N. Thaker *et al.*, "Identifying producers of antibacterial compounds by screening for antibiotic resistance," *Nature Biotechnology*, vol. 31, no. 10, pp. 922–927, Sep. 2013, doi: 10.1038/nbt.2685.
[40] M. J. Kwun and H.-J. Hong, "Genome Sequence of Streptomyces toyocaensis NRRL 15009, Producer of the Glycopeptide Antibiotic A47934," *Genome Announc*, vol. 2, no. 4, pp. e00749-14, Jul. 2014, doi: 10.1128/genomeA.00749-14.
[41] J. J. Banik, J. W. Craig, P. Y. Calle, and S. F. Brady, "Tailoring Enzyme-Rich Environmental DNA Clones: A Source of Enzymes for Generating Libraries of Unnatural Natural Products," *J. Am. Chem. Soc.*, vol. 132, no. 44, pp. 15661–15670, Nov. 2010, doi: 10.1021/ja105825a.
[42] 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*, vol. 110, no. 29, pp. 11797–11802, Jul. 2013, doi: 10.1073/pnas.1222159110.
[43] B. Nazari, C. C. Forneris, M. I. Gibson, K. Moon, K. R. Schramma, and M. R. Seyedsayamdost, "Nonomuraea sp. ATCC 55076 harbours the largest actinomycete chromosome to date and the kistamicin biosynthetic gene cluster," *Medchemcomm*, vol. 8, no. 4, pp. 780–788, Apr. 2017, doi: 10.1039/c6md00637j.
[44] M. J. Kwun, J. Cheng, S. H. Yang, D.-R. Lee, J.-W. Suh, and H.-J. Hong, "Draft Genome Sequence of Ristocetin-Producing Strain Amycolatopsis sp. Strain MJM2582 Isolated in South Korea," *Genome Announc*, vol. 2, no. 5, pp. e01091-14, Oct. 2014, doi: 10.1128/genomeA.01091-14.
[45] J. J. Banik and S. F. Brady, "Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary," *Proceedings of the National Academy of Sciences*, vol. 105, no. 45, pp. 17273–17277, Nov. 2008, doi: 10.1073/pnas.0807564105.
[46] K. Blin *et al.*, "antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation," *Nucleic Acids Research*, vol. 51, no. W1, pp. W46–W50, Jul. 2023, doi: 10.1093/nar/gkad344.
[47] J. C. Navarro-muñoz *et al.*, "A computational framework to explore large-scale biosynthetic diversity," *Nature Chemical Biology*, vol. 16, no. 1, pp. 60–68, 2020, doi: 10.1038/s41589-019-0400-9.A.
[48] C. L. M. Gilchrist and Y.-H. Chooi, "clinker & clustermap.js: Automatic generation of gene cluster comparison figures," *bioRxiv*, p. 2020.11.08.370650, 2020.

[49] P. Shannon *et al.*, “Cytoscape: a software environment for integrated models of biomolecular interaction networks,” *Genome Res*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.
[50] D. Bajusz, A. Rácz, and K. Héberger, “Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?,” *Journal of Cheminformatics*, vol. 7, no. 1, p. 20, May 2015, doi: 10.1186/s13321-015-0069-3.
[51] R. Salamzade *et al.*, “zol & fai: large-scale targeted detection and evolutionary investigation of gene clusters.” bioRxiv, p. 2023.06.07.544063, Jun. 10, 2023. doi: 10.1101/2023.06.07.544063.
[52] M. G. Chevrette, A. Gavrilidou, S. Mantri, N. Selem-Mojica, N. Ziemert, and F. Barona-G ´ Omez, “The confluence of big data and evolutionary genome mining for the discovery of natural products,” 2021, doi: 10.1039/d1np00013f.
[53] T. Stachelhaus, H. D. Mootz, and M. A. Marahiel, “The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases,” *Chem Biol*, vol. 6, no. 8, pp. 493–505, Aug. 1999, doi: 10.1016/S1074-5521(99)80082-9.
[54] M. H. Hansen *et al.*, “Resurrecting ancestral antibiotics: unveiling the origins of modern lipid II targeting glycopeptides,” *Nat Commun*, vol. 14, no. 1, p. 7842, Nov. 2023, doi: 10.1038/s41467-023-43451-4.
[55] K. Blin *et al.*, “antiSMASH 6.0: improving cluster detection and comparison capabilities,” *Nucleic Acids Res*, vol. 49, no. W1, pp. W29–W35, Jul. 2021, doi: 10.1093/nar/gkab335.
[56] A. K. Chavali and S. Y. Rhee, “Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites,” *Brief. Bioinform.*, vol. 19, no. 5, pp. 1022–1034, 2018, doi: 10.1093/bib/bbx020.
[57] N. Waglechner, A. G. McArthur, and G. D. Wright, “Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance,” *Nature Microbiology*, vol. 4, no. 11, pp. 1862–1871, 2019, doi: 10.1038/s41564-019-0531-5.
[58] M. Baunach, S. Chowdhury, P. Stallforth, and E. Dittmann, “The Landscape of Recombination Events That Create Nonribosomal Peptide Diversity,” *Molecular Biology and Evolution*, Jan. 2021, doi: 10.1093/molbev/msab015.
[59] M. H. Medema, P. Cimermancic, A. Sali, E. Takano, and M. A. Fischbach, “A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis,” *PLoS Computational Biology*, vol. 10, no. 12, Dec. 2014, doi: 10.1371/journal.pcbi.1004016.
[60] H. Jenke-Kodama and E. Dittmann, “Bioinformatic perspectives on NRPS/PKS megasynthases: Advances and challenges,” *Natural Product Reports*, vol. 26, no. 7, pp. 874–883, Jun. 2009, doi: 10.1039/b810283j.
[61] N. Ziemert, S. Podell, K. Penn, J. H. Badger, E. Allen, and P. R. Jensen, “The natural product domain seeker NaPDoS: A phylogeny based bioinformatic tool to classify secondary metabolite gene diversity,” *PLoS ONE*, vol. 7, no. 3, Mar. 2012, doi: 10.1371/journal.pone.0034064.
[62] M. Adamek *et al.*, “Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species,” *BMC Genomics*, vol. 19, no. 1, Jun. 2018, doi: 10.1186/s12864-018-4809-4.
[63] D. H. Huson, R. Rupp, and C. Scornavacca, Eds., “Phylogenetic networks from trees,” in *Phylogenetic Networks: Concepts, Algorithms and Applications*, Cambridge: Cambridge University Press, 2010, pp. 265–299. doi: 10.1017/CBO9780511974076.015.
[64] R. A. Barco, G. M. Garrity, J. J. Scott, J. P. Amend, K. H. Nealson, and D. Emerson, “A genus definition for bacteria and archaea based on a standard genome relatedness index,” *mBio*, vol. 11, no. 1, pp. 1–20, 2020, doi: 10.1128/MBIO.02475-19.
[65] H. Seto, T. Fujioka, K. Furihata, I. Kaneko, and S. Takahashi, “Structure of complestatin, a very strong inhibitor of protease activity of complement in the human complement system,” *Tetrahedron Letters*, vol. 30, no. 37, pp. 4987–4990, Jan. 1989, doi: 10.1016/S0040-4039(01)80562-1.
[66] N. Naruse *et al.*, “New antiviral antibiotics, kistamicins A and B. I. Taxonomy, production, isolation, physico-chemical properties and biological activities,” *J Antibiot (Tokyo)*, vol. 46, no. 12, pp. 1804–1811, Dec. 1993, doi: 10.7164/antibiotics.46.1804.
[67] N. Naruse, M. Oka, M. Konishi, and T. Oki, “New antiviral antibiotics, kistamicins A and B. II. Structure determination,” *J Antibiot (Tokyo)*, vol. 46, no. 12, pp. 1812–1818, Dec. 1993, doi: 10.7164/antibiotics.46.1812.
[68] B. R. Terlouw *et al.*, “MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters,” *Nucleic Acids Research*, vol. 51, no. D1, pp. D603–D610, Jan.

2023, doi: 10.1093/nar/gkac1049.
[69] T. Tatusova *et al.*, “NCBI prokaryotic genome annotation pipeline,” *Nucleic Acids Res*, vol. 44, no. 14, pp. 6614–6624, Aug. 2016, doi: 10.1093/nar/gkw569.
[70] D. H. Haft *et al.*, “RefSeq: An update on prokaryotic genome annotation and curation,” *Nucleic Acids Research*, vol. 46, no. D1, pp. D851–D860, 2018, doi: 10.1093/nar/gkx1068.
[71] K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, “GenBank,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D67–D72, Jan. 2016, doi: 10.1093/nar/gkv1276.
[72] I.-M. A. Chen *et al.*, “IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes,” *Nucleic Acids Res*, vol. 47, no. Database issue, pp. D666–D677, Jan. 2019, doi: 10.1093/nar/gky901.
[73] L. Richardson *et al.*, “MGnify: the microbiome sequence data analysis resource in 2023,” *Nucleic Acids Research*, vol. 51, no. D1, pp. D753–D759, Jan. 2023, doi: 10.1093/nar/gkac1080.
[74] A. M. Sharrar, A. Crits-Christoph, R. Méheust, S. Diamond, E. P. Starr, and J. F. Banfield, “Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type,” *mBio*, vol. 11, no. 3, pp. e00416-20, May 2020, doi: 10.1128/mBio.00416-20.
[75] “Microbial communities across a hillslope‐riparian transect shaped by proximity to the stream, groundwater table, and weathered bedrock - Lavy - 2019 - Ecology and Evolution - Wiley Online Library.” Accessed: Jan. 31, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/ece3.5254
[76] S. Nayfach *et al.*, “Author Correction: A genomic catalog of Earth’s microbiomes,” *Nat. Biotechnol.*, vol. 39, no. 4, p. 521, 2021, doi: 10.1038/s41587-021-00898-4.
[77] K. Haslinger, M. Peschke, C. Brieke, E. Maximowitsch, and M. J. Cryle, “X-domain of peptide synthetases recruits oxygenases crucial for glycopeptide biosynthesis,” *Nature*, vol. 521, no. 7550, pp. 105–109, 2015, doi: 10.1038/nature14141.
[78] O. Schwengers, L. Jelonek, M. A. Dieckmann, S. Beyvers, J. Blom, and A. Goesmann, “Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification,” *Microbial Genomics*, vol. 7, no. 11, Nov. 2021, doi: 10.1099/mgen.0.000685.
[79] T. Kieser, M. J. Bibb, M. J. Buttner, K. F. Chater, D. A. Hopwood, and others, *Practical streptomyces genetics*, vol. 291. John Innes Foundation Norwich, 2000.
[80] “Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads | PLOS Computational Biology.” Accessed: Jan. 31, 2023. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005595
[81] S. L. Heald, L. Mueller, and P. W. Jeffs, “Actinoidins A and A2: structure determination using 2D NMR methods,” *J Antibiot (Tokyo)*, vol. 40, no. 5, pp. 630–645, May 1987, doi: 10.7164/antibiotics.40.630.
[82] L. Vértesy, H. W. Fehlhaber, H. Kogler, and M. Limbert, “New 4-oxovancosamine-containing glycopeptide antibiotics from Amycolatopsis sp. Y-86,21022,” *J Antibiot (Tokyo)*, vol. 49, no. 1, pp. 115–118, Jan. 1996, doi: 10.7164/antibiotics.49.115.
[83] G. M. Sheldrick, P. G. Jones, O. Kennard, D. H. Williams, and G. A. Smith, “Structure of vancomycin and its complex with acetyl-D-alanyl-D-alanine,” *Nature*, vol. 271, no. 5642, pp. 223–225, Jan. 1978, doi: 10.1038/271223a0.
[84] S. B. Christensen *et al.*, “Parvodicin, a novel glycopeptide from a new species, Actinomadura parvosata: discovery, taxonomy, activity and structure elucidation,” *J Antibiot (Tokyo)*, vol. 40, no. 7, pp. 970–990, Jul. 1987, doi: 10.7164/antibiotics.40.970.
[85] F. Sztaricskai, C. M. Harris, A. Neszmelyi, and T. M. Harris, “Structural studies of ristocetin A (ristomycin A): carbohydrate-aglycone linkages,” *J. Am. Chem. Soc.*, vol. 102, no. 23, pp. 7093–7099, Nov. 1980, doi: 10.1021/ja00543a035.
[86] M. J. Zmijewski, B. Briggs, R. Logan, and L. D. Boeck, “Biosynthetic studies on antibiotic A47934,” *Antimicrob Agents Chemother*, vol. 31, no. 10, pp. 1497–1501, Oct. 1987, doi: 10.1128/AAC.31.10.1497.
[87] F. Parenti, “Structure and mechanism of action of teicoplanin,” *Journal of Hospital Infection*, vol. 7, pp. 79–83, Mar. 1986, doi: 10.1016/0195-6701(86)90011-3.
[88] S. B. Singh *et al.*, “The complestatins as HIV-1 integrase inhibitors. Efficient isolation, structure elucidation, and inhibitory activities of isocomplestatin, chloropeptin I, new complestatins, A and B, and acid-hydrolysis products of chloropeptin I,” *J Nat Prod*, vol. 64, no. 7, pp. 874–882, Jul. 2001, doi: 10.1021/np000632z.
[89] W. J. McGahren *et al.*, “Structure of avoparcin components,” *J. Am. Chem. Soc.*, vol. 102, no. 5, pp. 1671–1684, Feb. 1980, doi: 10.1021/ja00525a036.
[90] M. L. Sanchez, R. P. Wenzel, and R. N. Jones, “In vitro activity of decaplanin (M86-1410), a new

glycopeptide antibiotic," *Antimicrob Agents Chemother*, vol. 36, no. 4, pp. 873–875, Apr. 1992, doi: 10.1128/AAC.36.4.873.
[91] M. Röttig, M. H. Medema, K. Blin, T. Weber, C. Rausch, and O. Kohlbacher, "NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity," *Nucleic Acids Res*, vol. 39, no. Web Server issue, pp. W362-367, Jul. 2011, doi: 10.1093/nar/gkr323.
[92] K. Blin *et al.*, "AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline," *Nucleic Acids Research*, vol. 47, no. W1, pp. W81–W87, 2019, doi: 10.1093/nar/gkz310.
[93] S. F. Altschul *et al.*, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.," *Nucleic Acids Res*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.
[94] D. Megrian, N. Taib, A. L. Jaffe, J. F. Banfield, and S. Gribaldo, "Ancient origin and constrained evolution of the division and cell wall gene cluster in Bacteria," *Nat Microbiol*, pp. 1–14, Nov. 2022, doi: 10.1038/s41564-022-01257-y.
[95] K. Katoh and D. M. Standley, "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability," *Mol Biol Evol*, vol. 30, no. 4, pp. 772–780, Apr. 2013, doi: 10.1093/molbev/mst010.
[96] B. Q. Minh *et al.*, "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era," *Molecular Biology and Evolution*, vol. 37, no. 5, pp. 1530–1534, May 2020, doi: 10.1093/molbev/msaa015.
[97] S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin, "ModelFinder: fast model selection for accurate phylogenetic estimates," *Nat Methods*, vol. 14, no. 6, Art. no. 6, Jun. 2017, doi: 10.1038/nmeth.4285.
[98] D. H. Huson, T. Dezulian, T. Klopper, and M. A. Steel, "Phylogenetic super-networks from partial trees," *IEEE/ACM Trans. Comput. Biol. and Bioinf.*, vol. 1, no. 4, pp. 151–158, Oct. 2004, doi: 10.1109/TCBB.2004.44.
[99] D. H. Huson, "SplitsTree: analyzing and visualizing evolutionary data.," *Bioinformatics*, vol. 14, no. 1, pp. 68–73, Jan. 1998, doi: 10.1093/bioinformatics/14.1.68.
[100] D. H. Huson and D. Bryant, "Application of Phylogenetic Networks in Evolutionary Studies," *Molecular Biology and Evolution*, vol. 23, no. 2, pp. 254–267, Feb. 2006, doi: 10.1093/molbev/msj030.
[101] D. H. Parks *et al.*, "Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life," *Nat Microbiol*, vol. 2, no. 11, pp. 1533–1542, 2017, doi: 10.1038/s41564-017-0012-7.
[102] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, "trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses," *Bioinformatics*, vol. 25, no. 15, pp. 1972–1973, Aug. 2009, doi: 10.1093/bioinformatics/btp348.
[103] P. J. A. Cock *et al.*, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, Jun. 2009, doi: 10.1093/bioinformatics/btp163.
[104] "FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix | Molecular Biology and Evolution | Oxford Academic." Accessed: Nov. 17, 2022. [Online]. Available: https://academic.oup.com/mbe/article/26/7/1641/1128976
[105] F. D. K. Tria, G. Landan, and T. Dagan, "Phylogenetic rooting using minimal ancestor deviation," *Nat Ecol Evol*, vol. 1, no. 7, Art. no. 7, Jun. 2017, doi: 10.1038/s41559-017-0193.
[106] I. Letunic and P. Bork, "Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation," *Nucleic Acids Res.*, vol. 49, no. W1, pp. W293–W296, 2021, doi: 10.1093/nar/gkab301.

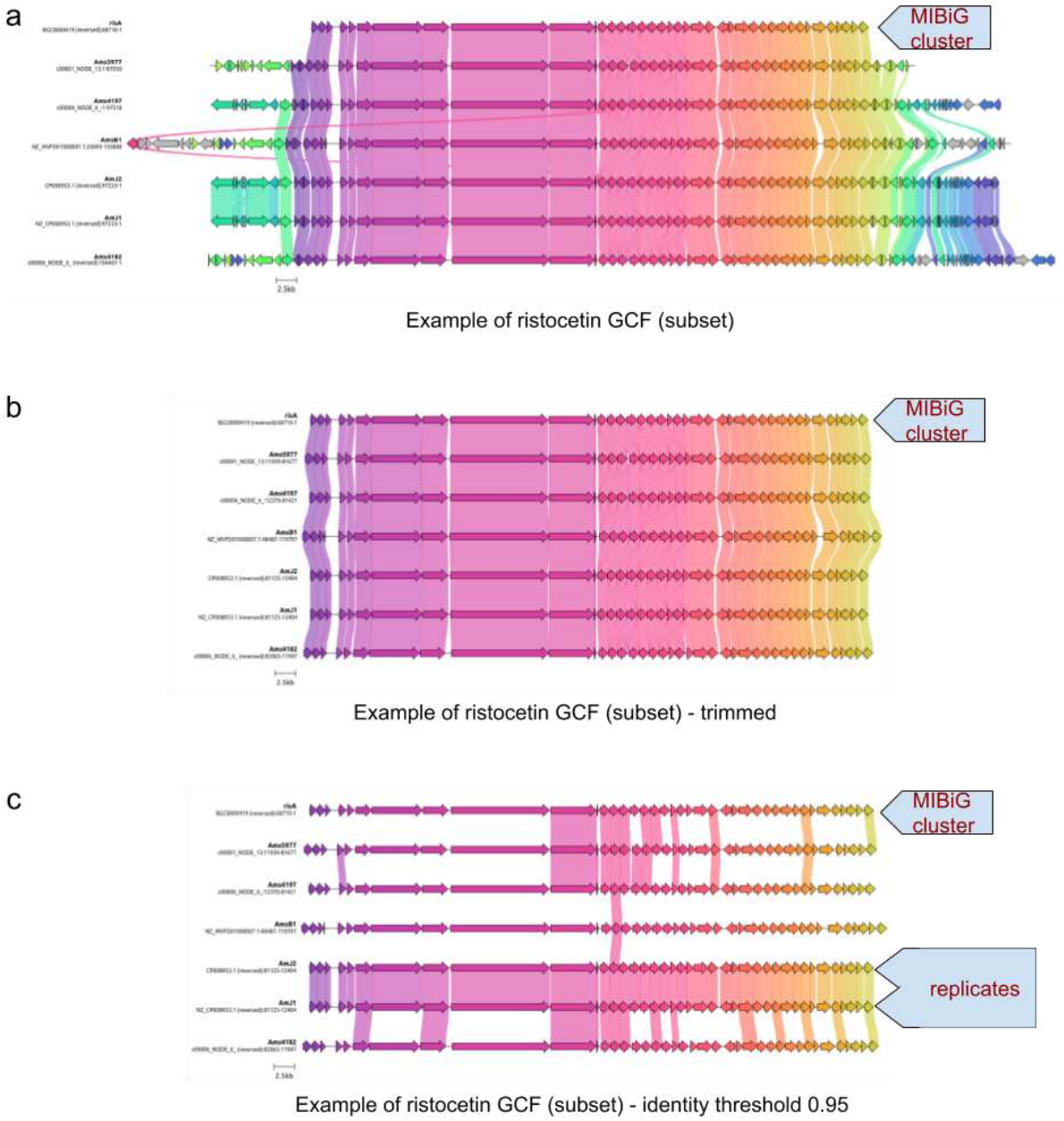
## Supplementary Material

All supplementary material of unpublished projects are available for download (upon request) from a zenodo repository: https://doi.org/10.5281/zenodo.10879735

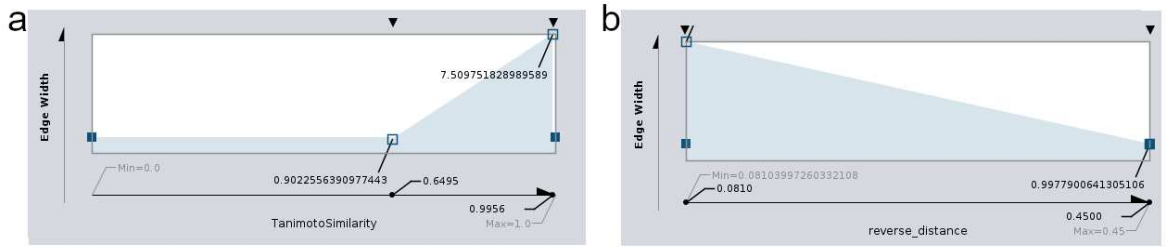
Supplementary Table 1: SMILES of selected known GPA structures. Also, pairwise tanimoto similarity values.
Supplementary Table 2: extended Table 1 including genes and domains present in GRPs, as well as references.
Supplementary Table 3: dataset of GPA BGCs and related metadata.
Supplementary Table 4: results of Stachelhaus code analysis of GPA BGC A-domains.

Supplementary Table 5: table of orthologous groups and gene presence/absence table from Supplementary Figure 3.
Supplementary Table 6: results of congruence analysis.
Supplementary Table 7: table of ML models used for the phylogenetic trees of genes and domains.
Supplementary Data 1: SplitsTree6 session file, containing all (154) calculated phylogenetic trees, as well as the super network of Figure 3.
Supplementary Data 2: fasta file of the concatenated MSA and newick files (unrooted with bootstrap values and MAD rooted) containing the concatenated phylogeny of true GPAs/GRPs .
Supplementary Data 3: Stachelhaus code-based concatenated MSA.
Supplementary Data 4: tables of accession numbers of NCBI RefSeq and Genbank entries included in the X-domain search and the result of the search.
Supplementary Data 5: custom scripts used for the analysis.
Supplementary Data 6: sequencing report of *S. varsoviesnsis*.
Supplementary Data 7: fasta files of all gene and domain MSAs (and trimmed MSAs) and newick files of the phylogenetic trees they were used to build.

## Supplementary Figures

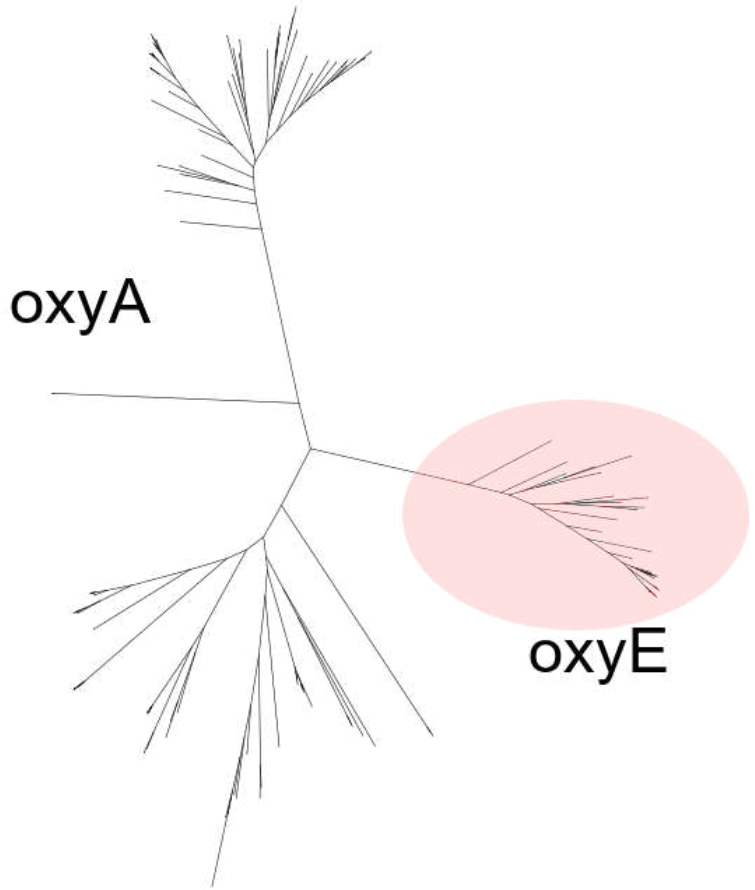



**Supplementary Figure 1:** Example of clinker-based trimming and dereplication. Panel a: subset of the ristocetin gene cluster family (GCF), visualised with clinker. Genes are coloured the same if they were placed in the same gene group based on similarity by clinker and are connected by coloured bands if their identity is higher than 0.5. The MIBIG cluster is the top one (marked) and the trimming will be based on it. Panel b: the same visualisation as panel a, after the trimming is finished. Now all BGCs belonging to this GCF are uniformly trimmed. Panel c:

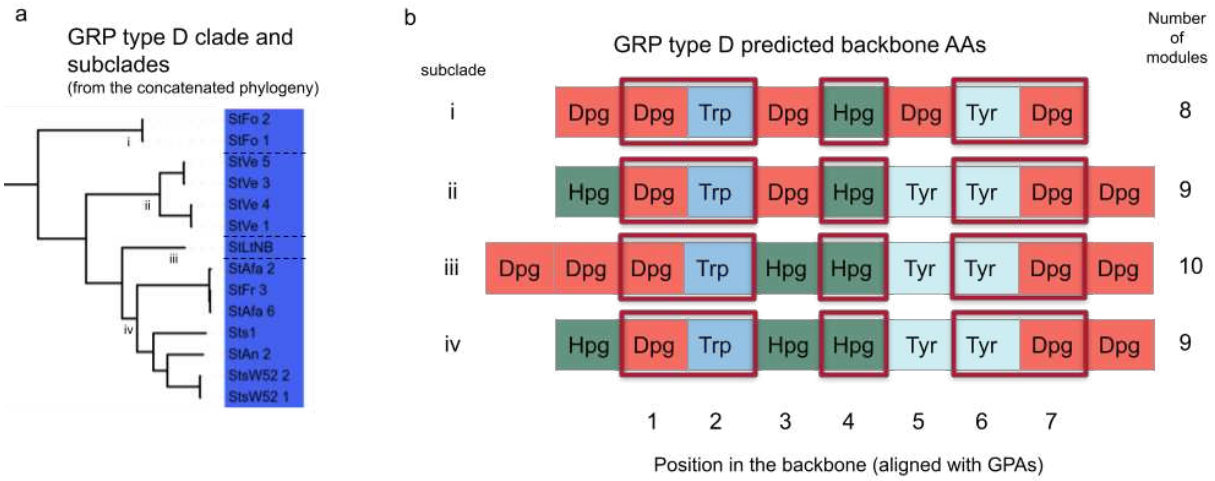



**Supplementary Figure 2:** Edge-width legends for Figure 2. Panel a: relation of edge width to Tanimoto Similarity metric in Figure 2a. Panel b: relation of edge width to DSS (labelled reverse_distance) in Figure 2b.

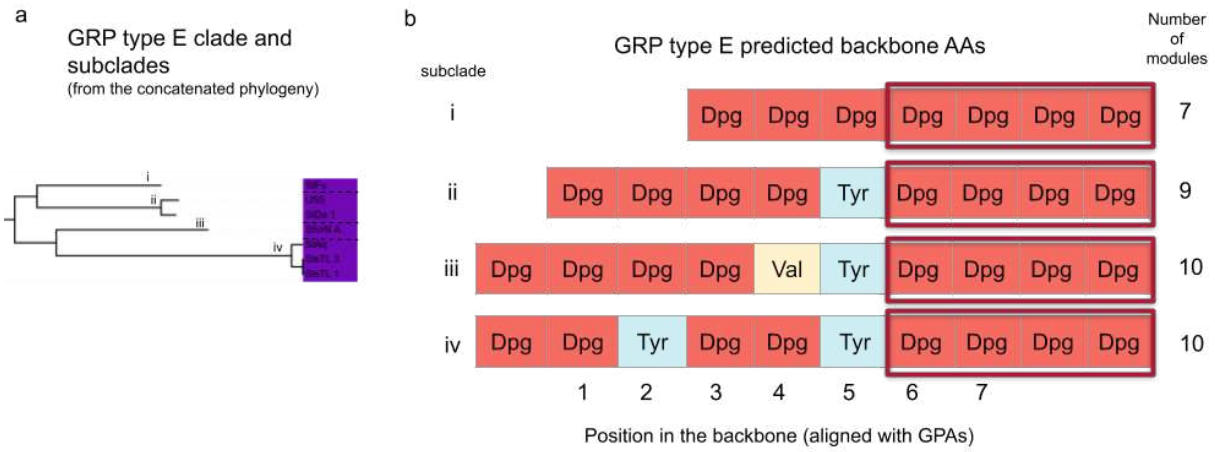
**Supplementary Figure 3:** High-resolution Figure 4 and Figure 5, provided as a PDF file.




**Supplementary Figure 4:** Phylogenetic tree (unrooted) of the mixed zol orthogroup OG3, which includes p450 monooxygenases oxyA and oxyE. The clades of the two genes can be separated after annotation of the corresponding enzymes with known function. The known oxyE genes all belong to the clade in the coloured circle and were removed from OG3 and assigned to the artificial group OG3e. Trees of both OG3 and OG3e can be explored in **Supplementary Data 1**.




**Supplementary Figure 5**: GRP type D subclades and backbone compositions. Panel a: clade of GRP type D, with marked subclades i-iv. BGC IDs are shown as leaf labels. Panel b: predicted backbone connected to its corresponding subclade. The red squares highlight the characteristic motif of the type.




**Supplementary Figure 6**: GRP type E subclades and backbone compositions. Panel a: clade of GRP type D, with marked subclades i-iv. BGC IDs are shown as leaf labels. Panel b: predicted backbone connected to its corresponding subclade.The red squares highlight the characteristic motif of the type.

# Chapter 5: BGC-aware gene coincidence analysis of the *Amycolatopsis* pangenome using the Goldfinder tool

*(Initial manuscript; Analysis in progress)*

Athina Gavriilidou[1,2], Franz Baumdicker[2,3,4], Nadine Ziemert[1,2,5]

1: Translational Genome Mining for Natural Products, Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), University of Tübingen, Tübingen, Germany
2: Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Tübingen, Germany
3: Cluster of Excellence "Controlling Microbes to Fight Infections", Mathematical and Computational Population Genetics, University of Tübingen, Germany
4: Cluster of Excellence "Machine Learning in Science", University of Tübingen, Germany
5: German Centre for Infection Research (DZIF), Partnersite Tübingen, Tübingen, Germany

Personal contributions:

Scientific ideas/data generation/analysis & interpretation/writing: 60/100/80/100%

## Abstract

Given the escalating emergence of antimicrobial resistance among pathogens, the ongoing pursuit of novel antibiotic candidates is imperative. Since the majority of known antibiotics are manufactured by bacteria, we need to understand which of them are capable of production and how they manage it. In this study, we investigated the adaptation mechanisms employed by bacteria of the Amycolatopsis genus in the production of glycopeptide antibiotics (GPAs), focusing on their regulatory frameworks and primary metabolic pathways. Through a comprehensive analysis of the genus' pangenome, we elucidated the distribution patterns of Biosynthetic Gene Clusters (BGCs) responsible for GPA biosynthesis. Additionally, we conducted gene coincidence analyses to discover patterns among known GPA producers and non-producers, using the Goldfinder tool. Our results represent some of the first identified adaptation mechanisms to acquisition of GPA encoding BGCs.

## Introduction

Bacteria are an established source of discovery of new bioactive compounds, which can include antibiotics[1]. This is especially relevant under the prism of the ever-growing threat of multi-drug resistant pathogens[2], which underlines the dire need for discovering new antibiotics. However, what is not yet explained is the taxonomic distribution of the bacterial producers. Namely, most taxa encode few biosynthetic gene clusters (BGCs) - the genetic elements connected to the biosynthesis of specialised metabolites - and others seem to be very rich in BGC quantity and diversity[1]. Depending on the bacterial taxon and the type of BGC, these dissimilarities can be connected to the evolutionary history of the gene clusters[3,4] and this knowledge can be helpful in the development of genome mining methods[5] for the detection of new BGCs - and new bioactive compounds, like antibiotics.

The horizontal transmission of gene clusters is a phenomenon that is relatively common under the prism of bacterial specialised metabolism and, in many cases, horizontal gene transfer (HGT) events can explain the distribution of a BGC among distant taxa[5]. However, it is not clear why these events happen in some taxa more than others. The acquisition of a BGC via HGT is an event that is assumed cause a great deal of changes in the host organism, as the microorganism would need to adapt to the additional metabolic stress that the expression of such a gene cluster is associated with, while keeping all vital cellular processes unaffected. These changes could be focused on the primary metabolism but it is also possible that other specialised metabolic pathways (other BGCs) are affected as well. Bacterial genomes can harbour multiple BGCs, related to various types of biosynthetic pathways[1], and it has not been established so far if their presence is independent to each other or if there is some association between them that is not obvious due to their genomic distance. For example, one possibility is that two specific BGCs are both dependent on the same primary metabolic pathway, making it energetically favourable for a bacterium to have both in its genome, while others may be mutually exclusive.

To study this kind of interplay and the putative genetic changes that a vertically transmitted BGC would bring about, both producer (encoding a certain BGC) and non-producer strains (not encoding this BGC, but perhaps others) need to be analysed using comparative genomic methods. Having phylogenetically closely related candidate genomes would make the comparison easier, as the level of conservation among other genomic areas should be high, enabling the detection of changes related to more recent HGT events, instead of earlier speciation events.

Glycopeptide antibiotics (GPAs) are last-resort antibiotics used in the clinics against gram-positive pathogenic bacteria[6,7]. GPA-encoding BGCs have been detected in the genomes of multiple genera, which is an indication that HGT events took place in their history[3]. This class of compounds were discovered a few decades ago and their biosynthesis has been studied extensively. There are only a few tens of structures known and connected to their bacterial producers as products of their secondary (or specialised) metabolism. Among the producers, the most prolific ones come from the *Streptomyces* and *Amycolatopsis* genera[4,8–14].

In the current manuscript, we apply genome mining methods to show that GPA-encoding BGCs are present in only some of the genomes of *Amycolatopsis* bacteria, constituting the genus a suitable candidate to further examine the adaptation mechanisms to HGT. A comparative analysis among predicted GPA producer genomes and non producer genomes was conducted through the new Goldfinder tool, which detects strongly coinciding or disassociating gene pairs in a pangenome. The *Amycolatopsis* genus' pangenome was used for a BGC-aware gene coincidence analysis using Goldfinder, with the purpose of identifying any significant relationships between genes involved in GPA biosynthesis and those involved in other secondary or even in primary metabolic pathways.

## Results

### Specialised metabolism encoded in the Amycolatopsis pangenome

All publicly available assembled genomes of the *Amycolatopsis* genus were downloaded from NCBI RefSeq. In order to investigate the biosynthetic capacity of the genus, the strains' biosynthetic gene clusters (BGCs) were detected using antiSMASH[15]. The BGCs responsible for glycopeptide antibiotics (GPAs) biosynthesis were identified by the characteristic X-domain[16] present in the last module of their core biosynthetic genes, encoding for non-ribosomal peptide synthetases (NRPS). Both known producers and several strains not reported in the literature had a hit, and their distribution in the genus was visualised in **Figure 1**.

Many clades including various species are not GPA producers, though the genetic capacity to biosynthesize GPAs seems conserved at a species level in *Amycolatopsis*. Apart from one clade rich in producers, BGCs encoding the biosynthesis of GPAs (GPA BGCs) are found in distant species in the genus, phylogenetically interspersed by non-producers. These observations underline the highly likely horizontal gene transfer (HGT) events that have taken place in the evolutionary history of these BGCs, confirming the suitability of the *Amycolatopsis* pangenome as a candidate for comparative analysis.

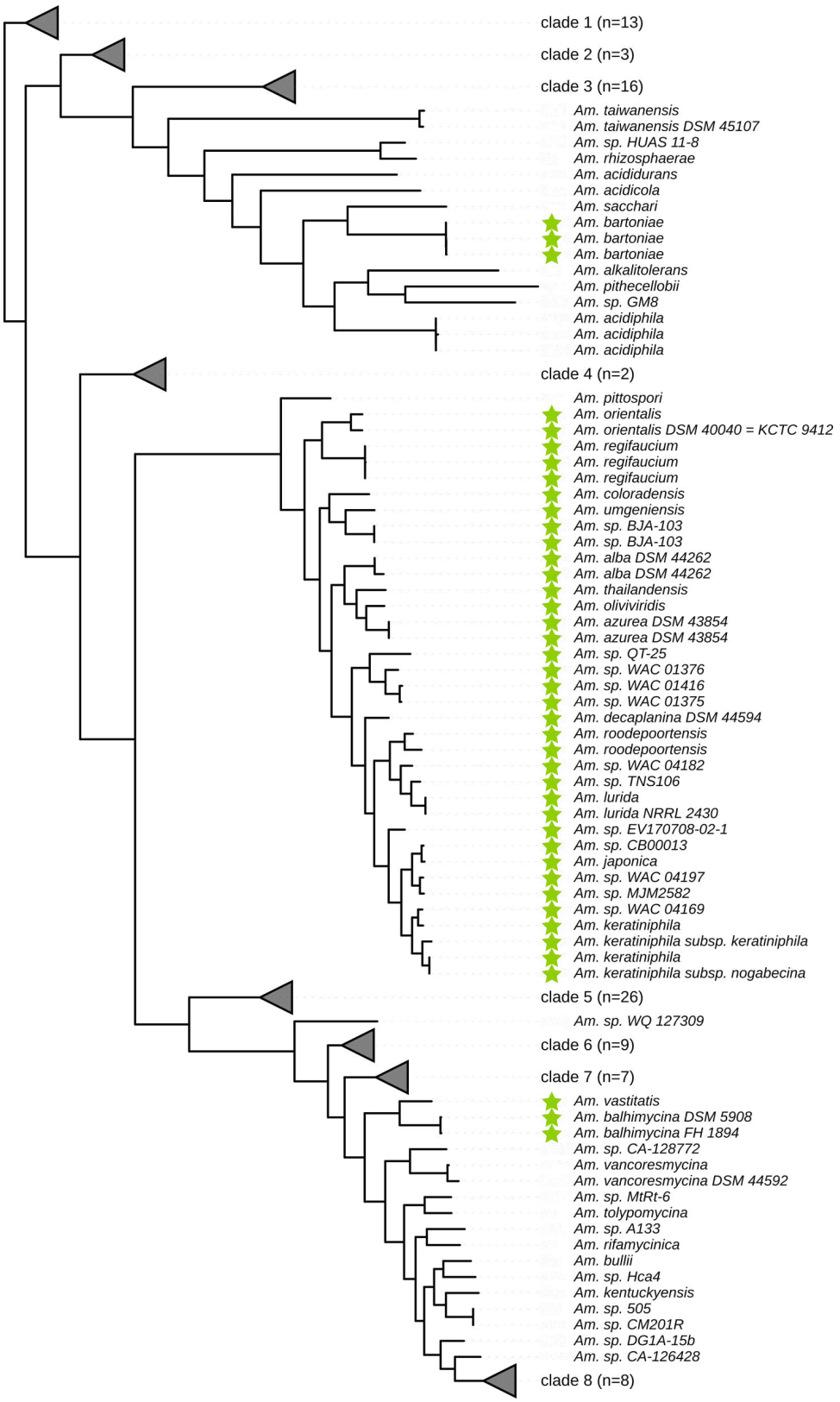



**Figure 1: Phylogenetic distribution of GPA-encoding BGCs in *Amycolatopsis* genomes**. Species tree of the *Amycolatopsis* genus was inferred by Orthofinder[17] and visualised with iTOL[18]. The tree was decorated with information on the presence of GPA-encoding BGCs (green stars). The collapsed clades include no putative GPA producers and their leaves (n) can be seen in **Supplementary Figures 1-4**.

.

Apart from the GPA-encoding BGCs, a large variety of other types were detected by antiSMASH (5,116 BGCs in total). To measure their diversity, BiG-SCAPE[19] was applied to cluster them based on similarity into gene cluster families (GCFs) and clans. The result was 1,793 GCFs (1,096 singletons) and 96 clans, which can be visualised in a similarity network (**Figure 2**). Knowing which categories of BGCs, using different definitions (BGC class, GCF, clan) is important to this analysis, as it is not known at which level of organisation possible associating or dissociating relationships may exist. For that reason, this information was included in the coincidence analysis that followed.

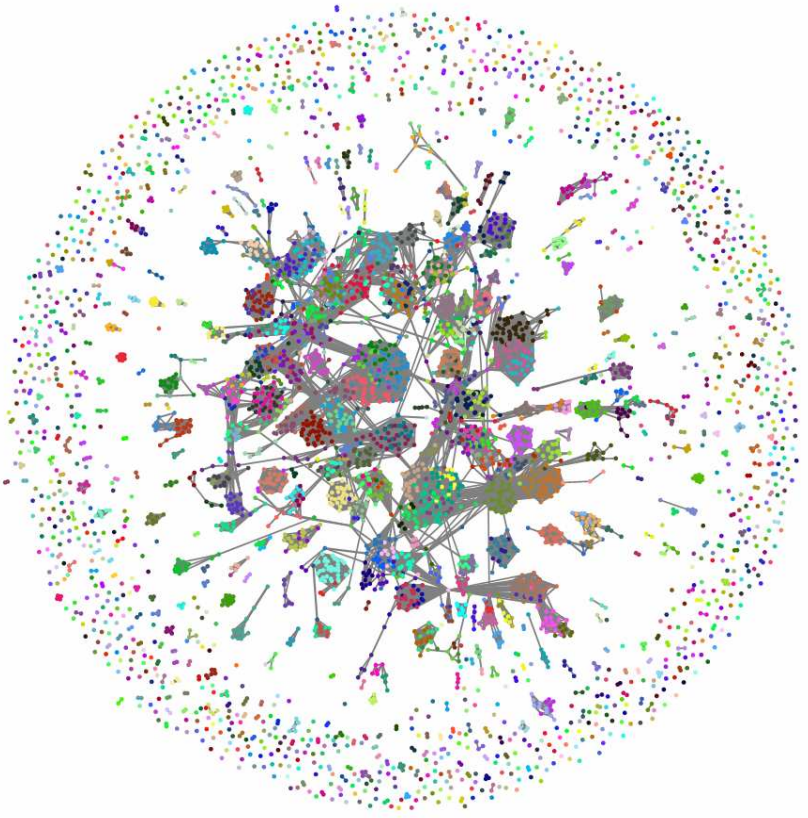


**Figure 2: Diversity of BGCs encoded in *Amycolatopsis* genomes**. Similarity network of the 5,116 BGCs detected in *Amycolatopsis* genomes as generated by BiG-SCAPE. Every node represents a BGC and similar BGCs are connected with an edge. BGCs belonging to the same gene cluster family (GCF) have the same colour (n=1,793). Connected elements (subnetworks) represent clans (n=96).

### BGC-aware gene coincidence analysis using Goldfinder

It is not known how bacteria adapt to acquiring a new BGC. Therefore the magnitude of this effect is not clear, meaning it could be single genes whose encoded products are involved in the primary metabolism, genes only present in BGCs, or groups of genes that are affected. In order to cover multiple hypotheses, three datasets were used in the search for significant patterns (**Table 1**).

**Table 1: Datasets of pangenomes used for Goldfinder analysis**. Description of the different methods and data that were involved in the generation of each dataset.

| Dataset name | Method | Input data |
|---|---|---|
| pangenome | orthofinder OGs | Full proteomes |
| BGC | zol OGs | Protein sequences of enzymes involved in specialised metabolism (their genes were part of a BGC) |
| BiG-SCAPE | pseudo-OGs | BiG-SCAPE based P/A table of GCFs |

To perform a gene coincidence analysis, it was necessary to create a gene presence/absence (P/A) table. This data was firstly generated with an orthology inference tool, Orthofinder[17], which clusters the genes into orthologous groups (OGs) based on their translated protein sequences (pangenome dataset).

Additionally, to incorporate the possibility that certain genes involved in specialised metabolism are related to cell adaptation mechanisms, though their homologues from the primary metabolism may not be, a BGC-gene-focused dataset was created. In this instance, orthology inference was conducted with zol[20], which is designed to create OGs from genes involved in gene clusters, and the resulting P/A table was also considered in the search for significant gene-gene relationships in the *Amycolatopsis* pangenome (BGC dataset).

The last approach was based on the BiG-SCAPE result described earlier, to cover the chance that whole BGCs are associated, even if common genes involved in specialised metabolism do not display such patterns. The focus was on the GPA BGCs, which is why an additional analysis was conducted to ensure their accurate placement into GCFs (see methods). The BGC to GCF assignments generated by BiG-SCAPE were transformed into GCF presence/absence patterns for each genome. The resulting P/A table was used as input for the gene coincidence analysis (BiG-SCAPE dataset).

The gene coincidence analysis was performed with the new Goldfinder tool (see methods), which detects associations or avoidance relationships between pairs of genes (OGs or pseudo-OGs). Goldfinder incorporates phylogenetic information to filter out associations due to evolutionary proximity of the host genomes. This increases the likelihood that, when a relationship between two genes is deemed significant by the algorithm, it is because they are interacting in some way and not because they were both vertically inherited from a common ancestor.

Out of the so far considered three approaches described in **Table 1**, significant results were found only in the third one, which was based on the BiG-SCAPE analysis. We expect to find more interesting gene co-occurrences in more sophisticated generated P/A data in the future, as argued in the Discussion section. Goldfinder showed that in the BiG-SCAPE dataset, 106 out of the 1,793 GCFs were

involved in associating relationships (**Figure 3**). Among the 283 pairwise associations, 8 subnetworks were formed. 5 subnetworks comprised pairs of GCFs, 1 involved 6 GCFs, while the last two included tens of GCFs and were much more complicated.

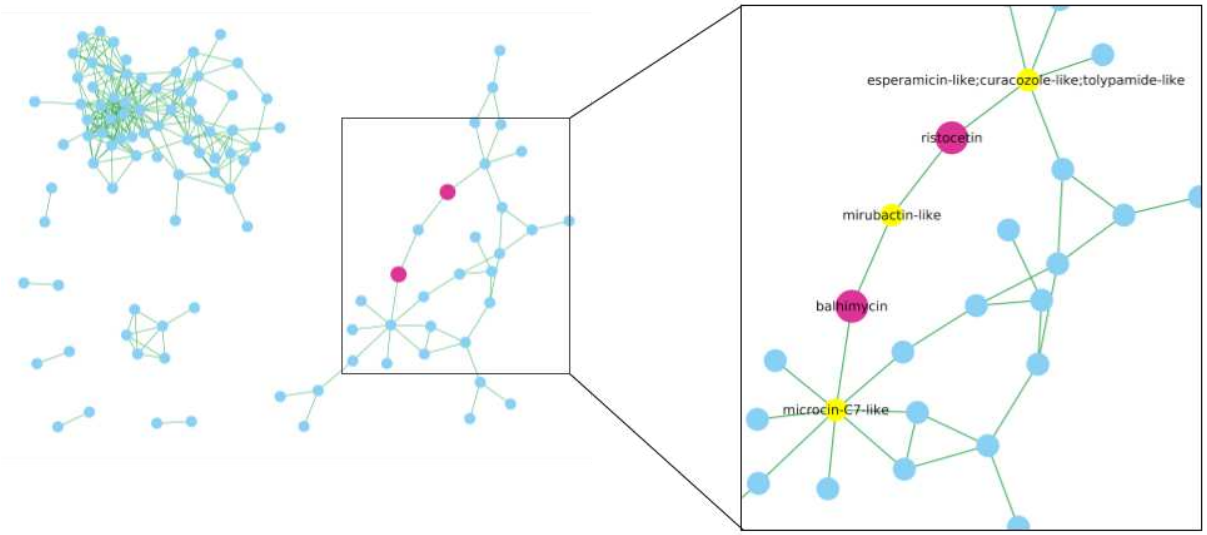



**Figure 3: Network of coinciding pseudo-OGs**. The nodes (n=106) represent pseudo-OGs (GCFs) and coinciding pairs calculated by Goldfinder are connected with an edge (n=283). The GCFs which include the GPA BGCs are larger and coloured pink (balhimycin and ristocetin GCFs). The interesting region of the network is marked with a square and on the right it is shown in higher resolution. The yellow nodes represent the GCFs associated with GPAs and they are labelled by their probable products (based on MIBiG[21] similarity as reported by antiSMASH[15]).

In this dataset, Goldfinder detected three significant relationships involving GCFs formed from GPA BGCs, one including the balhimycin-associated BGCs and the other the ristocetin ones (**Figure 3**). The two GPA GCFs had one shared coinciding GCF and each had an additional associated GCF that they did not share. The latter included, for the case of balhimycin, an RRE-containing RiPP GCF, whose BGCs showed moderate similarity with MIBiG[21] BGCs associated with the production of microcin C7[22] and quinolobactin[23]. The ristocetin GCF was co-occurring with a GCF of hybrid BGCs, which include regions dedicated to thiopeptide[24], type I polyketide synthases (PKS)[25] and Linear azol(in)e-containing peptides (LAP)[26] biosynthetic pathways. The most closely related BGCs from MIBiG were responsible for the biosynthesis of esperamicin[27], curacozole[28] and tolypamide[29], though their similarity was low.

Finally, the GCF which coincided both with balhimycin and with ristocetin belongs in the NRP-metallophore class, showing very high similarity to the BGC of the siderophore antibiotic mirubactin[30]. At first glance, there does not seem to be a connection between the GPAs and the mirubactin GCF. However, this compound, or rather its degraded product mirubactin C, was recently proven to have a unique property: it can make a bacterium resistant to mutations of genes involved in cell wall biosynthesis[31,32]. This constitutes its coincidence with the glycopeptide GCFs very

interesting, considering the resistance mechanism of balhimycin and ristocetin, which involves altering the tail of lipid II, a cell wall precursor molecule[33].

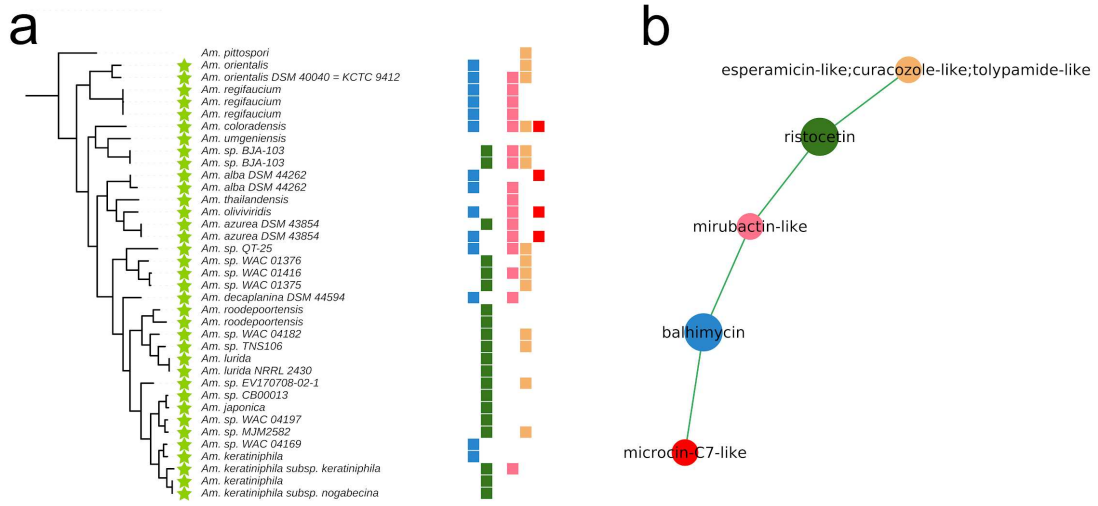



**Figure 4: GCF presence in the GPA-producer-rich clade**. Panel a: the clade including almost exclusively GPA producers from the tree in Figure 1 is shown and next to it coloured rectangles represent the presence absence pattern of the GCFs of either GPAs or their associated GCFs. Panel b: the subnetwork of Goldfinder showing these relationships is drawn on the right. The colours of the nodes in panel b match the colours of the presence/absence rectangles next to the tree in panel a.

Finally, it is worth noting that the observations listed above were relevant only for a single clade in the *Amycolatopsis* species tree (**Figure 4**), as none of these GCFs were present in any other genomes, with the exception of balhimycin which was found in two more unrelated strains (from the *Am. balhimycina* species), that did not encode any of the associated GCFs.

**Distribution of mirubactin-related BGCs in GPA producer genomes**

Having identified a potentially interesting association between some GPA BGCs and mirubactin BGCs in *Amycolatopsis* genomes, we went on to investigate how common the presence of the latter is among all GPA producers. For that purpose, a verified dataset of GPA producers (see methods) was screened for BGCs with similarity to mirubactin BGCs. Indeed, putative mirubactin producers were not uncommon among GPA producers, though they were not the norm (**Figure 5a**).

The GPA BGCs had GCF and type designations (see methods), which were used to determine if the presence of mirubactin BGCs was type-specific, which was not the case (**Figure 5b**). The presence of mirubactin BGCs in the genomes of GCFs that had a hit was quite conserved among the GCF members (**Figure 5c**).

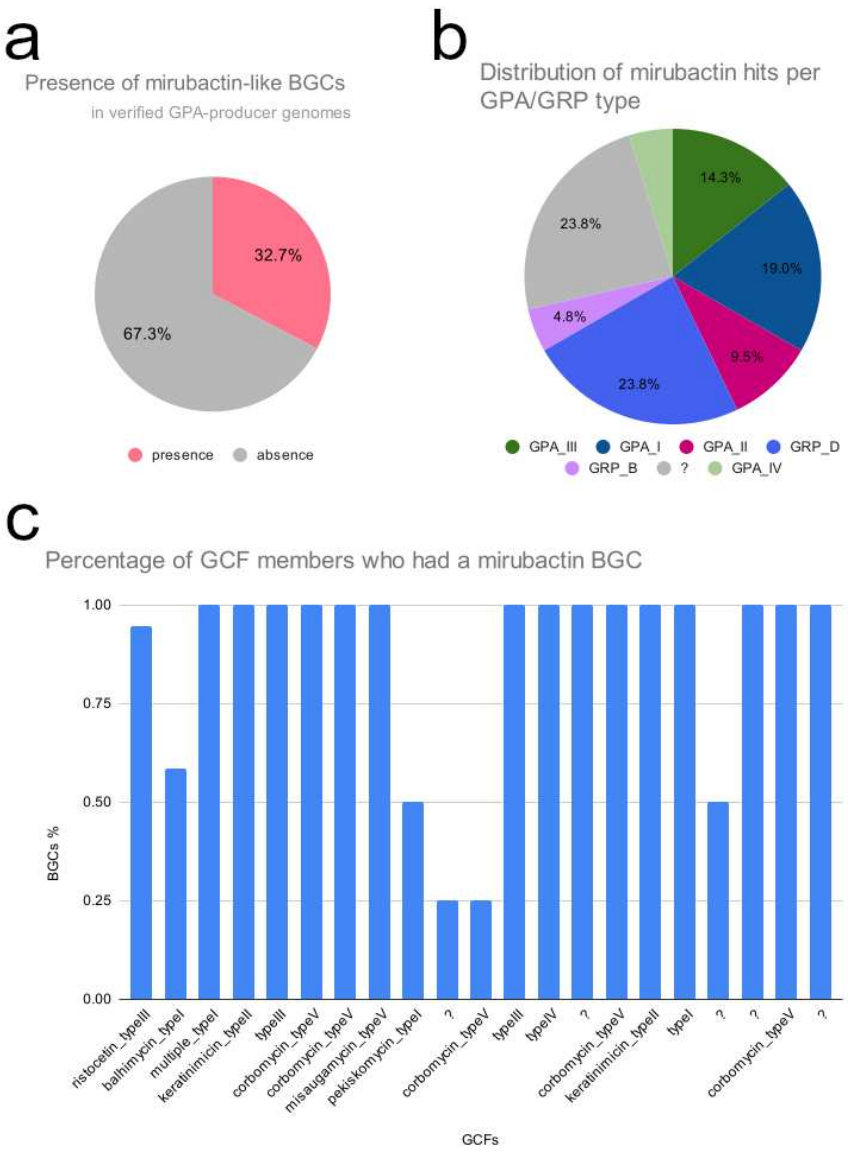



**Figure 5: Distribution of mirubactin-related BGCs in GPA producer genomes.** Panel a: Pie chart depicting the percentage of verified GPA-producer genomes (211 verified out of 291 candidates from chapter 4) which also contained a BGC with similarity to the one encoding for mirubactin biosynthesis (mirubactin BGCs) in MIBiG[21]. If the unverified GPA producers are included in the search, the mirubactin is present in 30,2% of them. Panel b: Pie chart depicting the percentage of GCFs (n=21) whose BGCs are found in genomes that also contain mirubactin BGCs per GPA or GRB type (see chapter 4). All types of GPAs (I-IV) but only two (types B, D) among the GRP types (former GPA type V, current GRP types A-E) had hits in the genome-wide mirubactin BGC search. The question mark (?) corresponds to GCFs comprising incomplete BGCs whose type could not be determined. Panel c: Bar chart depicting wthin-GCF presence of mirubactin BGC hits.The 21 GCFs are the same as in panel b. They are labelled according to similarity to a known BGC (if applicable) and their type designation (GPA types I-V). The question mark (?) corresponds to GCFs comprising incomplete BGCs whose type could not be determined.

However, it is important to highlight that mirubactin BGCs were detected predominantly in *Amycolatopsis* and *Streptomyces* genomes, among all GPA producers, which are also the most common GPA producers (**Supplementary Figure 5**).

## Discussion

In the present manuscript, the genomic adaptation to horizontal gene transfer (HGT) of biosynthetic gene clusters (BGCs) was studied using the *Amycolatopsis* pangenome. The genomes in the dataset were mined for their biosynthetic capacities and the distribution of GPA producers revealed no pattern consistent with speciation events. Orthology inference analyses were conducted on a genome scale but also focused only on specialised metabolism (BGCs). This information, combined with the clustering of the BGCs into groups into gene cluster families (GCFs) was used for the creation of three presence/absence (P/A) tables, suitable for gene coincidence analysis. The latter was conducted with the new Goldfinder tool and presented in this manuscript, providing the first indication of a cluster of genes seemingly unrelated to a specific BGC possibly being implicated in the host's adaptation to its horizontal inheritance.

The first step in the analysis was the download of genomic sequences as well as the PGAT[34] annotated proteomes of *Amycolatopsis* bacterial strains from the NCBI RefSeq database. The genetic content of the pangenome was inferred based on these proteomes and is therefore highly dependent on the accuracy of the algorithm used, both for gene calling and for protein annotation. Even so, the use of the same tool for all genomes in the dataset ensures uniformity, which minimises any bias that may arise from inaccurate open reading frame (ORF) limits, since the same logic for choosing these locations is applied throughout the dataset.

Similarly, the choice of the genome mining tool that conducted the detection of the BGCs in the *Amycolatopsis* genomes can affect the precision of the analysis. AntiSMASH is a rule-based tool, which is known to be generous with the proposed limits of the BGCs, possibly including neighbouring genes that are not related to the corresponding metabolic pathway. Considering that the within-BGC orthologous groups (BGC dataset) are also considered in this analysis, this could create "noise" in the gene patterns of the pangenome. However, the possible effect any unrelated genes would have if they are falsely considered to belong to the specialised metabolism should be mitigated by the different combinations of gene patterns chosen for the datasets analysed by Goldfinder (**Table 1**).

The decision to not only examine one version of the observed gene patterns but multiple, was rooted on the fact that it is not known what could be correlated with the presence of a BGC in a genome. It could be a gene encoding for an enzyme that is part of the primary metabolism, but it could also be a gene that is part of another BGC, or a whole other BGC, or a type of BGC. The datasets in **Table 1** were chosen with this uncertainty in mind and to examine as many possibilities as possible.

The pangenome dataset (**Table 1**) failed to give interpretable results (data not shown), most likely due to the inaccurate orthogroup (OG) definitions. The tool that calculated the OGs, Orthofinder[17], though fast and easy to use, is known to "over-cluster" genes, leading to large OGs including higher variety than they

should[35]. This had a direct effect on the P/A table that was created from the gene-OG assignments. For example, consider homologous genes A, B and C that have evolved separately and are related to different functions. If these genes are placed in one OG, e.g. OG_1, then a genome that contains all three genes would have the same P/A pattern in OG_1 as another genome that contains B and C, and another that contains only A. Therefore, interpretations of any associations involving OG_1 would be problematic, especially if the true P/A patterns of gene A, B and C are different. In the following iterations of this analysis, we plan to overcome this issue by recreating the pangenome dataset with a more suitable tool, such as panX[35] or panaroo[36]. Only after ensuring the quality of the P/A table can we focus on the identification of significant relationships among genes. Additionally, to capture any coincidence or avoidance between genes involved in primary and specialised metabolism, we plan to analyse a dataset they are considered separately (distinct OGs).

Gene coincidence analysis of the BGC dataset (**Table 1**) also ran into a dead end. In this case, zol[20] was employed for the definition of OGs from genes included in BGCs. This tool is designed for context-specific orthology inference of genes included in gene clusters and also performs functional annotation of the OGs, a very useful characteristic for an analysis such as ours. The fault with this approach was the universal application to all types of BGCs at once. As demonstrated in the full pipeline presented in the zol paper, the intended input data for zol are homologous or orthologous gene clusters. However in this first attempt of the analysis, the full biosynthetic capacity of the *Amycolatopsis* pangenome, captured in its BGCs, was indiscriminately given to zol for OG calculation. The resulting P/A table displayed the same problem as the pangenome dataset - genes from very different BGCs were placed in the same OGs, making any associations between them uninterpretable. In the future recalculation of the BGC dataset, we plan to get around this issue by applying zol to BGCs belonging to the same gene cluster family (GCF).

The third and final dataset of **Table 1**, the BiG-SCAPE dataset, though reliant on the similarity metrics and threshold used[19], did not have the problems described above, and allowed the identification of some interesting relationships between GCFs, including GPA GCFs. The balhimycin GCF (GPA type I) and the ristocetin GCF (GPA type III) each had a separate associated GCF (**Figure 4**), which were only listed in the current manuscript and their significance will be explored in detail in the future. The most compelling observation though, was their shared associated GCF, which includes BGCs very similar to the MIBiG[21] BGC associated with the biosynthesis of mirubactin[30].

To comprehend the possible implications between the associations of two types of BGCs, they need to be studied in detail, with a focus on possible connections between them. Mirubactin A is a known iron siderophore, whose role is to be exported into the environment, where it binds to ferric iron, and then be transported back into the cell, where it can be exploited for the host's cellular needs[30]. However,

its (spontaneous) degradation product mirubactin C has been proven to protect *Bacillus subtilis* cells from cell wall biosynthesis mutations, which normally cause cell death[31]. This unusual secondary bioactivity has been linked to specific enzymatic complexes involved in cell wall biosynthesis[32]. More specifically, this concerns two types of assembly machineries for peptidoglycan (PG), a cell wall component, seem to be affected. The first is the “rod complex”, occurring in rod-shaped bacteria such as *Bacillus subtilis*, which includes a glycosyltransferase and transpeptidases known as class B penicillin-binding proteins (bPBPs). The second is part of an alternative, optional, cell wall biosynthetic pathway that is activated when the first cannot function properly and includes bifunctional class A penicillin-binding proteins (aPBPs). It is worth noting that homologous pathways are found in many bacteria[37], including *Amycolatopsis*. If bPBPs are affected due to mutations, aPBPs still lead to the formation of a cell wall, but with morphological abnormalities, such as a rounded shape[32]. More importantly, mutations in bPBPs can, through a cascade of affected enzymes and regulators, result in the production of reactive oxygen species (ROS), which cause cell death due to peroxidation of cell wall lipids (LPO). LPO takes place under the presence of redox-active iron, which is prevented due to the effect of mirubactin C. The latter protects the mutants from this effect through the following chain of events[32]. This compound is a degraded form of mirubactin A, a siderophore, and does not retain all the functions necessary for siderophore bioactivity. It can bind to environmental ferric iron but it cannot be transported back into the host cell, causing iron limitation, which in turn reduces LPO and prevents cell death.

On the other hand, (self) resistance to GPAs (types I-IV) involves alteration of the PG precursor lipid II, replacing the D-Ala-D-Ala tail with a D-Ala-D-Lac tail[33,38]. This change does not allow GPAs to bind to the cell wall of the producer cells, but it also causes a number of morphological and functional adjustments, such as a spherical shape, which has been attributed to affected PBP activities[39]. It has also been suggested that not all bacteria are capable of adapting to PG precursor modifications and that this may be connected to the observed distribution of vancomycin resistance, which includes bacteria that can not biosynthesize GPAs (vancomycin-resistant)[40].

Based on the descriptions above, it appears that there is some overlap between the functions of the enzymes whose mutation effects mirubactin C can protect from and the self-resistance mechanism of balhimycin and ristocetin producers. The latter is a necessary adaptation to the acquisition of their corresponding BGCs, since lack of self-resistance would cause death and the HGT event would be deleterious to the recipient cell. Therefore, we formulate the following hypothesis: it is possible that the presence of a mirubactin BGC increases the host’s tolerance to changes in the cell wall biosynthetic pathway, which in turn increases the chances of successfully integrating a horizontally transferred GPA BGC.

Naturally, extensive additional investigations need to take place to support such a hypothesis. The fact that mirubactin-like BGCs were found in multiple GPA

producers (**Figure 5**) is a positive indication, which has to be supported by further gene coincidence analyses. Additionally, it would be useful to study the evolutionary origin of mirubactin, the same way as has been conducted for GPAs[3,41]. Knowing the relative order in which the two BGCs appeared in *Amycolatopsis*, for example, would be beneficial to the elucidation of their relationship. The same kind of examination with a focus on the emergence of the resistance genes could possibly support an association between them and the mirubactin BGC, should any be detected by additional gene coincidence analyses (including GPA resistant non-producers). Furthermore, in the brief investigation of the mirubactin BGC distribution among GPA producers, two types of GRP (former GPA type V, see chapter 4) GCFs seemed to coincide with the siderophore. If such a relationship is confirmed, it would be interesting to draw possible connections between mirubactin and GRPs, whose mode of action also affects cell wall biosynthesis but in a different way[42]. All of these approaches will be considered in future analyses.

The present manuscript describes the first attempt at a BGC-aware gene coincidence analysis. Even at its preliminary state, a potentially important association was drawn between two different kinds of BGCs: GPA (balhimycin & ristocetin) and mirubactin BGCs. The exact nature of their relationship will be the focus of further research, while future attempts with an improved study design are expected to unearth additional clues into the adaptation mechanisms of horizontally acquired BGCs.

## Materials and Methods

### Sequencing data

All the assembled genomes (protein fasta files and genbank files) of the genus *Amycolatopsis* (NCBI taxon ID: 1813) available from the NCBI RefSeq database were downloaded via the Datasets resource[43]. Custom IDs were generated for each genome, which were used instead of their accession numbers for the rest of the analysis. All information on the genomes is reported in **Supplementary Table 1**.

### Orthology inference of the pangenome

The protein fasta files downloaded from NCBI were used as input for the Orthofinder tool (v2.3.11, run with default settings)[17], which infers orthologous relationships and places genes in groups accordingly (orthologous groups - OGs). The tool generates a species phylogeny as part of its pipeline[44,45], which was imported to iTOL[18] and annotated to produce **Figure 1**.

### BGC dataset

Detection of putative biosynthetic gene clusters (BGCs) was accomplished with the antiSMASH tool (v7.0.0)[15]. The tool was run with default settings and the additional comparison to the MIBiG[21] dataset (--cc-mibig). Information on the BGCs detected is reported in **Supplementary Table 2**.

**BGC similarity clustering**

The total set of BGCs were analysed with BiG-SCAPE 1.0.1 (2020-01-27)[19], which clusters BGCs into gene cluster families (GCFs) and clans based on their similarity. The tool was run on auto mode with default settings except with the option to mix all BGC classes (--mix) and the singleton flag (--include_singletons). The resulting network was visualised in **Figure 2**. The BGCs encoding for the biosynthesis of GPAs were analysed separately, with an adapted BiG-SCAPE script that considered them all in the NRPS class (instead of some in NRPS and some in NRPS-PKS hybrids), and a threshold of 0.2 that is known to always separate the GPA types into GCFs (based on a previous analysis [4]). The GCF and clan assignment of all BGCs is included in **Supplementary Table 2**.

**Orthology inference of the genus' specialised metabolism**

The protein sequences of the genes involved in specialised metabolism (included in the putative BGCs' range) were used as input for the zol tool (v1.3.9, run with default settings)[20], which performs orthology inference on gene clusters. The orthologous groups defined by this tool were used as additional information for the gene coincidence analysis.

**Generation of pseudo-OGs**

The results of BiG-SCAPE assigned a GCF value to each BGC, while the separate handling of the GPAs ensured that their GCFs were "pure" and did not include any non-GPA BGCs. The information for the GCF assignment of each BGC was transformed into a presence/absence table that matched the presence of each GCF in each genome in the *Amycolatopsis* genus. This table was adapted in a suitable format and these "pseudo-OGs" (the GCFs) were used as input for the Goldfinder tool. All datasets are included in **Supplementary Data 1**.

**Gene coincidence analysis with Goldfinder**

The OGs and pseudoOGs described above were used as input for the Goldfinder tool (unpublished), which applies a phylogeny-aware method to infer coinciding or dissociating pairs of genes. Goldfinder is a new gene coincidence analysis tool, which incorporates elements of the phylogenetic genome-wide association study tool treeGWAS[46] in order to detect the truly significant relationships between genes in a pangenome (**Supplementary Figure 6**). The implemented approach includes a step reconstructing the ancestral state (presence/absence) of the internal nodes, using a user-given species tree and using this information to calculate the association scores among all pairs of genes. This way, the association of genes that can be attributed to common ancestry is scored lower and the truly significant associations are the ones the tool outputs. For all Goldfinder analyses conducted (default settings), the Orthofinder-generated species tree was used as input, and each of the P/A tables of the datasets in **Table 1**. The results can be found in **Supplementary Data 2**.

The output of gene coincidence analysis tools, such as Goldfinder, can be overwhelming due to its sheer size[47]. In order to improve the interpretability of the

results, we developed a visualisation script for the result tables of the Goldfinder tool (**Supplementary Data 3**). The goal was to highlight the most significant relationships and to effectively include metadata into the visualisation, aiding in the prioritisation of relationships for further study. The commands for the generation of the images can be executed through a jupyter notebook, which will eventually be incorporated into the tools' files. The visualisation requires Cytoscape (v3.10.0)[48] and the py4cytocsape library[49]. The most significant relationships from the BiG-SCAPE dataset, as seen in **Figures 3** and **4**, were visualised using this method.

### Distribution analysis of mirubactin-related BGCs

For the investigation of the presence of mirubactin-related BGCs in GPA producer genomes, the dataset presented in chapter 4 was used. The antiSMASH results of the genomes that were considered candidate GPA producers were text-mined for "mirubactin". The type designation from chapter 4 was also used to analyse the distribution of the hits among types (**Supplementary Table 3**). GPA type V is used interchangeably with GRP (term first introduced in chapter 4). All statistics and charts for this part of the analysis were calculated using Google Sheets (Google Inc., Mountain View, CA).

## References


1. Gavriilidou, A. *et al.* Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nat Microbiol* **7**, 726–735 (2022).
2. Bragg, R. R., Meyburgh, C. M., Lee, J.-Y. & Coetzee, M. Potential Treatment Options in a Post-antibiotic Era. *Adv Exp Med Biol* **1052**, 51–61 (2018).
3. Hansen, M. H. *et al.* Resurrecting ancestral antibiotics: unveiling the origins of modern lipid II targeting glycopeptides. *Nat Commun* **14**, 7842 (2023).
4. Gavriilidou, A. *et al.* Phylogenetic distance and structural diversity directing a reclassification of glycopeptide antibiotics. 2023.02.10.526856 Preprint at https://doi.org/10.1101/2023.02.10.526856 (2023).
5. Chevrette, M. G. *et al.* The confluence of big data and evolutionary genome mining for the discovery of natural products. *Nat. Prod. Rep.* (2021) doi:10.1039/d1np00013f.
6. Stegmann, E., Frasch, H. J. & Wohlleben, W. Glycopeptide biosynthesis in the context of basic cellular functions. *Current Opinion in Microbiology* **13**, 595–602 (2010).
7. Butler, M. S., Hansford, K. A., Blaskovich, M. A. T., Halai, R. & Cooper, M. A. Glycopeptide antibiotics: Back to the future. *J Antibiot* **67**, 631–644 (2014).
8. Mccormick, M. H., Mcguire, J. M., Pittenger, G. E., Pittenger, R. C. & Stark, W. M. Vancomycin, a new antibiotic. I. Chemical and biologic properties. *Antibiot Annu* **3**, 606–611 (1955).
9. Hansen, M. H., Stegmann, E. & Cryle, M. J. Beyond vancomycin: recent advances in the modification, reengineering, production and discovery of improved glycopeptide antibiotics to tackle multidrug-resistant bacteria. *Curr Opin Biotechnol* **77**, 102767 (2022).
10. Greule, A. *et al.* Kistamicin biosynthesis reveals the biosynthetic requirements for production of highly crosslinked glycopeptide antibiotics. *Nature Communications 2019 10:1* **10**, 1–15 (2019).
11. Schoppet, M. *et al.* The biosynthetic implications of late-stage condensation domain selectivity during glycopeptide antibiotic biosynthesis. *Chemical Science* **10**, 118–133 (2019).
12. Kittilä, T. *et al.* Halogenation of glycopeptide antibiotics occurs at the amino acid level during non-ribosomal peptide synthesis. *Chemical Science* **8**, 5992–6004 (2017).
13. Wohlleben, W., Stegmann, E. & Süssmuth, R. D. Chapter 18. Molecular genetic approaches to analyze glycopeptide biosynthesis. *Methods Enzymol* **458**, 459–486 (2009).
14. Truman, A. W. *et al.* Antibiotic resistance mechanisms inform discovery: identification and characterization of a novel amycolatopsis strain producing ristocetin. *Antimicrob Agents Chemother* **58**, 5687–5695 (2014).
15. Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Research* **51**, W46–W50 (2023).

16. Nicolaou, K. C., Boddy, C. N. C., Bräse, S. & Winssinger, N. Chemistry, Biology, and Medicine of the Glycopeptide Antibiotics. *Angewandte Chemie International Edition* **38**, 2096–2152 (1999).
17. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).
18. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
19. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
20. Salamzade, R. *et al.* zol & fai: large-scale targeted detection and evolutionary investigation of gene clusters. 2023.06.07.544063 Preprint at https://doi.org/10.1101/2023.06.07.544063 (2023).
21. Terlouw, B. R. *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Research* **51**, D603–D610 (2023).
22. González-Pastor, J. E., San Millán, J. L., Castilla, M. A. & Moreno, F. Structure and organization of plasmid genes required to produce the translation inhibitor microcin C7. *J Bacteriol* **177**, 7131–7140 (1995).
23. Matthijs, S. *et al.* The Pseudomonas siderophore quinolobactin is synthesized from xanthurenic acid, an intermediate of the kynurenine pathway. *Mol Microbiol* **52**, 371–384 (2004).
24. Bennallack, P. R. & Griffitts, J. S. Elucidating and engineering thiopeptide biosynthesis. *World J Microbiol Biotechnol* **33**, 119 (2017).
25. Keatinge-Clay, A. T. The structures of type I polyketide synthases. *Nat Prod Rep* **29**, 1050–1073 (2012).
26. Acedo, J. Z., Chiorean, S., Vederas, J. C. & van Belkum, M. J. The expanding structural variety among bacteriocins from Gram-positive bacteria. *FEMS Microbiol Rev* **42**, 805–828 (2018).
27. Liu, W. *et al.* Rapid PCR amplification of minimal enediyne polyketide synthase cassettes leads to a predictive familial classification model. *Proc Natl Acad Sci U S A* **100**, 11959–11963 (2003).
28. Kaweewan, I. *et al.* Isolation and structure determination of a new cytotoxic peptide, curacozole, from Streptomyces curacoi based on genome mining. *J Antibiot* **72**, 1–7 (2019).
29. Purushothaman, M. *et al.* Genome-Mining-Based Discovery of the Cyclic Peptide Tolypamide and TolF, a Ser/Thr Forward O-Prenyltransferase. *Angewandte Chemie International Edition* **60**, 8460–8465 (2021).
30. Giessen, T. W. *et al.* Isolation, Structure Elucidation, and Biosynthesis of an Unusual Hydroxamic Acid Ester-Containing Siderophore from Actinosynnema mirum. *J. Nat. Prod.* **75**, 905–914 (2012).
31. Kepplinger, B. *et al.* Mirubactin C rescues the lethal effect of cell wall biosynthesis mutations in Bacillus subtilis. *Front Microbiol* **13**, 1004737 (2022).
32. Kawai, Y. *et al.* On the mechanisms of lysis triggered by perturbations of bacterial cell wall biosynthesis. *Nat Commun* **14**, 4123 (2023).
33. Kilian, R., Frasch, H.-J., Kulik, A., Wohlleben, W. & Stegmann, E. The VanRS Homologous Two-Component System VnlRSAb of the Glycopeptide Producer Amycolatopsis balhimycina Activates Transcription of the vanHAXSc Genes in Streptomyces coelicolor, but not in A. balhimycina. *Microb Drug Resist* **22**, 499–509 (2016).
34. PGAT: a multistrain analysis resource for microbial genomes | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/27/17/2429/224993.
35. Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. *Nucleic Acids Research* **46**, (2017).
36. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020).
37. Dion, M. F. *et al.* Bacillus subtilis cell diameter is determined by the opposing actions of two distinct cell wall synthetic systems. *Nat Microbiol* **4**, 1294–1305 (2019).
38. Yushchuk, O., Binda, E. & Marinelli, F. Glycopeptide Antibiotic Resistance Genes: Distribution and Function in the Producer Actinomycetes. *Front Microbiol* **11**, 1173 (2020).
39. Deghorain, M. *et al.* Functional and Morphological Adaptation to Peptidoglycan Precursor Alteration in Lactococcus lactis*. *Journal of Biological Chemistry* **285**, 24003–24013 (2010).
40. Pootoolal, J., Neu, J. & Wright, G. D. Glycopeptide Antibiotic Resistance. *Annual Review of Pharmacology and Toxicology* **42**, 381–408 (2002).
41. Waglechner, N., McArthur, A. G. & Wright, G. D. Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance. *Nature Microbiology* **4**, 1862–1871 (2019).

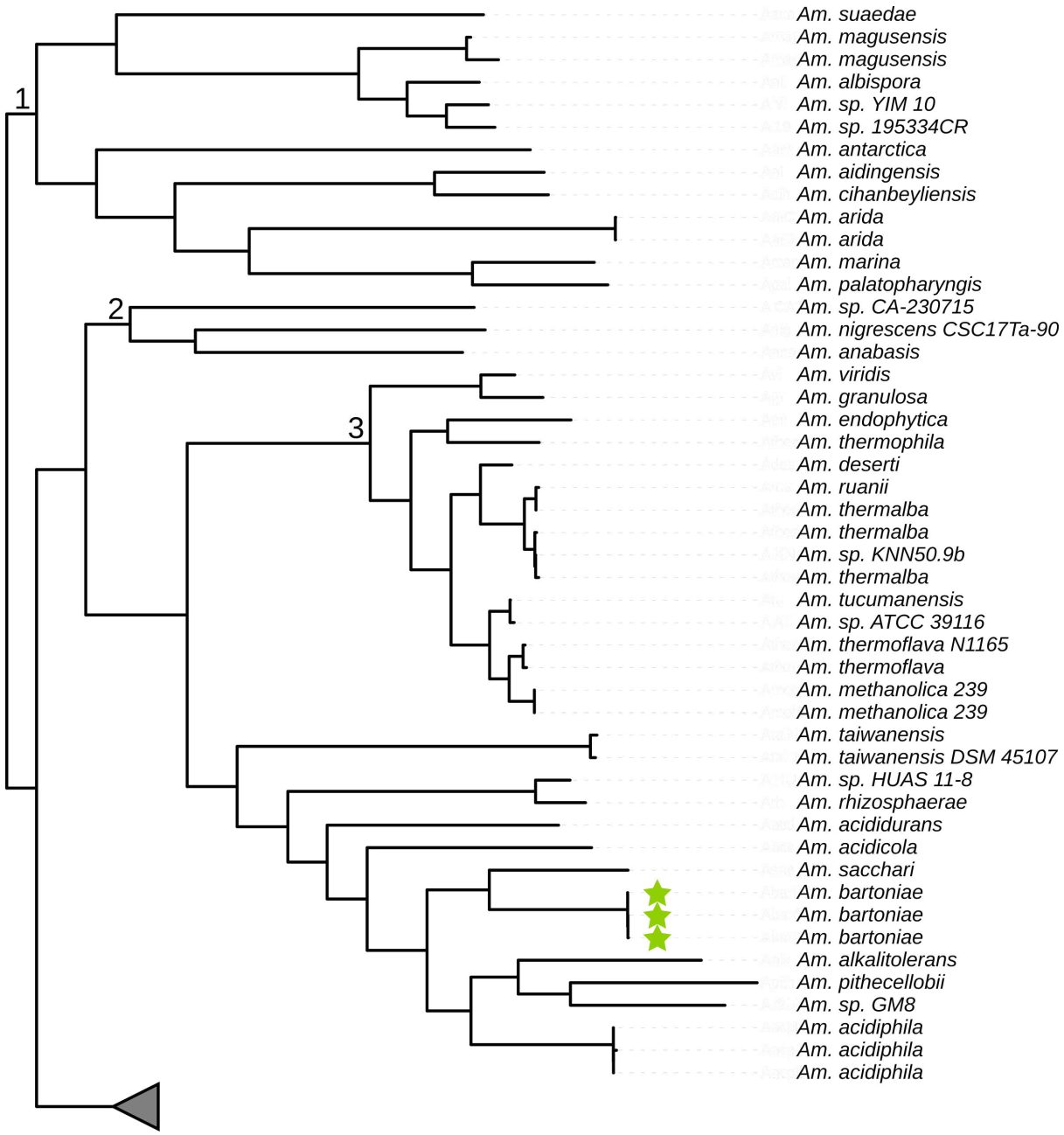
42. Xu, M. *et al.* GPAHex-A synthetic biology platform for Type IV–V glycopeptide antibiotic production and discovery. *Nature Communications 2020 11:1* **11**, 1–12 (2020).
43. Datasets: Genome [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; accessed on 6th November 2023. https://www.ncbi.nlm.nih.gov/datasets/genome.
44. Emms, D. M. & Kelly, S. STAG: Species Tree Inference from All Genes. 267914 Preprint at https://doi.org/10.1101/267914 (2018).
45. Emms, D. M. & Kelly, S. STRIDE: Species Tree Root Inference from Gene Duplication Events. *Molecular Biology and Evolution* **34**, 3267–3278 (2017).
46. Collins, C. & Didelot, X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLOS Computational Biology* **14**, e1005958 (2018).
47. Whelan, F. J., Rusilowicz, M. & McInerney, J. O. Coinfinder: Detecting significant associations and dissociations in pangenomes. *Microbial Genomics* **6**, e000338 (2020).
48. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003).
49. Otasek, D., Morris, J. H., Bouças, J., Pico, A. R. & Demchak, B. Cytoscape Automation: empowering workflow-based network analysis. *Genome Biol* **20**, 185 (2019).

## Supplementary Material

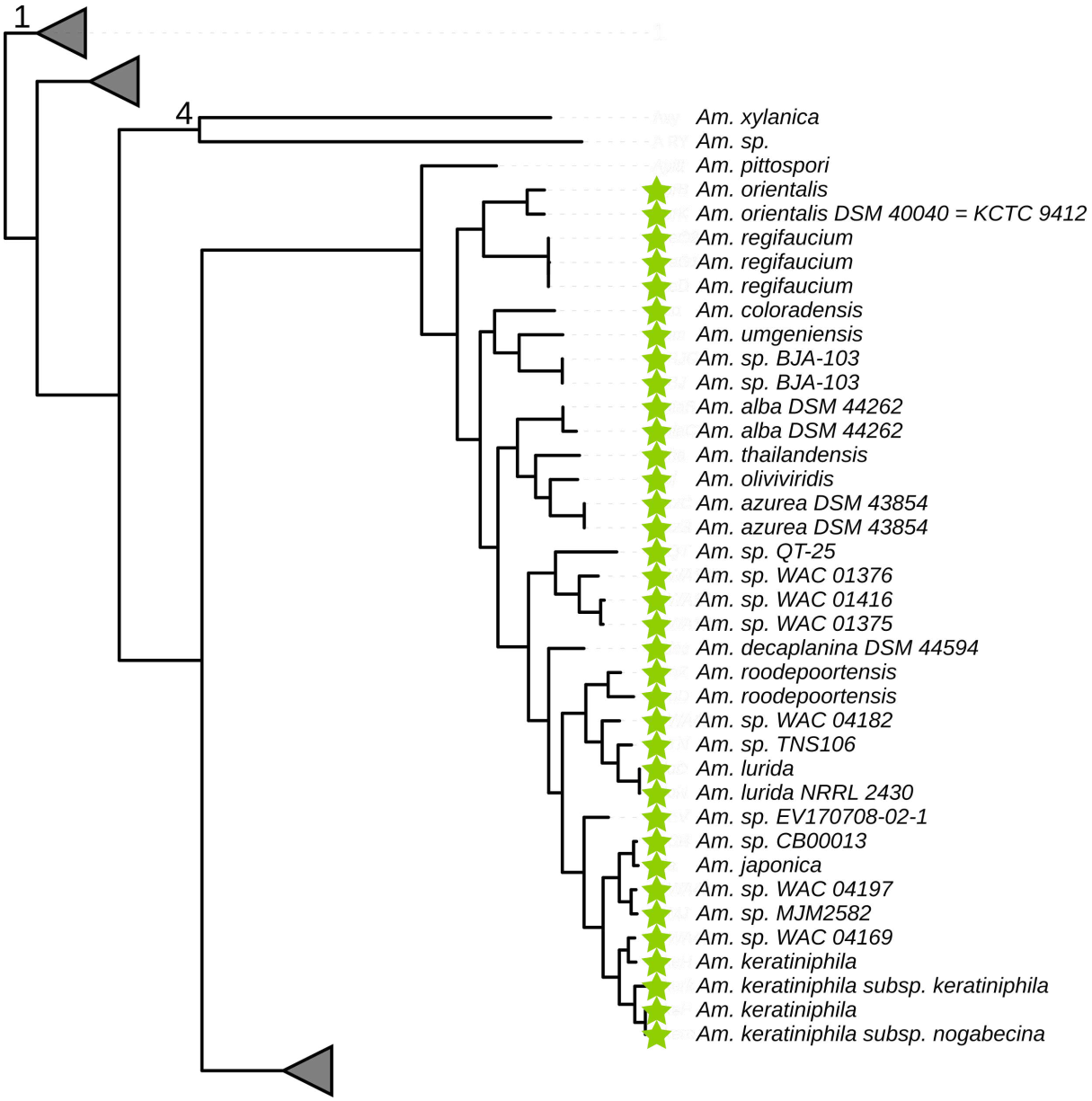
All supplementary material of unpublished projects are available for download (upon request) from a zenodo repository: https://doi.org/10.5281/zenodo.10879735

Supplementary Table 1: Accession numbers and metadata on the genomes involved in the study.
Supplementary Table 2: Metadata on the detected putative BGCs abd table of mirubactin search hits.
Supplementary Data 1: Generated datasets of OGs and pseudoOGs that were used as input for Goldfinder.
Supplementary Data 2: Goldfinder results.
Supplementary Data 3: code for visualisation of Goldfinder results through Cytoscape.

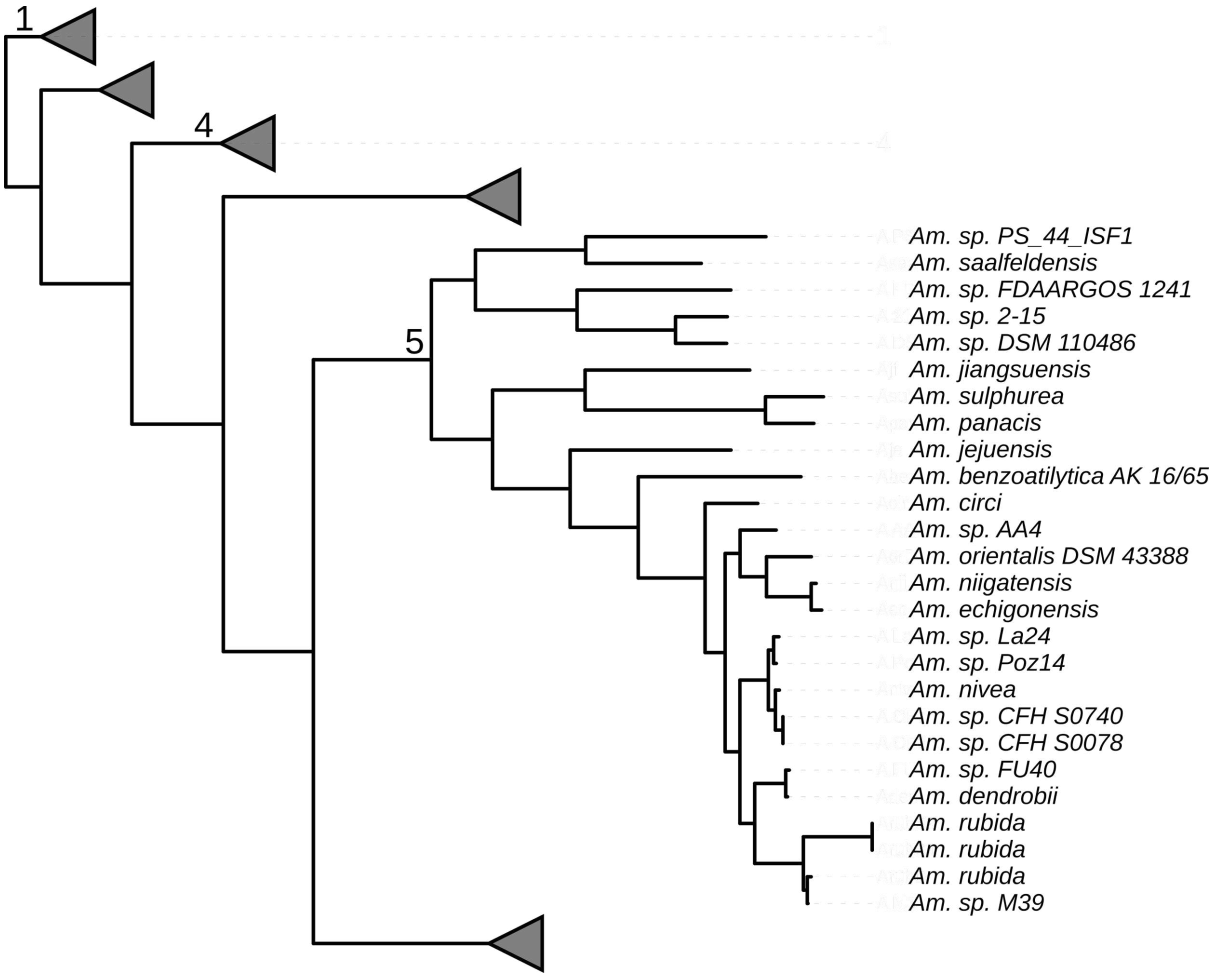
## Supplementary Figures



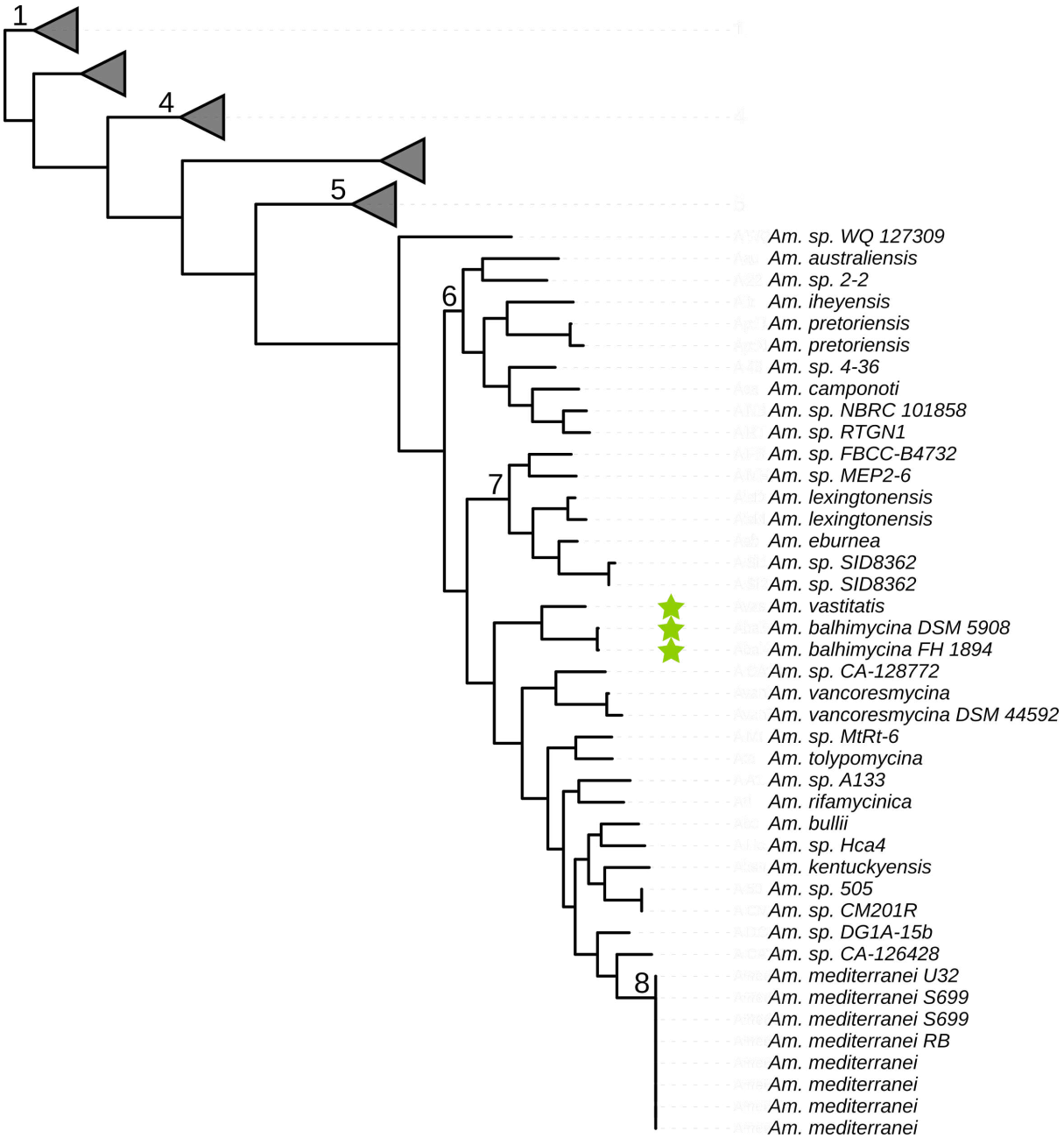

**Supplementary Figure 1**: species tree of the *Amycolatopsis* genus. Clades 1-3 (Figure 1) are expanded, while the rest of the tree is collapsed for better visibility.



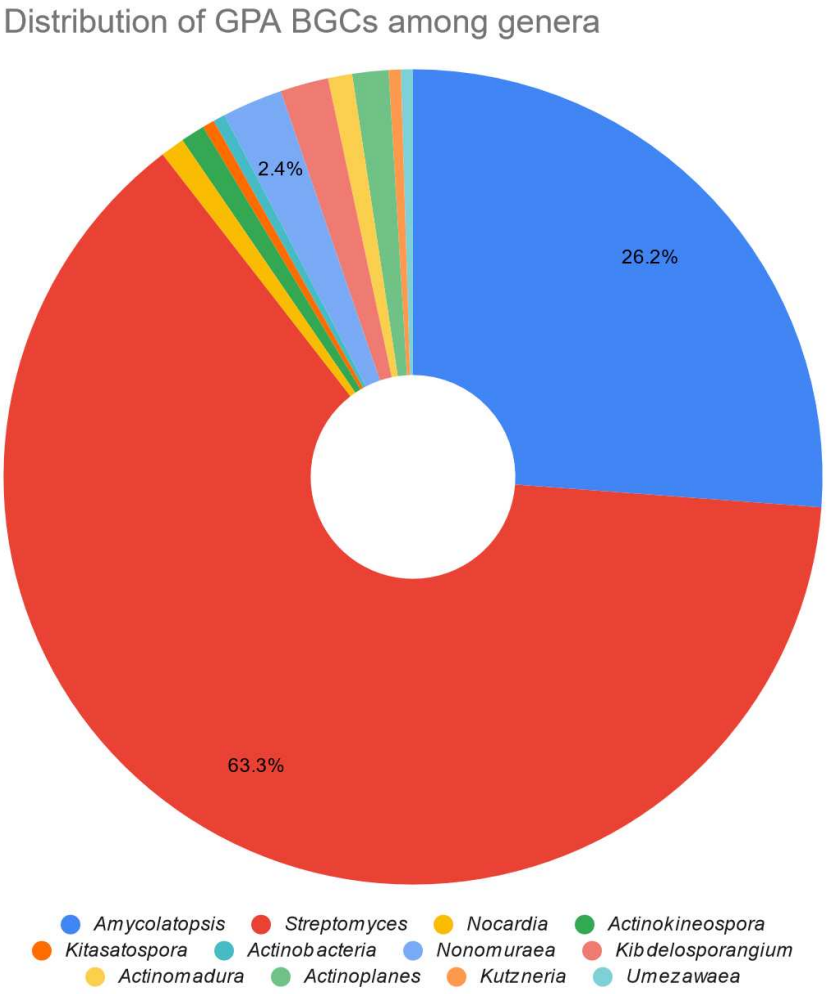

**Supplementary Figure 2**: species tree of the *Amycolatopsis* genus. Clade 4 (Figure 1) is expanded, while the rest of the tree is collapsed for better visibility.



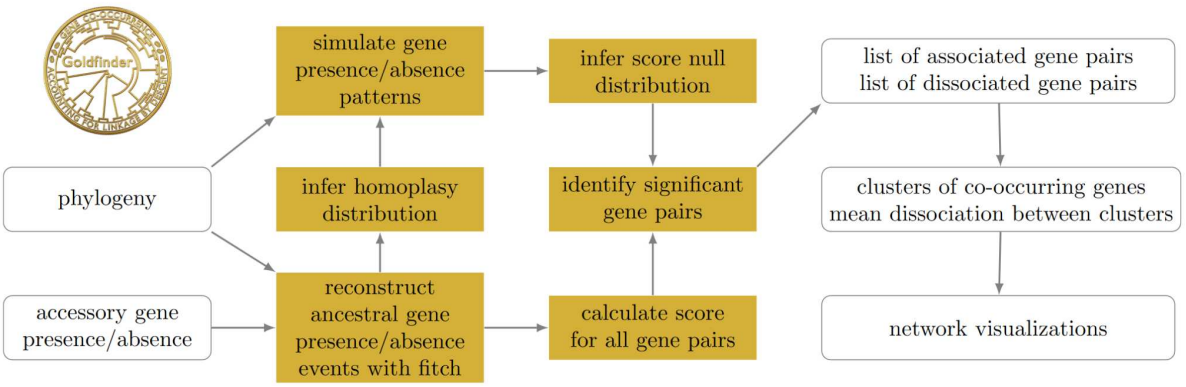

**Supplementary Figure 3**: species tree of the *Amycolatopsis* genus. Clade 5 (Figure 1) is expanded, while the rest of the tree is collapsed for better visibility.




**Supplementary Figure 4**: species tree of the *Amycolatopsis* genus. Clades 6-8 (Figure 1) is expanded, while the rest of the tree is collapsed for better visibility.




**Supplementary Figure 5**: Pie chart depicting the percentage of GPA producers (n=211) detected in each genus (based on data from chapter 4). *Streptomyces* and *Amycolatopsis* are the dominant genera in the dataset. Percentages of small slices in order from left to right: *Nocardia* (1%), *Actinokineospora* (2%), *Kitasatospora* (0.5%), Actinobacteria (undetermined genus - 0.5%), *Kibdelosporangium* (1.9%), *Actinomadura* (1%), *Actinoplanes* (1.4%), *Kutzneria* (0.5%), *Umezawaea* (0.5%).




**Supplementary Figure 6**: The Goldfinder workflow. Figure adapted from the Goldfinder manuscript (unpublished, see 'List of publications not included in the thesis').

# Discussion and conclusions

These five chapters unfold the narrative of an evolutionary investigation of bacterial BGCs across multiple scales. The latter term refers both to the volume and the type of data analysed in each project (e.g., genomes, BGCs, GCFs). As far as genomes are concerned, the data ranged from hundreds of thousands of bacterial genomes from all taxa in chapter 2, to a few hundred genomes (i.e., putative GPA producers) in chapter 4, and to 150 genomes from a single genus in chapter 5. When it comes to biosynthetic capacity and diversity, the data included more than a million BGCs of great variety in chapter 2, a couple of hundred BGCs belonging to a very specific system (i.e. GPA and GRP BGCs) in chapter 4, and a few thousands BGCs clustered into hundreds of GCFs in chapter 5. Each of these approaches has contributed in a different way to the elucidation of BGC distribution and evolution in bacterial hosts.

## The panoramic view: Biosynthetic diversity in bacterial genomes

The advances in sequencing technologies have greatly increased the publicly available data suitable for BGC detection and analysis, as was discussed in chapter 1. The latter was only possible thanks to the development of algorithms capable of handling such volumes of data, like the BiG-SLICE tool for similarity clustering of BGCs, which was one of the main tools included in the analysis of chapter 2. This project was the first to study the diversity of specialised metabolism encoded in bacterial genomes on a domain-wide scale. It was possible to quantify the diversity of BGCs encoded in the available genomes and metagenome-assembled genomes (MAGs), while ensuring that the metric used (i.e., GCFs) is comparable to chemical classes, a measure of chemical diversity. It was also extrapolated how much more diversity can be expected when the volume of sequenced genomes is multiplied, concluding that there is a lot more to discover, especially from metagenomic data.

The potential of environmental samples as sources of new biosynthetic diversity has been realised by other studies as well[12,28–30]. Having established that, there are still hurdles to be overcome with metagenomics-based discovery of new specialised metabolites. One of the most common issues with this approach is the

inadequate sequencing and assembly quality[30]. This is being addressed by the development of new methodologies. For example, from the experimental side, which includes elegant combinations of existing sequencing methods and DNA libraries to efficiently capture BGCs from metagenomes[31], and conduct culture-free expression[32]. Simultaneously, various computational approaches are investigated to overcome the shortcomings of the current sequencing and assembly methods, such as reconstructing BGCs directly from assembly graphs[33], which would additionally sidestep the possible underpredictions due to binning processes being biassed against HGT regions - a common occurrence in BGCs[30]. Other approaches include employing artificial intelligence[12,34] and other innovative schemes[35,36] to capture the elusive new-to-nature compounds and the BGCs encoding their biosynthesis. The focus of the genome mining community seems to be shifting towards metagenomes and it is exciting to think what will come out of it in the near future.

Additionally, another obstacle in comparative analyses of BGCs was overcome in chapter 2: the genus was identified as the taxonomic rank most suitable for examining the biosynthetic diversity of different bacteria in contrast to each other. The uniformity of the GCFs within genera had been observed before, but never confirmed within the scope of the entire bacterial domain. A newer study also reaffirmed this conclusion for a large metagenomic dataset[37], which was not attempted in this project due to a lack of fine-scale taxonomic placement of the MAGs that were analysed. Nevertheless, chapter 2 introduced the concept of the REDgroups, a genus-equivalent taxon characterised by equal evolutionary distance among the members. A compendium of information on the biosynthetic diversity and potential across the entire bacterial domain by comparison of GCF values among REDgroups was presented.

The methodology presented in chapter 2 has since been adapted and applied in other studies, with similar goals. Beck and coauthors[38] combined this approach with machine learning classifiers to predict the activity of the compounds whose biosynthesis is encoded in BGCs, identifying promising genera for various applications[38]. Since one of the bottlenecks in the discovery of new bioactive compounds is dereplication of the candidate targets, studies have also focused on means to identify true biosynthetic novelty[39]. Another article describes analogous methods aimed at capturing the distribution of anti-phage defence systems encoded in the phylum Actinomycetota, in which there is genus-specific variance[40], in

agreement with chapter 2's findings on BGCs. More research projects like these are anticipated, which make use of the now-established taxonomic rank for comparisons, likely making the projects themselves commensurate as well.

Making use of this finding, some rare candidates were identified in the study's results whose biosynthetic capacities had not been explored yet, which will be the target of future bioprospecting efforts[39]. Surprisingly, it was also concluded that a lot of known gifted producer taxa will continue to be the source of novel biosynthetic diversity. The most unexpected finding was the recognition that a REDgroup corresponding to a subset of the genus *Streptomyces* was by far the most promising group. Already a new study has discovered a new *Streptomyces* species whose members display high biosynthetic diversity[41], in agreement with these results. Due to the high interest of the community in this taxon not only for its own biosynthetic capacity, but for its potential to accommodate various BGCs, efforts are being made to harness its potential fully, for example through the development of heterologous expression protocols[42] or cell-free gene expression systems[43].

While the importance of *Streptomyces* for the field of specialised metabolites is clear, an explanation behind its ability to synthesise so many and diverse compounds has not been presented. In the context of the study, it is hypothesised that the high number of various BGCs may be related to increased HGT events, which possibly constitutes an adaptive mechanism that allows these bacteria to take over new ecological niches. Other analyses, such as one conducted in a rhizosphere population of *Streptomyces* strains[44], confirm the high occurrence of conjugative events, one of the means of HGT, which supports this notion. Another view suggests that mobile genetic elements in *Streptomyces* contribute to the plasticity of their genome by promoting chromosomal rearrangements. These bacteria dedicate a high proportion of their genome to genes encoding non-essential metabolic pathways, a phenomenon which is still being investigated[40]. It is important to note that HGT events are rare, even if they are observed more often in BGCs than the rest of the bacterial genomes[12], while the rate of their occurrences does not appear uniform within the *Streptomyces* clade[41]. Several studies have been conducted on a grand scale, attempting to explain the frequency and means of HGT events[30], as well as the taxa that are involved[45], but a definitive answer on the specific case of *Streptomyces* is still lacking. The investigation of the reasons that constitute

streptomycetes such gifted producers was not conducted in the scope of the analysis in chapter 2, but it is expected to be the focus of future studies.

Another aspect to the distribution of biosynthetic diversity that was only considered in part in the project, was the biogeographical aspect. In the subset of the metagenomic data that was suitable for such an investigation, habitat-specific GCFs were observed. This was later confirmed by another study on a metagenomic dataset isolated from the Greenland Ice Sheet[28] and also on symbiotic bacteria, such as the ones inhabiting the human body, which have been more systematically examined[46,47]. Identifying which specialised metabolic pathways are widely distributed and which ones are associated with specific environmental niches will support the study of their ecological role in bacterial communities, which is poorly understood at the moment. To that end, a comprehensive analysis of the biosynthetic diversity, on a global scale and encompassing as many environmental biomes as possible, would be necessary. Thankfully, such an effort is already in progress in the BGC-atlas project (hosted on https://bgc-atlas.ziemertlab.com/). The next step would be to connect yet more variables to this information, such as the effect of time, as was done *in situ* for a specific biome[48]. Naturally, this would require a massive collaborative effort from researchers around the world in order to be achievable on a comparable scale as the data we now have on specialised metabolism. A more realistic goal is the normalisation of multi-omics approaches to microbiome analysis, which would generate the information necessary to study community dynamics in each sample[49,50].

## The monocladic standpoint: Phylogenetic and coincidence analysis of BGCs encoding the biosynthesis of GPAs

Having explored the biosynthetic diversity encoded in the bacterial kingdom, this dissertation moved from a grand scale to the perspective of a single well-known specialised metabolic pathway, that of GPAs. The motivation here was to investigate the distribution and evolutionary events that shaped the associated BGCs, in an effort to formulate some hypotheses that would explain the findings of chapter 2.

Naturally, in order to study their evolution, it is vital to firstly understand what is already known about the biosynthesis of GPAs. They are compounds that have been studied for a long time and from scientists of many different fields, each focusing on

another aspect. For example, in the field of microbiology, they are relevant for their ecological role in the bacterial communities. Biochemists are interested in the exact steps of their biosynthetic pathways, and there are bioengineering efforts to produce semi-synthetic compounds based on the natural assembly mechanism. Bioinformaticians are interested in their evolution and any information that can improve genome mining efforts. At the same time, each field relies on the other for advancing their own goals. Microbiologists are interested in the genetic elements involved because they can be exploited for heterologous expression of the BGCs. Knowing the evolutionary history of the genes involved can be useful to biochemists for elucidating the function of unknown enzymes. Developing bioinformatic algorithms and pipelines to address any of these questions requires familiarisation with the established background, depending on the field. The study of BGCs, including GPA BGCs, is a very interdisciplinary endeavour, which requires effective communication between scientists of different backgrounds. To this end, chapter 3 constitutes an innovative approach to convey the current state of research in regards to the biosynthesis of GPAs, in a manner approachable for a broad scientific audience. It is the hope of the authors that, in the future, such endeavours will become routine and expected from articles describing biosynthetic pathways.

The successive chapter 4 describes an extensive study on the evolutionary history of GPA BGCs, rendered possible after the conduction of the bibliographic review of chapter 3. The first step of that project was the creation of an extensive dataset of BGCs through scanning publicly available databases. This led to the identification of a large number of BGCs in bacterial genomes of various taxonomic origin. Not all of these BGCs could be included in the study, due to low quality, but the results of this search did prove that many bacterial taxa not associated with GPAs encode the genes for their biosynthesis. In addition, not only was the first case of a single bacterium encoding two such clusters found, but also several very unusual BGCs were detected (e.g., GRP type E), whose compound structure and mode of action elucidation are completely unknown for now.

The original goal of the analysis in chapter 4 was to investigate the phylogeny of these BGCs, but, due to some interesting observations on the differences of type V, it transformed into a presentation of a new classification system, supported by structural and phylogenetic information. Having an informative and dependable classification system is important for research on any bioactive compounds, but

especially so for antibiotics. There are several efforts in place, both on national[51,52] and international[53] levels, to control antibiotic usage in an effort to combat the emergence of multi-resistant pathogens. The World Health Organisation (WHO) includes GPAs, both natural and semi-synthetic variants, in its antibiotic stewardship program, intended to support monitoring activities. They are placed in the “Watch” and “Reserve” categories, which means they are considered at high risk of resistance emerging against them and WHO suggests to restrict their use as a ‘last resort’ in order to preserve their effectiveness[53]. These categorisations are important for the allowed usage of such compounds in relation to human[54], animal[55], even environmental[56] health, as well as in the food industry[57,58]. For the design of related policies, the collection of knowledge on each drug is required, which allows informed decisions on the risks and benefits of use[59]. This is where the importance of the classification system of GPAs comes into play. Mislabeling of certain compounds as GPAs could, for example, lead to associating patterns of GPA resistance with glycopeptide related peptides (GRPs), whose mode of action - and quite possibly their resistance mechanism - differs. Studying each type of GPA and especially each type of GRP separately will allow researchers to reach accurate conclusions, which will translate to better-informed policy-building.

Besides the suggestion for reclassification, chapter 4 describes the extensive phylogenetic analysis that supports it, conducted on the related BGCs. The value of increasing our understanding of the evolutionary events that led to the BGCs we detect today, such as the development of genome mining tools, has been discussed in chapter 1. Additionally, apart from the bioinformatic field, there are further applications of such information, as in bioengineering of new-to-nature compounds[60]. Preexisting knowledge on the evolution of specific genes related to GPA biosynthesis[61] was already helpful for the elucidation of certain events in the extended dataset of chapter 4. It is therefore expected that the trove of phylogenetic trees, for every gene and domain related to GPAs and GRPs, generated in this study, will support future investigations concentrated on specific enzymes or functions, possibly aiding in the elucidation of their role or chemical structure.

Furthermore, the rationale of the pipeline applied for conducting the phylogenetic analysis on a BGC level can be introduced to other BGC systems as well. Special attention needs to be dedicated to the evolutionary events taking place in the various genomic components of a BGC. The standard methods for

phylogenetic inference are designed for vertically transmitted genes and therefore the assumptions they are based on do not always apply to genes involved in specialised metabolism. In chapter 4, the modular genes encoding non-ribosomal peptide synthetases (NRPSs) needed to be studied on the level of the functional domain, as it has been established that evolutionary pressure can be applied to different degrees in each of them. Additionally, this study was the first to make use of super networks based on partial trees of all well-populated genes and domains. This uncommon phylogenetic method has been mathematically proven to be able to handle incongruence in the dataset, making its application well suited for the study of BGCs. The calculation of a concatenated phylogeny on the other hand, ensured that the information from the different sources (genes, domains) was combined, when appropriate (if they passed a congruence check), into a representative phylogeny for the whole gene cluster. A comparison of the latter with the species tree of the producers would give some valuable insights on the origin of the BGCs[62], but this was not conducted in the scope of this analysis. Studying gene to species tree resemblance can highlight possible HGT events[30,63], which is now possible for a myriad of genes and domains involved in GPA and GRP BGCs, and will be conducted in future projects.

Chapter 4 elucidated the evolutionary history of GPA and GRP BGCs, but that did not yet explain why certain bacteria are more prolific producers of specialised metabolites than others. It was hypothesised that the answer may be related to the presence of specific adaptation mechanisms that ameliorate the metabolic cost of incorporating a horizontally acquired BGC, leading to such an investigation in chapter 5. In that project, the entire genomic repertoire of the *Amycolatopsis* genus was studied in a first BGC-aware coincidence analysis. The ability to incorporate both producers and non producers in such a study was made possible thanks to the development of appropriate efficient tools like Goldfinder (unpublished), since the inclusion of both genomes that contained GPA BGCs and those that did not required a genus-sized pangenome. The study of typical pangenomes (all genomes belonging to one species) and the search for associations among genes is not a new concept[64,65], but has not been applied in the scope of BGCs.

Though preliminary, these results pointed to a potentially significant association between the ability to produce GPAs and the ability to produce mirubactin, a known siderophore antibiotic. In chapter 6, a hypothesis for the nature of their relationship

was formulated, which could be directly related to an adaptation mechanism for HGT of BGCs encoding GPA biosynthesis, but further analyses need to be conducted to confirm or disprove this. In any case, through the attempts described in chapter 5, a lot of observations were made on possible flaws of the current methodology. An improvement of those in the next iteration of the analysis is expected to reveal more interesting relationships, of association or of avoidance, between BGCs or genes involved in the primary or specialised metabolism. Furthermore, it will be possible to apply the pipeline to any well-studied system of BGCs and any genus of appropriate membership size. Perhaps a comparison of observed relationships among datasets will eventually be possible as well. Adopting such approaches for the, much larger, *Streptomyces* genus may ultimately shed light to what sets it apart from the rest of the bacteria as a gifted producer.

It is worth noting that the quality of the analysis presented in the last two chapters is dependent on a common factor: the accuracy of the orthology inference. Indeed the categorization of the genes studied into orthologous groups (OGs) was vital both for the evolutionary exploration of the GPA and GRP BGCs in chapter 4 and for the coincidence analysis of the *Amycolatopsis* pangenome in chapter 5. The authors are grateful for the success of the zol[66] orthology inference tool in the case of the GPA-related genes. However, for the larger and more diverse dataset of BGCs used for the gene coincidence analysis, a more creative approach is required and will be attempted, such as a pre-clustering step of closely related BGCs prior to the zol analysis. The distinction of orthologs and paralogs is not a trivial problem and there are a number of tools attempting to tackle it, as well as a dedicated consortium (Quest for Orthologs[67]). Though the focus of existing tools and databases is biassed towards eukaryotic organisms, there is a shift towards bacteria[68] that is expected to soon bridge that gap. As presented in the chapters above, the evolutionary study of specialised metabolism has much to gain from such advancements.

Given the opportunity arising from the last statement, it is important to mention that though this dissertation was focused on bacteria, the search for new-to-nature bioactive compounds is being conducted on multiple domains of life[69,70]. Several analyses described in the chapters above would be very informative if applied on different datasets. The distribution of biosynthetic diversity could be investigated in fungi or plants, though it would be harder to achieve due to the complexity of eukaryotic genomes[71]. Similarly, evolutionary investigations can be conducted on

BGCs from any source, provided the underlying assumptions are carefully considered. Even more interesting would be the combined investigation of different organisms, as is the case in the study of symbiotic microbes capable of specialised metabolism[3]. Association analyses, like the ones presented in chapter 5, could reveal unknown connections between host and symbiont genetic elements. The latter may highlight gene patterns conserved across hosts or across symbionts or it could expose any dependencies of the symbiont's biosynthetic capacity on the host environment. Naturally, the hypothetical studies mentioned here would require a great deal of effort to become possible, both to generate the necessary datasets and to design the appropriate methodology. However, it is still exciting to think about the directions that the field may take in the future.

## Outlook

This dissertation serves to highlight how important it is to approach a topic as complex as specialised metabolism from multiple perspectives. Having started from a global view of the bacterial biosynthetic potential, we moved to the evolution of a specific system and then to gene dynamics within a single genus. Each chapter contributed a different morsel of knowledge to the quest for discovering new bioactive compounds from bacterial sources. Having successfully increased our understanding of the distribution and evolutionary history of BGCs, it is exciting to see this new information being exploited in further research.

# Bibliography

Note: the publications listed in this chapter were referenced in the *Introduction* and the *Discussion and conclusion* chapters of this dissertation. For each of the main chapters (1-5), a dedicated list of references is present at the end of the corresponding chapter.


1. Madigan, M. T., Martinko, J. M. & Parker, J. I. Principles of Microbiology: 1. Microorganisms and Microbiology. in *Brock Biology of Microorganisms* 4–6 (Prentice Hall/Pearson Education, 2003).
2. Arimura, G., Matsui, K. & Takabayashi, J. Chemical and Molecular Ecology of Herbivore-Induced Plant Volatiles: Proximate Factors and Their Ultimate Functions. *Plant and Cell Physiology* **50**, 911–923 (2009).
3. Berasategui, A. *et al.* The leaf beetle Chelymorpha alternans propagates a plant pathogen in exchange for pupal protection. *Curr Biol* **32**, 4114-4127.e6 (2022).
4. Tobias, N. J. & Bode, H. B. Heterogeneity in Bacterial Specialized Metabolism. *Journal of Molecular Biology* **431**, 4589–4598 (2019).
5. Crozier, A., Ashihara, H. & Tomás-Barbéran, F. *Teas, Cocoa and Coffee: Plant Secondary Metabolites and Health*. (Wiley, 2011).
6. Bernardini, S., Tiezzi, A., Laghezza Masci, V. & Ovidi, E. Natural products for human health: an historical overview of the drug discovery approaches. *Natural Product Research* **32**, 1926–1950 (2018).
7. Atanasov, A. G. *et al.* Natural products in drug discovery: advances and opportunities. *Nature Reviews Drug Discovery* **20**, 200–216 (2021).
8. Divekar, P. A. *et al.* Plant Secondary Metabolites as Defense Tools against Herbivores for Sustainable Crop Protection. *Int J Mol Sci* **23**, 2690 (2022).
9. Kallscheuer, N., Classen, T., Drepper, T. & Marienhagen, J. Production of plant

metabolites with applications in the food industry using engineered microorganisms. *Current Opinion in Biotechnology* **56**, 7–17 (2019).

10. Bhadra, S., Chettri, D. & Kumar Verma, A. Biosurfactants: Secondary Metabolites Involved in the Process of Bioremediation and Biofilm Removal. *Appl Biochem Biotechnol* **195**, 5541–5567 (2023).

11. Terlouw, B. R. *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Research* **51**, D603–D610 (2023).

12. Chevrette, M. G. *et al.* The confluence of big data and evolutionary genome mining for the discovery of natural products. *Nat. Prod. Rep.* (2021) doi:10.1039/d1np00013f.

13. Medema, M. H. *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research* **39**, W339–W346 (2011).

14. Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Research* **51**, W46–W50 (2023).

15. Mullowney, M. W. *et al.* Artificial intelligence for natural product drug discovery. *Nat Rev Drug Discov* **22**, 895–916 (2023).

16. Hannigan, G. D. *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res* **47**, e110 (2019).

17. Kautsar, S. A., J van der Hooft, J. J., de Ridder, D. & Medema, M. H. BiG-SLiCE: A Highly Scalable Tool Maps the Diversity of 1.2 Million Biosynthetic Gene Clusters. *bioRxiv* 2020.08.17.240838 (2020).

18. Müller, A., Klöckner, A. & Schneider, T. Targeting a cell wall biosynthesis hot

spot. *Natural Product Reports* **34**, 909–932 (2017).

19. Gaynes, R. The Discovery of Penicillin—New Insights After More Than 75 Years of Clinical Use. *Emerg Infect Dis* **23**, 849–853 (2017).

20. Berdi, J. Bioactive Microbial Metabolites: A Personal View. *Journal of Antibiotics. Antibiotics* **58**, 1–26 (2005).

21. Hegemann, J. D., Birkelbach, J., Walesch, S. & Müller, R. Current developments in antibiotic discovery. *EMBO Rep* **24**, e56184 (2022).

22. Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature* **529**, 336–343 (2016).

23. Hansen, M. H., Stegmann, E. & Cryle, M. J. Beyond vancomycin: recent advances in the modification, reengineering, production and discovery of improved glycopeptide antibiotics to tackle multidrug-resistant bacteria. *Curr Opin Biotechnol* **77**, 102767 (2022).

24. Donadio, S., Sosio, M., Stegmann, E., Weber, T. & Wohlleben, W. Comparative analysis and insights into the evolution of gene clusters for glycopeptide antibiotic biosynthesis. *Molecular Genetics and Genomics* **274**, 40–50 (2005).

25. Alanjary, M. *et al.* The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Research* **45**, W42–W48 (2017).

26. Mungan, M. D. *et al.* ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. *Nucleic acids research* **48**, W546–W552 (2020).

27. Kapli, P., Yang, Z. & Telford, M. J. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics* **21**, 428–444 (2020).

28. Jaarsma, A. H. *et al.* The undiscovered biosynthetic potential of the Greenland Ice Sheet microbiome. *Front. Microbiol.* **14**, (2023).

29. Loureiro, C., Medema, M. H., van der Oost, J. & Sipkema, D. Exploration and exploitation of the environment for novel specialized metabolites. *Current Opinion in Biotechnology* **50**, 206–213 (2018).

30. Seshadri, R. *et al.* Expanding the genomic encyclopedia of Actinobacteria with 824 isolate reference genomes. *Cell Genomics* **2**, (2022).

31. Negri, T. *et al.* A rapid and efficient strategy to identify and recover biosynthetic gene clusters from soil metagenomes. *Appl Microbiol Biotechnol* **106**, 3293–3306 (2022).

32. Xu, Y. *et al.* Recent Advances in the Heterologous Expression of Biosynthetic Gene Clusters for Marine Natural Products. *Mar Drugs* **20**, 341 (2022).

33. Meleshko, D. *et al.* BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Res.* **29**, 1352–1362 (2019).

34. Deep self-supervised learning for biosynthetic gene cluster detection and product classification | PLOS Computational Biology. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011162.

35. Gupta, V. K. *et al.* TaxiBGC: a Taxonomy-Guided Approach for Profiling Experimentally Characterized Microbial Biosynthetic Gene Clusters and Secondary Metabolite Production Potential in Metagenomes. *mSystems* **7**, e00925-22 (2022).

36. Wang, Z., Forelli, N., Hernandez, Y., Ternei, M. & Brady, S. F. Lapcin, a potent dual topoisomerase I/II inhibitor discovered by soil metagenome guided total chemical synthesis. *Nat Commun* **13**, 842 (2022).

37. Zhang, D. *et al.* A systematically biosynthetic investigation of lactic acid

bacteria reveals diverse antagonistic bacteriocins that potentially shape the human microbiome. *Microbiome* **11**, 91 (2023).

38. Beck, M. L., Song, S., Shuster, I. E., Miharia, A. & Walker, A. S. Diversity and taxonomic distribution of bacterial biosynthetic gene clusters predicted to produce compounds with therapeutically relevant bioactivities. *Journal of Industrial Microbiology and Biotechnology* **50**, kuad024 (2023).

39. González-Salazar, L. A. *et al.* Biosynthetic novelty index reveals the metabolic potential of rare actinobacteria isolated from highly oligotrophic sediments. *Microbial Genomics* **9**, 000921 (2023).

40. Georjon, H., Tesson, F., Shomar, H. & Bernheim, A. Genomic characterization of the antiviral arsenal of Actinobacteria. *Microbiology* **169**, 001374 (2023).

41. Williams, S. E. *et al.* Discovery and biosynthetic assessment of ‘Streptomyces ortus’ sp. nov. isolated from a deep-sea sponge. *Microbial Genomics* **9**, 000996 (2023).

42. Building Streptomyces albus as a chassis for synthesis of bacterial terpenoids - Chemical Science (RSC Publishing). https://pubs.rsc.org/en/content/articlelanding/2023/sc/d2sc06033g.

43. Moore, S. J., Lai, H.-E., Li, J. & Freemont, P. S. Streptomyces cell-free systems for natural product discovery and engineering. *Nat. Prod. Rep.* **40**, 228–236 (2023).

44. Choufa, C. *et al.* Prevalence and mobility of integrative and conjugative elements within a Streptomyces natural population. *Front. Microbiol.* **13**, (2022).

45. Redondo-Salvo, S. *et al.* Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat Commun* **11**, 3602 (2020).

46. Stubbendieck, R. M., Zelasko, S. E., Safdar, N. & Currie, C. R. Biogeography

of Bacterial Communities and Specialized Metabolism in Human Aerodigestive Tract Microbiomes. *Microbiology Spectrum* **9**, (2021).

47. Hirsch, P. *et al.* ABC-HuMi: the Atlas of Biosynthetic Gene Clusters in the Human Microbiome. *Nucleic Acids Research* **52**, D579–D585 (2024).

48. Bech, P. K. *et al.* Succession of microbial community composition and secondary metabolism during marine biofilm development. *ISME Communications* **4**, ycae006 (2024).

49. Chevrette, M. G. *et al.* Microbiome composition modulates secondary metabolism in a multispecies bacterial community. *Proceedings of the National Academy of Sciences* **119**, e2212930119 (2022).

50. Chase, A. B., Bogdanov, A., Demko, A. M. & Jensen, P. R. Biogeographic patterns of biosynthetic potential and specialized metabolites in marine sediments. *The ISME Journal* **17**, 976–983 (2023).

51. Holloway, K. A., Rosella, L. & Henry, D. The Impact of WHO Essential Medicines Policies on Inappropriate Use of Antibiotics. *PLoS One* **11**, e0152020 (2016).

52. Rogers Van Katwyk, S. *et al.* Government policy interventions to reduce human antimicrobial use: A systematic review and evidence map. *PLoS Med* **16**, e1002819 (2019).

53. 2021 AWaRe classification. Accessed on 28/03/2024. https://www.who.int/publications-detail-redirect/2021-aware-classification.

54. Huemer, M., Mairpady Shambat, S., Brugger, S. D. & Zinkernagel, A. S. Antibiotic resistance and persistence—Implications for human health and treatment perspectives. *EMBO Rep* **21**, e51034 (2020).

55. Jacob, M. E. *et al.* Opinions of clinical veterinarians at a US veterinary

teaching hospital regarding antimicrobial use and antimicrobial-resistant infections. *Journal of the American Veterinary Medical Association* **247**, 938–944 (2015).

56. Aslam, B. *et al.* Antibiotic Resistance: One Health One World Outlook. *Front. Cell. Infect. Microbiol.* **11**, (2021).

57. Halawa, E. M. *et al.* Antibiotic action and resistance: updated review of mechanisms, spread, influencing factors, and alternative approaches for combating resistance. *Front Pharmacol* **14**, 1305294 (2024).

58. Kasabova, S. *et al.* Antibiotic Usage Pattern in Broiler Chicken Flocks in Germany. *Front Vet Sci* **8**, 673809 (2021).

59. Shrestha, J., Zahra, F. & Cannady, J. Antimicrobial Stewardship. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2024).

60. Bozhüyük, K. A. J. *et al.* Evolution-inspired engineering of nonribosomal peptide synthetases. *Science* **383**, eadg4320 (2024).

61. Hansen, M. H. *et al.* Resurrecting ancestral antibiotics: unveiling the origins of modern lipid II targeting glycopeptides. *Nat Commun* **14**, 7842 (2023).

62. Hoogendoorn, K. *et al.* Evolution and Diversity of Biosynthetic Gene Clusters in Fusarium. *Front Microbiol* **9**, 1158 (2018).

63. Wu, D., Jiang, B., Ye, C.-Y., Timko, M. P. & Fan, L. Horizontal transfer and evolution of the biosynthetic gene cluster for benzoxazinoids in plants. *Plant Communications* **3**, 100320 (2022).

64. Page, A. J. *et al.* Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).

65. Whelan, F. J., Rusilowicz, M. & McInerney, J. O. Coinfinder: Detecting significant associations and dissociations in pangenomes. *Microbial Genomics* **6**,

e000338 (2020).

66. Salamzade, R. *et al.* zol & fai: large-scale targeted detection and evolutionary investigation of gene clusters. 2023.06.07.544063 Preprint at https://doi.org/10.1101/2023.06.07.544063 (2023).

67. Linard, B. *et al.* Ten Years of Collaborative Progress in the Quest for Orthologs. *Molecular Biology and Evolution* **38**, 3033–3045 (2021).

68. Uchiyama, I., Mihara, M., Nishide, H., Chiba, H. & Kato, M. MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons. *Nucleic Acids Research* **47**, D382–D389 (2019).

69. Yılmaz, T. M., Mungan, M. D., Berasategui, A. & Ziemert, N. FunARTS, the Fungal bioActive compound Resistant Target Seeker, an exploration engine for target-directed genome mining in fungi. *Nucleic Acids Res* **51**, W191–W197 (2023).

70. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research* **45**, W55–W63 (2017).

71. Cooper, G. M. The Complexity of Eukaryotic Genomes. in *The Cell: A Molecular Approach. 2nd edition* (Sinauer Associates, 2000).