# Genome Mining Tools for Secondary Metabolites in Bacteria

## Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Mehmet Direnç Mungan
aus Eskişehir/Türkei

Tübingen
2022

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

| | |
|---|---|
| Tag der mündlichen Qualifikation: | 21.11.2022 |
| Dekan: | Prof. Dr. Thilo Stehle |
| 1. Berichterstatterin: | Prof. Dr. Nadine Ziemert |
| 2. Berichterstatterin: | Prof. Dr. Kay Nieselt |
| 3. Berichterstatter: | Prof. Dr. Wolfgang Wohlleben |

# Abstract

As products of billions of years of evolution, secondary metabolites perform a wide range of activities ensuring the survival of organisms in competitive environments. These natural products synthesized by diverse living beings throughout the tree of life have been a valuable resource for many industrial applications. Specifically, in pharmaceutical ventures, natural products are used profoundly against cancer, pests and microorganisms. Peaked in the golden era of antibiotics, drug discovery against infectious diseases was mainly centered around natural products from fungi and bacteria. Consequently however, microbes have made impressive and frightening progress in gaining resistance against antimicrobials fueled by their improper usage. Coupled with the stagnation in discovery rates of novel natural products, antimicrobial resistance has become a destructive phenomenon damaging humanity financially and healthwise. To fight off such resistant microbes, it is of paramount importance that we find and produce novel secondary metabolites with antimicrobial features. With the vast improvements in sequencing technologies and analysis algorithms, we possess repositories swarming with "multiomics"-based data, ready to be mined. Now, a crucial thing to do is to enable the prioritization of such data for the subsequent processes in wet-lab applications.

In this thesis, I have built command line tools as well as web-based databases and pipelines to I) detect genes conferring antibiotic resistance in order to find promising biosynthetic gene clusters that might encode for novel antibiotics and II) prioritize target genes for genetic manipulation that could be used to increase the production of secondary metabolites.

# Kurzfassung

Als Ergebnis eines Milliarden Jahre langen Evolutionsprozesses, üben Sekundärmetabolite eine Vielzahl von Aktivitäten aus, die ihr Überleben in einer konkurrenzfähigen Umgebung sichern. Diese Naturstoffe, die von verschiedenen Organismen im gesamten evolutionären Stammbaum synthetisiert werden, sind eine wertvolle Ressource für viele industrielle Anwendungen. Insbesondere in der pharmazeutischen Industrie werden Naturstoffe in großem Umfang zur als Antibiotika, Antiparasitika, Immunsuppresiva oder zur Krebsbekämpfung eingesetzt. Die Entdeckung von Medikamenten gegen Infektionskrankheiten hat im goldenen Zeitalter der Antibiotika ihren Höhepunkt erreicht. Diese umfassten hauptsächlich aus Naturstoffe, die aus Pilzen und Bakterien gewonnen wurden. Infolgedessen haben Mikroben, angetrieben durch die falsche Verwendung von antimikrobiellen Wirkstoffen zunehmend, auf beeindruckende und beängstigende Art und Weise, Resistenzen gegen antimikrobielle Wirkstoffe entwickelt. Verbunden mit einem Rückgang bei der Entdeckung neuer Naturstoffe, hat sich die antimikrobielle Resistenz zu einem zerstörerischen Phänomen entwickelt, das die Menschheit in finanzieller und gesundheitlicher Hinsicht beschädigt. Um solche resistenten Mikroben zu bekämpfen, ist die Entdeckung und Produktion neuer Sekundärmetabolite mit antimikrobiellen Eigenschaften von größter Bedeutung. Durch die enorme Verbesserung der Sequenziertechnologien und Analysealgorithmen verfügen wir über eine Vielzahl von "MultiomicsDaten, die nur noch ausgewertet werden müssen. Nun kommt es darauf an, die Priorisierung dieser Daten für die nachfolgenden Prozesse im Labor zu ermöglichen.

In dieser/meiner Doktorarbeit habe ich command-line Tools sowie webbasierte Datenbanken und Pipelines entwickelt um I) Antibiotika-Resistenzgene zu detektieren und somit vielversprechende Biosynthese-Gencluster zu finden die für neue Antibiotika kodieren könnten, und II) Zielgene für genetische Manipulationen zu identifizieren, die für eine Erhöhung der Sekundärmetabolit-Produktion optimiert werden können.

# Acknowledgments

I would like to thank my advisors Prof. Dr. Nadine Ziemert and Prof. Dr. Kay Nieselt for their support, patience and guidance. It was a didactic and joyous ride for the past few years in Tübingen, most likely the nicest place I've ever been in Germany. I'm very grateful to all of my colleagues especially in the Ziemertlab and my co-authors as they have taught so much and helped me all along. Also, I would like to thank my committee members for taking the time and in their efforts evaluating this work. For their financial and computational support I would like to thank DZIF, BinAC and de.NBI.

My undying gratitude to my family both in Turkey and Germany. They have been enduring me for quite some time now. As dedication, this thesis is for anyone who somehow feels the need to read it.

# Contents

# Chapter 1

# Introduction

What does a newborn baby covered in the hot ashes of a plant, a man smoking the root of a tree and a woman applying milk-stewed fungi on her wound all have in common? A rather baffling question, asked by scientists for quite a long time. The path to illumination lay in the valley of the shadow of death since these seemingly unrelated acts were all done for the purpose of increasing the longevity of humankind and fighting against the deadly infections [1, 2, 3]. Even though it was known in the 17th century that microorganisms do exist [4], it was not proven until the late 19th century that they can cause such infectious diseases [5]. In the earlier days of scientific research, various organic compounds were synthesized such as Salvarsan [6] in order to fight the microorganisms and their infections however, the high amount of detrimental side effects they caused rendered their usage rather counterproductive [7]. In 1928 came the discovery of penicillin, an antibacterial compound naturally produced by the fungus *Penicillium chrysogenum* [8]. It was one of the biggest breakthroughs in antibiotic research however, penicillin could not be produced in high quantities until the '40s [9]. Afterward, with the successful application of penicillin, most pharmaceutical companies launched their own microorganism-based natural product discovery ventures [10]. This brings us to our initial question. In this so-called the "golden era" of antibiotics, researchers began to fully understand that the natural products were indeed responsible for a wide range of functionalities, most useful to humankind [11].

Throughout this chapter, I highlight the importance of natural products, review the challenges of the novel antibiotic discovery efforts and give an outline of my Ph.D. work for the improvement of the current state of the field.

## 1.1 Natural Products and Antibiotics

Produced as secondary (or specialized) metabolites, natural products are responsible for a wide range of biological activities, profoundly affecting organisms throughout the tree of life [12]. The various feelings we get, e.g. enjoyment, from a variety of products that humans use every day such as coffee, tobacco or cocaine is a direct cause of natural products and their derivatives [13, 14, 15]. Crucially, such bioactive compounds are also used as drugs functioning as antimicrobials, anticancers, statins, painkillers, etc. As estimated by Newman and Cragg, at least 3 out of 4 approved drugs for the past 40 years were derived from natural products. Concordantly, the best-selling drug for the past 25 years used to treat cardiovascular diseases was developed from a direct descendant of a fungal natural product [16, 17]. Since the mentioned “golden era”, antibacterial drugs have saved millions of lives however, pharmaceutical companies have gradually lost interest in the development pipelines [18]. One of the reasons is that developing antibiotics, much like any other medicine, is difficult, expensive and expected to become a failed attempt around 95% of the time [19, 20]. However, specifically for antibacterials, another reason is that the potential profits in terms of money rarely outweigh the risks. Thus creating the current stagnation period in the antibacterial drug research. This low profitability issue is mainly caused by the fact that the current antibiotics, some of which originated from natural products discovered more than half a century ago, are still usable to quite some extent [21]. The question is, for how long?

### 1.1.1 Houston, we have a problem

Even though antibacterial agents have had a revolutionary effect on medicine, they are also doomed to become ineffective given enough time [22]. This phenomenon, also known as antibiotic resistance, is simply the microorganism's ability to continue its activities in an environment treated with lethal doses of antibiotics [23]. Driven by various evolutionary processes such as mutation, organisms can naturally gain resistance and adapt to the environment through selective pressure [24]. However, inappropriate, inadequate or overuse of antibiotics whilst fighting infections as well as extensive antibiotic usage in agricultural applications, hasten the selective pressure process, in turn, stimulating antibiotic resistance [25].

As every action has a reaction, antibiotic resistance can be viewed as organisms' gene-specific way of response to the assassination attempts made on them. Since humans are not the only ones trying to kill microorganisms, the evolution of the resistance genes can be a reaction to the antibiotics made by other organisms or to the products made by the organism itself, creating a self-resistance mechanism [26]. Concordantly, resistance conferred by groups of beta-lactamases has been shown to originate way before the golden age of antibiotics, dating up to 2 billion years ago [27]. In order to effectively solve the problem at hand, we must first understand the modes of action of antibiotics and the resistance mechanisms as bacterial countermeasures.

#### The good

Antibiotics are mostly originated from natural products produced by bacteria and fungi and work through various modes of action (Table 1.1). Mainly, they block the essential processes, in turn inhibiting microbial growth or killing the microorganism, without killing the host [28]. This issue of specificity is one of the core challenges when it comes to developing an antibiotic. For example, one of the differences between the human cells and bacteria is that cells belonging to the latter are surrounded by a wall [29]. Beta-lactams such as penicillins disrupt the cell wall formation by interfering with the biosynthesis of its main content peptidoglycan [30]. Even though certain side effects can be seen through the usage of beta-lactams (e.g. allergy to penicillin exposure), the cell wall remains an antibacterial target with a low risk of impacting the mammalian host tissue [31].

### The bad

"The bear knows ways to flee as many as the traps of the hunter." This, thousands of years old Turkish proverb can be used the describe the fact that so far, the clinically used antibiotics were matched with the resistance mechanisms developed or acquired by the targeted bacteria [32]. When streptomycin was introduced in 1944 in order to fight "The Great White Plague", tuberculosis, it took only a few years for scientists to discover the mutant strains of *Mycobacterium tuberculosis* showing resistance to the therapeutic concentrations of the drug [33]. In the case of many antibiotics, the emergence of a resistant strain was discovered most of the time within the next decade [34]. Among many others shown in Table 1.1, a common tactic for negating the effect of an antibiotic on the host is pumping out the toxic compounds. For example, RND (Resistance-Nodulation-Division) family efflux pumps are known for their broad-spectrum substrate profiles, in turn, contributing to the multidrug resistance capabilities of Gram-negative bacteria [35].

### The ugly

When awarded his Nobel prize in 1945, Alexander Fleming cautioned: "There is the danger that the ignorant man may easily underdose himself and by exposing his microbes to non-lethal quantities of the drug make them resistant." Unfortunately, this presage has soon come to pass after years of worldwide use as *Staphylococci* strains resistant to penicillin emerged. To fight off these re-

Table 1.1: Commonly used antibiotic classes, producers and targets

| Class | Antibiotic | Producer | Target | Resistance Mechanism(s) |
|---|---|---|---|---|
| $\beta$-Lactams | Amoxicillin | *Penicillium chrysogenum** | Peptidoglycan biosynthesis | hydrolysis, efflux, target modification |
| Sulfonamides | Mafenide | Synthetic | Folate synthesis | efflux, target modification |
| Aminoglycosides | Kanamycin A | *Streptomyces kanamyceticus* | Protein synthesis: 30S ribosomal subunit | acetylation, efflux, target modification |
| Tetracyclines | Tetracycline | *Streptomyces aureofaciens* | Protein synthesis: 30S ribosomal subunit | efflux, target modification, oxygenation |
| Glycopeptides | Vancomycin | *Amycolatopsis orientalis* | Peptidoglycan biosynthesis | Reprogramming peptidoglycan biosynthesis |
| Macrolides | Erythromycin | *Saccharopolyspora erythraea* | Protein synthesis: 50S ribosomal subunit | Hydrolysis, glycosylation, phosphorylation, efflux, altered target |
| Lincosamides | Clindamycin | *Streptomyces lincolnensis* | Protein synthesis: 50S ribosomal subunit | Nucleotidylation, efflux, altered target |
| Streptogramins | Pristinamycin | *Streptomyces pristinaespiralis* | Protein synthesis: 50S ribosomal subunit | C-O lyase (type B streptogramins),acetylation (type A streptogramins), efflux, altered target |
| Oxazolidinones | Linezolid | Synthetic | Protein synthesis: 50S ribosomal subunit | Efflux, altered target |
| (Fluoro)quinolones | Ciprofloxacin | Synthetic | DNA synthesis: inhibition of DNA gyrase, and topoisomerase IV | Acetylation, efflux, altered target |
| Pyrimidines | Trimethoprim | Synthetic | Folate synthesis: inhibition of dihydrofolate reductase | Efflux, altered target |
| Ansamycins | Rifamycin SV | *Amycolatopsis rifamycinica** | Nucleic acid synthesis: RNA polymerase | ADP-ribosylation, efflux, altered target |
| Lipopeptides | Daptomycin | *Streptomyces roseosporus* | Cell membrane disruption | Altered target |

* Semi-synthetic antibiotic derived from the natural product of the producer strain.

sistant bacteria, the semi-synthetic antibiotic methicillin was introduced however, it took around 3 years for the methicillin-resistant *Staphylococcus aureus*

(MRSA) to emerge, renowned as the first superbug in history [36]. The fight between antibiotics and bacterial resistance went back and forth for decades, eventually resulting in the emergence of the ESKAPE pathogens. Comprised of the bacteria *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Enterobacter spp.*, these pathogens are equipped with extensive antimicrobial resistance (AMR) capabilities and are regarded as a fundamental threat by the United Nations [37, 38]. Furthermore, the World Health Organization recently recognized the urgent need for new antibiotics for the treatment of the infections caused by 12 bacterial species including ESKAPE pathogens [39]. In light of the information and the estimations we have on AMR so far, it is evident that we are in dire and urgent need of novel antibiotics.

### 1.1.2 Sources and traditional discovery of natural products

As mentioned earlier, various types of drugs were originated from natural products, and produced by organisms from different branches of the tree of life such as fungi, bacteria and plants [40]. Such small molecules, especially with antibacterial functions are mainly produced by bacteria [41]. In order to survive in complex ecosystems, microorganisms have evolved to compete with their neighbours. For example, the production of iron-chelating agents like siderophores allows microbes to scavenge valued metals for themselves from the environment [42]. Another type of competition involves killing or the impairment of other neighbouring organisms leading to the further growth and survival of the dominant bacterium [43]. Consequently, soil-dwelling terrestrial organisms were extensively studied by researchers in order to find natural products that can be used against bacteria. These efforts showed that members of the phylum Actinobacteria had a huge potential to produce antibiotics [44]. Specifically, the genus *Streptomyces* has been regarded as the richest source so far, inspiring more than 70% of the clinically used antibacterial drugs [45]. Furthermore, as recently shown by Gavriilidou and colleagues, organisms belonging to other taxa such as *Bacteroidota* or *Myxococcota* can also be considered as sources of diverse biosynthetic potential [46].

By the time of the golden era of antibiotics, more than half a century ago, searching for natural products was mainly centered around bioactivity-screening efforts (Figure 1.1). These so-called “top-down” approaches, begin with acquiring biological samples in order to isolate and cultivate a microorganism that can generate valuable natural products. However, it can prove extremely difficult to effectively cultivate a microbe in laboratory conditions, let alone force it to produce the metabolites it instinctively produces guided by various environmental effects [47, 48]. In order to isolate compounds using such a roadmap, researchers often try to mimic the settings of the natural environment for the microorganisms such as introducing external stresses or co-culturing with other species [49]. Uncommon methods are also applied. As seen from a research conducted by Cichewicz et al., using a commercial breakfast cereal “Cheerios” in media formulations, promoted the production of two novel compounds (a diarylcyclopentendione and a biphenyl metabolite) from *Preussia typharum* isolate [50]. Alternatively, another common effort is to collect the extracts directly from the sample, without any prior cultivation of a specific strain. Subsequently, numerous collected extracts are then subjected to a screening process, where the compounds are tested for their functionality (e.g. antibacterial activities) guided by bioassays. Provided that interesting bioactivity is observed through screening, present compounds are then isolated for their structural characterization. A number of advancements in sampling and cultivation efforts, High-Throughput Screening (HTS) methodologies and the characterization of metabolites have been described for the increased efficiency [51]. However, the traditional bioactivity-guided approach commonly suffers from the facts that the total workflow is laborious, time-consuming and quite susceptible to rediscoveries of the already described compounds. Since it is evident that resources can’t be thrown away for high-risk, low-reward initiatives, drug discovery efforts are often guided or complemented by the computational methods which are continuously improved each day [52].

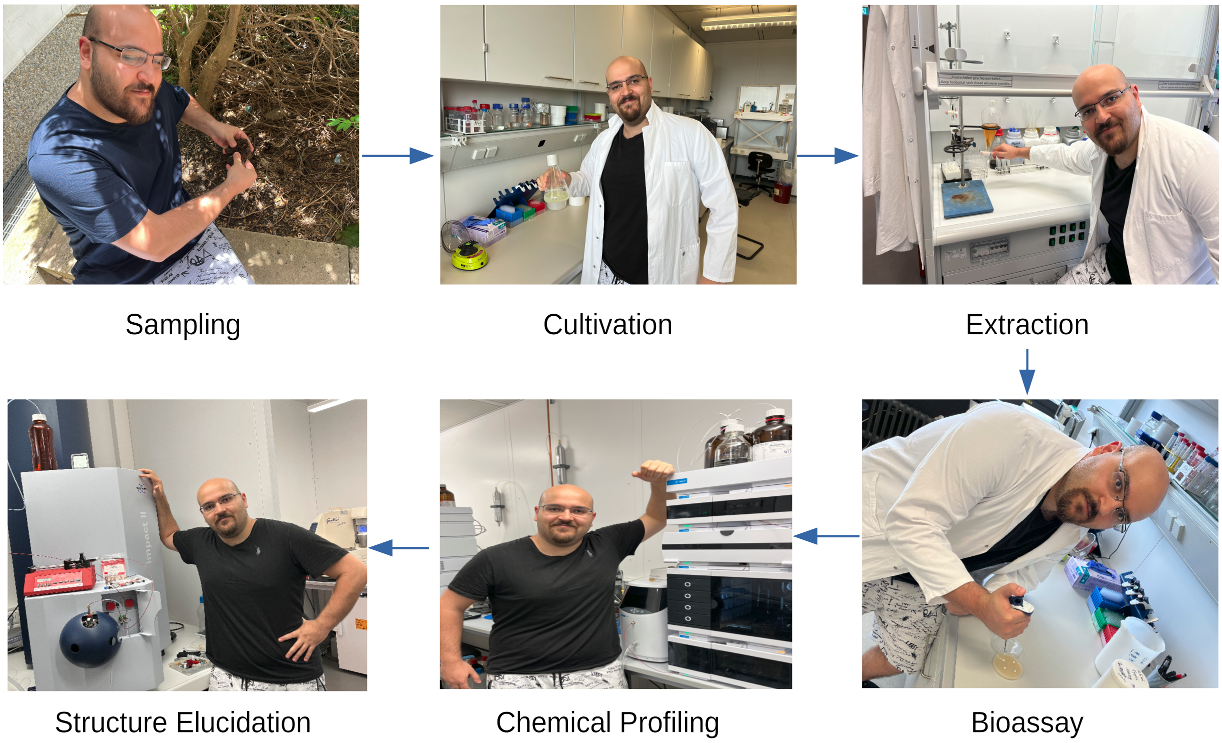


Figure 1.1: Illustration of the traditional natural product discovery process. After the acquisition of samples, one can try to cultivate the microbes or directly extract the contents. The extracts are then subjected to bioassays to investigate bioactivity. Afterward, if an interesting activity is observed, the final steps are obtaining a high amount of the compound and elucidating its structure (e.g. by using the liquid chromatography–mass spectrometry technique).

## 1.2 Enter Informatics

Much like any other industry, biotechnology-based industries have been mobilizing toward digitalized environments, since the beginning of the Information Age. With Frederick Sanger's work on discovering amino acid sequences of insulin from different hosts in the 1950s [53], it soon became evident that sequence databases were needed to store valuable information. In the mid-1960s

Margaret Oakley Dayhoff enabled the usage of computerized systems in order to store and analyze amino acid sequences, pioneering the field of bioinformatics [54, 55, 56]. Later on, making use of the DNA-sequencing methods, scientists began to generate large nucleotide sequences, even a full genome of a bacteriophage in 1977 [57]. A year later, computational algorithms were used on such nucleotide sequences in order to search for patterns through the usage of Markov chains [58]. By 1979, several tools were developed in order to determine the molecular structures of natural products [59]. Now, after decades full of technological advancements, computer-guided "bottom-up" approaches are crucial to the discovery of natural products and the design of novel drugs [60].

### 1.2.1 Genetics of natural products - *to find them all* -

As mentioned, *Streptomyces* species serve as a treasure trove for useful natural products. Unsurprisingly, belonging organisms such as *Streptomyces coelicolor*, also served as important bacteria paving the way for understanding the mechanisms of the biosynthesis of bioactive compounds [61]. In particular, David A. Hopwood and his colleagues extensively investigated the bacterium, and showed that the *S. coelicolor* strain A3(2) is potentially a rich source of antibiotics, producing methylenomycin A and actinorhodin [62, 63]. Further decoding the genetics of the biosynthesis, he reported that the enzymes necessary for the production of these compounds were encoded by the genes closely positioned together, forming a cluster [64]. Eventually, these genes in close proximate to each other were termed Biosynthetic Gene Clusters (BGCs) (Figure 1.2). Apart from the main biosynthetic enzymes, it soon became evident that BGCs incorporate genes with other functionalities. For example in the novobiocin production mechanism, it was shown that tailoring enzymes such as glycosyltransferases were responsible for the introduction of deoxysugars to the compound, shaping its biological activity [65]. Certain regulators are also commonly observed within the BGC, controlling when or if the compound should be produced, depending on various factors such as environmental stress [66]. Furthermore, these clusters are often accompanied by transporter encoding genes to pump the product out of the cell and resistance genes for the protection of the organism against its own medicine [67, 68].

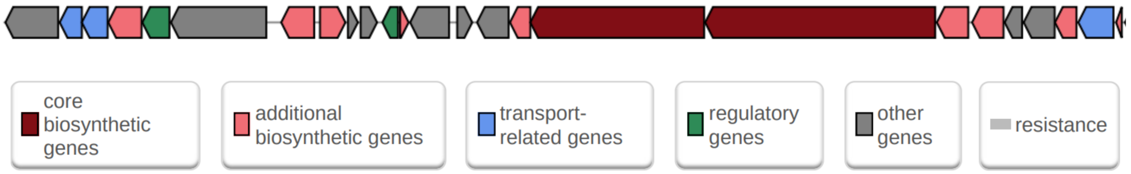


Figure 1.2: An example PKS type BGC predicted by antiSMASH. Once the core domains for a type (herein KS and AT domains) are detected, genes with such domains are labeled as "core biosynthetic genes". Afterward comes the "extension" step, where the algorithm looks for "additional biosynthetic genes" e.g., tailoring enzymes. Additional gene types are also searched, such as genes regulating the proposed cluster or transporting the final compound, leading to the final formation of the predicted BGC boundaries.

Understanding the organization of BGCs, coupled with their highly conserved machinery, greatly improved our ability to detect various classes of secondary metabolites. Many BGC classes exist such as terpenes, bacteriocins, ribosomally and post-translationally modified peptides (RiPPs) however, non-ribosomally synthesized peptides (NRPs) and polyketides (PKs) are the popular types since they are known for their diverse structures and uses in pharmaceutical applications. The variety of such structures is explained through their architectural design of so-called "assembly lines" [69]. For example, in the case of PK production, the organism needs certain primary metabolite monomers activated by thioesters such as malonyl-CoA for the initiation phase. After that, the elongation of the assembly chain commences by the linkage of the substrates and intermediates. Throughout the elongation of the assembly line, further post-translational modifications take place which creates a specialized functionality of the secondary metabolite. In the end, the assembly line is finalized by terminating enzymes, playing an important part in the tailoring of the final product and its release to make use of its bioactivity [70].

As in the above example, these structural transformations and final termination are catalyzed by polyketide synthase (PKS) modules with specific functional domains for each enzymatic reaction. Domains like acyltransferase (AT) and the acyl carrier protein (ACP) are involved in the starting point of the process, while domains such as ketoreductase (KR) and ketosynthase (KS) are involved in tailoring the structure at each step [71]. Obviously, this is overly-summarized information about one BGC class, but it provides an important point. With each solved part of the puzzle in the biosynthesis of these compounds, we are one step closer to finding and (re)engineering BGCs, increasing their production, finding useful homologue sequences of key enzymes, exploring the diversity of BGCs leading to novel discoveries and devise many more potential avenues in order to make use of new natural products. As our understanding of the biosynthesis of various BGC classes evolved, increasingly accurate bioinformatic tools to detect, characterize and (over)produce their corresponding products soon followed.

### 1.2.2 The age of "-omics" *- to rule them all -*

Even before the completely sequenced genomes of two bacteria in 1995, sequencing and manipulation of biosynthetic genes have played crucial roles in the discovery and production of natural products. Cloning a whole BGC in 1984, Malpartida and Hopwood were able to produce actinorhodin in the heterologous host of *Streptomyces parvulus* [64]. With the increased amount of generated DNA sequences coupled with the improvements in genetic engineering techniques, researchers were able to further determine the diversity of core genes required for the synthesis of natural products from a variety of microorganisms, paving the way for the creation of bioinformatics tools capable of mining genomes [72].

Like many fields, natural product research was also revolutionized by the advent of Next Generation Sequencing (NGS). With the rapid increase of bacterial (meta)genomes in public repositories and the amount of identified BGCs not correlated to known natural products, genome mining-based discovery gained profound popularity [73]. Such genomics-based workflow starts with the detection (and prioritization, as discussed in the later sections) of BGCs accompanied by a de-replication process in order to avoid nonnovel products [74]. Initially, even before the full impact of NGS-based technologies, detection workflows were highly dependent on applying the Basic Local Alignment Searching Tool (BLAST) [75] on DNA and amino acid sequences in order to find locations of core biosynthetic genes such as non-ribosomal peptide synthetase (NRPS) in genomes of interest. This method of BGC detection worked relatively well for several years, aided by a number of automated analysis tools like DECIPHER or NRPS-PKS [76, 77]. However, especially with the updated algorithms belonging to the HMMER3 package and the generation of Hidden Markov Model (HMM) profiles of conserved domains of core biosynthetic enzymes, BGC detection workflows shifted towards pHMM-based searching methods [78]. A number of prediction tools using pHMMs were made available to researchers for the analysis of fungal and bacterial organisms [79]. First published by Medema et al. in 2011 [80] and continuously improved, antiSMASH enabled the prediction of several classes of BGCs in one web server. Now in its sixth iteration [81], antiSMASH is widely considered the gold standard as it has analyzed more than 1.2 million input sequences online, achieving $\approx$7500 citations in total as of June 2022.

As pointed out, understanding the mechanisms of various aspects of natural products (synthesis, modes of action, structure, etc.), greatly enhances our ability to develop and further improve the discovery pipelines. In that regard, another powerful method in the genomics toolbox is the phylogenetic analysis of sequences of interest. With the growing genetic data on BGCs and their hosting genomes, phylogeny-based approaches have become crucial since they can be used to infer the diversity and prediction of BGCs, as well as the evolutionary processes driving BGC development in microorganisms such as horizontal gene transfer (HGT) or duplication [82]. Driven by these principles, EvoMining was created to search for enzymes functioning in primary metabolic pathways which have gone through "gene expansion" phases and conduct subsequent phylogenetic analysis to detect clusters that might have diverged from primary metabolism [83]. Another widely used method, with metagenomic data which is comprised of DNA harboured from environmental samples, is to screen input sequences for domains of interest to detect phylogenetic markers which can be used to track down a cluster's evolutionary path [84]. Using such a roadmap, Brady and his colleagues were able to discover structurally diverse novel secondary metabolites; antibiotics such as Fasamycin A [85] and anticancers such as Arimetamycin A shown to be more potent than natural Anthracyclines used in clinics [86]. A variety of pipelines have been created in order to aid researchers in phylogeny-guided (meta)genome mining. Tools such as NapDos [87], FunOrder [88] and eSNaPD [89] can be used for the detection of biosynthetic genes and assessing their diversity, while CORASON and BiG-SCAPE can be used to investigate evolutionary links between a large scale of BGCs and their sequence similarity networks [90].

Once an interesting BGC is detected based on its novelty and potential function, an important step is to promote its expression. As organisms do not want to spend the time, energy, or any precious precursor compound on an unnecessary metabolite, many BGCs stay in a "silent" mode, waiting to be activated when need be [91]. One efficient way to identify the product of a corresponding BGC is its heterologous expression in a genetically tractable host [92]. Since it is estimated that at least 90% of bacterial organisms can't be cultured in laboratory conditions, moving specific biosynthetic genes or a complete BGC to another host (e.g. *Escherichia coli* for bacteria or *Saccharomyces cerevisiae* for fungi) for metabolite production is shown to be a fruitful approach [93]. A variety of novel antibacterial compounds are described using this method including a sulfo-glycopeptide by Owen et al. [94], scleric acid by Fabrizio et al. [95] and several more by the Brady group, guided by metagenomic approaches, such as malacidin and borregomycin (which also shows anticancer bioactivity) [96, 97].

Another strategy for the activation of a silent cluster involves focusing on the expression of the BGC in its natural strain termed native host expression. Initially, upon the discovery of an interesting BGC which is naturally evolved at some point of the bacterium's evolutionary timeline for a specific purpose such as increasing its survivability when facing specific environmental stress, this method mostly relies on genetically engineering the host or applying different culturing conditions to induce expression [98]. In a study by Hertweck and his colleagues, various nutrients, chemicals and stress triggers were applied to the culturing conditions of an anaerobic bacterium *Clostridium cellulolyticum* in order to induce secondary metabolite production, but to no avail. However, when they added an aqueous soil extract, in an effort to mimic its natural habitat, they discovered closthioamide, exhibiting antibacterial activity [99]. By inducing mutations on genes *RPSL* and *RPOB* (encoding ribosomal protein s12 and RNA polymerase $\beta$-subunit, respectively), Takeshi et al. were able to discover a novel antibiotic piperidamycin from its native producer *Streptomyces mauvecolor* [100].

An invaluable approach to studying silent clusters, which is often used to complement genomics-based strategies, is mining the transcriptome of a potential BGC producer organism. Increasing our knowledge about the regulatory mechanisms of BGCs continuously improves our ability to discover and produce new compounds. Using microarrays, early examples of transcriptome mining led to the improvement of genetic engineering strategies [101] and the detection of specific genes for the overproduction of natural products [102]. Developed more than a decade ago, RNA-Sequencing (RNA-Seq) was hailed as a method that revolutionized the field of transcriptomics. This method is widely used to take a "snapshot" of the available mRNA sequences of the current state of a sample, which is then reversely transcribed to cDNA fragments, afterward subjected to High-Throughput Sequencing (HTS) to generate DNA sequences (reads) that can be used to infer expression level of corresponding genes [103]. Especially with the continuous improvements of the algorithms used in the subsequent data analysis workflows, RNA-Seq has also become a crucial part of the discovery and (over)production of secondary metabolites. An extensive comparative transcriptomic study by Amos et al. on four strains of marine *Salinispora* linked predicted BGCs to their (novel) products, also describing their putative functionalities [104]. In one of the newest examples using the RNA-Seq approach, improvement of erythromycin yield by 71% is achieved by simultaneous overexpression of the genes *sucB* and *sucA* which encode enzymes for the provision of precursor compounds [105].

Starting from a mere DNA sequence of a genome to the final characterization of a new compound, each step of the entire multi-omics analysis yields a piece of information most valuable for further research. Consequently, an important component of natural product research is finding data of interest.

### 1.2.3 Data sources *- to bring them all -*

With the technological advancements in sequencing methods coupled with the developments of data mining algorithms, databases have become a crucial part of natural product discovery and production efforts. There are dozens of databases that may prove useful to natural product research [106] however, several of them are mentioned for the purposes of this thesis. Being one of the oldest and most comprehensive data sources, The National Center for Biotechnology Information (NCBI) hosts a variety of databases holding a large amount of sequence data representing organisms from all around the tree of life. With a non-stop increase over the years (approximately gets bigger twice in size every 18 months), NCBI's GenBank database [107] contains more than 1.1 million bacterial genomes of which $\approx$32500 are tagged as complete as of June 2022. However, with the great number of sequences, comes great redundancy. First introduced in 2000, NCBI's RefSeq database was designed to address this problem by offering manually curated non-redundant sequences from GenBank, that would only keep the-up-to date version of each entry. Covering 68260 bacterial species, RefSeq contains ≈27000 bacterial complete genomes [108]. Additionally, repositories such as European Nucleotide Archive and Sequence Read Archive (maintained by the European Bioinformatics Institute and NCBI, respectively) provide HTS data from a wide range of organisms and sequencing platforms [109, 110]. Operable through multiple programming access tools, such databases enable the exchange of a large amount of HTS-generated reads. Acquisition of reads of interest makes a number of research objectives possible such as the reassembly of the (meta)genomes for specific purposes [111], generating RNA-Seq based transcriptomic profiles from multiple samples of an organism and even the reconstruction of BGCs[112].

With the logarithmic increase of genomic data and the illumination of BGC biosynthesis and formation, it became evident that BGC-centered databases were needed. Mainly focusing on NRPS and PKS types, early databases include doBISCUIT [113], ClusterMine360 [114] and StreptomeDB [115] with the latter focusing specifically on *Streptomyces* strains. Taking advantage of the second version of antiSMASH [116] and another BGC prediction tool, ClusterFinder [117], which in contrast to antiSMASH was built to include the prediction of the BGCs belonging to unknown classes as well, Hadjithomas and his colleagues developed the Atlas of Biosynthetic Clusters within the Integrated Microbial Genomes system (IMG-ABC) [118]. Initially, the database held over a million predicted BGCs however, with the updated version [119], the authors abandoned the "unknown territory" and used only the fifth version of antiSMASH resulting in more than half a million drop in total predicted BGC count but with higher confidence. Additionally, now in its third iteration, the antiSMASH-DB provided researchers with an easy-to-use database allowing the construction of comprehensive queries to search BGCs by their types, including domains, most similar MIBiG clusters, etc. Also, they dereplicated the genomes by filtering them based on their average nucleotide identity (%99.6 similarity cutoff), resulting in 165084 predicted BGCs from bacteria [81]. Given the high amount of BGCs only labelled as "predicted", standardization and evaluation of BGCs have essentially become an important task. In 2015, a collaborative effort of scientists addressed this issue by creating the Minimum Information about a Biosynthetic Gene Cluster (MIBiG), offering the community a repository of manually curated, experimentally characterized set of BGCs, linking them to their produced compounds [120]. Since most of the mentioned (and many more) databases and prediction tools concerning natural products are consistently improved, Secondary Metabolite Bioinformatics Portal was introduced as a community-updated web server offering a "catalog" of major tools used in natural product research [121].

### Prioritization

As is the major point of this thesis so far, each day comes with another scientific work that improves our knowledge of natural products, brings another technological development, and one more batch of data to be mined. Even though there are millions of public or private predicted BGCs, the majority of them remains orphan with a small fraction being experimentally verified and linked to its compound. With the immense wealth of information generated by "omics" approaches, prioritization of such data has become crucial to reducing the costs of time, money and resources for further wet-lab applications [122]. To date, several methods were designed to prioritize BGCs and strains for the discovery of novel compounds. For example, some tools use chemical structure information to infer the enzymatic mechanism of natural product biosynthesis [123], some combine metabolomic and genomic data [124, 125, 126], and some use genome mining methods to find promising strains which can possess a rich and diverse source of natural products [127, 128]. Another prioritization method relies on the fact that any bacterium that creates an antibiotic, has to encode a form of self-defense mechanism to avoid suicide [129]. In one of the mechanisms providing self-resistance, the BGC may include a duplicated or modified version of a housekeeping gene that acts as a resistant target to the encoded product. Key evolutionary mechanisms to this process are discussed in detail in publications 1-3 [130, 131, 132]. Screening such genes and BGCs, this so-called target-directed genome mining (TDGM) or self-resistance based genome mining approach allowed the prioritization and in turn, the discovery of many natural products from both bacterial and fungal organisms in the past decade [133]. However, none of the mentioned methods are focusing on increasing the yield of a compound encoded by the predicted BGC. To that aim, an invaluable method is to alter the expression of genes playing a pivotal role in the production of natural products (discussed in detail in publication 4). A high number of genes might play such a role however, with the use of comparative transcriptomic approaches, researchers were able to prioritize genes for manipulation purposes which led to the overproduction of known compounds as well as the discovery of novel metabolites [134].

## 1.3 Aim of the thesis - *to bind them all* -

The main objectives of this work can be summarized in two parts: designing tools for the discovery of novel BGCs and the (over)production of their corresponding compounds.

For the discovery part:

- With ARTS 2.0, we have extended the target-directed based prioritization methodologies to encapsulate the entire bacterial kingdom, introduced functionalities for the comparative mining of input genomes
- With SYN-View, we proposed a phylogeny-based method to increase the efficiency of TDGM approaches by investigating the synteny of clusters of interest amongst the closest relatives of the input genomes
- ARTS-DB was created to enable easy exploration of TDGM mechanisms throughout the bacterial kingdom

For the (over)production part:

- We have created SeMa-Trap to allow for promising experimental design and RNA-Seq based transcriptome mining in order to find promising target genes for the (over)expression experiments of BGCs of interest

# Chapter 2

# Publications

## 2.1 ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining

### Contributions

With valued discussions and guidance from Prof. Dr. Nadine Ziemert, I did everything concerning the new reference sets making ARTS available for the whole bacterial kingdom, updating the incorporated algorithms and the base code, testing and validating the updated pipeline, writing the original paper draft and carrying the server to de.NBI cloud services. Dr. Mohammad Alanjary was responsible for all of the visualization of the results and adapting the "front-end" to the updated pipeline. All the authors spent time reviewing and editing the final manuscript.

*W546–W552* *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* *Published online 19 May 2020*
*doi: 10.1093/nar/gkaa374*

# ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining

**Mehmet Direnç Mungan[1,2,†], Mohammad Alanjary[3,†], Kai Blin[4], Tilmann Weber[4], Marnix H. Medema[3] and Nadine Ziemert[1,2,*]**

[1]Interfaculty Institute of Microbiology and Infection Medicine, University of Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany, [2]German Centre for Infection Research (DZIF), Partner Site Tübingen, Germany, [3]Bioinformatics Group, Wageningen University, Droevendaalsesteeg 1, 6708PB Wageningen, the Netherlands and [4]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet Bygning 220, 2800 Kgs. Lyngby, Denmark

Received February 28, 2020; Revised April 19, 2020; Editorial Decision April 29, 2020; Accepted April 29, 2020

## ABSTRACT

**Multi-drug resistant pathogens have become a major threat to human health and new antibiotics are urgently needed. Most antibiotics are derived from secondary metabolites produced by bacteria. In order to avoid suicide, these bacteria usually encode resistance genes, in some cases within the biosynthetic gene cluster (BGC) of the respective antibiotic compound. Modern genome mining tools enable researchers to computationally detect and predict BGCs that encode the biosynthesis of secondary metabolites. The major challenge now is the prioritization of the most promising BGCs encoding antibiotics with novel modes of action. A recently developed target-directed genome mining approach allows researchers to predict the mode of action of the encoded compound of an uncharacterized BGC based on the presence of resistant target genes. In 2017, we introduced the 'Antibiotic Resistant Target Seeker' (ARTS). ARTS allows for specific and efficient genome mining for antibiotics with interesting and novel targets by rapidly linking housekeeping and known resistance genes to BGC proximity, duplication and horizontal gene transfer (HGT) events. Here, we present ARTS 2.0 available at http://arts.ziemertlab.com. ARTS 2.0 now includes options for automated target directed genome mining in all bacterial taxa as well as metagenomic data. Furthermore, it enables comparison of similar BGCs from different genomes and their putative resistance genes.**

## INTRODUCTION

Due to the continuous increase of drug-resistant bacteria, antibiotic resistance is regarded as a global public health threat (1). The lack of new antibiotics with novel modes of action in the current drug development pipeline, makes finding new compounds to fight off resistant pathogens a critical task (2). Since the discovery of penicillin, secondary metabolites (SMs) produced by various living organisms have been foundational to the development of antimicrobial drugs (3). The majority of antibiotic compounds are isolated as natural products, from fungi and bacteria (4). For many decades, screening biological samples for desired bioactivity has been the traditional methodology for natural product discovery (5). Due to the high rediscovery rates and labor-intensive nature of the process, *in silico* methods have become a promising way to guide modern drug discovery efforts (6,7). Gene-centered methods, such as genome mining, enable researchers nowadays to computationally detect the biosynthetic gene clusters (BGCs) encoding enzymes necessary for the biosynthesis of antibiotics and predict encoded compounds (8). Over the last decade, greatly improved genome mining tools such as antiSMASH (9), EvoMining (10), PRISM (11) or DeepBGC (12) use methods like Hidden Markov Models, phylogeny or deep learning to highlight a variety of natural product classes. Combined with databases such as MIBiG (13), Natural Product Atlas (14) and the antiSMASH database (15), these tools allow for fast and efficient mining and dereplication of thousands of bacterial genomes and BGCs. According to the latest version of the Atlas of Biosynthetic Gene Clusters (IMG-ABC) (16) there currently are ∼400 000 predicted BGCs sequenced. Moreover, <1% of total clusters are experimentally verified, which leads to an important question:

[*]To whom correspondence should be addressed. Tel: +49 7071 2978841; Email: nadine.ziemert@uni-tuebingen.de
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Which of these clusters should be further examined with wet lab experiments?

Recently, researchers adopted a prioritization approach for antibiotic discovery that is based on the observation that antibiotic producers have to be resistant against their own products to avoid suicide (17). This so called target-directed or self resistance based genome mining approach allows the prediction of the mode of action of the encoded compound of an uncharacterized BGC based on resistance genes, in some cases co-located within the antibiotic BGC (18). Multiple resistance mechanisms exist, such as inactivation and export of antibiotics as well as target modification. In the latter case, a duplicated and antibiotic-resistant homologue of an essential housekeeping gene is detectable within the antibiotic BGC and allows the prediction of the mode of action of the encoded compound even without knowing a chemical structure (19–21). Moore *et al.*, for example, were able to identify a fatty acid synthase inhibiting antibiotic by screening for duplicated fatty acid synthase genes within orphan BGCs (22).

In 2017, we introduced the first version of the 'Antibiotic Resistant Target Seeker' (ARTS) (23), a user-friendly web server that automates target-directed genome mining to prioritize promising strains that produce antibiotics with new mode of actions. Since a resistant copy of the antibiotic target gene is typically detectable in the genome, can be observed within the BGC of the antibiotic and horizontally acquired with the BGC (23), ARTS automatically detects possible resistant housekeeping genes based on three criteria: duplication, localization within a biosynthetic gene cluster, and evidence of Horizontal Gene Transfer (HGT). One previous limitation of the ARTS pipeline was its focus on actinobacterial genomes. Although natural product discovery historically was highly focusing on the phylum Actinobacteria, prominent families from other phyla such as Proteobacteria or Firmicutes are known to have high natural product biosynthetic potential (24–26). Here, we introduce a greatly improved version 2 of the ARTS webserver, now allowing the analysis of the entire kingdom of bacteria, metagenomic data, and the comparison of multiple genomes. This update therefore will facilitate natural product prioritization and antibiotic discovery efforts beyond actinomycetes.

## NEW FEATURES AND UPDATES

The workflow of the ARTS pipeline involves a few key steps: First, query genomes are screened for BGCs using antiSMASH (9). At the same time essential housekeeping (core) genes within the genome are determined using TIGRFAM models that have been identified by comparing a reference set of similar genomes (27) (Figure 1B). During the next steps the identified core and known resistance genes are screened for their location within BGCs. Duplication thresholds are determined for each core gene model, based on their respective frequencies among the reference set. Finally, possible HGT events are detected via phylogenetic screening with the help of constructed species trees and gene trees. All the results are summarized into interactive output tables.

### Reference sets of organisms and core genes

Since the determination of core gene content and the construction of phylogenetic trees is more specific and accurate when query genomes are compared with genomes from similar organisms, we aimed to generate phylum specific reference sets. However, since the number of genomes in the different phyla varied significantly, reference sets were sometimes also created by class or a group of closely related phyla (Supplementary Table S1).

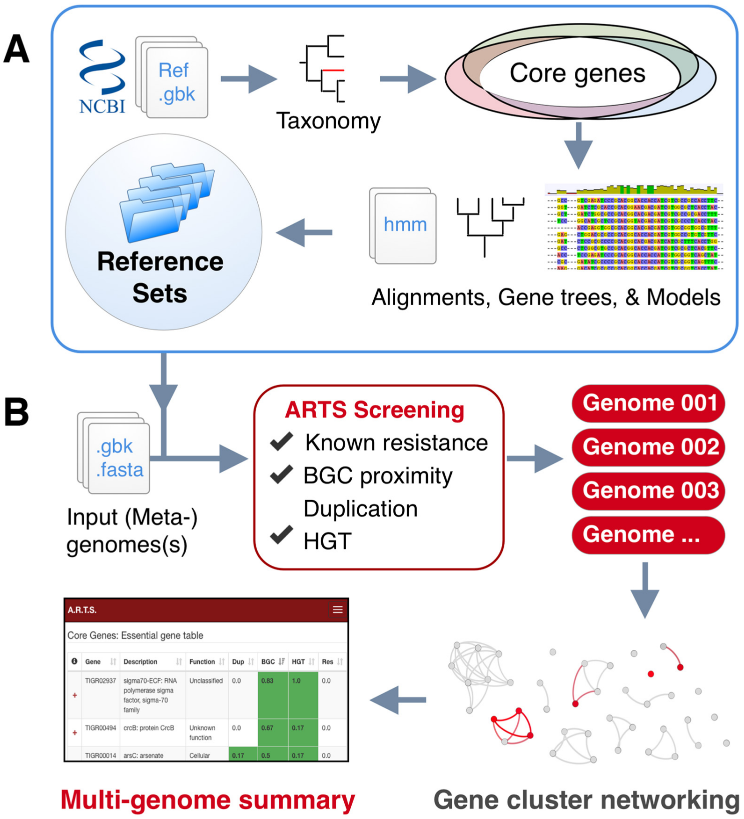
In a first step, sequences of all classified bacteria were downloaded through NCBI's RefSeq database (28) for further evaluation (Figure 1A). Redundant sequences were filtered with MASH (29) with a +95% similarity cut off. Where applicable, only complete genomes were used in a reference set. If the number and diversity of complete genomes within a phylum was not sufficient (distributed among a genus or two with <100 sequences), contig-level assemblies were also taken into consideration to expand the particular reference. Around 330 genome sequences were used for the creation of each individual reference set, which sum up to 4936 genomes in total.

Based on the number of genomes for each reference set, different boundaries were then selected for phyla with different levels of diversity. Given the diversity and large number of proteobacterial genomes deposited in Refseq (30), four different reference sets were created for proteobacterial genomes (Alpha, Beta, Gamma, Delta-Epsilon). In cases where a phylum does not comprise sufficient sequenced genome sequences (less than 100 genomes), multiple phyla were grouped into one reference set. In that way, 22 phyla were grouped into three reference sets. Groupings were based on phylogenetic distances in the tree of life (31) and the NCBI Lifemap (32). Another feature of the grouped sets is the high coverage of bacteria from harsh environments, allowing the analysis of extremophiles. For example, group 2, which was created from 214 organisms, is mainly comprised of the phyla Thermotogae and Chloroflexi (Supplementary Table S1), which are known to be mostly thermophilic (33,34).

### Reference set and core gene analysis

*Determination of core genes.* Core genes were determined for each reference set using the method developed for the previous version of ARTS (23). Subsequently, the core genes from each set were compared with sequences from the Database of Essential Genes (DEG)v 1.5 (46). On average, 85% of genes had a match to one or more records (Supplementary Table S2). The majority of the genes that are not found in DEG belong to the gene categories 'unclassified', 'unknown function' or 'energy metabolism'. Furthermore, functional classification of each reference set revealed that, on average, genes with functions such as protein and amino acid synthesis, energy and metabolism were the most abundant as would be expected from essential genes (Supplementary Figure S1). The importance of individual reference sets is highlighted by the fact that one set only accounts for ∼40% of the total unique core genes from all sets (Supplementary Table S4).

Additionally, the reliability of the generated gene trees for each reference set were estimated by branch support (Sup-



A.R.T.S.

Core Genes: Essential gene table

|  | Gene | Description | Function | Dup | BGC | HGT | Res |
|---|---|---|---|---|---|---|---|
| + | TIGR02937 | sigma70-ECF: RNA polymerase sigma factor, sigma-70 family | Unclassified | 0.0 | 0.83 | 1.0 | 0.0 |
| + | TIGR00494 | crcB: protein CrcB | Unknown function | 0.0 | 0.67 | 0.17 | 0.0 |
|  | TIGR00014 | arsC: arsenate | Cellular | 0.17 | 0.5 | 0.17 | 0.0 |

**Figure 1.** Outline representation of the ARTS pipeline. (**A**) Basic machinery of creating reference sets. Housekeeping core genes and duplication thresholds are detected per clade of organisms and gene alignments and trees are created for fast HGT detection. (**B**) Workflow with multi-genome comparative analysis. Input data is screened for ARTS selection criteria. All found BGCs are then subjected to BiG-SCAPE clustering algorithm. Finally, interactive output tables are presented for comparative analysis.

plementary Figure S2) and comparison to taxonomically correct species trees generated by the Accurate Species TRee ALgorithm (ASTRAL) (47) (Supplementary data).

*Positive controls and detection frequencies.* In order to test ARTS' ability to detect resistant targets in non-actinobacterial genomes using the new reference sets, we analyzed known examples of self-resistance mechanisms. We identified several known non-actinobacterial examples as positive controls (Table 1). Out of 11 antibiotic natural products with identified resistance mechanisms, five of them had available genome sequences regarding specific isolates that contained respective BGCs. All of these cases showed at least two ARTS hits when run in normal mode with default cutoffs. To detect the *accA* gene, a known transferase, exploration mode had to be used. Otherwise, ARTS 2.0 predicted resistance genes in almost all control BGCs except one. The CoA reductase resistant gene was not detected because specific CoA reductase models were missing in both the core and known resistance set. We also analyzed ∼5000 genomes belonging to all reference sets for statistical evaluation (Supplementary Table S3). On average, only one gene model shows positive hits for three or more ARTS criteria. Also, most of the core genes from the respective sets are found in each analyzed genome. Around 2–5% of core genes are highlighted for each criterion. The percent of core genes that went through HGT is in conformity with the HGT estimate levels in the literature (48,49).

### Reference sets for metagenomic data

Since metagenomic approaches are becoming increasingly popular in natural product research (50,51), submissions of whole metagenomes to the ARTS webserver are also showing a significant increase. Therefore, we have built an additional reference set available for metagenome analysis, which does not include phylogeny and duplications. Given that metagenomes are usually quite diverse and comprise more than one single phylum, core genes are defined as genes belonging to the Database of Essential Genes (DEG) (Supplementary Table S3).

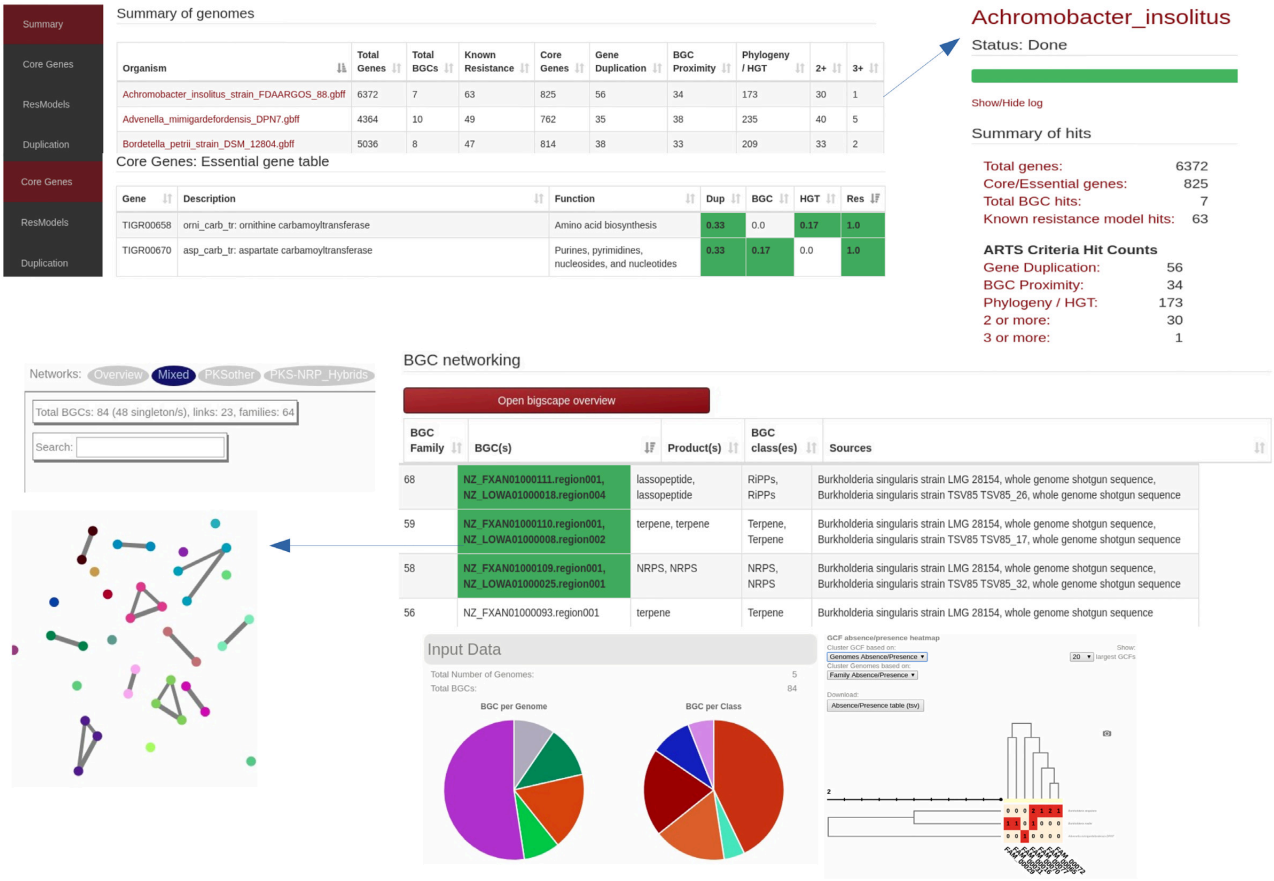
### Comparative analysis

ARTS 2.0 now makes it easier for users to analyze multiple genomes and applies a comparative analysis of provided organisms (Figure 2). Throughout the analysis, individual ARTS results are accessible upon completion of each run. Once all the sequences of interest are analyzed, an interactive summary table representing all genomes with each resulting criterion is provided. In addition, shared core genes with their respective hits and their observed frequen-

Table 1. Default ARTS analysis for positive examples of genomes and BGCs with known self-resistance mechanisms

| Product | Resistance gene | Organism | ARTS hits | Criteria hits (>2, >3) | Genes (core, total) |
|---|---|---|---|---|---|
| Thiocillin | ribosomal protein L11(35) | *Bacillus cereus* ATCC 14579 | D,B,P | 9,1 | 472, 5231 |
| Myxovirescin | *lspa*: signal peptidase II(36) | *Myxococcus xanthus* DK 1622 | D,B,P | 15,2 | 372, 7267 |
| Thailandamide | *accA*: acetyl-CoA carboxylase(37) | *Burkholderia thailandensis* E264 | D,B,P,R* | 42, 5 | 838, 6347 |
| Indolmycin | *trypS*: tryptophan-tRNA synthetase(38) | *Pseudoalteromonas luteoviolacea* | D,B | 13, 2 | 540, 4963 |
| Agrocin 84 | leu tRNA synthase(39) | *Agrobacterium radiobacter* K84 | D,P | 41, 2 | 470, 6876 |
| Bengamide | methionine aminopeptidase(40) | *Myxococcus virescens* DSM 15898 | Core | N/A | 1, 18 |
| Mupirocin | Ile-tRNA synthetase(41) | *Pseudomonas fluorescens* NCIMB 10586 | Core | N/A | 1, 36 |
| Andrimid | *accD*: acetyl-CoA carboxylase(42) | *Pantoea agglomerans* Eh335 | Core | N/A | 1, 18 |
| Cystobactamid | Pentapeptide repeat protein(43) | *Cystobacter* sp. Cbv34 | R* | N/A | 0, 24 |
| Phaseolotoxin | ornithine carbamoyltransferase(44) | *Pseudomonas savastanoi* pv. *phaseolicola* | Core, R* | N/A | 3, 26 |
| Kalimantacin | *fabI*: enoyl reductase(45) | *Pseudomonas fluorescens* BCCM ID9359 | No hits | N/A | 3, 29 |

Hits to ARTS criteria are shown as; D: duplication, B: BGC proximity, P: phylogeny, R: resistance model. Rows in gray indicate only complete gene cluster as input rather than whole genome. Stars indicate exploration mode.



Summary of genomes

| Organism | Total Genes | Total BGCs | Known Resistance | Core Genes | Gene Duplication | BGC Proximity | Phylogeny / HGT | 2+ | 3+ |
|---|---|---|---|---|---|---|---|---|---|
| Achromobacter_insolitus_strain_FDAARGOS_88.gbff | 6372 | 7 | 63 | 825 | 56 | 34 | 173 | 30 | 1 |
| Advenella_mimigardefordensis_DPN7.gbff | 4364 | 10 | 49 | 762 | 35 | 38 | 235 | 40 | 5 |
| Bordetella_petrii_strain_DSM_12804.gbff | 5036 | 8 | 47 | 814 | 38 | 33 | 209 | 33 | 2 |

Core Genes: Essential gene table

| Gene | Description | Function | Dup | BGC | HGT | Res |
|---|---|---|---|---|---|---|
| TIGR00658 | orni_carb_tr: ornithine carbamoyltransferase | Amino acid biosynthesis | 0.33 | 0.0 | 0.17 | 1.0 |
| TIGR00670 | asp_carb_tr: aspartate carbamoyltransferase | Purines, pyrimidines, nucleosides, and nucleotides | 0.33 | 0.17 | 0.0 | 1.0 |

Achromobacter_insolitus

Status: Done

Show/Hide log

Summary of hits

| | |
|---|---|
| Total genes: | 6372 |
| Core/Essential genes: | 825 |
| Total BGC hits: | 7 |
| Known resistance model hits: | 63 |

ARTS Criteria Hit Counts

| | |
|---|---|
| Gene Duplication: | 56 |
| BGC Proximity: | 34 |
| Phylogeny / HGT: | 173 |
| 2 or more: | 30 |
| 3 or more: | 1 |

Total BGCs: 84 (48 singleton/s), links: 23, families: 64

Search:

BGC networking

| BGC Family | BGC(s) | Product(s) | BGC class(es) | Sources |
|---|---|---|---|---|
| 68 | NZ_FXAN01000111.region001,<br>NZ_LOWA01000018.region004 | lassopeptide,<br>lassopeptide | RiPPs,<br>RiPPs | Burkholderia singularis strain LMG 28154, whole genome shotgun sequence,<br>Burkholderia singularis strain TSV85 TSV85_26, whole genome shotgun sequence |
| 59 | NZ_FXAN01000110.region001,<br>NZ_LOWA01000008.region002 | terpene, terpene | Terpene,<br>Terpene | Burkholderia singularis strain LMG 28154, whole genome shotgun sequence,<br>Burkholderia singularis strain TSV85 TSV85_17, whole genome shotgun sequence |
| 58 | NZ_FXAN01000109.region001,<br>NZ_LOWA01000025.region001 | NRPS, NRPS | NRPS,<br>NRPS | Burkholderia singularis strain LMG 28154, whole genome shotgun sequence,<br>Burkholderia singularis strain TSV85 TSV85_32, whole genome shotgun sequence |
| 56 | NZ_FXAN01000093.region001 | terpene | Terpene | Burkholderia singularis strain LMG 28154, whole genome shotgun sequence |

Input Data

| | |
|---|---|
| Total Number of Genomes: | 5 |
| Total BGCs: | 84 |

Cluster GCF based on: Genomes Absence/Presence ▾

Cluster Genomes based on: Family Absence/Presence ▾

Show: 20 ▾ largest GCFs

**Figure 2.** Example output of multi-genome ARTS analysis. Top part of the page represents the summaries of individual arts runs and shared core genes throughout the whole analysis with respective ARTS hits. At the bottom, shared BGCs and resistance models can easily be navigated and an interactive BiG-SCAPE graph output can also be found via 'Open BiG-SCAPE overview" option.

cies among all genomes can be inspected via dynamic output tables. This aids in further prioritizing ARTS hits for those that are detected in multiple contexts or related BGCs and therefore are more likely to be involved in resistance. For example, users can now narrow HGT hits by inspecting those that are shared across multiple organisms. In addition to these data, the BiG-SCAPE algorithm (52) is applied on all detected BGCs, allowing users to investigate similar BGCs from multiple sources by constructing gene cluster sequence similarity networks and identifying gene cluster families inside these networks. Furthermore, each of the BGCs in a gene cluster family can be examined in order to assess whether they have core or resistance models as shared hits, as well as whether a cluster stands out with unique hits compared to its relatives from other species.

### Server-side updates and speed up

In order to keep the ARTS pipeline at high standards, third party tools used in the workflow were updated. ARTS 2.0 now uses antiSMASH v5 and is able to analyze antiSMASH results from their newest JSON format. The most time consuming part of the ARTS pipeline is the creation of species and gene trees for phylogenetic analysis via ASTRAL. By updating antiSMASH and ASTRAL, the average runtime of the whole pipeline could now be cut down to half. Also, in order to satisfy the increasing demand, ARTS 2.0 is now hosted at the highly scalable de.NBI cloud system with seven times the computational power. With these hardware and software updates, the ARTS 2.0 webserver is now capable of analyzing multiple inputs up to 100MB and depending on the genomes and selected parameters, 3-8 times faster than the previous version.

## CONCLUSIONS AND FUTURE PERSPECTIVES

Currently, ARTS is the only platform to automate resistance and putative drug-target based genome mining in bacteria via a user-friendly webserver. By design, ARTS aims to survey a wide scope of potential genes as drug targets while minimizing manual inspection by using the dynamic output and multiple screening criteria for more confident target predictions. Thus it is incumbent on the user to examine potential hits with provided metadata and contextual framing. Some of the ARTS hits might be more likely involved in biosynthesis and not associated with resistance. Although we removed common biosynthesis genes from the core gene sets to avoid false positives (23), it is currently not possible to automatically distinguish if genes are more likely involved in biosynthesis or resistance,  for example fatty acid synthases are involved in both (22). The occasional high counts of positive hits in exploration mode, largely due to undefined cluster boundaries, can be easily and rapidly filtered in the interactive output page. As shown previously, this inspection can even serve to help define the true boundaries of clusters, which remains a largely unresolved challenge when dealing with bacterial BGCs (23). Newly introduced features now make ARTS 2.0 a fast and comprehensive pipeline allowing users to: analyze sequences from all bacterial genomes as well as metagenomic samples, apply comparative analysis on multiple genomes, and interrogate similar BGCs for shared resistant genes. For future applications, we are working on increasing ARTS' availability by making it directly accessible through other webservers such as antiSMASH. This will enable researchers to easily apply target-directed genome mining approaches on sequences from different databases as a plugin. Furthermore, we are currently in process of creating the ARTS database, which will contain preanalyzed ARTS results for all bacterial genomes within the Refseq database, and will allow global analysis and comparisons of resistant targets within BGC. We hope that with this update, ARTS 2.0 will now provide an even broader access to resistance based genome mining methods and facilitate the discovery of competitive antibiotics.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The authors acknowledge the use of de.NBI cloud and the support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen and the Federal Ministry of Education and Research (BMBF) through grant no 031 A535A. We also acknowledge all ARTS users for helpful comments and feedback.

## FUNDING

High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen; State of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) [INST 37/935-1 FUGG]; N.Z., M.A. and M.D.M. acknowledge the German Center for Infection Research [DZIF TTU09.704]; T.W. and K.B. were supported by grants from the Novo Nordisk Foundation [NNF10CC1016517, NNF16OC0021746]; M.A. and M.H.M. were supported by an ERA NET CoBiotech ('Bestbiosurf') grant through the Netherlands Organization for Scientific Research (NWO) [053.80.739]. Funding for open access charge: BMBF [DZIF TTU09.704].

*Conflict of interest statement.* None declared.

## REFERENCES

1. Michael,C.A., Dominey-Howes,D. and Labbate,M. (2014) The antimicrobial resistance crisis: causes, consequences, and management. *Front. Public Health*, **2**, 145.
2. Cragg,G.M. and Newman,D.J. (2013) Natural products: a continuing source of novel drug leads. *Biochim. Biophys. Acta*, **1830**, 3670–3695.
3. Harvey,A.L., Edrada-Ebel,R. and Quinn,R.J. (2015) The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug. Discov*, **14**, 111–129.
4. Newman,D.J. and Cragg,G.M. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.*, **75**, 311–335.
5. Ziemert,N., Alanjary,M. and Weber,T. (2016) The evolution of genome mining in microbes–a review. *Nat. Prod. Rep.*, **33**, 988–1005.
6. Stokes,J.M., Yang,K., Swanson,K., Jin,W., Cubillos-Ruiz,A., Donghia,N.M., MacNair,C.R., French,S., Carfrae,L.A., Bloom-Ackerman,Z. *et al.* (2020) A deep learning approach to antibiotic discovery. *Cell*, **180**, 688–702.

7. Li,Z., Zhu,D. and Shen,Y. (2018) Discovery of novel bioactive natural products driven by genome mining. *Drug Discov. Ther.*, **12**, 318–328.
8. Bachmann,B.O., Van Lanen,S.G. and Baltz,R.H. (2014) Microbial genome mining for accelerated natural products discovery: is a renaissance in the making?. *J. Ind. Microbiol. Biot.*, **41**, 175–184.
9. Blin,K., Shaw,S., Steinke,K., Villebro,R., Ziemert,N., Lee,S.Y., Medema,M.H. and Weber,T. (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.*, **47**, W81–W87.
10. Sélem-Mojica,N., Aguilar,C., Gutiérrez-García,K., Martínez-Guerrero,C.E. and Barona-Gómez,F. (2019) EvoMining reveals the origin and fate of natural product biosynthetic enzymes. *Microb. Genom.*, **5**, e000260.
11. Skinnider,M.A., Merwin,N.J., Johnston,C.W. and Magarvey,N.A. (2017) PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res.*, **45**, W49–W54.
12. Hannigan,G.D., Prihoda,D., Palicka,A., Soukup,J., Klempir,O., Rampula,L., Durcak,J., Wurst,M., Kotowski,J., Chang,D. *et al.* (2019) A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.*, **47**, e110.
13. Kautsar,S.A., Blin,K., Shaw,S., Navarro-Muñoz,J.C., Terlouw,B.R., van der Hooft,J.J., Van Santen,J.A., Tracanna,V., Suarez Duran,H.G., Pascal Andreu,V. *et al.* (2020) MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.*, **48**, D454–D458.
14. Van Santen,J.A., Jacob,G., Singh,A.L., Aniebok,V., Balunas,M.J., Bunsko,D., Neto,F.C., Castaño-Espriu,L., Chang,C., Clark,T.N. *et al.* (2019) The natural products atlas: an open access knowledge base for microbial natural products discovery. *ACS Cent. Sci.*, **5**, 1824–1833.
15. Blin,K., Pascal Andreu,V., de los Santos,E. L.C., Del Carratore,F., Lee,S.Y., Medema,M.H. and Weber,T. (2019) The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.*, **47**, D625–D630.
16. Palaniappan,K., Chen,I.-M.A., Chu,K., Ratner,A., Seshadri,R., Kyrpides,N.C., Ivanova,N.N. and Mouncey,N.J. (2020) IMG-ABC v. 5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Res.*, **48**, D422–D430.
17. Almabruk,K.H., Dinh,L.K. and Philmus,B. (2018) Self-resistance of natural product producers: Past, present, and future focusing on self-resistant protein variants. *ACS Chem. Biol.*, **13**, 1426–1437.
18. Yan,Y., Liu,Q., Zang,X., Yuan,S., Bat-Erdene,U., Nguyen,C., Gan,J., Zhou,J., Jacobsen,S.E. and Tang,Y. (2018) Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action. *Nature*, **559**, 415–418.
19. Brochet,M., Couvé,E., Zouine,M., Poyart,C. and Glaser,P. (2008) A naturally occurring gene amplification leading to sulfonamide and trimethoprim resistance in Streptococcus agalactiae. *J. Bacteriol.*, **190**, 672–680.
20. Freel,K.C., Millán-Aguiñaga,N. and Jensen,P.R. (2013) Multilocus sequence typing reveals evidence of homologous recombination linked to antibiotic resistance in the genus Salinispora. *Appl. Environ. Microbiol.*, **79**, 5997–6005.
21. Thaker,M.N., Wang,W., Spanogiannopoulos,P., Waglechner,N., King,A.M., Medina,R. and Wright,G.D. (2013) Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nat. Biotechnol.*, **31**, 922.
22. Tang,X., Li,J., Millán-Aguiñaga,N., Zhang,J.J., O'Neill,E.C., Ugalde,J.A., Jensen,P.R., Mantovani,S.M. and Moore,B.S. (2015) Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem. Biol.*, **10**, 2841–2849.
23. Alanjary,M., Kronmiller,B., Adamek,M., Blin,K., Weber,T., Huson,D., Philmus,B. and Ziemert,N. (2017) The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Res.*, **45**, W42–W48.
24. Cimermancic,P., Medema,M.H., Claesen,J., Kurita,K., Brown,L. C.W., Mavrommatis,K., Pati,A., Godfrey,P.A., Koehrsen,M., Clardy,J. *et al.* (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, **158**, 412–421.
25. Li,Y., Li,Z., Yamanaka,K., Xu,Y., Zhang,W., Vlamakis,H., Kolter,R., Moore,B.S. and Qian,P.-Y. (2015) Directed natural product biosynthesis gene cluster capture and expression in the model bacterium Bacillus subtilis. *Sci. Rep.-UK*, **5**, 9383.
26. Weissman,K.J. and Müller,R. (2010) Myxobacterial secondary metabolites: bioactivities and modes-of-action. *Nat. Prod. Rep.*, **27**, 1276–1295.
27. Haft,D.H., Selengut,J.D., Richter,R.A., Harkins,D., Basu,M.K. and Beck,E. (2012) TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.*, **41**, D387–D395.
28. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
29. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
30. Gupta,R.S. (2000) The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiol. Rev.*, **24**, 367–402.
31. Hug,L.A., Baker,B.J., Anantharaman,K., Brown,C.T., Probst,A.J., Castelle,C.J., Butterfield,C.N., Hernsdorf,A.W., Amano,Y., Ise,K. *et al.* (2016) A new view of the tree of life. *Nat. Microbiol.*, **1**, 16048.
32. de Vienne,D.M. (2016) Lifemap: exploring the entire tree of life. *PLoS Biol.*, **14**, e2001624.
33. Gupta,R.S. and Bhandari,V. (2011) Phylogeny and molecular signatures for the phylum Thermotogae and its subgroups. *Anton. Leeuw.*, **100**, 1.
34. Gregoire,P., Bohli,M., Cayol,J.-L., Joseph,M., Guasco,S., Dubourg,K., Cambar,J., Michotey,V., Bonin,P., Fardeau,M.-L. *et al.* (2011) Caldilinea tarbellica sp. nov., a filamentous, thermophilic, anaerobic bacterium isolated from a deep hot aquifer in the Aquitaine Basin. *Int. J. Syst. Evol. Micr.*, **61**, 1436–1441.
35. Brown,L. C.W., Acker,M.G., Clardy,J., Walsh,C.T. and Fischbach,M.A. (2009) Thirteen posttranslational modifications convert a 14-residue peptide into the antibiotic thiocillin. *Proc. Natl. Acad. Sci.*, **106**, 2549–2553.
36. Xiao,Y., Gerth,K., Müller,R. and Wall,D. (2012) Myxobacterium-produced antibiotic TA (myxovirescin) inhibits type II signal peptidase. *Antimicrob. Agents Chemother*, **56**, 2014–2021.
37. Wozniak,C.E., Lin,Z., Schmidt,E.W., Hughes,K.T. and Liou,T.G. (2018) Thailandamide, a fatty acid synthesis antibiotic that is coexpressed with a resistant target gene. *Antimicrob. Agents Chemother*, **62**, e00463-18.
38. Du,Y.-L., Alkhalaf,L.M. and Ryan,K.S. (2015) In vitro reconstitution of indolmycin biosynthesis reveals the molecular basis of oxazolinone assembly. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 2717–2722.
39. Ryder,M., Slota,J., Scarim,A. and Farrand,S. (1987) Genetic analysis of agrocin 84 production and immunity in Agrobacterium spp. *J. Bacteriol.*, **169**, 4184–4189.
40. Wenzel,S.C., Hoffmann,H., Zhang,J., Debussche,L., Haag-Richter,S., Kurz,M., Nardi,F., Lukat,P., Kochems,I., Tietgen,H. *et al.* (2015) Production of the bengamide class of marine natural products in myxobacteria: biosynthesis and structure–activity relationships. *Angew. Chem. Int. Ed.*, **54**, 15560–15564.
41. El-Sayed,A.K., Hothersall,J., Cooper,S.M., Stephens,E., Simpson,T.J. and Thomas,C.M. (2003) Characterization of the mupirocin biosynthesis gene cluster from Pseudomonas fluorescens NCIMB 10586. *Chem. Biol.*, **10**, 419–430.
42. Liu,X., Fortin,P.D. and Walsh,C.T. (2008) Andrimid producers encode an acetyl-CoA carboxyltransferase subunit resistant to the action of the antibiotic. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 13321–13326.
43. Baumann,S., Herrmann,J., Raju,R., Steinmetz,H., Mohr,K.I., Hüttel,S., Harmrolfs,K., Stadler,M. and Müller,R. (2014) Cystobactamids: myxobacterial topoisomerase inhibitors exhibiting potent antibacterial activity. *Angew. Chem. Int. Ed.*, **53**, 14605–14609.
44. Chen,L., Li,P., Deng,Z. and Zhao,C. (2015) Ornithine transcarbamylase ArgK plays a dual role for the self-defense of phaseolotoxin producing Pseudomonas syringae pv. phaseolicola. *Sci. Rep.-UK*, **5**, 12892–12892.

45. Mattheus,W., Masschelein,J., Gao,L.-J., Herdewijn,P., Landuyt,B., Volckaert,G. and Lavigne,R. (2010) The kalimantacin/batumin biosynthesis operon encodes a self-resistance isoform of the FabI bacterial target. *Chem. Biol.*, **17**, 1067–1071.
46. Luo,H., Lin,Y., Gao,F., Zhang,C.-T. and Zhang,R. (2014) DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res.*, **42**, D574–D580.
47. Zhang,C., Rabiee,M., Sayyari,E. and Mirarab,S. (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, **19**, 153.
48. Jeong,H., Sung,S., Kwon,T., Seo,M., Caetano-Anollés,K., Choi,S.H., Cho,S., Nasir,A. and Kim,H. (2016) HGTree: database of horizontally transferred genes determined by tree reconciliation. *Nucleic Acids Res.*, **44**, D610–D619.
49. Nakamura,Y. (2018) Prediction of horizontally and widely transferred genes in prokaryotes. *Evol. Bioinform.*, **14**, doi:10.1177/1176934318810785.
50. Trindade,M., van Zyl,L.J., Navarro-Fernández,J. and Abd Elrazak,A. (2015) Targeted metagenomics as a tool to tap into marine natural product diversity for the discovery and production of drug candidates. *Front. Microbiol.*, **6**, 890.
51. Garcia,R., La Clair,J.J. and Müller,R. (2018) Future directions of marine myxobacterial natural product discovery inferred from metagenomics. *Mar. Drugs*, **16**, 303.
52. Navarro-Muñoz,J.C., Selem-Mojica,N., Mullowney,M.W., Kautsar,S.A., Tryon,J.H., Parkinson,E.I., De Los Santos,E.L., Yeong,M., Cruz-Morales,P., Abubucker,S. *et al.* (2020) A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.*, **16**, 60–68.

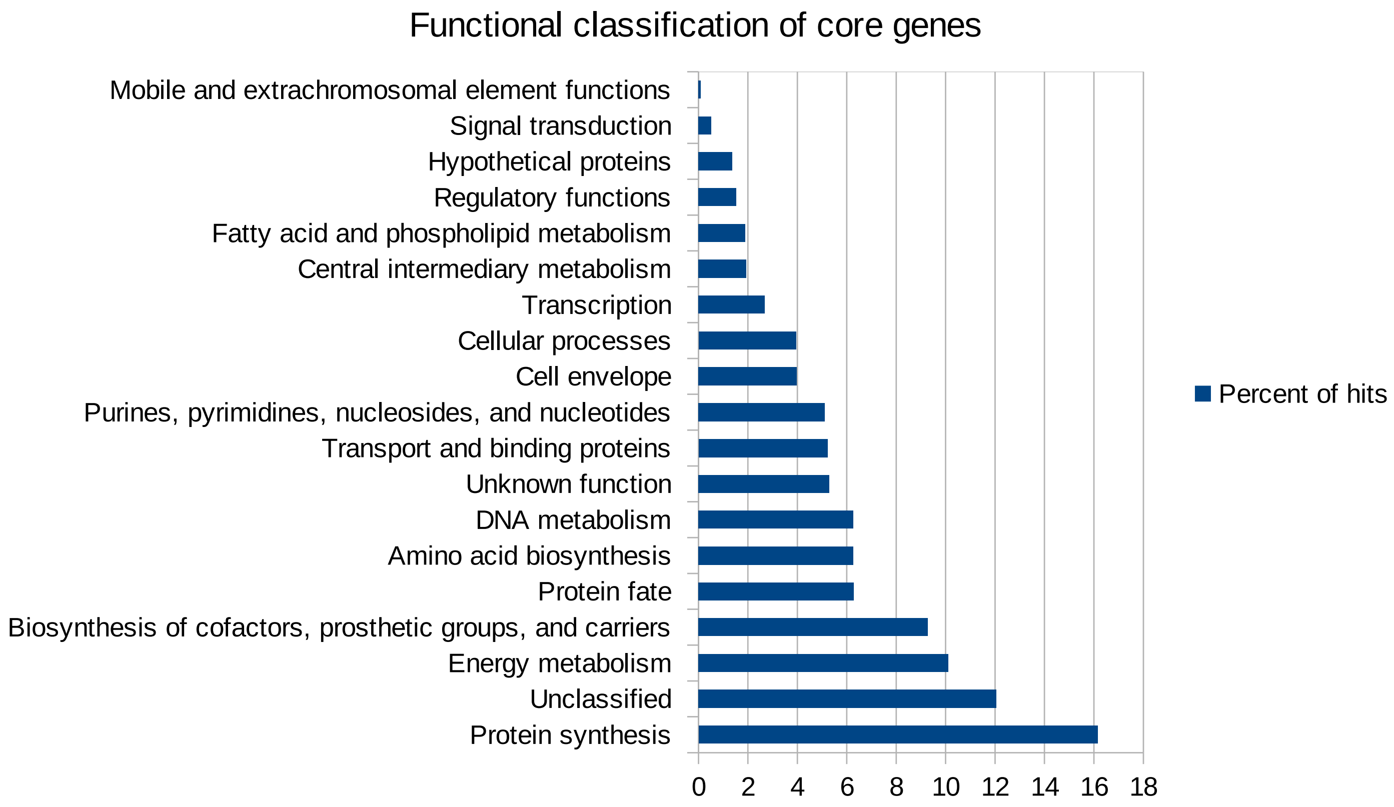
## Supplementary Table 1

Distribution of sequence counts for each created reference set and their assembly qualities. Phylum contents of the assembled groups are also noted at the bottom

| Reference Set | Sequence Count | Sequence Type |
|---|---:|---|
| Gammaproteobacteria | 987 | Complete Genomes |
| Firmicutes | 738 | Complete Genomes |
| Alphaproteobacteria | 516 | Complete Genomes |
| Bacteroidetes | 404 | Complete Genomes |
| Chlamydiae | 391 | Complete Genomes + Contig Assemblies |
| Betaproteobacteria | 359 | Complete Genomes |
| Group3 | 229 | Complete Genomes + Contig Assemblies |
| Fusobacteria | 227 | Complete Genomes + Contig Assemblies |
| Group2 | 220 | Complete Genomes + Contig Assemblies |
| Actinobacteria | 190 | Complete Genomes |
| Spirochaetes | 165 | Complete Genomes + Contig Assemblies |
| Delta_Epsilon-proteobacteria | 158 | Complete Genomes |
| Deinococcus-thermus | 155 | Complete Genomes + Contig Assemblies |
| Verrucomicrobia | 129 | Complete Genomes + Contig Assemblies |
| Cyanobacteria | 118 | Complete Genomes |
| Tenericutes | 94 | Complete Genomes |
| Group1 | 46 | Complete Genomes + Contig Assemblies |
| Group1: Calditrichaeota, Aquificae, Coprothermobacterota, Deferribactes, Chrysiogenetes<br>Group2: Thermotogae, Synergistetes, Dictyoglomi, Chloroflexi, Armatimonadetes<br>Group3: Ignavibacteriae, Chlorobi, Gemmatimonadetes, Lentisphaerae, Nitrospirae, Thermodesulfobacteria, Planctomycetes, Acidobacteria, Elusimicrobia, Fibrobacteres, Balneolaeota, Rhodothermaeota | | |

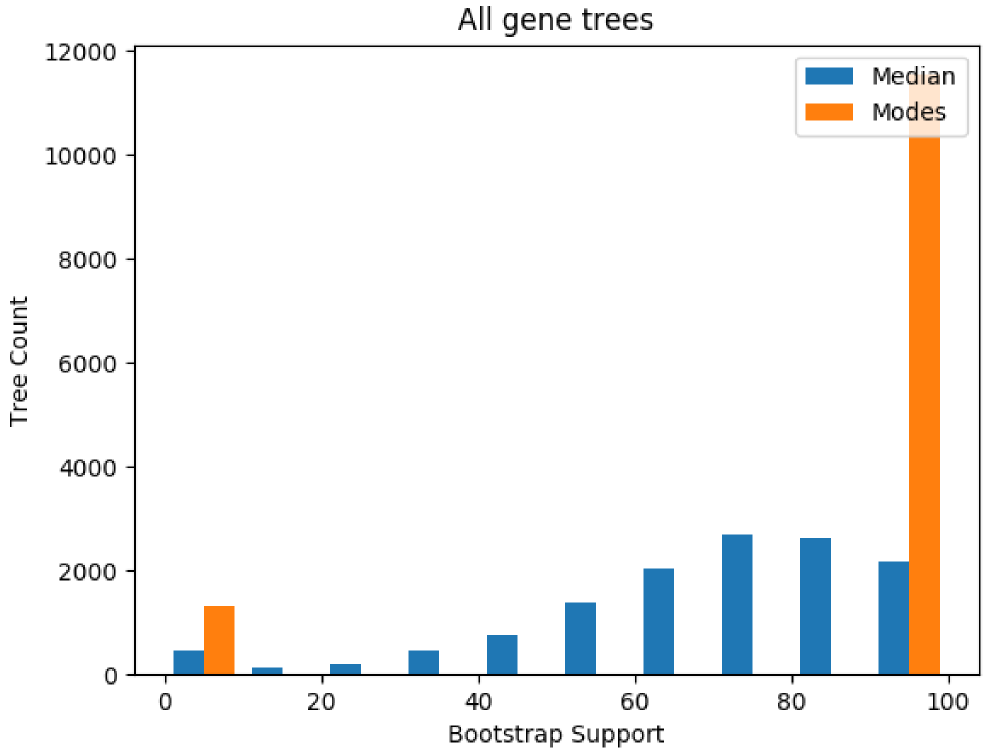
**Supplementary Figure 1**

Average distribution of detected core genes from all reference sets



**Supplementary Figure 2**

Median and mode distributions of all bootstrap support values from all reference sets



**Supplementary Table 2**

Amount of core genes that are also found in the Database of Essential Genes (DEG) v1.5

| Reference Set | Default Mode | | Exploratory Mode | |
|---|---|---|---|---|
| | Deg | Core | Deg | Core |
| Chlamydiae | 254 | 274 | 412 | 445 |
| Group1 | 227 | 234 | 383 | 398 |
| Group2 | 402 | 504 | 753 | 968 |
| Tenericutes | 209 | 219 | 308 | 321 |
| Fusobacteria | 331 | 358 | 587 | 650 |
| Verrucomicrobia | 349 | 405 | 660 | 776 |
| Deinococcus-thermus | 298 | 320 | 537 | 599 |
| Alphaproteobacteria | 448 | 547 | 878 | 1098 |
| Betaproteobacteria | 438 | 496 | 805 | 923 |
| Spirochaetes | 364 | 415 | 659 | 765 |
| Firmicutes | 427 | 560 | 788 | 1019 |
| Bacteroidetes | 394 | 450 | 726 | 860 |
| Actinobacteria | 367 | 432 | 540 | 664 |
| Delta_Epsilon-proteobacteria | 398 | 464 | 739 | 866 |
| Group3 | 412 | 478 | 770 | 921 |
| Cyanobacteria | 366 | 447 | 720 | 882 |
| Gammaproteobacteria | 562 | 740 | 1101 | 1478 |
| Metagenome | 1568 | 1568 | 1568 | 4507 |

## Supplementary Table 3

Detection frequency statistics of reference organisms totaling 4919 sequences, used to create all reference sets

| | **Percentage of hits found** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Reference Set** | HGT | BGC Proximity | Duplication | Known Resistance | Core Genes | 2+ Hits | 3+ Hits | #Total Core Genes |
| Verrucomicrobia | 13.8 | 4.09 | 1.97 | 7.39 | 321.45 | 1.26 | 0.11 | 405 |
| Tenericutes | 5.63 | 0.09 | 0.83 | 6.11 | 92.4 | 0.44 | 0 | 219 |
| Spirochaetes | 4.9 | 0.61 | 3.14 | 6.67 | 305.2 | 0.67 | 0.01 | 415 |
| Group3 | 24.37 | 3.22 | 4.69 | 7.43 | 324.19 | 2.61 | 0.12 | 478 |
| Group2 | 14.44 | 0.64 | 4.06 | 5.92 | 322.5 | 2.13 | 0.07 | 504 |
| Group1 | 16.58 | 1.8 | 2.16 | 7.71 | 227.67 | 1.68 | 0.05 | 234 |
| Gammaproteobacteria | 10.25 | 2.28 | 5.47 | 7.06 | 527.79 | 2.25 | 0.1 | 740 |
| Fusobacteria | 3.17 | 2.42 | 1.93 | 5.09 | 314.76 | 0.57 | 0.03 | 358 |
| Firmicutes | 11.39 | 2.41 | 4.7 | 7.41 | 406.75 | 2.14 | 0.12 | 560 |
| Delta_Epsilon-proteobacteria | 14.43 | 2.16 | 4.52 | 6.89 | 343.9 | 2.62 | 0.19 | 464 |
| Deinococcus-thermus | 2.3 | 1.39 | 1.78 | 7.14 | 292.49 | 0.24 | 0.004 | 320 |
| Cyanobacteria | 22.74 | 4.6 | 3.57 | 11.75 | 394.97 | 3.27 | 0.28 | 447 |
| Chlamydiae | 14.32 | 0.15 | 0.68 | 5.77 | 248.62 | 0.22 | 0.004 | 274 |
| Betaproteobacteria | 15.54 | 3.25 | 4.05 | 9.26 | 438.09 | 2.69 | 0.17 | 496 |
| Bacteroidetes | 28.92 | 2.52 | 2.64 | 7.97 | 318.13 | 2.19 | 0.1 | 450 |
| Alphaproteobacteria | 17.4 | 2.68 | 4.64 | 8.67 | 411.25 | 3.38 | 0.18 | 547 |

**Supplementary Table 4**

Amount of core genes that are unique for a respective reference set or shared by multiple sets.

| Number of reference sets | Shared core genes in normal mode | Shared core genes in exploratory mode | Percentage of shared core genes | |
|---|---|---|---|---|
| 1 | 528 | 778 | 39.55 | 32.07 |
| 2 | 167 | 322 | 12.51 | 13.27 |
| 3 | 88 | 210 | 6.59 | 8.66 |
| 4 | 45 | 131 | 3.37 | 5.4 |
| 5 | 48 | 114 | 3.6 | 4.7 |
| 6 | 30 | 74 | 2.25 | 3.05 |
| 7 | 36 | 75 | 2.7 | 3.09 |
| 8 | 33 | 57 | 2.47 | 2.35 |
| 9 | 20 | 46 | 1.5 | 1.9 |
| 10 | 24 | 55 | 1.8 | 2.27 |
| 11 | 36 | 83 | 2.7 | 3.42 |
| 12 | 32 | 61 | 2.4 | 2.51 |
| 13 | 24 | 55 | 1.8 | 2.27 |
| 14 | 28 | 76 | 2.1 | 3.13 |
| 15 | 40 | 74 | 3 | 3.05 |
| 16 | 55 | 91 | 4.12 | 3.75 |
| 17 | 101 | 124 | 7.57 | 5.11 |
| Total Core Genes | 1335 | 2426 | | |

## 2.2 SYN-View: A Phylogeny-Based Synteny Exploration Tool for the Identification of Gene Clusters Linked to Antibiotic Resistance

### Contributions

Also written as a part of the manuscript, contributions are specified as below. J.S. Jason Stahlecker, N.Z. Nadine Ziemert, M.D.M Mehmet Direnc Mungan, E.M. Erik Mingyar.

Conceptualization, E.M. and N.Z.; methodology, M.D.M. and N.Z.; software, J.S. and M.D.M.; validation, J.S. and M.D.M.; formal analysis, J.S.; investigation, J.S. and M.D.M.; resources, N.Z.; data curation, J.S. and M.D.M.; writing—original draft preparation, J.S. and M.D.M.; writing—review and editing, M.D.M., N.Z., and E.M.; visualization, J.S.; supervision, M.D.M. and N.Z.; project administration, N.Z.; funding acquisition, N.Z. All authors have read and agreed to the published version of the manuscript.

molecules

MDPI

*Communication*

# SYN-View: A Phylogeny-Based Synteny Exploration Tool for the Identification of Gene Clusters Linked to Antibiotic Resistance

**Jason Stahlecker [1], Erik Mingyar [1], Nadine Ziemert [1,2] and Mehmet Direnç Mungan [1,2,*]**

[1] Interfaculty Institute of Microbiology and Infection Medicine, University of Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany; wilhelm-jason.stahlecker@student.uni-tuebingen.de (J.S.); erik.mingyar@uni-tuebingen.de (E.M.); nadine.ziemert@uni-tuebingen.de (N.Z.)

[2] German Centre for Infection Research (DZIF), Partner Site Tübingen, 38124 Tübingen, Germany

[*] Correspondence: mehmet-direnc.mungan@uni-tuebingen.de

**Abstract:** The development of new antibacterial drugs has become one of the most important tasks of the century in order to overcome the posing threat of drug resistance in pathogenic bacteria. Many antibiotics originate from natural products produced by various microorganisms. Over the last decades, bioinformatical approaches have facilitated the discovery and characterization of these small compounds using genome mining methodologies. A key part of this process is the identification of the most promising biosynthetic gene clusters (BGCs), which encode novel natural products. In 2017, the Antibiotic Resistant Target Seeker (ARTS) was developed in order to enable an automated target-directed genome mining approach. ARTS identifies possible resistant target genes within antibiotic gene clusters, in order to detect promising BGCs encoding antibiotics with novel modes of action. Although ARTS can predict promising targets based on multiple criteria, it provides little information about the cluster structures of possible resistant genes. Here, we present SYN-view. Based on a phylogenetic approach, SYN-view allows for easy comparison of gene clusters of interest and distinguishing genes with regular housekeeping functions from genes functioning as antibiotic resistant targets. Our aim is to implement our proposed method into the ARTS web-server, further improving the target-directed genome mining strategy of the ARTS pipeline.

**Keywords:** biosynthetic gene clusters; natural products; genome mining; antibiotic resistance

check for updates

**Citation:** Stahlecker, J.; Mingyar, E.; Ziemert, N.; Mungan, M.D. SYN-View, a Phylogeny Based Synteny Exploration Tool for the Identification of Gene Clusters Linked to Antibiotic Resistance. *Molecules* **2021**, *26*, 144. https://doi.org/10.3390/molecules26010144

Academic Editors: Daniel Krug and Lena Keller

Received: 14 December 2020
Accepted: 25 December 2020
Published: 31 December 2020

## 1. Introduction

With the increasing number of drug-resistant bacteria, antimicrobial resistance has become a global health threat [1]. As the number of approved drugs have been decreasing over the past few decades, finding new compounds to feed the antibiotic discovery pipeline has become a crucial task [2]. Most of the antibiotics are derived from secondary metabolites (SMs) produced by fungal and bacterial organisms [3]. Many of these so-called natural products were found by labor-intensive methods such as screening biological samples for desired bioactivities. However, these traditional methods have been losing their efficiency, due to their high rediscovery rates [4]. Ever since the cost of DNA sequencing technologies has decreased substantially, in silico methods such as genome mining have gained an increased amount of popularity among researchers [5,6]. As a result, a number of computational tools such as antiSMASH [7] and PRISM [8] have been developed, in order to detect gene clusters encoding for natural products. The main approach of these tools is the identification of locally clustered groups of genes called biosynthetic gene clusters (BGCs), which are in conjunction responsible for the synthesis of secondary metabolites [9]. Using those BGC prediction tools, a large number of BGCs have been deposited in public databases. The newest version of Atlas of Biosynthetic Gene Clusters (IMG-ABC) [10], the largest database containing predicted BGCs, contains roughly 400,000 clusters, from which less than 1% have been experimentally verified. This large discrepancy emphasizes the need for new and updated tools as well as the importance of prioritization of predicted BGCs for downstream processes. In order to address this issue, in 2017, Alanjary et al. developed the Antibiotic Resistant Target Seeker (ARTS) [11] to detect most promising BGCs with potential new modes of action by automating the resistance based genome mining technique also called target directed genome mining. This approach is based on the notion that the antibiotic-producing bacteria have to be resistant to their own products [12]. Resistance genes can be encoded within the BGC of the respective compound. Additionally, in case of a resistance mechanism that is provided by a resistant target, this kind of genome mining method not only provides insights into the mode of action of the encoded antibiotics, but in turn also allows screening BGCs for natural products with promising and putatively novel targets [13]. ARTS links essential housekeeping genes to evolution driven events such as duplication, horizontal gene transfer (HGT), or co-localization within the BGC, which have been extensively shown to be the key processes in target-based strategies [14–16]. Although ARTS rapidly screens essential genes of an entire genome, the number of potential resistant targets can become quite large, especially when the BGC boundaries are set too far. In such cases, the distinction of a resistance gene and a regular housekeeping gene is hard to make. As stated by O'Neill and his colleagues, inferring such distinctions may be possible, by comparing gene ortholog neighbors of the putative resistance genes and the context of the clusters they lie within in related organisms. Regular housekeeping genes often show synteny in their cluster structure, whereas the resistant target genes within antibiotic gene clusters are often only randomly present in closely related taxa [17]. Following up on this hypothesis, we analyzed the novobiocin producer *Streptomyces niveus* NCIMB 11891, with duplicated *gyrB* gene as known self-resistance mechanism, yielding a large number of false positives by an initial ARTS search shown in the first ARTS paper [11]. Visualized in Figure 1, the comparison of the neighborhoods of gene of interest (NGIs) to the NGIs from closely related organisms, clearly shows that the neighborhood of the housekeeping gene is almost identical, whereas the resistant target gene shows no orthologous genes in the neighborhood.

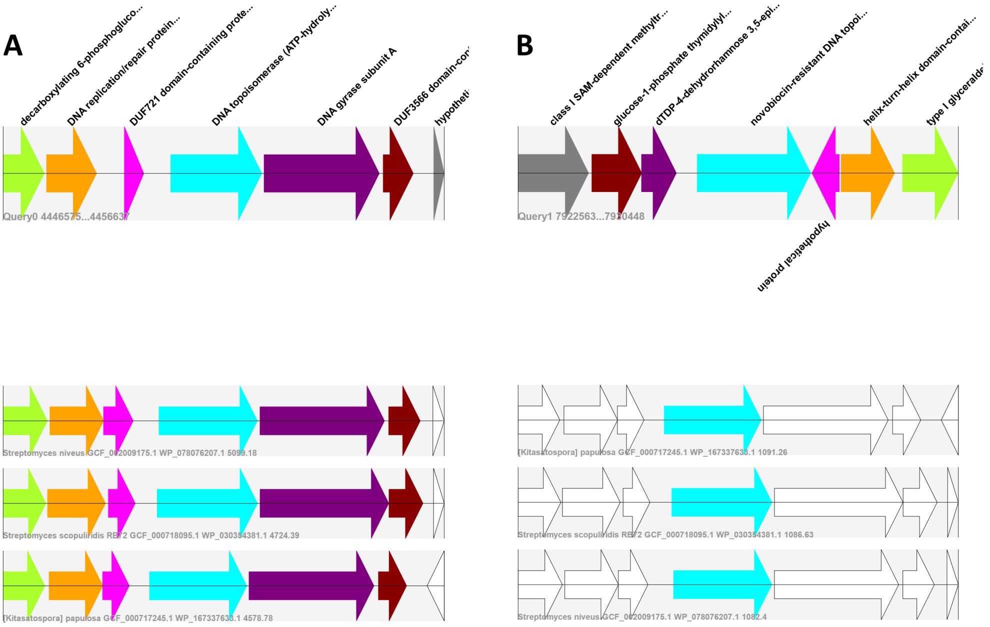

A
decarboxylating 6-phosphogluco...
DNA replication/repair protein...
DUF721 domain-containing prote...
DNA topoisomerase (ATP-hydroly...
DNA gyrase subunit A
DUF3566 domain-con
hypotheti
Query0 4446575...4456637
Streptomyces niveus GCF_002009175.1 WP_078076207.1 5099.18
Streptomyces scopuliridis RB72 GCF_000718095.1 WP_030354381.1 4724.39
[Kitasatospora] papulosa GCF_000717245.1 WP_167337633.1 4578.78
B
class I SAM-dependent methyltr...
glucose-1-phosphate thymidylyl...
dTDP-4-dehydrorhamnose 3,5-epi...
novobiocin-resistant DNA topoi...
helix-turn-helix domain-contai...
type I glyceralde
hypothetical protein
Query1 7922563...7930448
[Kitasatospora] papulosa GCF_000717245.1 WP_167337633.1 1091.26
Streptomyces scopuliridis RB72 GCF_000718095.1 WP_030354381.1 1086.63
Streptomyces niveus GCF_002009175.1 WP_078076207.1 1082.4

**Figure 1.** Exemplary result of SYN-view. The figure shows two alignments of NGIs throughout the closest relatives of *Streptomyces nivues* NCIMB 11891 (Table 1). Note that for a clear comparison, only two NGI alignments are shown, while three were found (Supplementary Data). (**A**) NGI of DNA topoisomerase which is regularly observed in close relatives with the structure of the NGI is well conserved. (**B**) The NGI of the duplicated, resistant *gyrB* is unique to the antibiotic producing strain and can easily be distinguished.

While the housekeeping genes play an important role in target-directed genome mining approaches and BGC prioritization, the context of the gene neighborhood has not yet been focused on. In order to address this issue, here we introduce SYN-view, for further improvement of prioritization of the BGCs, based on a self-resistance approach. With the aid of phylogenetic methods such as autoMLST [18], which provides a high-resolution species tree of a strain of interest, SYN-view compares NGIs, based on user-provided target genes to homologous NGIs from closest relatives. Unlike other tools such as MultiGeneBlast [19], which blasts a complete cluster to a specific database to find similar clusters, our pipeline aims to distinguish a potential target resistance gene from regular housekeeping genes, by rapidly comparing NGIs from closely related taxa.

## 2. Results and Discussion

Here, we present SYN-view, an easy-to-use pipeline in order to make rapid comparison of NGIs and provide an additional way to detect putative novel antibiotic resistant targets. SYN-view allows for easy to interpret visualizations of NGIs in order to distinguish genes of interest with different functions. Using an external tool such as autoMSLT [18], SYN-view uses homology search tools to find the input protein and its surrounding genes from closest taxa, in order to perform a synteny search for easy detection of unique gene cluster structures. SYN-view can be easily installed using conda packages [20] and is publicly available at https://bitbucket.org/jstahlecker/syn-view/. An overview of the workflow is illustrated in Figure 2.

### *2.1. Positive Controls*

For the proof of concept of our proposed method, first we examined bacterial strains reported for antibiotic production with known resistance mechanisms shown in Table 1, to test if there is a significant difference between NGI structures of regular housekeeping genes and genes responsible for self-resistance. Results suggested that when the resistance mechanism includes a duplication event, difference in respective NGIs can be easily recognized. In certain cases where resistance genes have been mutated instead of duplicated (Table 1, *A. mediteranei* S699, *rpoB*), differences in NGIs could not be observed. Nevertheless, it would be possible to detect a difference in NGIs even if there is no duplication of self resistance genes but if they are unique to a certain bacterial genome. All of the corresponding results are visualized in detail in the Supplementary Results.

**Table 1.** SYN-view analysis of example antibiotic producing strains with identified self-resistance genes. For comparison, respective ARTS hits are also provided from previous papers [11,21] (D: Duplication, B: BGC proximity, R: Resistance, P: Phylogeny). "Search Type" column indicates how the search was performed: H stands for HMM mode while B stands for blastp and the following indicates the corresponding TIGRFAM model and gene accession number, respectively. Easily identifiable differences are denoted as "Yes", if no difference is visible marked as "No".

| Organism | Resistance Gene | Search Type | ARTS Hits | Identifiable |
|---|---|---|---|---|
| *Streptomyces niveus* NCIMB 11891 | *gyrB* | H: TIGR01059 | D,B,R,P | Yes |
| *Streptomyces roseochromogenes* DS 12.976 | *gyrB* | H: TIGR01059 | D,B,R,P | Yes |
| *Burkholderia thailandensis* E264 | *accA* | H: TIGR00513 | D,B,R,P | Yes [a] |
| *Salinospora tropica* CNB-440 | beta-proteasome subunit | H: TIGR03690 | D,B,R,P | Yes [a] |
| *Myxococcus xanthus* DK 1622 | *lspa*: signal peptidase II | H: TIGR00077 | D,B,P | Yes |
| *Bacillus cereus* ATCC 14579 | duplicated RL11 | H: TIGR01632 | D,P | Yes |
| *Nordica farnica* IFM 10152 | *rpoB* | H: TIGR02013 | D | Yes |
| *Agrobacterium radiobacter* K84 | Leu-tRNA synthase | H: TIGR00396 | D,P | Yes |
| *Streptomyces viridochromogenes* Tue57 | 23S rRNA methyltransferase | B: AAG32066.1 | No Hits | Yes |
| *Amycolatopsis mediterranei* S699 | *rpoB* | H: TIGR02013 | R | No |

[a] A difference was better observed after using 50 rather than the default 10 closest genomes.

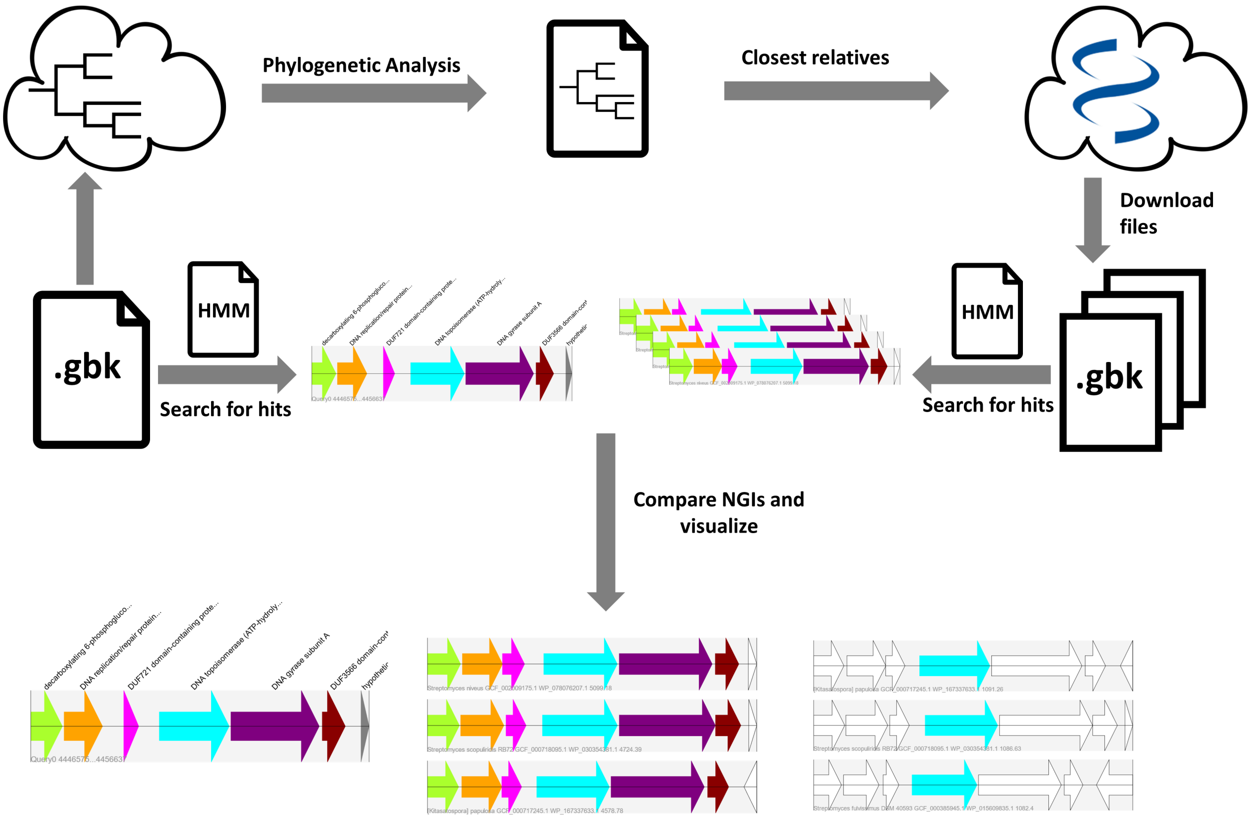

Phylogenetic Analysis
Closest relatives
Download files
HMM
.gbk
Search for hits
HMM
Search for hits
.gbk
Compare NGIs and visualize
decarboxylating 6-phosphogluco...
DNA replication/repair protein...
DUF721 domain-containing prote...
DNA topoisomerase (ATP-hydroly...
DNA gyrase subunit A
DUF3566 domain-con
hypothetic

**Figure 2.** Schematic workflow. A phylogeny file needs to be created using autoMLST or an appropriate folder must be specified. Based on that, the 10 closest relatives are downloaded from the NCBI refseq database. Using an appropirate hmm or protein fasta file, NGIs are created, scored, and sorted. Finally, the results are saved as an svg file.

### 2.2. *SYN-View as a Complementary Method*

In order to prove that SYN-view can further improve the current ARTS pipeline as a complementary method, we employed a final test case where ARTS could not find hits for a known self resistance mechanism. As stated in the first ARTS paper, 23S rRNA methyltransferase, which confers resistance for Avilamycin, was undetected by hmmsearch due to its short sequence length and low homology score. As HMMs are dependent on profiles built from multiple sequence alignment [22], it may fail to represent sequences which are not fully reflecting specific domains characterized from respective proteins. For such cases, SYN-view supports homology search using blastp algorithm, which makes it possible for users to analyze shorter sequences or proteins without an accurate HMM model. As shown in Supplementary Figure S1, the synteny among closest relatives of the NGI of 23S rRNA methyltransferase, conferring self-resistance, is significantly different than the NGI with regular housekeeping function.

## 3. Materials and Methods

*Input Options and Workflow*

An overview of the workflow is illustrated in Figure 2. First, SYN-view needs an annotated genome file in GenBank format (gbff, gbk). Additionally, an HMM or protein fasta file for a gene/protein is required, which is used to either run hmmsearch [23] or blastp [23], against the input genome to find similar proteins. SYN-view uses default cut-off values for hmmsearch and blastp algorithms, which can be redefined by the user. Using Biopython [24], the input genome is parsed and per hit, a query NGI is created based on the proximity of the respective hit. By default, this proximity setting is three surrounding genes in both sides of the gene of interest; however, it can be changed to decrease/increase the size of the NGI. Finally, close relatives of the input genome must be set, for the synteny search. For this purpose, the user can either provide the result file of an autoMLST job (mash_distances.txt, recommended) or provide a custom folder with specified genomes in GenBank format. If an autoMLST result is provided, the 10 closest organisms are, by default, downloaded from NCBIs RefSeq database [25]. As stated earlier, increasing the number of closest organisms would also increase the quality of the result. For the purposes of speed, it was set as 10 for default but can be changed via command line arguments. After downloading genomes, the next part of the SYN-view pipeline is detecting the input protein sequences from given genome, creating query NGIs. Afterwards, a database is created, containing only the NGIs from the closest relatives based on the input protein. The NGIs of the input are then blasted against the database and the NGI hits are scored by cumulative blast bit score. Higher bit scores suggest higher sequence similarity, while being independent of the database size. Therefore, summing over all individual bit scores of a NGI gives an indication of the sequence similarity of the whole NGI with respect to the query. In the results folder, all NGI hits per query can be analyzed using the corresponding visualization as an svg file as explained in Supplementary Results and can be compared to other hits using a standard web browser (Figure 1). The color coding makes it easier to identify similar hits to unique gene cluster structures. Same color indicates similarity to the query protein, while white suggests no hits, with the exception of being white colored in the query NGI, which indicates that the protein does not have a defined translated sequence.

## 4. Conclusions

With SYN-view, we developed a program that allows a rapid and easy to interpret overview about the gene neighborhoods of genes of interest. This can be used as an additional criterium to detect putative antibiotic resistant targets. However, SYN-view can also be used for the exploration of cluster formations of specific genes in phylogenetically similar bacteria. A preceding prioritization of genes of interest such as an ARTS run is recommended, as both tools utilize self-resistance. As it is impossible to identify resistance genes based on a single criterion, SYN-view is meant to be used as a complementary tool to help researchers in their efforts for the prioritization of their targets. As the genomic content of a NGI is specific for different cases, it is incumbent on the user to further analyze results. In order to further automate this workflow and increase efficiency our aim is to implement this functionality in ARTS web-server.

**Supplementary Materials:** The following are available online: Figure S1: SYN-view result of *Streptomyces viridochromogenes* Tue57, Figure S2: SYN-view result of *Streptomyces roseochromogenes* DS 12976, Figure S3: SYN-view result of *Burkholderia thailandensis* E264, Figure S4: SYN-view result of *Salinospora tropica* CNB-440, Figure S5: SYN-view result of *Myxococcus xanthus* DK 1622, Figure S6: SYN-view result of *Bacillus cereus* ATCC 14579, Figure S7: SYN-view result of *Nordica farnica* IFM 10152, Figure S8: SYN-view result of *Agrobacterium radiobacter* K84, Figure S9: SYN-view result of *Amycolatopsis mediterranei* S699.

**Author Contributions:** Conceptualization, E.M. and N.Z.; methodology, M.D.M. and N.Z.; software, J.S. and M.D.M.; validation, J.S. and M.D.M.; formal analysis, J.S.; investigation, J.S. and M.D.M.; resources, N.Z.; data curation, J.S. and M.D.M.; writing—original draft preparation, J.S. and M.D.M.; writing—review and editing, M.D.M., N.Z., and E.M.; visualization, J.S.; supervision, M.D.M. and N.Z.; project administration, N.Z.; funding acquisition, N.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** The authors acknowledge the German Center for Infection Research (DZIF TTU09.704) for funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data is contained within the article or supplementary material.

**Acknowledgments:** The authors acknowledge the use of de.NBI cloud and the support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen and the Federal Ministry of Education and Research (BMBF) through grant no 031 A535A.

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

**Sample Availability:** Samples of the compounds are not available from the authors.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| BGC | Biosynthetic gene cluster |
| NGI | neighborhood of gene of interest |
| hmm | Hidden Markov Model |
| ARTS | Antibiotic Resistant Target Seeker |
| BLAST | Basic Local Alignment Search Tool |

## References

1. Michael, C.A.; Dominey-Howes, D.; Labbate, M. The Antimicrobial Resistance Crisis: Causes, Consequences, and Management. *Front. Public Health* **2014**, *2*. [CrossRef] [PubMed]
2. Cragg, G.M.; Newman, D.J. Natural products: A continuing source of novel drug leads. *Biochim. Biophys. Acta (BBA)-Gen. Subj.* **2013**, *1830*, 3670–3695. [CrossRef] [PubMed]
3. Newman, D.J.; Cragg, G.M. Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* **2016**, *79*, 629–661. [CrossRef] [PubMed]
4. Zhang, M.M.; Qiao, Y.; Ang, E.L.; Zhao, H. Using natural products for drug discovery: The impact of the genomics era. *Expert Opin. Drug Discov.* **2017**, *12*, 475–487. [CrossRef] [PubMed]
5. Ziemert, N.; Alanjary, M.; Weber, T. The evolution of genome mining in microbes—A review. *Nat. Prod. Rep.* **2016**, *33*, 988–1005. [CrossRef] [PubMed]
6. Bachmann, B.O.; Van Lanen, S.G.; Baltz, R.H. Microbial genome mining for accelerated natural products discovery: Is a renaissance in the making? *J. Ind. Microbiol. Biotechnol.* **2014**, *41*, 175–184. [CrossRef]
7. Blin, K.; Shaw, S.; Steinke, K.; Villebro, R.; Ziemert, N.; Lee, S.Y.; Medema, M.H.; Weber, T. antiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **2019**, *47*, W81–W87. [CrossRef]
8. Skinnider, M.A.; Merwin, N.J.; Johnston, C.W.; Magarvey, N.A. PRISM 3: Expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res.* **2017**, *45*, W49–W54. [CrossRef]
9. Medema, M.H.; Kottmann, R.; Yilmaz, P.; Cummings, M.; Biggins, J.B.; Blin, K.; De Bruijn, I.; Chooi, Y.H.; Claesen, J.; Coates, R.C.; et al. Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **2015**, *11*, 625–631. [CrossRef]
10. Palaniappan, K.; Chen, I.M.A.; Chu, K.; Ratner, A.; Seshadri, R.; Kyrpides, N.C.; Ivanova, N.N.; Mouncey, N.J. IMG-ABC v. 5.0: An update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Res.* **2020**, *48*, D422–D430. [CrossRef]
11. Alanjary, M.; Kronmiller, B.; Adamek, M.; Blin, K.; Weber, T.; Huson, D.; Philmus, B.; Ziemert, N. The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Res.* **2017**, *45*, W42–W48. [CrossRef] [PubMed]
12. Almabruk, K.H.; Dinh, L.K.; Philmus, B. Self-resistance of natural product producers: Past, present, and future focusing on self-resistant protein variants. *ACS Chem. Biol.* **2018**, *13*, 1426–1437. [CrossRef] [PubMed]
13. Yan, Y.; Liu, Q.; Zang, X.; Yuan, S.; Bat-Erdene, U.; Nguyen, C.; Gan, J.; Zhou, J.; Jacobsen, S.E.; Tang, Y. Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action. *Nature* **2018**, *559*, 415–418. [CrossRef] [PubMed]
14. Tang, X.; Li, J.; Millán-Aguiñaga, N.; Zhang, J.J.; O'Neill, E.C.; Ugalde, J.A.; Jensen, P.R.; Mantovani, S.M.; Moore, B.S. Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem. Biol.* **2015**, *10*, 2841–2849. [CrossRef]
15. Freel, K.C.; Millán-Aguiñaga, N.; Jensen, P.R. Multilocus sequence typing reveals evidence of homologous recombination linked to antibiotic resistance in the genus Salinispora. *Appl. Environ. Microbiol.* **2013**, *79*, 5997–6005. [CrossRef]
16. Thaker, M.N.; Wang, W.; Spanogiannopoulos, P.; Waglechner, N.; King, A.M.; Medina, R.; Wright, G.D. Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nat. Biotechnol.* **2013**, *31*, 922–927. [CrossRef]
17. O'Neill, E.C.; Schorn, M.; Larson, C.B.; Millán-Aguiñaga, N. Targeted antibiotic discovery through biosynthesis-associated resistance determinants: Target directed genome mining. *Crit. Rev. Microbiol.* **2019**, *45*, 255–277. [CrossRef]
18. Alanjary, M.; Steinke, K.; Ziemert, N. AutoMLST: An automated web server for generating multi-locus species trees highlighting natural product potential. *Nucleic Acids Res.* **2019**, *47*, W276–W282. [CrossRef]
19. Medema, M.H.; Takano, E.; Breitling, R. Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol. Biol. Evol.* **2013**, *30*, 1218–1223. [CrossRef]

20. Grüning, B.; Dale, R.; Sjödin, A.; Chapman, B.A.; Rowe, J.; Tomkins-Tinch, C.H.; Valieris, R.; Köster, J. Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **2018**, *15*, 475–476. [CrossRef]
21. Mungan, M.D.; Alanjary, M.; Blin, K.; Weber, T.; Medema, M.H.; Ziemert, N. ARTS 2.0: Feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. *Nucleic Acids Res.* **2020**. [CrossRef] [PubMed]
22. Eddy, S.R. Hidden markov models. *Curr. Opin. Struct. Biol.* **1996**, *6*, 361–365. [CrossRef]
23. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [CrossRef] [PubMed]
24. Cock, P.J.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423. [CrossRef]
25. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2015**, *44*, D733–D745. [CrossRef]

# Supplementary Materials: SYN-view, a Phylogeny Based Synteny Exploration Tool for the Identification of Gene Clusters Linked to Antibiotic Resistance

**Jason Stahlecker[1], Erik Mingyar[1], Nadine Ziemert[1,2], Mehmet Direnç Mungan[1,2]***

## 1. Contents of the Result Folder

In the results directory (default is the current directory) the "*.faa" file of the corresponding "*.gbk/gbff" file is created. Additionally, a folder named "SYN-view_results" is created in which all other files are saved. Most importantly the subfolder "RESULTS" contains all results as *svg* and *txt* files. The folder "protein_faa" contains all ".faa" files of either the genome_gbff folder (default) or the specified genomes. For each hit in the input.gbk a query folder is created. In hmm mode all hmm runs are saved in the folder "SYN-view_results", while in protein mode the blast databases of the genomes of the relatives are saved in the folder "blast_databases" and of the query in "SYN-view_results". Moreover, temporary files, needed for transfer of information are saved in "SYN-view_results". The query folders contain the blast results and all svg files. The supplementary folder only contains the input parameters and the "RESULTS" subfolder.

## 2. Figures of the Results of Table 1

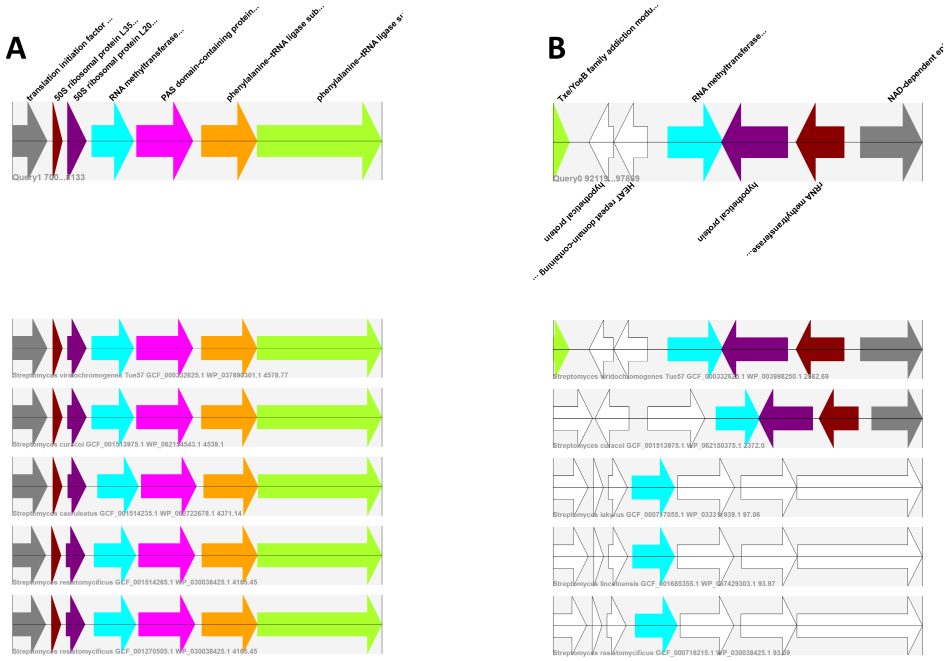

A
translation initiation factor ...
50S ribosomal protein L35...
50S ribosomal protein L20...
RNA methyltransferase...
PAS domain-containing protein...
phenylalanine--tRNA ligase sub...
phenylalanine--tRNA ligase su
Query1 700...8133
Streptomyces viridochromogenes Tue57 GCF_000332625.1 WP_037890301.1 4578.77
Streptomyces curacoi GCF_001513975.1 WP_062154543.1 4539.1
Streptomyces caeruleatus GCF_001514235.1 WP_062722678.1 4371.14
Streptomyces resistomycificus GCF_001514265.1 WP_030038425.1 4165.45
Streptomyces resistomycificus GCF_001270505.1 WP_030038425.1 4165.45
B
Txe/YoeB family addiction modu...
RNA methyltransferase...
NAD-dependent en
Query0 92119...97869
hypothetical protein
... HEAT repeat domain-containing
hypothetical protein
...rRNA methyltransferase
Streptomyces viridochromogenes Tue57 GCF_000332625.1 WP_003998250.1 2582.69
Streptomyces curacoi GCF_001513975.1 WP_062150375.1 2372.0
Streptomyces iakyrus GCF_000717055.1 WP_033311939.1 97.06
Streptomyces lincolnensis GCF_001685355.1 WP_067429303.1 93.97
Streptomyces resistomycificus GCF_000716215.1 WP_030038425.1 93.69

**Figure S1.** SYN-view result of *Streptomyces viridochromogenes* Tue57. The figure shows two alignments of NGIs throughout the closest relatives of *Streptomyces viridochromogenes* Tue57 (Table 1). Note that for a clear comparison, only two NGI alignments are shown, while three were found (Supplementary data). **A**: NGI of rRNA methyltransferase which is regularly observed in close relatives. **B**: The NGI of the resistant rRNA methyltransferase is unique to the antibiotic producing strain and can easily be distinguished. The first resulted NGI is identical since its the NGI from the same organism, found in autoMLST search.

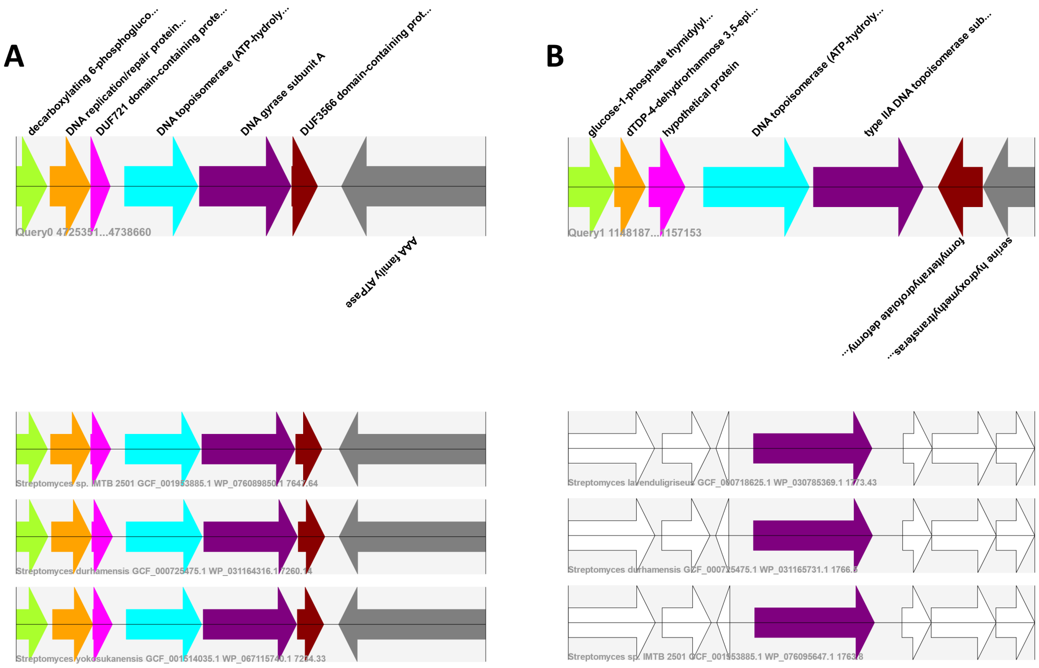

A
decarboxylating 6-phosphogluco...
DNA replication/repair protein...
DUF721 domain-containing prote...
DNA topoisomerase (ATP-hydroly...
DNA gyrase subunit A
DUF3566 domain-containing prot...
Query0 4725351...4738660
AAA family ATPase
Streptomyces sp. IMTB 2501 GCF_001953885.1 WP_076089850.1 7647.64
Streptomyces durhamensis GCF_000725475.1 WP_031164316.1 7260.14
Streptomyces yokosukanensis GCF_001514035.1 WP_067115740.1 7264.33
B
glucose-1-phosphate thymidylyl...
dTDP-4-dehydrorhamnose 3,5-epi...
hypothetical protein
DNA topoisomerase (ATP-hydroly...
type IIA DNA topoisomerase sub...
Query1 1148187...1157153
formyltetrahydrofolate deformy...
serine hydroxymethyltransferas...
Streptomyces lavenduligriseus GCF_000718625.1 WP_030785369.1 1773.43
Streptomyces durhamensis GCF_000725475.1 WP_031165731.1 1766.5
Streptomyces sp. IMTB 2501 GCF_001953885.1 WP_076095647.1 1763.8

**Figure S2.** SYN-view result of *Streptomyces roseochromogenes* DS 12.976. Only two of four NGIs are displayed (supplementary data). **A**: NGI of *gyrB* which is regularly observed in close relatives. **B**: The NGI is unique to the strain and can easily be distinguished.

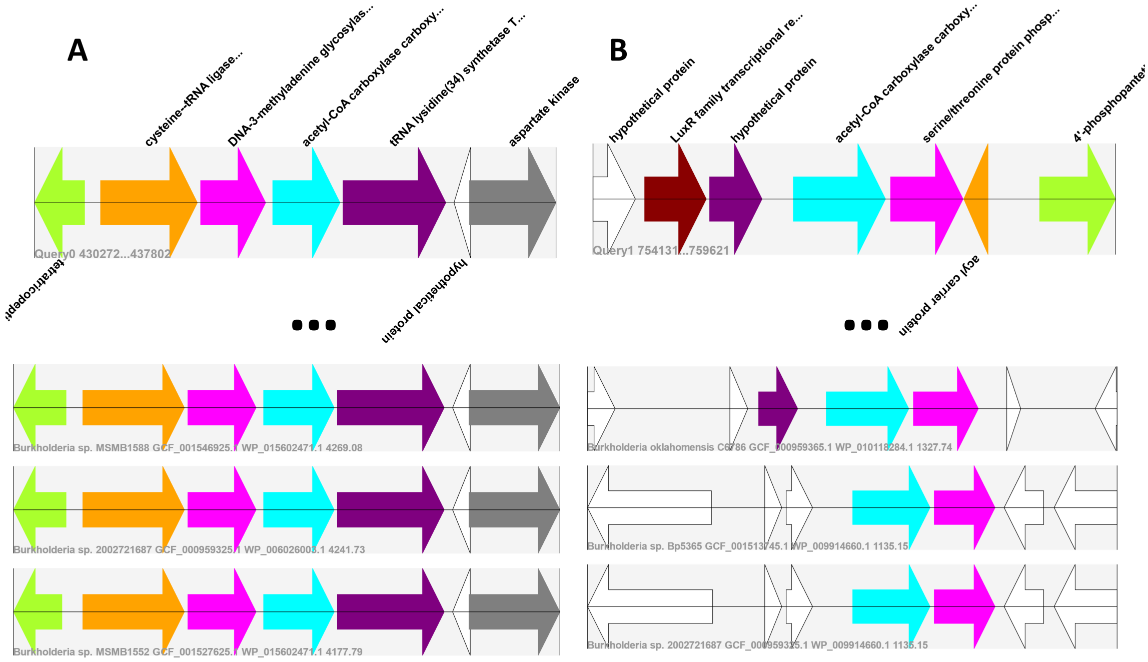

A
cysteine--tRNA ligase...
DNA-3-methyladenine glycosylas...
acetyl-CoA carboxylase carboxy...
tRNA lysidine(34) synthetase T...
aspartate kinase
Query0 430272...437802
tetratricopept
hypothetical protein
Burkholderia sp. MSMB1588 GCF_001546925.1 WP_015602471.1 4269.08
Burkholderia sp. 2002721687 GCF_000959325.1 WP_006026003.1 4241.73
Burkholderia sp. MSMB1552 GCF_001527625.1 WP_015602471.1 4177.79
B
hypothetical protein
LuxR family transcriptional re...
hypothetical protein
acetyl-CoA carboxylase carboxy...
serine/threonine protein phosp...
4'-phosphopante
Query1 754131...759621
acyl carrier protein
Burkholderia oklahomensis C6786 GCF_000959365.1 WP_010118284.1 1327.74
Burkholderia sp. Bp5365 GCF_001513745.1 WP_009914660.1 1135.15
Burkholderia sp. 2002721687 GCF_000959325.1 WP_009914660.1 1135.15

**Figure S3.** SYN-view result of *Burkholderia thailandensis* E264. Please note that the first displayed NGI of the close relatives is the eleventh NGI in total. All NGIs before are from different *B. thailandensis* strains and show no difference to the query. **A**: NGI of accA which is regularly observed in close relatives. **B**: The NGI is less frequent and differences are observed after evaluation of the results.

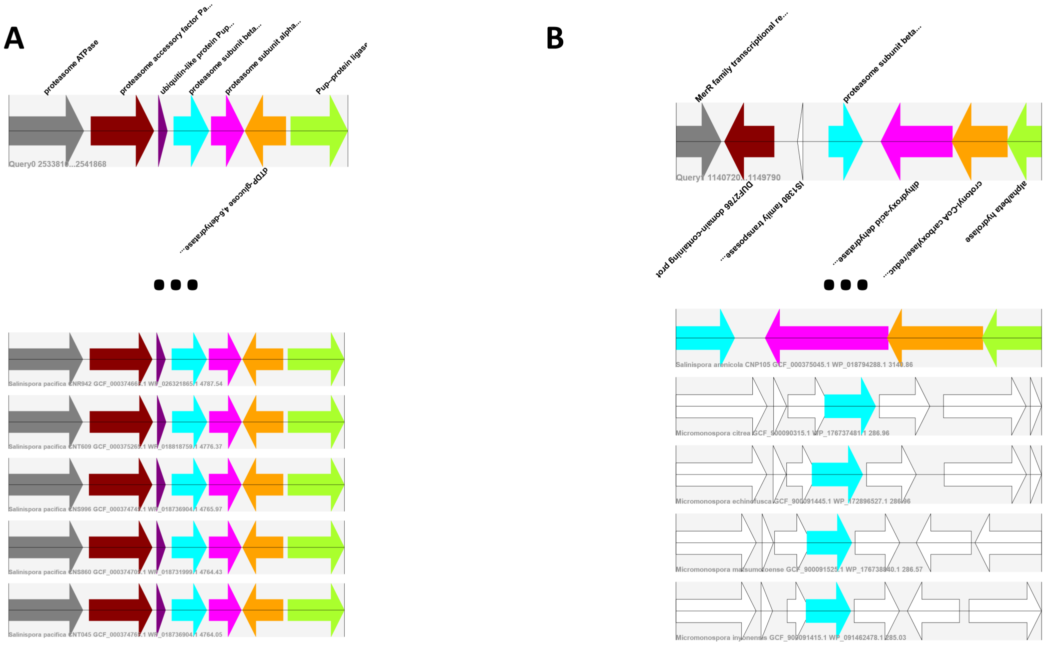

A
proteasome ATPase
proteasome accessory factor Pa...
ubiquitin-like protein Pup...
proteasome subunit beta...
proteasome subunit alpha...
Pup--protein ligase
Query0 2533810...2541868
dTDP-glucose 4,6-dehydratase...
Salinispora pacifica CNR942 GCF_000374665.1 WP_026321865.1 4787.54
Salinispora pacifica CNT609 GCF_000375265.1 WP_018818759.1 4776.37
Salinispora pacifica CNS996 GCF_000374745.1 WP_018736904.1 4765.97
Salinispora pacifica CNS860 GCF_000374705.1 WP_018731999.1 4764.43
Salinispora pacifica CNT045 GCF_000374765.1 WP_018736904.1 4764.05
B
MerR family transcriptional re...
proteasome subunit beta...
Query1 1140720...1149790
DUF2786 domain-containing prot
IS1380 family transposase...
dihydroxy-acid dehydratase...
crotonyl-CoA carboxylase/reduc...
alpha/beta hydrolase
Salinispora arenicola CNP105 GCF_000375045.1 WP_018794288.1 3140.86
Micromonospora citrea GCF_900090315.1 WP_176737481.1 286.96
Micromonospora echinofusca GCF_900091445.1 WP_172896527.1 286.96
Micromonospora matsumotoense GCF_900091525.1 WP_176738840.1 286.57
Micromonospora inyonensis GCF_900091415.1 WP_091462478.1 285.03

**Figure S4.** SYN-view result of *Salinospora tropica* CNB-440. Please note that the first displayed NGI of the close relatives is the tenth NGI in total. All NGIs before are from different *S. tropica* strains and show no difference to the query. **A**: NGI of the beta-proteasome subunit which is regularly observed in close relatives. **B**: The NGI is less frequent and differences are observed after evaluation of the results.

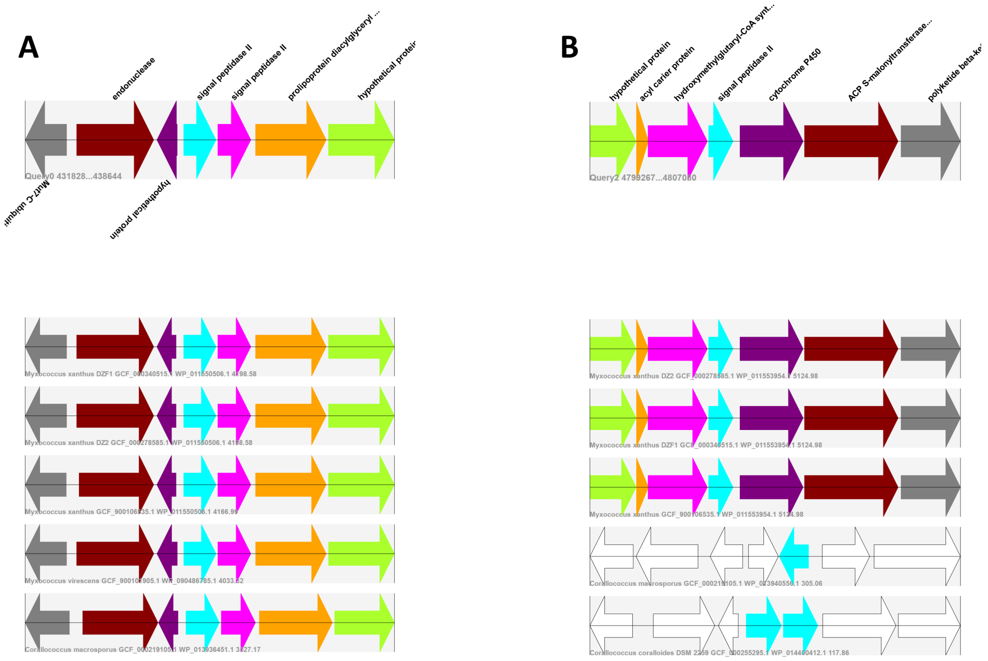

A
endonuclease
signal peptidase II
signal peptidase II
prolipoprotein diacylglyceryl ...
hypothetical protein
Query0 431828...438644
Mut7-C ubiquit
hypothetical protein
Myxococcus xanthus DZF1 GCF_000340515.1 WP_011550506.1 4198.58
Myxococcus xanthus DZ2 GCF_000278585.1 WP_011550506.1 4198.58
Myxococcus xanthus GCF_900106535.1 WP_011550506.1 4166.99
Myxococcus virescens GCF_900101905.1 WP_090486785.1 4033.52
Corallococcus macrosporus GCF_000219105.1 WP_013936451.1 3527.17
B
hypothetical protein
acyl carier protein
hydroxymethylglutaryl-CoA synt...
signal peptidase II
cytochrome P450
ACP S-malonyltransferase...
polyketide beta-ke
Query2 4799267...4807080
Myxococcus xanthus DZ2 GCF_000278585.1 WP_011553954.1 5124.98
Myxococcus xanthus DZF1 GCF_000340515.1 WP_011553954.1 5124.98
Myxococcus xanthus GCF_900106535.1 WP_011553954.1 5124.98
Corallococcus macrosporus GCF_000219105.1 WP_013940556.1 305.06
Corallococcus coralloides DSM 2259 GCF_000255295.1 WP_014400412.1 117.86

**Figure S5.** SYN-view result of *Myxococcus xanthus* DK 1622. Only two of three NGIs are displayed (supplementary data). **A**: NGI of the signal peptidase II which is regularly observed in close relatives. **B**: Few *M. xanthus* strains contain the query NGI, but starting at the fouth NGI no similar NGIs can be observed.

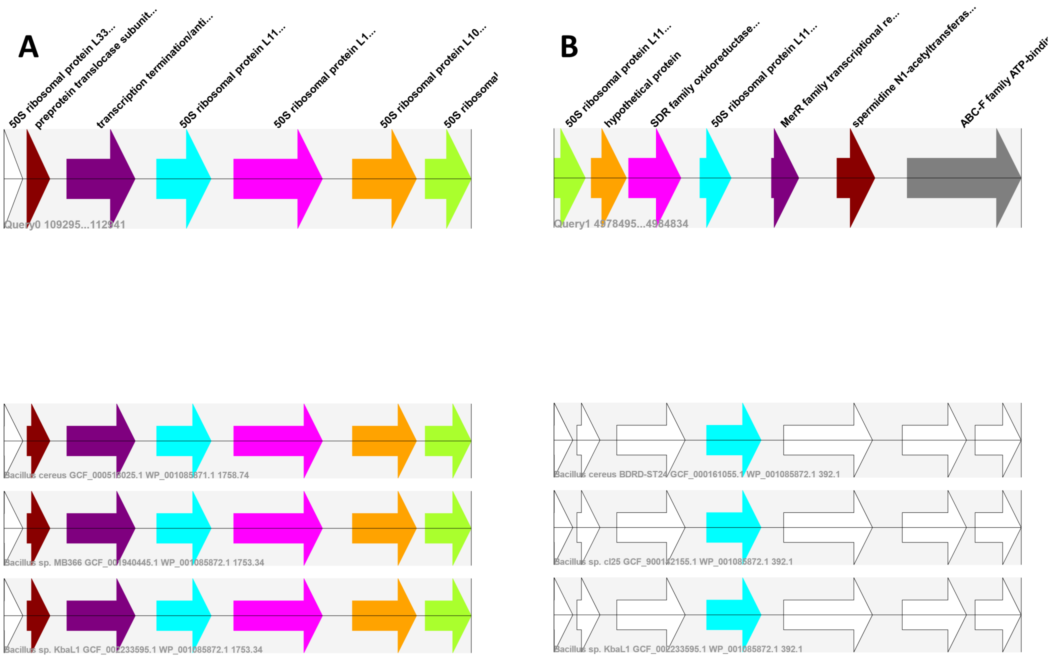

A
50S ribosomal protein L33...
preprotein translocase subunit...
transcription termination/anti...
50S ribosomal protein L11...
50S ribosomal protein L1...
50S ribosomal protein L10...
50S ribosomal
Query0 109295...112841
Bacillus cereus GCF_000518025.1 WP_001085871.1 1758.74
Bacillus sp. MB366 GCF_001940445.1 WP_001085872.1 1753.34
Bacillus sp. KbaL1 GCF_002233595.1 WP_001085872.1 1753.34
B
50S ribosomal protein L11...
hypothetical protein
SDR family oxidoreductase...
50S ribosomal protein L11...
MerR family transcriptional re...
spermidine N1-acetyltransferas...
ABC-F family ATP-bindi
Query1 4978495...4984834
Bacillus cereus BDRD-ST24 GCF_000161055.1 WP_001085872.1 392.1
Bacillus sp. cl25 GCF_900182155.1 WP_001085872.1 392.1
Bacillus sp. KbaL1 GCF_002233595.1 WP_001085872.1 392.1

**Figure S6.** SYN-view result of *Bacillus cereus* ATCC 14579. **A**: NGI of RL11 which is regularly observed in close relatives. **B**: The NGI is unique to the strain and can easily be distinguished.

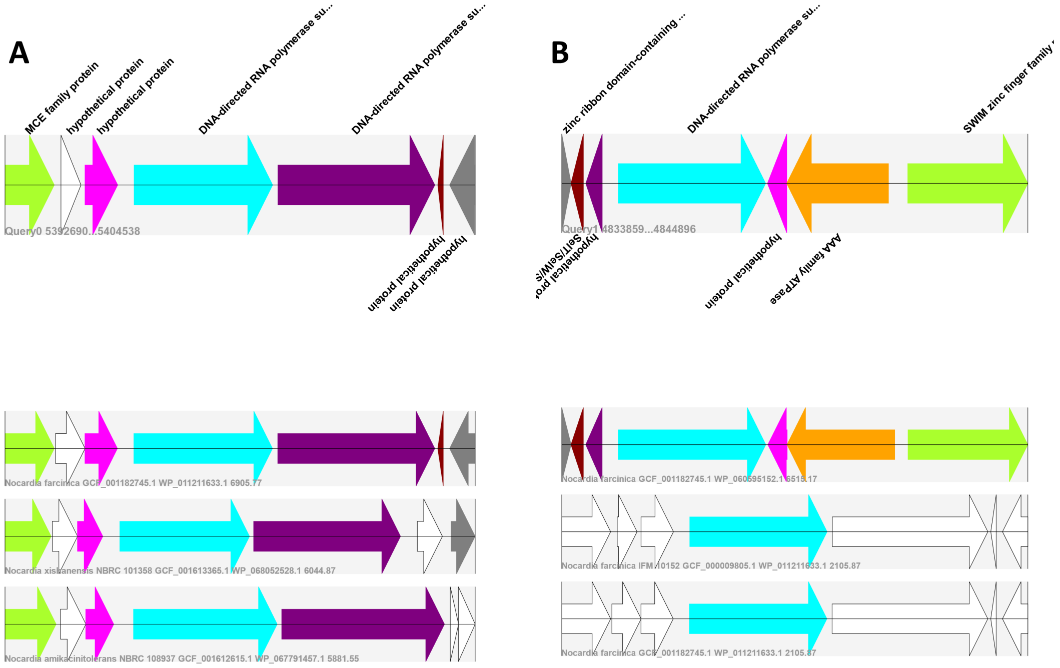

A
MCE family protein
hypothetical protein
hypothetical protein
DNA-directed RNA polymerase su...
DNA-directed RNA polymerase su...
Query0 5392690...5404538
hypothetical protein
hypothetical protein
Nocardia farcinica GCF_001182745.1 WP_011211633.1 6905.77
Nocardia xishanensis NBRC 101358 GCF_001613365.1 WP_068052528.1 6044.87
Nocardia amikacinitolerans NBRC 108937 GCF_001612615.1 WP_067791457.1 5881.55
B
zinc ribbon domain-containing ...
DNA-directed RNA polymerase su...
SWIM zinc finger family
Query1 4833859...4844896
SsrA/SelW/S
hypothetical pro
hypothetical protein
AAA family ATPase
Nocardia farcinica GCF_001182745.1 WP_060595152.1 5515.17
Nocardia farcinica IFM 10152 GCF_000009805.1 WP_011211633.1 2105.87
Nocardia farcinica GCF_001182745.1 WP_011211633.1 2105.87

**Figure S7.** SYN-view result of *Nordica farnica* IFM 10152. **A**: NGI of *rpoB* which is regularly observed in close relatives. **B**: One *N. farnica* strain contains the same NGI. The NGI is unique to all other close relatives.

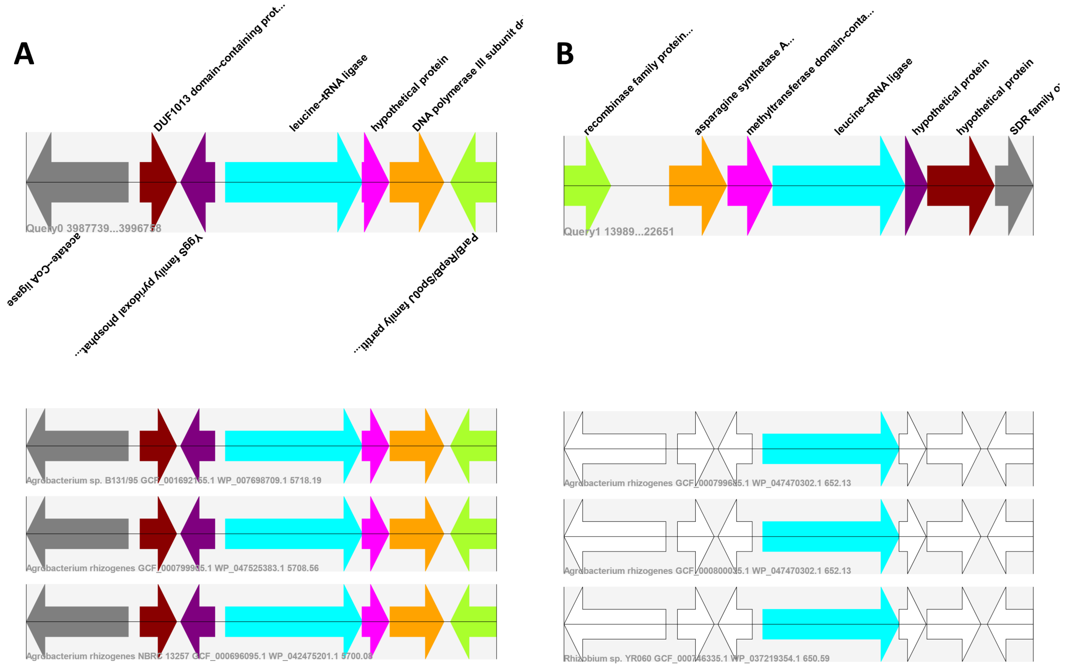

A
DUF1013 domain-containing prot...
leucine--tRNA ligase
hypothetical protein
DNA polymerase III subunit de
Query0 3987739...3996758
acetate--CoA ligase
YggS family pyridoxal phosphat...
ParB/RepB/Spo0J family partiti...
Agrobacterium sp. B131/95 GCF_001692155.1 WP_007698709.1 5718.19
Agrobacterium rhizogenes GCF_000799905.1 WP_047525383.1 5708.56
Agrobacterium rhizogenes NBRC 13257 GCF_000696095.1 WP_042475201.1 5700.08
B
recombinase family protein...
asparagine synthetase A...
methyltransferase domain-conta...
leucine--tRNA ligase
hypothetical protein
hypothetical protein
SDR family o
Query1 13989...22651
Agrobacterium rhizogenes GCF_000799685.1 WP_047470302.1 652.13
Agrobacterium rhizogenes GCF_000800035.1 WP_047470302.1 652.13
Rhizobium sp. YR060 GCF_000746335.1 WP_037219354.1 650.59

**Figure S8.** SYN-view result of *Agrobacterium radiobacter* K84. Only two of five NGIs are displayed (supplementary data). **A**: NGI of the Leu-tRNA synthase which is regularly observed in close relatives. **B**: The NGI is unique to the strain and can easily be distinguished.

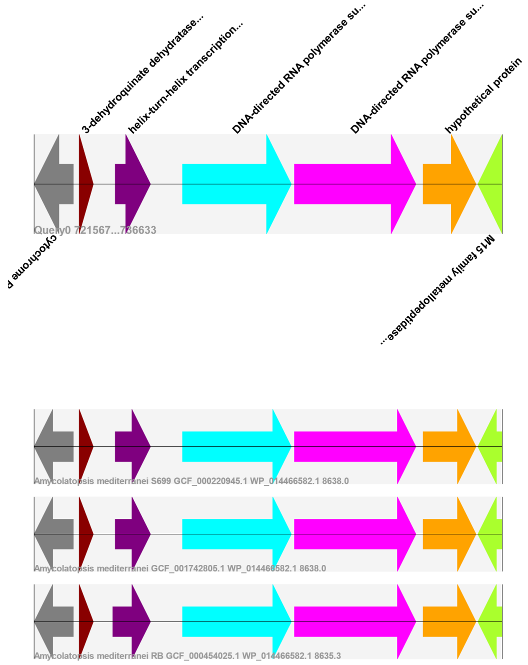

3-dehydroquinate dehydratase...
helix-turn-helix transcription...
DNA-directed RNA polymerase su...
DNA-directed RNA polymerase su...
hypothetical protein
Query0 721567...736633
cytochrome P
M15 family metallopeptidase...
Amycolatopsis mediterranei S699 GCF_000220945.1 WP_014466582.1 8638.0
Amycolatopsis mediterranei GCF_001742805.1 WP_014466582.1 8638.0
Amycolatopsis mediterranei RB GCF_000454025.1 WP_014466582.1 8635.3

**Figure S9.** SYN-view result of *Amycolatopsis mediterranei* S699. The NGI of *rpoB* does not show differences of close relatives

## 2.3 ARTS-DB: a database for antibiotic resistant targets

### Contributions

Dr. Kai Blin provided us with an easy route to download antiSMASH results. All the authors spent time reviewing and editing the final manuscript. Other than these, I did everything related to this paper with valued discussions from Prof. Dr. Nadine Ziemert.

*D736–D740 Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 28 October 2021*
*https://doi.org/10.1093/nar/gkab940*

# ARTS-DB: a database for antibiotic resistant targets

**Mehmet Direnç Mungan [1,2], Kai Blin [3] and Nadine Ziemert [1,2,*]**

[1] Interfaculty Institute of Microbiology and Infection Medicine, Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany, [2] German Center for Infection Research (DZIF), Partner Site Tübingen, 72076 Tübingen, Germany and [3] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet Bygning 220, 2800 Kgs. Lyngby, Denmark

Received August 12, 2021; Revised September 14, 2021; Editorial Decision September 27, 2021; Accepted October 01, 2021

## ABSTRACT

**As a result of the continuous evolution of drug resistant bacteria, new antibiotics are urgently needed. Encoded by biosynthetic gene clusters (BGCs), antibiotic compounds are mostly produced by bacteria. With the exponential increase in the number of publicly available, sequenced genomes and the advancements of BGC prediction tools, genome mining algorithms have uncovered millions of uncharacterized BGCs for further evaluation. Since compound identification and characterization remain bottlenecks, a major challenge is prioritizing promising BGCs. Recently, researchers adopted self-resistance based strategies allowing them to predict the biological activities of natural products encoded by uncharacterized BGCs. Since 2017, the Antibiotic Resistant Target Seeker (ARTS) facilitated this so-called target-directed genome mining (TDGM) approach for the prioritization of BGCs encoding potentially novel antibiotics. Here, we present the ARTS database, available at https://arts-db.ziemertlab.com/. The ARTS database provides pre-computed ARTS results for >70,000 genomes and metagenome assembled genomes in total. Advanced search queries allow users to rapidly explore the fundamental criteria of TDGM such as BGC proximity, duplication and horizontal gene transfers of essential housekeeping genes. Furthermore, the ARTS database provides results interconnected throughout the bacterial kingdom as well as links to known databases in natural product research.**

## INTRODUCTION

Throughout history, humanity has been in a constant battle with bacteria causing infectious diseases (1). Especially in the last decades, due to the escalation of multi-drug resistant bacteria, these continuously evolving pathogens have become a serious threat to human health. Consequently, there is an urgent need for novel antibiotics with new modes of action (2,3). Secondary metabolites (SMs) are the key molecules feeding antimicrobial drug development pipelines (4). These so-called natural products, are profusely found and isolated from fungal and bacterial organisms (5). The discovery of natural products has traditionally been centered on bioactivity screening. With the advent of genome sequencing in the last decade or two, *in silico* methods can now be used to complement these approaches. Presently, genome mining offers a wide range of computational applications that predict the biosynthetic gene clusters (BGCs) encoding enzymes necessary for the formation of natural products (6,7). Adopting algorithmic architectures like deep learning and hidden markov models, BGC prediction tools such as antiSMASH (8), PRISM (9) or DeepBGC (10), have been used in natural product research for over a decade. As a result of the genome mining efforts, hundreds of thousands of BGCs are continuously deposited in publicly available databases such as antiSMASH-DB (8) and Atlas of Biosynthetic Gene Clusters (IMG-ABC). The total of experimentally verified genome-mined BGCs however, falls even below 1% (11). Since experimental validation of a BGC and its compound is a labour-intensive process (12), a crucial task now is the prioritization of BGCs for further downstream analysis.

A recently established technique adopts a BGC prioritization approach leveraging the idea that in order to avoid suicide, bacteria need to be evolved in such a way that they are resistant to the compounds they produce (13). One of the resistance mechanisms bacteria use to protect themselves from their own bioactive compounds is the modification of the antibiotics target (14). In such processes, the duplicated and modified antibiotic target gene can be found within the BGC, providing self resistance (15,16). This so-called target-directed genome mining (TDGM) approach allowed researchers to predict the mode of action of the compounds encoded by uncharacterized BGCs and led to the identification of new natural products (17–19). Since 2017, the Antibiotic Resistant Target Seeker (ARTS) facilitated TDGM approaches in order to prioritize promising strains producing antibiotics with putative novel modes of action by rapidly linking housekeeping and known resis-

[*] To whom correspondence should be addressed. Tel: +49 7071 2978841; Email: nadine.ziemert@uni-tuebingen.de

tance genes to BGC proximity, duplication and horizontal gene transfer (HGT) events (20,21). By design, the ARTS pipeline functions as a web-server, analyzing user supplied genomes individually with a 'one job at a time' mentality which takes a certain processing time. In order to further improve our work on self resistance genome mining, we have developed the ARTS database, a user-friendly web-server for the extensive exploration of the bacterial kingdom using TDGM approaches. The ARTS database provides a global picture of ARTS results interconnected with the whole kingdom of bacteria and provides connections between potential targets and relevant databases containing additional information about respective BGCs or existing drugs. Currently, the ARTS database contains pre-computed ARTS results for a total of 27,096 high quality bacterial genomes obtained from NCBI's RefSeq database (22), also present in the antiSMASH-DB. Given that there is an ever-increasing usage of metagenomic applications on natural product research, we have also included 43,130 metagenome assembled genomes (MAGs) in the ARTS database described by Nayfach *et al.* (23).

The ARTS database allows researchers to facilitate TDGM based exploration through two main search functions. One of them is the exploration of fundamental ARTS hits such as BGC proximity, duplication and HGT evidence by using a query builder. All of the returned sequences are linked to individual ARTS and antiSMASH results for closer inspection. Second, a target-oriented exploration can be made. Here, the user can search a gene of interest throughout the database, in order to find phylogenetical and statistical information about a potential resistant target with respect to bacterial kingdom.

## DATABASE DESIGN

Using a multi-layered setup, the ARTS database provides rapid execution of provided queries using SQLAlchemy toolkit (https://www.sqlalchemy.org/) for relational mapping on a Flask-based framework (https://flask.palletsprojects.com/). The whole database is originally stored using SQLite database engine (https://www.sqlite.org/). The front end is comprised of jquery, bootstrap and ajax for high compatibility between different devices and browsers. The web service layer allows for easy execution of SQL logic packed in a single page. All ARTS results can be linked via web application and are stored on a disk hosted by de.NBI cloud (24).

### Genomic sequence content

The ARTS database includes genomic sequences, fueled by two different repositories. One of them is NCBI's publicly available RefSeq database (22) whose bacterial genomes are also used by the antiSMASH-DB. Selection and filtering of the genomes are explained in detail in the latest version of the antiSMASH-DB described by Blin *et al.* (8). In summary, the ARTS database contains 27,096 high quality bacterial genomes (Figure 1A) which were selected according to their completeness level. To discard fragmented and low quality assemblies, genomes labeled as complete assembly or with contig count <100 were included in the database. Using MASH (25), redundant sequences were also filtered out with a similarity cutoff of 99.6%.

Additionally, the ARTS database covers sequences from metagenomes. In a recent study published in 2021, Nayfach and his colleagues explored microbiomes from a wide range of habitats all around the Earth as well as mammalian hosts, forming the Genomes from the Earth's Microbiomes (GEM) catalogue. GEM has supplied the community with >52,000 MAGs and their genome mining data regarding BGCs deposited in IMG/M (26), greatly increasing the existing knowledge about secondary metabolite biosynthetic potential of microorganisms. However, for an accurate housekeeping gene search, ARTS pipeline is dependent on reference sets which were built using closely related taxa. Therefore, it doesn't guarantee high accuracy for bacteria that are assigned to a candidate phylum. For ARTS database, we have selected >43,000 MAGs based on their taxonomic annotation via GTDB (27) (Figure 1B) that fit the ARTS reference sets.

## MAIN APPLICATIONS

As mentioned earlier, the ARTS database offers two search options: 'Query Building' and 'Target-Oriented Search'. Using a query builder, users can explore available data sources in the ARTS database through four main routes (Figure 2A). These routes allow for: generating statistical summaries of ARTS results for the initial filtering of genomes of interest, finding essential housekeeping genes that have hits for fundamental ARTS criteria, exploring duplication rates of a gene of interest based on its occurrence frequency in different phyla as well as an essential genes function and frequency in different BGCs. Complex queries can be easily built by using the 'Add Term' button and adding the conditions indicating advanced properties of the search. The resulting tables can also be filtered, sorted or searched dynamically, allowing easy navigation through the resulting potential targets.

In addition, the 'Target-Oriented Search' option gives a broader view about the characteristics of the selected gene such as its proximity to different BGC types or in which phyla it is considered as an essential housekeeping gene. In order to maintain a high level of inter-operability, the ARTS database offers cross-links to available repositories such as MIBiG (28) and BiGFAM (29) for exterior information about a predicted BGC and its cluster families, respectively. Furthermore, DrugBank (30) entries are provided where applicable, for additional information about a genes affiliation with existing drugs and their known modes of action.

### Building queries

*Case study.* In a recent study, Hoskisson *et al.* investigated how the expansion of primary metabolism plays a role in the biosynthesis of antibiotics (31). In order to find gene expansion events in primary metabolism pathways, they analyzed 612 actinobacterial genomes to generate gene frequencies for 60 genera. Of note, they were exclusively interested to gene expansions through duplication but not via HGT. After going through extensive bioinformatic pipeline

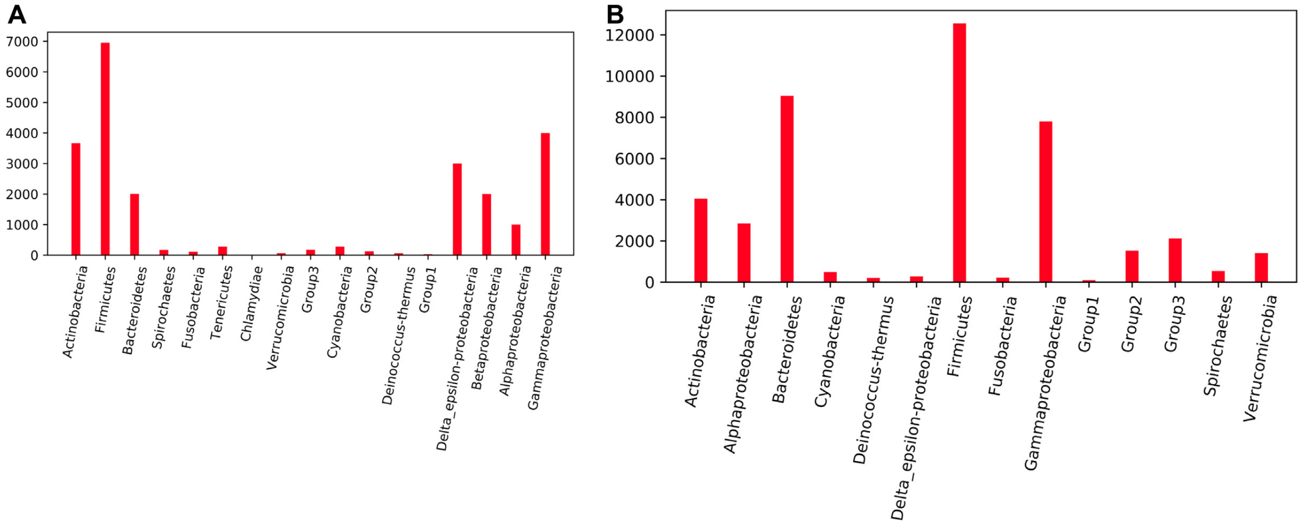
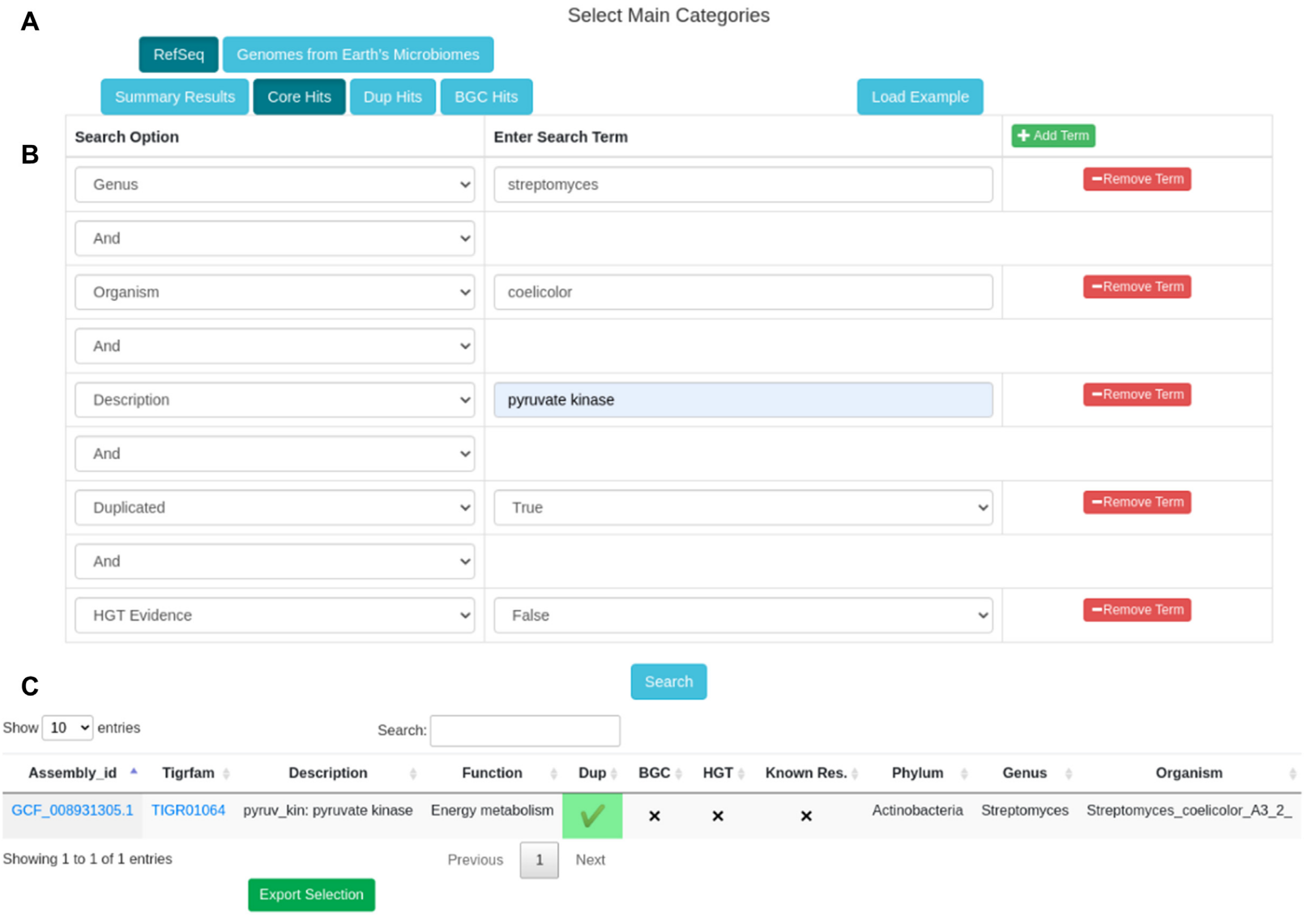


**Figure 1.** Included genome counts by reference set. Panels (**A**) and (**B**) show the phylogenetic distribution of genomes acquired from data sources NCBI RefSeq and GEM, respectively. Genome contents of the reference sets termed as ‘Group’ comprise underexplored phyla of the bacterial kingdom and described in detail in latest ARTS publication (21).



| Search Option | Enter Search Term |
|---|---|
| Genus | streptomyces |
| And | |
| Organism | coelicolor |
| And | |
| Description | pyruvate kinase |
| And | |
| Duplicated | True |
| And | |
| HGT Evidence | False |

Show 10 entries Search:

| Assembly_id | Tigrfam | Description | Function | Dup | BGC | HGT | Known Res. | Phylum | Genus | Organism |
|---|---|---|---|---|---|---|---|---|---|---|
| GCF_008931305.1 | TIGR01064 | pyruv_kin: pyruvate kinase | Energy metabolism | ✓ | × | × | × | Actinobacteria | Streptomyces | Streptomyces_coelicolor_A3_2_ |

Showing 1 to 1 of 1 entries Previous 1 Next

**Figure 2.** Query example in the ARTS database. (**A**) One of the available data sources ‘RefSeq’ and ‘Genomes from Earth’s Microbiomes’ and one of the four main routes below to explore selected data source must be selected. (**B**) After selecting main categories, search options and terms must be specified by using the ‘Add Term’ button. (**C**) The example output.

sessions to satisfy such requisites, their analysis pointed them towards a duplicated pyruvate kinase in *Streptomyces coelicolor* A3(2), for further evaluation. Using the ARTS database, such enquiries can be made in seconds.

*Duplication search.* In order to execute such a query, after selecting data source as 'RefSeq' and search category as 'Dup Hits', the user can click on the 'Add Term' button to start shaping the search. For example, after adding 'Genus' search option with the term 'streptomyces' and pressing 'Search', the user will have access to duplication rates of all essential genes from the genus streptomyces, including the gene counts in specific organisms, average gene counts for the reference set and its standard deviation. Afterward, dynamic filtering of the results for specific organisms or genes can easily be done by simply typing 'coelicolor pyruvate kinase' in the 'Search' box. However, it is advised to shape the initial search with parameters of interest since it will ease the browser's memory usage down and increase the execution speed of the query.

*Core hit attributes.* After detecting the gene of interest that shows statistical evidence for the duplication event, the user can easily check whether the gene fits in with other aspects of TDGM, here, an HGT event (Figure 2B). In our case, such query can be made using 'Core Hits' tab this time with the 'Genus' option with the term 'streptomyces' and adding the 'Description' option with the term 'pyruvate kinase' and simply adding the search option 'HGT Evidence' set to 'False'. Resulting table will only contain the gene of interest with direct links to individual ARTS result of the genome and HMM model of the gene for closer inspection (Figure 2C).

*Further examination.* The ARTS database provides opportunities for closer inspection of the resulting queries. For example, if the user is interested in BGCs that contain the gene of interest, the 'BGC Hits' tab can be used with the same search options to retrieve BGC specific results. Thereafter, the user can check the antiSMASH results of specific clusters, their gene cluster families in BiGFAM database consisting of closely related BGCs or the complete ARTS result, using the provided links. Items in the column 'Model Name' will lead to the target-oriented result page. Here, the user can explore the characteristics of a specific target gene and its fundamental ARTS criteria hits, with respect to the phyla where it is considered as an essential housekeeping gene. Moreover, commercially available drugs targeting the genes of interest are also shown through the links connected to the DrugBank database as well as the known BGCs that contain the gene via links to the MIBiG database. All of the resulting tables and individual ARTS results can be downloaded in order to feed in-house analysis pipelines.

## CONCLUSIONS AND FUTURE PERSPECTIVES

With the continuous advancements in genome sequencing techniques and BGC prediction algorithms, genome mining applications have become a vital factor in natural product research. A recently developed self resistance based approach, is progressively used by researchers for the discovery of natural products with novel modes of action. Since its first release in 2017, ARTS has been allowing researchers to rapidly mine their sequences with self resistance based genome mining approaches. Currently, to the best of our knowledge, ARTS is the only webserver enabling such method in all bacteria. Here, we present the ARTS database, a comprehensive repository containing a high quality bacterial genome set from NCBI's RefSeq and GEM catalogue processed with TDGM strategies. The ARTS database now allows researchers to quickly access pre-computed ARTS results and explore the bacterial kingdom via a broader view.

For future work, in order to further improve ARTS and the ARTS database, we are in the process of making ARTS analysis available for fungal genomes as well. We are also developing complementary tools such as SYN-view (32) for the enhancement of the ARTS pipeline and increasing its accuracy using additional criteria. Since the need for new antibiotics and the usage of genome mining methodologies increase on a daily base, we are confident that the ARTS database will be a resource of significant importance in the search for novel natural products.

## DATA AVAILABILITY

The ARTS database is publicly available online at https://arts-db.ziemertlab.com/ with no access restrictions. All of the source code involving Python and JS scripts as well as HTML content is available on Bitbucket at https://bitbucket.org/mehmetdirenc/arts_database/. All the accessions and queries are safely executed via HTTPS protocol.

## ACKNOWLEDGEMENTS

The authors acknowledge the use of de.NBI cloud and the support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen and the Federal Ministry of Education and Research (BMBF) through grant no 031 A535A. N.Z. and M.D.M thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2124 – 390838134 for the infrastructural support. We also thank Mohammad Alanjary for his valued discussions and all ARTS users for helpful comments and feedback.

## FUNDING

University of Tübingen; German Research Foundation (DFG) [INST 37/935-1 FUGG]; German Center for Infection Research [DZIF TTU09.716 to N.Z., M.D.M.]; Novo Nordisk Foundation [NNF20CC0035580, NNF16OC0021746 to K.B.]. Funding for open access charge: BMBF [DZIF TTU09.716].

*Conflict of interest statement.* None declared.

## REFERENCES

1. Harvey,A.L., Edrada-Ebel,R. and Quinn,R.J. (2015) The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.*, **14**, 111–129.
2. Iwu,C.D., Korsten,L. and Okoh,A.I. (2020) The incidence of antibiotic resistance within and beyond the agricultural ecosystem: a concern for public health. *Microbiologyopen*, **9**, e1035.

3. Newman,D.J. and Cragg,G.M. (2020) Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.*, **83**, 770–803.
4. Atanasov,A.G., Zotchev,S.B., Dirsch,V.M. and Supuran,C.T. (2021) Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.*, **20**, 200–216.
5. Scherlach,K. and Hertweck,C. (2020) Chemical mediators at the bacterial-fungal interface. *Annu. Rev. Microbiol.*, **74**, 267–290.
6. Ziemert,N., Alanjary,M. and Weber,T. (2016) The evolution of genome mining in microbes–a review. *Nat. Prod. Rep.*, **33**, 988–1005.
7. Scherlach,K. and Hertweck,C. (2021) Mining and unearthing hidden biosynthetic potential. *Nat. Commun*, **12**, 1–12.
8. Blin,K., Shaw,S., Kloosterman,A.M., Charlop-Powers,Z., van Wezel,G.P., Medema,M.H. and Weber,T. (2021) antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.*, **49**, W29–W35.
9. Skinnider,M.A., Johnston,C.W., Gunabalasingam,M., Merwin,N.J., Kieliszek,A.M., MacLellan,R.J., Li,H., Ranieri,M.R., Webster,A.L., Cao,M.P. *et al.* (2020) Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun*, **11**, 1–9.
10. Hannigan,G.D., Prihoda,D., Palicka,A., Soukup,J., Klempir,O., Rampula,L., Durcak,J., Wurst,M., Kotowski,J., Chang,D. *et al.* (2019) A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.*, **47**, e110.
11. Palaniappan,K., Chen,I.-M.A., Chu,K., Ratner,A., Seshadri,R., Kyrpides,N.C., Ivanova,N.N. and Mouncey,N.J. (2020) IMG-ABC v. 5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Res.*, **48**, D422–D430.
12. Katz,L. and Baltz,R.H. (2016) Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biot.*, **43**, 155–176.
13. Yan,Y., Liu,N. and Tang,Y. (2020) Recent developments in self-resistance gene directed natural product discovery. *Nat. Prod. Rep.*, **37**, 879–892.
14. Atkinson,G.C., Hansen,L.H., Tenson,T., Rasmussen,A., Kirpekar,F. and Vester,B. (2013) Distinction between the Cfr methyltransferase conferring antibiotic resistance and the housekeeping RlmN methyltransferase. *Antimicrob. Agents Ch.*, **57**, 4019–4026.
15. Almabruk,K.H., Dinh,L.K. and Philmus,B. (2018) Self-resistance of natural product producers: past, present, and future focusing on self-resistant protein variants. *ACS Chem. Biol.*, **13**, 1426–1437.
16. Thaker,M.N., Wang,W., Spanogiannopoulos,P., Waglechner,N., King,A.M., Medina,R. and Wright,G.D. (2013) Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nat. Biotechnol.*, **31**, 922.
17. Yan,Y., Liu,Q., Zang,X., Yuan,S., Bat-Erdene,U., Nguyen,C., Gan,J., Zhou,J., Jacobsen,S.E. and Tang,Y. (2018) Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action. *Nature*, **559**, 415–418.
18. Tang,X., Li,J., Millán-Aguiñaga,N., Zhang,J.J., O'Neill,E.C., Ugalde,J.A., Jensen,P.R., Mantovani,S.M. and Moore,B.S. (2015) Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem. Biol.*, **10**, 2841–2849.
19. Li,Y., Li,Z., Yamanaka,K., Xu,Y., Zhang,W., Vlamakis,H., Kolter,R., Moore,B.S. and Qian,P.-Y. (2015) Directed natural product biosynthesis gene cluster capture and expression in the model bacterium Bacillus subtilis. *Sci. Rep-UK*, **5**, 9383.
20. Alanjary,M., Kronmiller,B., Adamek,M., Blin,K., Weber,T., Huson,D., Philmus,B. and Ziemert,N. (2017) The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Res.*, **45**, W42–W48.
21. Mungan,M.D., Alanjary,M., Blin,K., Weber,T., Medema,M.H. and Ziemert,N. (2020) ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. *Nucleic Acids Res.*, **48**, W546–W552.
22. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
23. Nayfach,S., Roux,S., Seshadri,R., Udwary,D., Varghese,N., Schulz,F., Wu,D., Paez-Espino,D., Chen,I.-M., Huntemann,M. *et al.* (2021) A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.*, **39**, 499–509.
24. Belmann,P., Fischer,B., Krüger,J., Procházka,M., Rasche,H., Prinz,M., Hanussek,M., Lang,M., Bartusch,F., Gläßle,B. *et al.* (2019) de. NBI Cloud federation through ELIXIR AAI. *F1000Research*, **8**, 842.
25. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
26. Chen,I.-M.A., Chu,K., Palaniappan,K., Ratner,A., Huang,J., Huntemann,M., Hajek,P., Ritter,S., Varghese,N., Seshadri,R. *et al.* (2021) The IMG/M data management and analysis system v. 6.0: new tools and advanced capabilities. *Nucleic Acids Res.*, **49**, D751–D763.
27. Parks,D.H., Chuvochina,M., Chaumeil,P.-A., Rinke,C., Mussig,A.J. and Hugenholtz,P. (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.*, **38**, 1079–1086.
28. Kautsar,S.A., Blin,K., Shaw,S., Navarro-Muñoz,J.C., Terlouw,B.R., van der Hooft,J.J., Van Santen,J.A., Tracanna,V., Suarez Duran,H.G., Pascal Andreu,V. *et al.* (2020) MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.*, **48**, D454–D458.
29. Kautsar,S.A., Blin,K., Shaw,S., Weber,T. and Medema,M.H. (2021) BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res.*, **49**, D490–D497.
30. Wishart,D.S., Feunang,Y.D., Guo,A.C., Lo,E.J., Marcu,A., Grant,J.R., Sajed,T., Johnson,D., Li,C., Sayeeda,Z. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
31. Schniete,J.K., Cruz-Morales,P., Selem-Mojica,N., Fernández-Martínez,L.T., Hunter,I.S., Barona-Gómez,F. and Hoskisson,P.A. (2018) Expanding primary metabolism helps generate the metabolic robustness to facilitate antibiotic biosynthesis in Streptomyces. *MBio*, **9**, e02283-17.
32. Stahlecker,J., Mingyar,E., Ziemert,N. and Mungan,M.D. (2021) SYN-View: a phylogeny-based synteny exploration tool for the identification of gene clusters linked to antibiotic resistance. *Molecules*, **26**, 144.

## 2.4 Secondary Metabolite Transcriptomic Pipeline (SeMa-Trap), an expression-based exploration tool for increased secondary metabolite production in bacteria

### Contributions

With valued discussions and guidance of my advisors Prof. Dr. Nadine Ziemert and Prof. Dr. Kay Nieselt, I conceptualized SeMa-Trap and wrote the base code for the pipeline both in the back-end (also the creation of the cloud-based server) and front-end, was responsible for test and validation, original paper draft and finding positive examples. Theresa Anisja Harbig created the (web)pages for the visualization of all of the results. With the guidance of apl. Prof. Dr. Evi Stegmann, Naybel Hernandez Perez did all the wet-lab experiments for the overproduction of the complexing agent "[*S,S*]-EDDS" based on the RNA-Seq data initially generated from the isolated RNA by Dr. Simone Edenhart. All the authors spent time reviewing and editing the final manuscript.

*https://doi.org/10.1093/nar/gkac371*

# Secondary Metabolite Transcriptomic Pipeline (SeMa-Trap), an expression-based exploration tool for increased secondary metabolite production in bacteria

**Mehmet Direnç Mungan [1,2,3], Theresa Anisja Harbig[2], Naybel Hernandez Perez[1], Simone Edenhart[1], Evi Stegmann[1,3], Kay Nieselt[2] and Nadine Ziemert [1,2,3,*]**

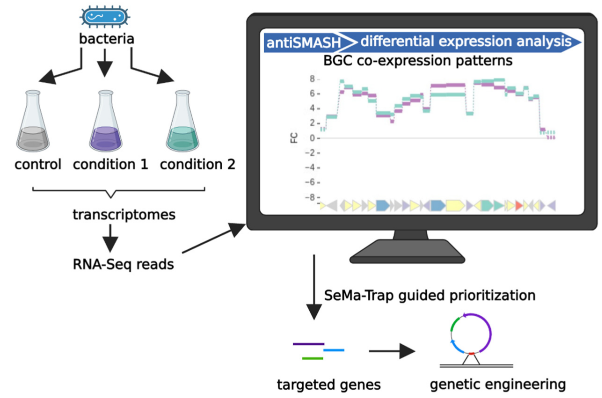
[1]Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), University of Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany, [2]Interfaculty Institute for Bioinformatics and Medical Informatics (IBMI), University of Tübingen, 72076 Tübingen, Germany and [3]German Center for Infection Research (DZIF), Partnersite Tübingen, 72076 Tübingen, Germany

Received March 07, 2022; Revised April 20, 2022; Editorial Decision April 27, 2022; Accepted April 29, 2022

## ABSTRACT

**For decades, natural products have been used as a primary resource in drug discovery pipelines to find new antibiotics, which are mainly produced as secondary metabolites by bacteria. The biosynthesis of these compounds is encoded in co-localized genes termed biosynthetic gene clusters (BGCs). However, BGCs are often not expressed under laboratory conditions. Several genetic manipulation strategies have been developed in order to activate or overexpress silent BGCs. Significant increases in production levels of secondary metabolites were indeed achieved by modifying the expression of genes encoding regulators and transporters, as well as genes involved in resistance or precursor biosynthesis. However, the abundance of genes encoding such functions within bacterial genomes requires prioritization of the most promising ones for genetic manipulation strategies. Here, we introduce the 'Secondary Metabolite Transcriptomic Pipeline' (SeMa-Trap), a user-friendly web-server, available at https://sema-trap.ziemertlab.com. SeMa-Trap facilitates RNA-Seq based transcriptome analyses, finds co-expression patterns between certain genes and BGCs of interest, and helps optimize the design of comparative transcriptomic analyses. Finally, SeMa-Trap provides interactive result pages for each BGC, allowing the easy exploration and comparison of expression patterns. In summary, SeMa-Trap allows a straightforward prioritization of genes that could be targeted via genetic engineering approaches to (over)express BGCs of interest.**

## GRAPHICAL ABSTRACT



## INTRODUCTION

By providing a wide range of biological functions, natural products have been foundational to the survival and evolutionary fitness of various organisms in the tree of life (1). Also known as secondary metabolites (SMs), these compounds are abundantly produced by plants and microorganisms (2). For decades, these molecules have been fueling various industries such as pharmaceutics as antimicrobial agents (3,4). However, the decrease in the discovery rates of novel antibiotics and the parallel increase in resistance towards the existing antibiotics make the identification of new bioactive compounds a task of paramount importance (5). By encoding the enzymes necessary for compound production, biosynthetic gene clusters (BGCs) represent the orga-

[*]To whom correspondence should be addressed. Tel: +49 7071 2978841; Email: nadine.ziemert@uni-tuebingen.de

nized groups of genes involved in the production of SMs (6). During the last decade an enormous number of genomic sequences have been made available, revolutionizing genome mining efforts in natural product research (7). Based on algorithmic concepts like hidden Markov models (HMMs), highly improved computational tools for BGC prediction such as antiSMASH (8) enable rapid mining of sequenced genomes. By using such tools, thousands of BGCs have been made available to researchers stored in public databases such as MIBiG (9), antiSMASH-DB (10) or The Natural Products Atlas (11). However, from the entire bacterial kingdom, it was recently shown that only 3% of its genomic potential for SMs has been experimentally verified (12). One of the main reasons for this phenomenon is that the expression of the BGCs is often tightly regulated and not observed under laboratory conditions. This non-expressed nature of the BGCs creates a major bottleneck in the identification of bioactive compounds with novel modes of action (13).

To activate silent BGCs and increase the production titers of SMs, several strategies have been devised such as altering the culturing conditions or heterologous expression of the BGCs (14,15). Additionally, genetically modifying global and local regulatory genes can enhance transcription levels of biosynthetic genes (16). Activation or disruption of positive and negative regulators, respectively, has led to the expression of many silent BGCs (17,18). Furthermore, it has been shown that increasing the expression of genes encoding transporters (19), conferring resistance (20), or involved in precursor supply (21) also increases SM production. However, major antibiotic producers like the organisms belonging to the genus *Streptomyces* (22) encode around 7000 genes on average (23). This raises the question: Which ones to genetically modify? Comparative transcriptomic analyses based on RNA-sequencing (RNA-seq) can help decipher the complex pathways that regulate the BGCs of interest and thereby, select the genes to prioritize (hereinafter referred to as target genes) (24,25). This strategy is mostly conducted by comparing the expression levels of BGCs from organisms with genetic variance or from the same strain cultured under different physiological conditions (26,27). The overwhelming number of possible experimental designs make the prioritization of promising culture conditions and target genes crucial for genetic manipulation approaches. To achieve this aim, we developed the 'Secondary Metabolite Transcriptomic Pipeline' (SeMa-Trap). Available at https://sema-trap.ziemertlab.com, SeMa-Trap allows for efficient transcriptome mining of BGCs in bacteria through a user-friendly web interface. The pipeline performs RNA-Seq based transcriptome analysis of BGCs predicted by antiSMASH, compares their fold-changes in various experiments, and allows for promising experimental design and prioritization of the target genes for BGC overexpression. Finally, SeMa-Trap provides interactive result pages for each BGC. This allows easy exploration of BGC expression under certain culturing conditions and the identification of co-regulated genes, which may be located elsewhere in the genome and display potentially interesting functions as defined by the KEGG database (28). Here we provide an overview of the pipeline, highlight the visualization of the interface and demonstrate the efficacy of SeMa-Trap through a case study.

## MATERIALS AND METHODS

### Workflow

The SeMa-Trap pipeline consists of 4 key steps (Figure 1). The first step is the acquisition of user provided genome and RNA-Seq data. Afterwards, genes involved in BGC expression regulation in the genome (e.g. transporters or regulators, referred to as genes of interest) are annotated, and BGCs are predicted by antiSMASH. BGC annotations in addition to those identified by antiSMASH can also be provided by the user by using the 'Defined clusters' option. To generate reference expression levels, essential housekeeping genes are also identified. In the third step, RNA-Seq analysis is performed to obtain expression levels and fold changes of the genes and BGCs of interest. Finally, results are presented by interactive visualizations and summarizing tables for easy exploration of the expression level changes. All results are kept in the server for 2 months. In addition, they can also be downloaded by saving the results page to the local machine. In case of larger data analysis, local installation and combining SeMa-Trap with in-house analysis pipelines is also possible using Anaconda.

### Input options and data acquisition

*Input form.* SeMa-Trap accepts user provided genomes in GenBank and FASTA format, however, the ideal input is the assembly accession number of the annotated GenBank file since that, in turn, will result in the automatic download of all annotation files from the NCBI FTP server. For efficient housekeeping gene identification, the corresponding taxonomic clade of the organism (e.g. Actinobacteria) should be selected through the 'Reference set' option. If the input genome is not represented by any available reference set, the 'Unknown' option offers HMM models acquired from the Database of Essential Genes (Supplementary Table S1) (29).

*RNA-Seq data.* For RNA-Seq based data options, allowed input types are run accession numbers from NCBI-SRA or EBI-ENA. Since it is imperative that the reads are downloaded in a fast and reliable fashion, SeMa-Trap utilizes multiple downloading options. IBM Aspera (https://www.ibm.com/products/aspera), a high-speed file transfer system, is the preferred and recommended way of data transfer (https://www.ncbi.nlm.nih.gov/books/NBK242621/). In case of any complications, SeMa-Trap will directly download from FTP servers or using fastq-dump (http://ncbi.github.io/sra-tools/). In case of pre-analyzed RNA-Seq data with other specific tools or parameters, the corresponding 'BAM' formatted files can also be uploaded. Limitations due to the current computational power and the implementation of the server are provided in the Supplementary Methods.

### RNA-Seq analysis

Once data acquisition is complete, SeMa-Trap utilizes several tools for analyzing the RNA-Seq data. Firstly, the fastp algorithm (30) is used to filter reads with low quality and for adapter trimming. Afterwards, filtered reads are mapped to

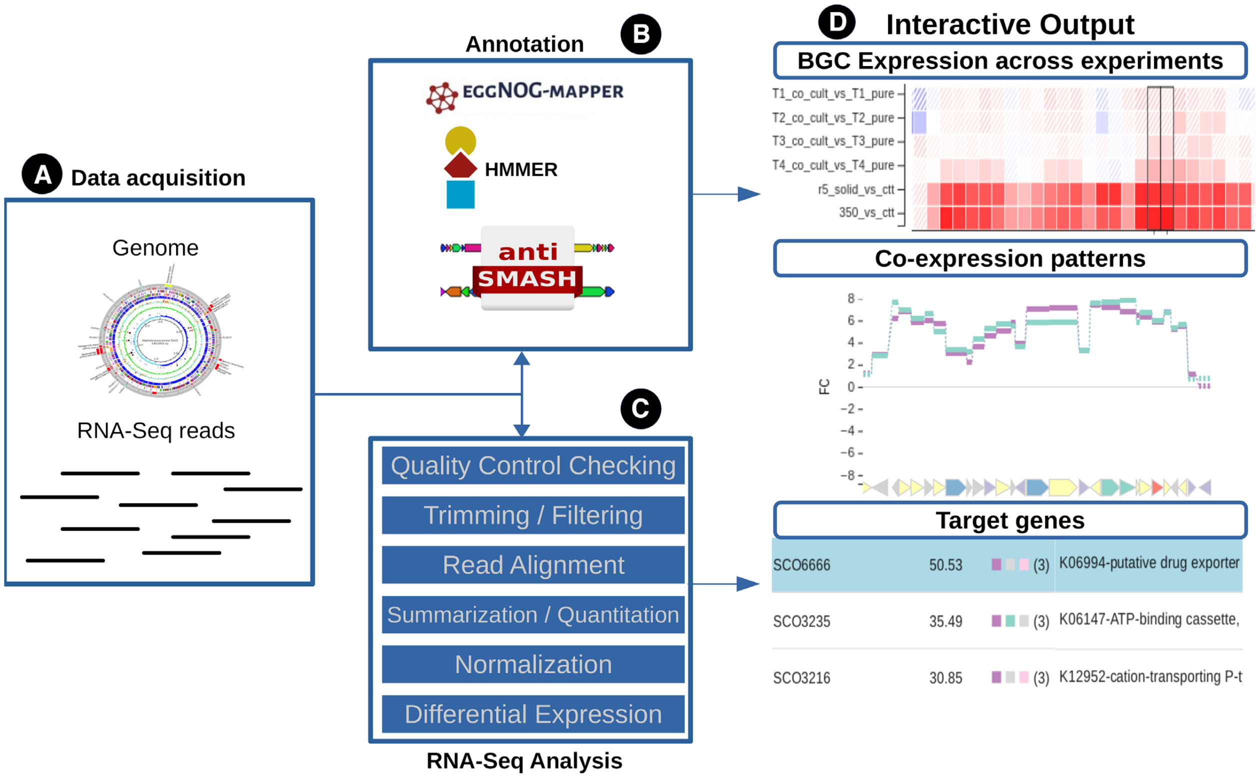


**Figure 1.** Overall workflow of the SeMa-Trap pipeline. First, the genomic and transcriptomic data provided by the user are acquired from relevant databases (**A**). Next step is the genome-wide annotation of the BGCs, essential housekeeping genes, secondary metabolite specific pathways and genes shown to have an impact on SM production (**B**). Final steps include a complete RNA-Seq analysis (**C**) and the generation of the interactive results (**D**).

the reference genome by Hisat2 (31) and sorted to generate corresponding BAM formatted files via samtools (32). Read count per gene is summarized by featureCounts (33). Finally, gene expression normalization takes place for each gene using the transcript per million (TPM) method described by Wagner et al. (34), and differential expression analysis is performed using DESeq2 (35), as detailed in Supplementary Methods. For the calculation of expression level or fold change of a BGC of interest, average expression of the 'core biosynthetic genes' (annotated by antiSMASH) is taken into account.

*Scoring.* In order to prioritize target genes, SeMa-Trap uses a scoring function dependent on the gene expression levels throughout the comparative transcriptomic experiments. To calculate such scores, fold changes of the selected BGC and the gene of interest are multiplied and then the calculated numbers from each selected experiment are added together (exemplified in Supplementary Table S2). However, it must be noted that a high score does  not necessarily prove an association between a BGC and a gene. It rather points to high expression changes in the different conditions relative to a BGC of interest. Only when using large amounts of expression data, credible associations can be effectively detected (36).

*Reference expression level.* In order to set meaningful thresholds to label a BGC as 'expressed', SeMa-Trap uses three different average expression levels of specific genes. One of them is the mean expression of housekeeping genes throughout the genome. These genes are annotated by hmmsearch (37) with specific TIGRFAM models (38) unique for each reference set (39,40). The idea here is that on average, a gene defined as 'essential housekeeping gene' should be expressed significantly to be used as a reference for expression (41). However, BGCs can be expressed at lower levels and still produce compounds (42). Since no exact threshold exists to define BGC expression, SeMa-Trap offers separate reference levels such as the mean of non-housekeeping genes or all of the existing genes.

## Annotation

Apart from antiSMASH's BGC prediction, the KnownClusterBlast algorithm is also applied to identify the compounds potentially produced by the BGC. If the provided genome is in FASTA format, an initial gene prediction step will take place using Prodigal (43). Since it is shown that certain types of genes actively control BGC expression, an extensive annotation of the genome is essential for prioritizing target genes to manipulate for BGC overexpression. For this purpose, the eggNOG-mapper (44) is used, particularly for the annotation of genes encoding transporters and genes residing in secondary metabolite specific KEGG pathways termed as 'biosynthesis of secondary metabolites' and 'biosynthesis of antibiotics'. Using hmmsearch, genes conferring antibiotic resistance or genes with regulatory functions are further defined via specific HMM models

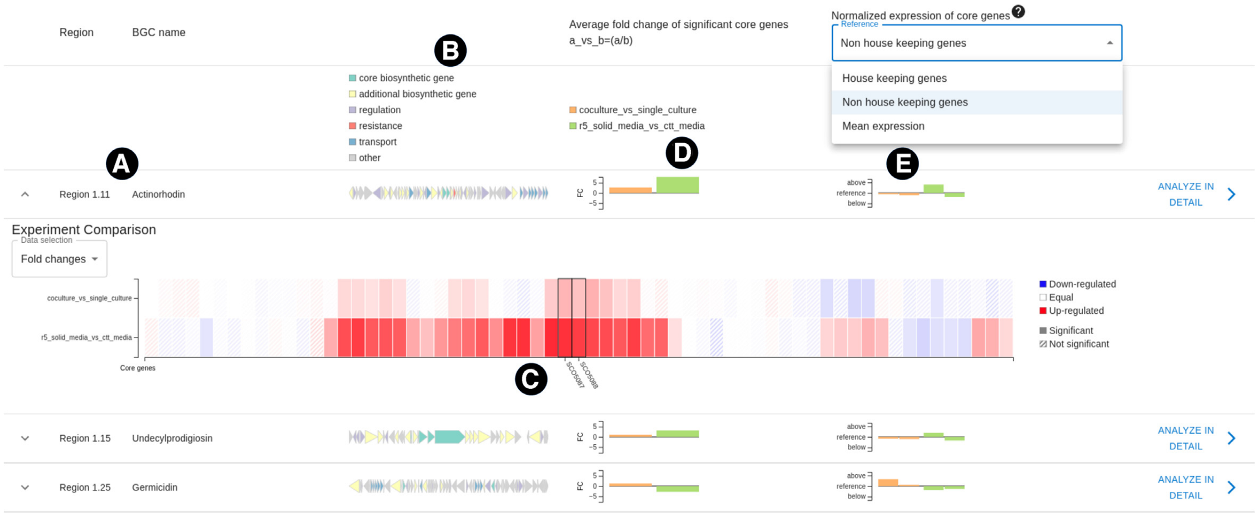


**Figure 2.** Overview result page of SeMa-Trap run for two comparative transcriptomic experiment designs. (**A**, **B**) The potential compound of the BGC and functional annotations of the genes within, respectively. (**C**) Heatmap of the BGC of interest, displaying each genes fold changes in different experiments. (**D**) Average fold change of the entire BGC, per experiment. (**E**) Expression (TPM) of a BGC relative to the selected, normalized reference expression level.

procured from PFAM (45), Resfams (46) and CARD (47) databases.

## RESULTS

### Overview

Once the analysis is complete, SeMa-Trap presents the overview of the overall fold changes of predicted BGCs and their expression levels relative to either of the mentioned reference expression levels (Figure 2). Various useful annotations of the genes in the BGCs are presented as well as the corresponding compound of the BGC if it is defined by KnownClusterBlast. Furthermore, a heatmap of the BGC content can be viewed in order to inspect fold changes of genes per experiment. BGCs can be further explored by clicking on the 'Analyze in detail' button.

### Case study

A recent study by Lee *et al.* demonstrated the various effects of microbial co-culturing on natural products biosynthesis at the transcriptome level (48). Using six different comparative experimental designs, the authors revealed that competition for iron increases the expression of specific genes leading to actinorhodin overproduction in *Streptomyces coelicolor* A3(2) when co-cultured with *Myxococcus xanthus*. In the following, by analyzing their publicly available RNA-Seq data, we illustrate how SeMa-Trap simplifies the entire analysis.

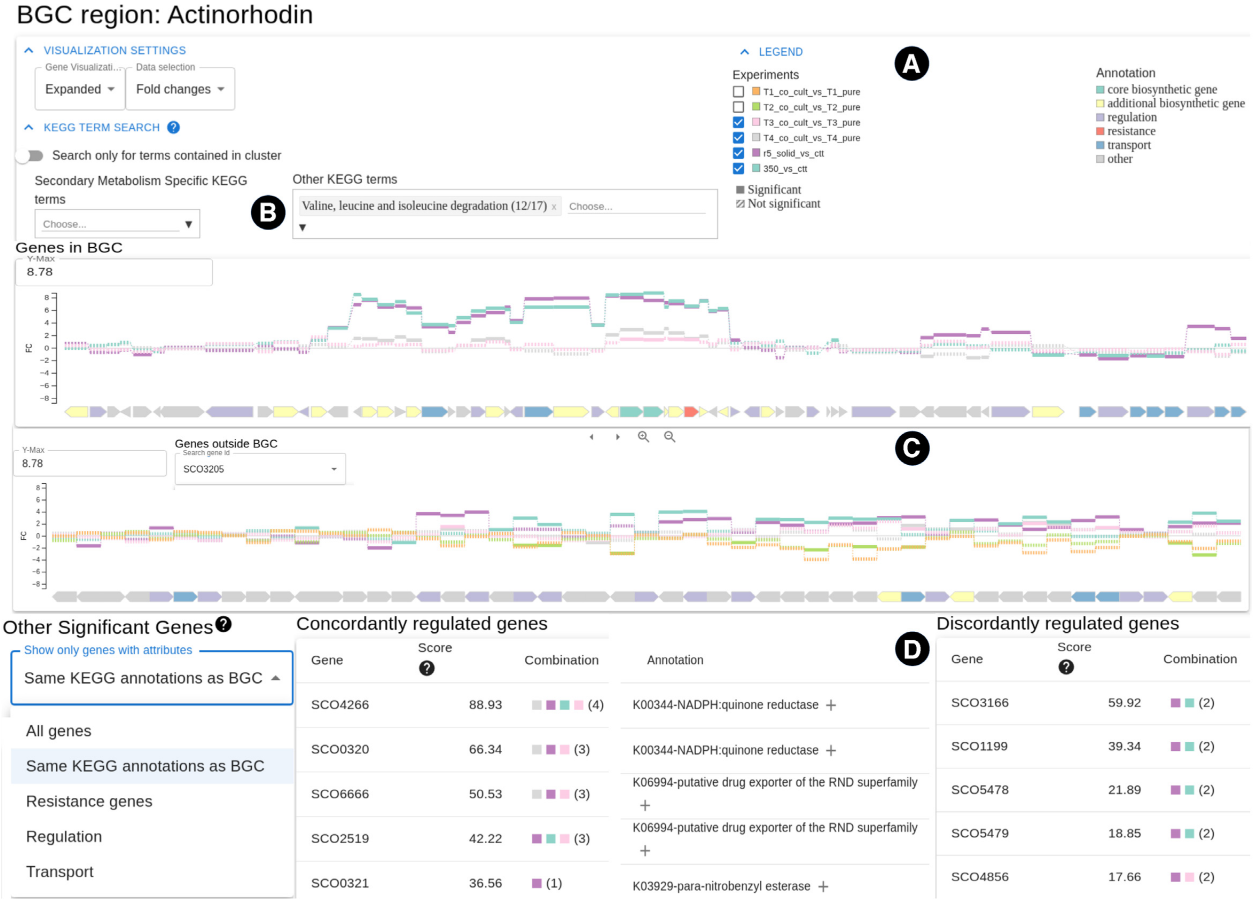
*Visualization options and pathway analysis.* The first part of the result page (Figure 3A) offers a range of options such as various displaying options for the presented genes, the selection of specific experiments, and visualization of RNA-Seq results by fold change or TPM based expression level. Furthermore, it is possible to analyze specific pathways more in detail and explore the amount of differentially expressed genes within. In the presented case study, genes involved in the leucine and isoleucine degradation pathways were shown to be overexpressed, which potentially provide precursors for the actinorhodin biosynthesis. Using Sema-Trap this can easily be highlighted (Figure 3B).

*Genome browser.* For the investigation of specific genes within the BGC or throughout the rest of the genome, a dynamic genome browser is available. Apart from efficient exploration of gene expression and annotation, the genome browser offers multiple options. Provided that the BGC of interest is significantly expressed, it is possible to set more accurate boundaries for the predicted BGC. Within the antiSMASH defined boundaries of a BGC (Figure 3C), a smaller, continuous succession of genes appears to be co-expressed, suggesting that those are regulated in an operon and represent the actual BGC boundaries.

*Target gene prioritization.* After thorough investigation, Lee and colleagues identified the SCO6666 gene encoding a transport system alternative to the one in the actinorhodin BGC, which is encoded by the genes SCO5083–5084. Furthermore, they found that the SCO6666 gene highly affected the production of actinorhodin in iron restricted conditions. Such prioritization can be easily made using the SeMa-Trap tables sorted by concordantly and discordantly co-regulated genes including scores (Figure 3D). Selection of the functional category 'Same KEGG annotations as BGC' further simplifies the investigation of the systems alternative to those encoded within the BGC of interest. The 'Combination' column denotes the selected experiments, thus providing information on which genes are co-regulated with the BGC of interest under which conditions.

### Proof of principle

As a proof of concept, we used SeMa-Trap to examine the transcriptome data of the actinomycete *Amycolatopsis*



BGC region: Actinorhodin

VISUALIZATION SETTINGS

Gene Visualizati...: Expanded

Data selection: Fold changes

KEGG TERM SEARCH

○ Search only for terms contained in cluster

Secondary Metabolism Specific KEGG terms: Choose...

Other KEGG terms: Valine, leucine and isoleucine degradation (12/17) x Choose...

LEGEND

Experiments

- □ T1_co_cult_vs_T1_pure
- □ T2_co_cult_vs_T2_pure
- ☑ T3_co_cult_vs_T3_pure
- ☑ T4_co_cult_vs_T4_pure
- ☑ r5_solid_vs_ctt
- ☑ 350_vs_ctt

■ Significant
▨ Not significant

Annotation

- core biosynthetic gene
- additional biosynthetic gene
- regulation
- resistance
- transport
- other

Genes in BGC

Y-Max: 8.78

Genes outside BGC

Y-Max: 8.78

Search gene id: SCO3205

Other Significant Genes

Show only genes with attributes: Same KEGG annotations as BGC

- All genes
- Same KEGG annotations as BGC
- Resistance genes
- Regulation
- Transport

Concordantly regulated genes

| Gene | Score | Combination | Annotation |
|---|---|---|---|
| SCO4266 | 88.93 | ⬜ 🟪 🟩 🩷 (4) | K00344-NADPH:quinone reductase + |
| SCO0320 | 66.34 | ⬜ 🟪 🩷 (3) | K00344-NADPH:quinone reductase + |
| SCO6666 | 50.53 | ⬜ 🟪 🩷 (3) | K06994-putative drug exporter of the RND superfamily + |
| SCO2519 | 42.22 | 🟪 🟩 🩷 (3) | K06994-putative drug exporter of the RND superfamily + |
| SCO0321 | 36.56 | 🟪 (1) | K03929-para-nitrobenzyl esterase + |

Discordantly regulated genes

| Gene | Score | Combination |
|---|---|---|
| SCO3166 | 59.92 | 🟪 🟩 (2) |
| SCO1199 | 39.34 | 🟪 🟩 (2) |
| SCO5478 | 21.89 | 🟪 🟩 (2) |
| SCO5479 | 18.85 | 🟪 🟩 (2) |
| SCO4856 | 17.66 | 🟪 🩷 (2) |

**Figure 3.** BGC centered results of SeMa-Trap. Initially, color codes for different annotations and multiple visualization settings are presented (**A**). Users can also highlight genes in specific pathways and choose to visualize the results based on the selected experiments (**B**). In section (**C**), two genome browsers are available in order to explore gene expressions from the selected experiments in the predicted cluster and throughout the genome. Finally, genes which are likely impacting the BGC expression based on transcriptomic data can be viewed through an interactive table **(D)**.

*japonicum*. *A. japonicum* is the producer of the complexing agent [*S,S*]-EDDS (49), a structural isomer of EDTA, which in contrast to EDTA is biodegradable and can replace EDTA in many industrial applications. However, [*S,S*]-EDDS production is inhibited by zinc at concentrations of 2 μM (50). Responsible for this regulation is the zinc uptake regulator *Zur*. To produce [*S,S*]-EDDS even in the presence of zinc the mutant *A. japonicum* Δ*zur* (referred to as zurko) was generated (51). To determine which genes to overexpress to increase [*S,S*]-EDDS production in *A. japonicum*, we performed transcriptomic analysis. For this purpose, RNA-Seq analyses of *A. japonicum* wild type (WT) and *A. japonicum* Δ*zur* cultured in the presence and absence of zinc for 24 h were performed. Thereby, a direct correlation between *zur* gene expression and the [*S,S*]-EDDS biosynthetic genes (BGs) could be observed. In particular, using SeMa-Trap we identified genes that exhibited high co-expression with the [*S,S*]-EDDS BG (concordantly regulated genes) and genes regulated in opposite manner (discordantly regulated genes). Since gene deletion is a multi-step, time-consuming process, we opted for a straightforward approach and overexpressed the targeted genes as a proof of concept. Thereby, we focused on genes with a regulatory function and those connected to secondary metabolism pathways. The target gene *bldC* (‘AJAP_RS36645’), with the second highest score in the category ‘regulation’, encodes a transcriptional regulator of differentiation which controls entry into development and the onset of antibiotic production in *Streptomyces* (52). The *lacI* gene, (‘AJAP_RS11995’), encodes a pleiotropic regulator (fifth highest score in the category ‘regulation’) which enhanced the production of antibiotics in *S. coelicolor* (53). From the pathways connected to secondary metabolism, we selected the glutamate synthase-encoding *glts* (‘AJAP_RS11230’) gene (with second best score) involved in glutamate biosynthesis. Since glutamate can be converted into L-aspartic acid, one of the precursors for EDDS biosynthesis, this gene was also taken into consideration. None of the selected genes have been experimentally shown to be linked to the [*S,S*]-EDDS production. Simultaneous overexpression of these genes resulted in an increased EDDS production by 3-fold compared to *A. japonicum* WT (Figure 4). Along with the experimental design, detailed methods (Supplementary Tables S3 and S4) and analysis (Supplementary Figures S1 and S2) can be further seen in the Supplementary Data.

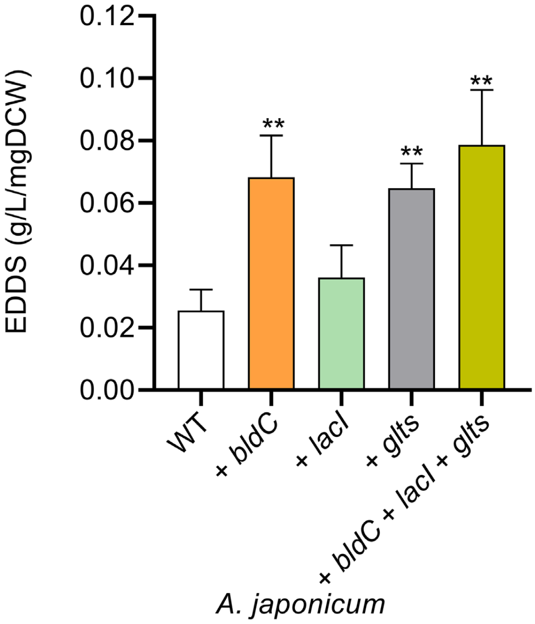


**Figure 4.** [*S,S*]-EDDS production in *A. japonicum* WT and recombinant strains. Strains were grown for 96 h in zinc depleted synthetic medium (SM). *A. japonicum* wild-type (WT); *A. japonicum* containing an additional copy of the genes *bldC*, *lacI* or glutamate synthase (*glts*), respectively and *A. japonicum* containing an additional copy of the three genes (*bldC* + *lacI* + *glts*).

## CONCLUSIONS AND FUTURE PERSPECTIVES

Leveraging on state-of-the-art sequencing techniques, comparative transcriptomic analyses have been continuously used to identify genes that are co-regulated with BGCs of interest and can be manipulated to activate silent BGCs. A variety of tools exists in order to annotate and effectively visualize biological functions of co-regulated genes such as KOBAS (54), conduct RNA-Seq analysis such as ProkSeq (55) or identify BGCs with co-expression data such as CASSIS (56). However, to the best of our knowledge, SeMa-Trap is the only public web server that combines genome mining and transcriptomic approaches for the identification of potential target genes for SM overproduction. The user-friendly graphical interface of the web server allows efficient and easy mining of RNA-Seq data, and was conceived for natural product researchers who are not acquainted with command line tools. Notably, SeMa-Trap also visualizes essential information about the cell response to the production of SMs on a transcriptomic level.

We showed herein that SeMa-Trap greatly facilitates the identification of co-regulated genes as illustrated on the actinorhodin-encoding BGC. However the limitations of the pipeline must be noted. The current scoring system is only designed to sort genes based on their similarity in transcription levels to a BGC of interest. It can not be used as an exclusive method for the selection of target genes. Thus, it is incumbent upon the users to further evaluate the hits returned by SeMa-Trap. For example, in the presented [*S,S*]-EDDS overproduction experiment, our literature search showed that the genes having the best co-expression score were unlikely to play a role in [*S,S*]-EDDS production. Consequently, three of the promising target genes were successfully overexpressed, leading to increased [*S,S*]-EDDS production. Especially when based on a few number of transcriptomic experiments, it becomes more likely that the SeMa-Trap analysis will include false positive target genes in the resulting tables. For future applications, by analyzing large amounts of publicly available RNA-Seq data, we are working on generating associations with certain gene types and classes of BGCs. Through co-expression networks, using statistical methods such as Pearson correlation coefficient, our aim is to reduce the number of false positives (57,58).

In summary, considering the ever-growing need for novel bioactive compounds, we believe that SeMa-Trap will serve as a helpful tool for the natural product community by facilitating the identification of specific co-expression patterns between different types of BGCs and genes with potential regulatory functions. Additionally, such analysis will also improve our ability to define expression thresholds above which the actual production of the encoded compound is observed. Last but not least, knowledge about the global cellular response to SM production may be the starting point to devise alternative strategies to optimize compound production and identify potential resistance mechanisms.

## DATA AVAILABILITY

SeMa-Trap is publicly available online at https://sema-trap.ziemertlab.com/ with no access restrictions. All of the source code is available on Bitbucket at https://bitbucket.org/mehmetdirenc/sematrap/. Source code for generating only the interactive HTML output is also available at https://github.com/Integrative-Transcriptomics/bgc-expression-viewer. Transcriptomic data files for EDDS overproduction and presented case study are available in the NCBI Bioproject database under the accession IDs PRJNA809550 and PRJEB25075, respectively.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Dr Libera do Presti for invaluable comments on the manuscript. We also acknowledge Quantitative Biology Center (QBiC) and all SeMa-Trap users for helpful comments and feedback.

## FUNDING

N.Z. and M.D.M. acknowledge the German Center for Infection Research [DZIF TTU09.716]; Germany's Excellence Strategy – EXC 2124 [390838134]; N.Z., T.H. and E.S. gratefully acknowledge financial support from the German Research Foundation (DFG) [TRR261, project ID 398967434]; The authors acknowledge the use of de.NBI cloud and the support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen through bwHPC and the German Research Foundation (DFG) [INST 37/935-1 FUGG] and the Federal Ministry of Education and Research (BMBF) [031 A535A]. Funding for open access charge: BMBF [DZIF TTU09.716].

*Conflict of interest statement.* None declared.

## REFERENCES

1. Newman,D.J. and Cragg,G.M. (2020) Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.*, **83**, 770–803.
2. Scherlach,K. and Hertweck,C. (2020) Chemical mediators at the bacterial-fungal interface. *Ann. Rev. Microbiol.*, **74**, 267–290.
3. Yan,Y., Liu,Q., Jacobsen,S.E. and Tang,Y. (2018) The impact and prospect of natural product discovery in agriculture: New technologies to explore the diversity of secondary metabolites in plants and microorganisms for applications in agriculture. *EMBO Rep.*, **19**, e46824.
4. Atanasov,A.G., Zotchev,S.B., Dirsch,V.M. and Supuran,C.T. (2021) Natural products in drug discovery: Advances and opportunities. *Nat. Rev. Drug Discov.*, **20**, 200–216.
5. Iwu,C.D., Korsten,L. and Okoh,A.I. (2020) The incidence of antibiotic resistance within and beyond the agricultural ecosystem: a concern for public health. *Microbiologyopen*, **9**, e1035.
6. Ziemert,N., Alanjary,M. and Weber,T. (2016) The evolution of genome mining in microbes—a review. *Nat. Prod. Rep.*, **33**, 988–1005.
7. Medema,M.H., de Rond,T. and Moore,B.S. (2021) Mining genomes to illuminate the specialized chemistry of life. *Nat. Rev. Genet.*, **22**, 553–571.
8. Blin,K., Shaw,S., Kloosterman,A.M., Charlop-Powers,Z., van Wezel,G.P., Medema,M.H. and Weber,T. (2021) antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.*, **49**, W29-W35.
9. Kautsar,S.A., Blin,K., Shaw,S., Navarro-Muñoz,J.C., Terlouw,B.R., van der Hooft,J.J., Van Santen,J.A., Tracanna,V., Suarez Duran,H.G., Pascal Andreu,V. *et al.* (2020) MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.*, **48**, D454–D458.
10. Blin,K., Shaw,S., Kautsar,S.A., Medema,M.H. and Weber,T. (2021) The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Res.*, **49**, D639–D643.
11. van Santen,J.A., Poynton,E.F., Iskakova,D., McMann,E., Alsup,T.A., Clark,T.N., Fergusson,C.H., Fewer,D.P., Hughes,A.H., McCadden,C.A. *et al.* (2022) The Natural Products Atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Res.*, **50**, D1317–D1323.
12. Gavriilidou,A., Kautsar,S.A., Zaburannyi,N., Krug,D., Müller,R., Medema,M.H. and Ziemert,N. (2022) Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nature Microbiology*, **7**, 726–735.
13. Chevrette,M.G., Gutiérrez-García,K., Selem-Mojica,N., Aguilar-Martínez,C., Yañez-Olvera,A., Ramos-Aboites,H.E., Hoskisson,P.A. and Barona-Gómez,F. (2020) Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat. Prod. Rep.*, **37**, 566–599.
14. Ambrosino,L., Tangherlini,M., Colantuono,C., Esposito,A., Sangiovanni,M., Miralto,M., Sansone,C. and Chiusano,M.L. (2019) Bioinformatics for marine products: an overview of resources, bottlenecks, and perspectives. *Mar. Drugs*, **17**, 576.
15. Zhang,J.J., Tang,X. and Moore,B.S. (2019) Genetic platforms for heterologous expression of microbial natural products. *Nat. Prod. Rep.*, **36**, 1313–1332.
16. Ochi,K. and Hosaka,T. (2013) New strategies for drug discovery: activation of silent or weakly expressed microbial gene clusters. *App. Microbiol. Biotechnol.*, **97**, 87–98.
17. Beck,C., Gren,T., Ortiz-López,F.J., Jørgensen,T.S., Carretero-Molina,D., Martín Serrano,J., Tormo,J.R., Oves-Costales,D., Kontou,E.E., Mohite,O.S. *et al.* (2021) Activation and identification of a griseusin cluster in *Streptomyces* sp. CA-256286 by employing transcriptional regulators and multi-omics methods. *Molecules*, **26**, 6580.
18. Mingyar,E., Mühling,L., Kulik,A., Winkler,A., Wibberg,D., Kalinowski,J., Blin,K., Weber,T., Wohlleben,W. and Stegmann,E. (2021) A regulator based ‘semi-targeted’ approach to activate silent biosynthetic gene clusters. *Int. J. Mol. Sci.*, **22**, 7567.
19. Severi,E. and Thomas,G.H. (2019) Antibiotic export: transporters involved in the final step of natural product production. *Microbiology*, **165**, 805–818.
20. Begani,J., Lakhani,J. and Harwani,D. (2018) Current strategies to induce secondary metabolites from microbial biosynthetic cryptic gene clusters. *Ann. Microbiol.*, **68**, 419–432.
21. Wang,W., Li,S., Li,Z., Zhang,J., Fan,K., Tan,G., Ai,G., Lam,S.M., Shui,G., Yang,Z. *et al.* (2020) Harnessing the intracellular triacylglycerols for titer improvement of polyketides in Streptomyces. *Nat. Biotechnol.*, **38**, 76–83.
22. Khadayat,K., Sherpa,D.D., Malla,K.P., Shrestha,S., Rana,N., Marasini,B.P., Khanal,S., Rayamajhee,B., Bhattarai,B.R. and Parajuli,N. (2020) Molecular identification and antimicrobial potential of *Streptomyces* species from Nepalese soil. *Int. J. Microbiol.*, **2020**, 8817467.
23. Lee,N., Kim,W., Hwang,S., Lee,Y., Cho,S., Palsson,B. and Cho,B.-K. (2020) Thirty complete *Streptomyces* genome sequences for mining novel secondary metabolite biosynthetic gene clusters. *Scientific Data*, **7**, 55.
24. Yi,J.S., Kim,M.W., Kim,M., Jeong,Y., Kim,E.-J., Cho,B.-K. and Kim,B.-G. (2017) A novel approach for gene expression optimization through native promoter and 5 UTR combinations based on RNA-seq, ribo-seq, and TSS-seq of *Streptomyces coelicolor*. *ACS Synt. Biol.*, **6**, 555–565.
25. Ferguson,N.L., Peña-Castillo,L., Moore,M.A., Bignell,D.R. and Tahlan,K. (2016) Proteomics analysis of global regulatory cascades involved in clavulanic acid production and morphological development in Streptomyces clavuligerus. *J. Ind. Microbiol. Biotechnol.*, **43**, 537–555.
26. Li,X., Wang,J., Li,S., Ji,J., Wang,W. and Yang,K. (2015) ScbR-and ScbR2-mediated signal transduction networks coordinate complex physiological responses in *Streptomyces coelicolor*. *Sci. Rep.*, **5**, 14831.
27. Ahmed,Y., Rebets,Y., Estévez,M.R., Zapp,J., Myronovskyi,M. and Luzhetskyy,A. (2020) Engineering of *Streptomyces lividans* for heterologous expression of secondary metabolite gene clusters. *Microb. Cell Fact.*, **19**, 5.
28. Kanehisa,M., Furumichi,M., Sato,Y., Ishiguro-Watanabe,M. and Tanabe,M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
29. Luo,H., Lin,Y., Gao,F., Zhang,C.-T. and Zhang,R. (2014) DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res.*, **42**, D574–D580.
30. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
31. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.
32. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A., Keane,T., McCarthy,S.A., Davies,R.M. *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**, giab008.
33. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
34. Wagner,G.P., Kin,K. and Lynch,V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theor. Biosci.*, **131**, 281–285.
35. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
36. Kwon,M.J., Steiniger,C., Cairns,T.C., Wisecaver,J.H., Lind,A.L., Pohl,C., Regner,C., Rokas,A. and Meyer,V. (2021) Beyond the biosynthetic gene cluster paradigm: genome-wide coexpression networks connect clustered and unclustered transcription factors to secondary metabolic pathways. *Microbiol. Spect.*, **9**, e00898-21.
37. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.

38. Haft,D.H., Selengut,J.D., Richter,R.A., Harkins,D., Basu,M.K. and Beck,E. (2012) TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.*, **41**, D387–D395.
39. Alanjary,M., Kronmiller,B., Adamek,M., Blin,K., Weber,T., Huson,D., Philmus,B. and Ziemert,N. (2017) The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Res.*, **45**, W42–W48.
40. Mungan,M.D., Alanjary,M., Blin,K., Weber,T., Medema,M.H. and Ziemert,N. (2020) ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. *Nucleic Acids Res.*, **48**, W546–W552.
41. Moureu,S., Caradec,T., Trivelli,X., Drobecq,H., Beury,D., Bouquet,P., Caboche,S., Desmecht,E., Maurier,F., Muharram,G. *et al.* (2021) Rubrolone production by *Dactylosporangium vinaceum*: biosynthesis, modulation and possible biological function. *Appl. Microbiol. Biotechnol.*, **105**, 5541–5551.
42. Amos,G.C., Awakawa,T., Tuttle,R.N., Letzel,A.-C., Kim,M.C., Kudo,Y., Fenical,W., Moore,B.S. and Jensen,P.R. (2017) Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E11121–E11130.
43. Hyatt,D., Chen,G.-L., LoCascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
44. Cantalapiedra,C.P., Hernández-Plaza,A., Letunic,I., Bork,P. and Huerta-Cepas,J. (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.*, **38**, 5825–5829.
45. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L., Tosatto,S.C., Paladin,L., Raj,S., Richardson,L.J. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
46. Gibson,M.K., Forsberg,K.J. and Dantas,G. (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.*, **9**, 207–216.
47. Alcock,B.P., Raphenya,A.R., Lau,T.T., Tsang,K.K., Bouchard,M., Edalatmand,A., Huynh,W., Nguyen,A.-L.V., Cheng,A.A., Liu,S. *et al.* (2020) CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, **48**, D517–D525.
48. Lee,N., Kim,W., Chung,J., Lee,Y., Cho,S., Jang,K.-S., Kim,S.C., Palsson,B. and Cho,B.-K. (2020) Iron competition triggers antibiotic biosynthesis in *Streptomyces coelicolor* during coculture with *Myxococcus xanthus*. *ISME J.*, **14**, 1111–1124.
49. Nishikiori,T., Okuyama,A., Naganawa,H., Takita,T., Hamada,M., Takeuchi,T., Aoyagi,T. and and Umezawa,H. (1984) Production by actinomycetes of (*S,S*)-N,N'-ethylenediamine-disuccinic acid, an inhibitor of phospholipase C.. *J. Antibiot. (Tokyo)*, **37**, 426–427.
50. Zwicker,N., Theobald,U., Zähner,H. and Fiedler,H. (1997) Optimization of fermentation conditions for the production of ethylene-diamine-disuccinic acid by *Amycolatopsis orientalis*. *J. Ind. Microbiol. Biotechnol.*, **19**, 280–285.
51. Spohn,M., Wohlleben,W. and Stegmann,E. (2016) Elucidation of the zinc-dependent regulation in *Amycolatopsisjaponicum* enabled the identification of the ethylenediamine-disuccinate (*S,S*-EDDS) genes. *Environ. Microbiol.*, **18**, 1249–1263.
52. Schumacher,M.A., den Hengst,C.D., Bush,M.J., Le,T., Tran,N.T., Chandra,G., Zeng,W., Travis,B., Brennan,R.G. and Buttner,M.J. (2018) The MerR-like protein BldC binds DNA direct repeats as cooperative multimers to regulate *Streptomyces* development. *Nat. Commun.*, **9**, 1139.
53. Meng,L., Yang,S.H., Kim,T.-J. and Suh,J.-W. (2012) Effects of two putative LacI-family transcriptional regulators, SCO4158 and SCO7554, on antibiotic pigment production of *Streptomyces coelicolor* and *Streptomyces lividans*. *J. Kor. Soc. Appl. Biol. Chem.*, **55**, 737–741.
54. Bu,D., Luo,H., Huo,P., Wang,Z., Zhang,S., He,Z., Wu,Y., Zhao,L., Liu,J., Guo,J. *et al.* (2021) KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.*, **49**, W317–W325.
55. Mahmud,A.F., Delhomme,N., Nandi,S. and Fällman,M. (2021) ProkSeq for complete analysis of RNA-seq data from prokaryotes. *Bioinformatics*, **37**, 126–128.
56. Wolf,T., Shelest,V., Nath,N. and Shelest,E. (2016) CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. *Bioinformatics*, **32**, 1138–1143.
57. Andersen,M.R., Nielsen,J.B., Klitgaard,A., Petersen,L.M., Zachariasen,M., Hansen,T.J., Blicher,L.H., Gotfredsen,C.H., Larsen,T.O., Nielsen,K.F. *et al.* (2013) Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *Proc. Nat. Acad. Sci. U.S.A.*, **110**, E99–E107.
58. Liesecke,F., Daudu,D., Dugé de Bernonville,R., Besseau,S., Clastre,M., Courdavault,V., De Craene,J.-O., Crèche,J., Giglioli-Guivarc'h,N., Glévarec,G. *et al.* (2018) Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Sci. Rep.*, **8**, 10885.

## SUPPLEMENTARY METHODS:

**Table S1.** Reference set specific HMM models used in housekeeping gene annotation

| Reference Set | HMM Count |
|---|---:|
| Chlamydiae | 445 |
| Group1 | 398 |
| Group2 | 968 |
| Tenericutes | 321 |
| Fusobacteria | 650 |
| Verrucomicrobia | 776 |
| Deinococcus-thermus | 599 |
| Alphaproteobacteria | 1098 |
| Betaproteobacteria | 923 |
| Spirochaetes | 765 |
| Firmicutes | 1019 |
| Bacteroidetes | 860 |
| Actinobacteria | 664 |
| Delta_Epsilon-proteobacteria | 866 |
| Group3 | 921 |
| Cyanobacteria | 882 |
| Gammaproteobacteria | 1478 |
| Unknown | 1568 |

### Methods

### Server Implementation

Hosted on highly scalable de.NBI cloud system, the SeMa-Trap web server runs on Ubuntu Linux (18.04.5 LTS) utilizing 2 TBs of hard drive space (1.5 in total available for a single run), 36 CPUs with approximately 1 to 1.4TB of RAM depending on the workload of cloud resources. Server-side application is based on Python3 Flask framework (https://flask.palletsprojects.com/) and Jinja2 templating language (https://jinja.palletsprojects.com) with JavaScript for user friendly input options. In combination, Redis (https://redis.io/), Nginx (https://www.nginx.com/), Gunicorn (https://gunicorn.org/) and Supervisor (http://supervisord.org/) tools are used for request handling and process control.

### RNA-Seq Analysis

Initially, a gene expression value is taken into account if it has the assigned padj value < 0.05, and the gene is considered as a differentially expressed gene (DEG) if its absolute fold change is > 2. A defined threshold (0.05) is set because of its wide acceptance as being “statistical significance” however, as such thresholds can be invalid for different experiments **(1)**, the cut-offs can be redefined by the user depending on the given data and how strict cut-offs are set in the experiment.

### Target gene prioritization

In order to calculate the given scores, fold changes of the selected BGC and the gene of interest are multiplied and then the calculated numbers from each selected experiment are added together. A visualized example of the scoring of SCO6666 gene (with score 50.53, shown in Figure 3 D), encoding the alternate transport system for actinorhodin BGC, can be seen in **Table S2**. SeMa-Trap result can also be seen for the whole analysis at https://sema-trap.ziemertlab.com/results/51e8fb5d-99e3-489b-aa57-2aa5cd2cc1cd.

**Table S2.** Example for scoring method.

| **Experiment** | **Actinorhodin Fold Change** | **SCO666 Fold Change** | **Multiplied** |
|---|---|---|---|
| T3_co_cult_vs_T3_pure | 1.37320261041672 | 7.6 | 10.436339839 |
| T4_co_cult_vs_T4_pure | 2.682977920151415 | 6.41060810700008 | 17.199520006 |
| r5_solid_vs_ctt | 7.824020991896 | 2.92558252110573 | 22.889819059 |
| Cumulative Final Score | | | ~50.53 |

### Strains, plasmids and oligonucleotides

The strains and plasmids are listed in **Table S3**. The oligonucleotides are listed in **Table S4.**

### Media and culture conditions

*Escherichia coli* strains were grown in Luria broth medium **(2)** at 37°C and were supplemented with 100 µg ml$^{-1}$ apramycin when necessary to maintain plasmids. Liquid cultures of *A. japonicum* were cultivated in 100 ml of R5 medium **(3)** in an orbital shaker (220 rpm) in 500-ml baffled Erlenmeyer flasks at 29°C. Liquid/solid media were supplemented with 100 µg ml$^{-1}$ apramycin to select for strains carrying integrated antibiotic resistance genes.

### Construction of the plasmids pRM4-*bldC*, pRM4-*lacI*, pRM4-*glts* and pRM4-*bldC-lacI-glts*

To construct the overexpression plasmids, *bldC* (AJAP_RS36645), *lacI* (AJAP_RS11995) and glutamate synthase (*glts*)( AJAP_RS11230) genes of *A. japonicum* were amplified via PCR with the primers listed in **Table S4** and purified using QIAquick gel extraction kit . The pRM4 vector **(4)**, containing the constitutive promoter *ermEp**, was linearized with the restriction enzyme NdeI and purified. Using NEBuilder HiFi DNA Assembly cloning kit (NEB, catalog no.E2621S) the linearized pRM4 was ligated with each of the amplified genes. In addition the pRM4-*bldC-lacI-glts* was constructed containing all the three genes. The plasmids were confirmed by enzymatic digestion and sequencing, and integrated into the genome of in *A. japonicum* WT. These steps allowed the generation of the overexpression strains *A. japonicum*::pRM4-*bldC*, *A. japonicum*::pRM4-*lacI, A. japonicum*::pRM4-*glts* and *A. japonicum*::pRM4-*bldC-lacI-glts.*

**[*S*,*S*]-EDDS production test**

Liquid culture of *A. japonicum* WT and recombinant strains was performed in 100 ml volume to determine [*S*,*S*]-EDDS production according to **(5), (6)**. The optimized synthetic medium (SM) consisted of glycerol (25 g $l^{-1}$ ), $MgSO_4 \times 7 H_2O$ (1.2 g $l^{-1}$), Ferric (III) citrate (60 mg $l^{-1}$), KH2PO4 (8 g $l^{-1}$ ), $Na_2HPO_4 \times 2 H_2O$ (12 g $l^{-1}$ ) and sodium glutamate monohydrate (11.3 g $l^{-1}$), which was used as the nitrogen source. Pre-cultures were grown on a rotary shaker (120 rpm) at 29°C in complex culture medium (glycerol (20 g $l^{-1}$); soybean meal (20 g $l^{-1}$) at pH 7.5) in 50 ml volume for 48 h. A total of 5 ml of this pre-culture was used to inoculate 95 ml of SM. The cultures were grown for further 96 h before the [*S*,*S*]-EDDS production was analysed.

**Detection of [*S*,*S*]-EDDS biosynthesis using HPLC-DAD**

[*S*,*S*]-EDDS measurement was performed as described by **(6).** The analysis was carried out on a HP1090M liquid chromatograph equipped with a thermostated autosampler, a diodearray detector and an HP Kayak XM 600 ChemStation (Agilent). A total of 10 µl of samples were injected onto a Hypersil ODS column (125 × 4 mm, 3 µm) fitted with a guard column (10 × 4 mm, 3 µm; Stagroma) and analysed by isocratic elution with solvent A – acetonitrile (96:4, v/v) at a flow rate of 1 ml $min^{-1}$. Solvent A consisted of 20 mM Sorensen's phosphate buffer (pH 7.2) with 5 mM tetrabutylammoniumhydrogensulfate. UV detection was performed at 253 nm. For data analysis, Chemstation LC3D software Rev. A.08.03 was used. Commercial [*S*,*S*]-EDDS in solution (Sigma Aldrich) was used as standard.

**Quantification**

The HPLC analysis was performed from 1 ml supernatant. In order to determine the production of [*S*,*S*]-EDDS of cells, the [*S*,*S*]-EDDS concentration was divided by the dry cell weight (DCW). The [*S*,*S*]-EDDS production was expressed by (g/l/mg DCW).

**Table S3: Bacterial strains and plasmids used in this study**

| Strain or plasmid | Description | Source of reference |
|---|---|---|
| *A. japonicum* MG17-CF17 | [*S,S*]-EDDS producing wild-type | **(7)** |
| pRM4 | pSET152*ermEp** with artificial RBS, Apra$^{r}$ | (4) |
| pRM4-*bldC* | pRM4 carrying *bldC* gene from *A. japonicum* WT | This study |
| pRM4-*lacI* | pRM4 carrying *lacI* gene from *A. japonicum* WT | This study |
| pRM4-*glts* | pRM4 carrying glutamate synthase (*glts*) gene from *A. japonicum* WT | This study |
| pRM4-*bldC-lacI-glts* | pRM4 carrying *bldC, lacI* and *glts* gene from *A. japonicum* WT | This study |

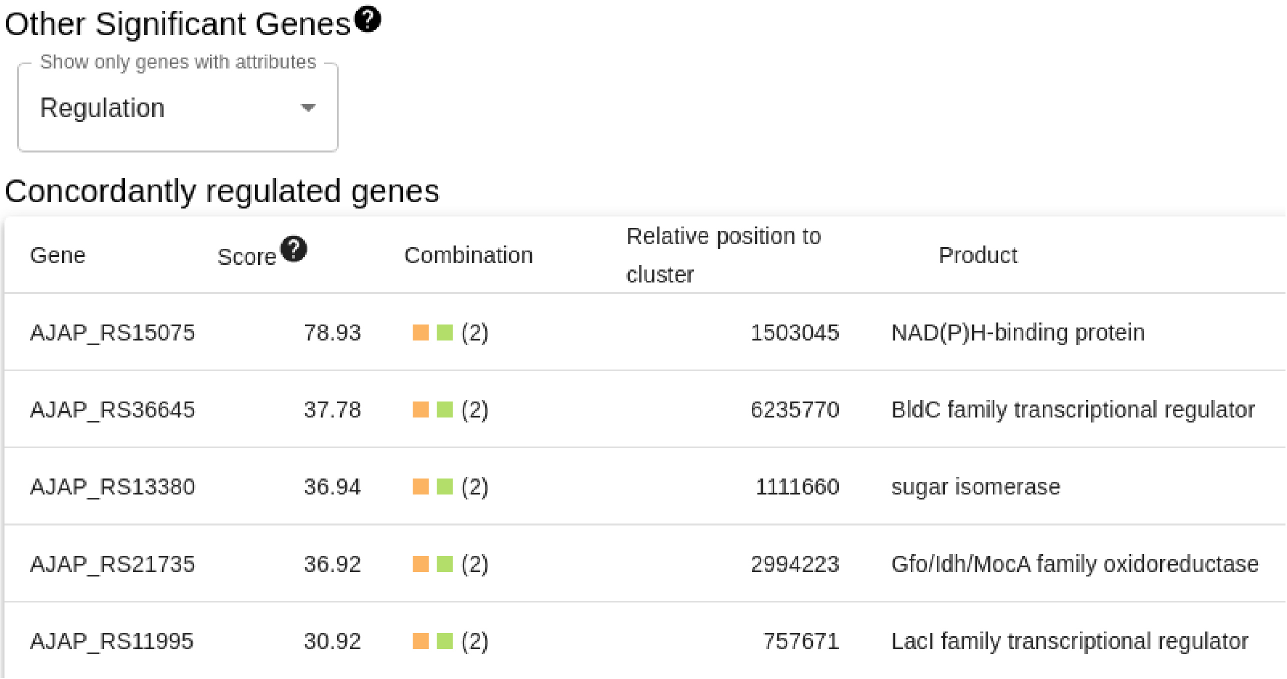
**Table S4. Oligonucleotides used in this study**

| Primer | Sequence (5´-3´) |
|---|---|
| Primers used for amplification of the *A. japonicum bldC* (AJAP_RS36645) coding region overlapping with pRM4 vector | |
| bldC_pRM4_F | CGACGGTATCGATAAGCTAGCCAGGGGAGGACCCAATGACCGCGACCATGGGCGGA |
| bldC_pRM4_R | GGGCTGCAGGAATTCGATATCAAGCTTAGATCTCATCAGACCTTGCGAGCGGGCTCG |
| Primers used for amplification of the *A. japonicum lacI* (AJAP_RS11995) coding region overlapping with pRM4 vector | |
| lacI_pRM4_F | GGGCTGCAGGAATTCGATATCAAGCTTAGATCTCATCATGCGGGGTACTCCTGGGTCGATTCG |
| lacI_pRM4-R | CGACGGTATCGATAAGCTAGCCAGGGGAGGACCCAATGTCGCTGGCGAAGGTGGCCC |
| Primers used for amplification of the *A. japonicum* glutamate synthase (*glts*) (AJAP_RS11230) coding region overlapping with pRM4 vector | |
| GS_pRM4_F | CGACGGTATCGATAAGCTAGCCAGGGGAGGACCCAGTGGCTGATCCGACGGGTTTCCTGA AGTACG |
| GS_pRM4_R | GGGCTGCAGGAATTCGATATCAAGCTTAGATCTCATCAGACCACCGCGAGCGGCA |
| Primers used for amplification of the *A. japonicum bldC*, *lacI*, *glts* coding region overlapping with pRM4 vector | |

| | |
|---|---|
| bldC_F_assem_pRM4 | CGACGGTATCGATAAGCTAGCCAGGGGAGGACCCAATGACCGCGACCATGG GCGGAAGG |
| bldC_R_assem_pRM4 | ACCTTCGCCAGCGACATTCAGACCTTGCGAGCGGGCTCGCT |
| lacI_F_assem_pRM4 | CCCGCTCGCAAGGTCTGAATGTCGCTGGCGAAGGTGGCCCG |
| lacI_R_assem_pRM4 | CCCGTCGGATCAGCCACTCATGCGGGGTACTCCTGGGTCGATTCGCG |
| GSsmall_F_assem_pRM4 | CAGGAGTACCCCGCATGAGTGGCTGATCCGACGGGTTTCCTGAAGTACGAC |
| GSsmall_R_assem_pRM4 | GGGCTGCAGGAATTCGATATCAAGCTTAGATCTCATCAGACCACCGCGAGCG GCAACG |

## Figure S1

Target genes with regulation attributes for [S,S]-EDDS overproduction. Note that, after a literature search, second and fifth most co-regulated genes were selected for overproduction experiments. Genes with putative regulatory domains were left out from our selection because their annotations were not strong enough to deduce an actual regulatory mechanism for a BGC. Entire analysis can be seen at https://sema-trap.ziemertlab.com/results/4985a8de-4c61-426f-8011-51b6aaa51350.



Other Significant Genes

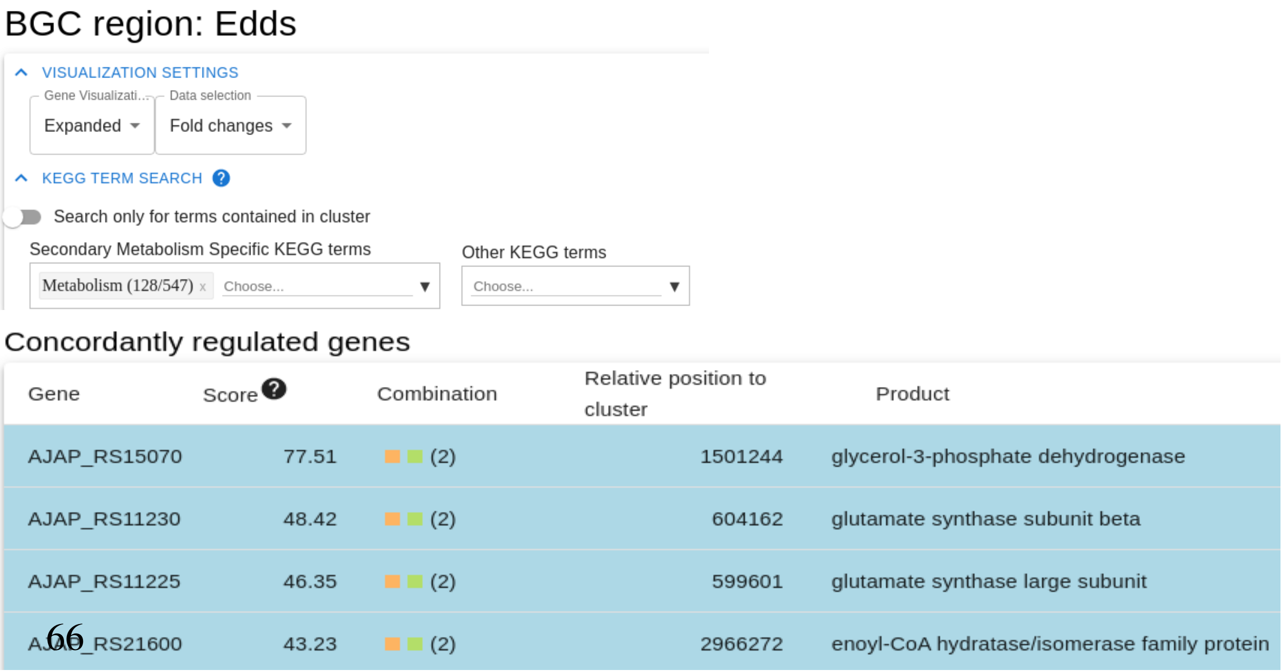
Show only genes with attributes: Regulation

Concordantly regulated genes

| Gene | Score | Combination | Relative position to cluster | Product |
|---|---|---|---|---|
| AJAP_RS15075 | 78.93 | 🟧 🟩 (2) | 1503045 | NAD(P)H-binding protein |
| AJAP_RS36645 | 37.78 | 🟧 🟩 (2) | 6235770 | BldC family transcriptional regulator |
| AJAP_RS13380 | 36.94 | 🟧 🟩 (2) | 1111660 | sugar isomerase |
| AJAP_RS21735 | 36.92 | 🟧 🟩 (2) | 2994223 | Gfo/Idh/MocA family oxidoreductase |
| AJAP_RS11995 | 30.92 | 🟧 🟩 (2) | 757671 | LacI family transcriptional regulator |

## Figure S2

Target genes that are defined in KEGG as specific to secondary metabolite “Metabolism” pathways for [S,S]-EDDS overproduction. Note that after a literature search involving precursor production mechanisms for [S,S]-EDDS, the second most co-regulated gene was selected for overproduction experiments.



BGC region: Edds

VISUALIZATION SETTINGS

Gene Visualizati...: Expanded

Data selection: Fold changes

KEGG TERM SEARCH

Search only for terms contained in cluster

Secondary Metabolism Specific KEGG terms: Metabolism (128/547)

Other KEGG terms: Choose...

Concordantly regulated genes

| Gene | Score | Combination | Relative position to cluster | Product |
|---|---|---|---|---|
| AJAP_RS15070 | 77.51 | 🟧 🟩 (2) | 1501244 | glycerol-3-phosphate dehydrogenase |
| AJAP_RS11230 | 48.42 | 🟧 🟩 (2) | 604162 | glutamate synthase subunit beta |
| AJAP_RS11225 | 46.35 | 🟧 🟩 (2) | 599601 | glutamate synthase large subunit |
| AJAP_RS21600 | 43.23 | 🟧 🟩 (2) | 2966272 | enoyl-CoA hydratase/isomerase family protein |

**References**

(1) Jafari, Mohieddin; Ansari-Pour, Naser (2019): Why, When and How to Adjust Your P Values? In: *Cell journal* 20 (4), S. 604–607. DOI: 10.22074/cellj.2019.5992.

(2) Sambrook J, Fritsch EF, Maniatis T. 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

(3) Kieser T, Bibb MJ, Buttner MJ, Chater KF, Hopwood DA. 2000. Practical Streptomyces genetics. John Innes Foundation, Norwich, United Kingdom

(4) Menges R, Muth G, Wohlleben W, Stegmann E. 2007. The ABC transporter Tba of Amycolatopsis balhimycina is required for efficient export of the glycopeptide antibiotic balhimycin. Appl. Microbiol. Biotechnol. 77: 125–134.

(5) Zwicker, N., Theobald, U., Zähner, H., and Fiedler, H.P. (1997) Optimization of fermentation conditions for the production of ethylene-diamine-disuccinic acid by Amycolatopsis orientalis. J Ind Microbiol Biotechnol 19: 280–285

(6) Spohn, M., Wohlleben, W., Stegmann, E. (2016) Elucidation of the zinc dependent regulation in Amycolatopsis japonicum enabled the identification of the ethylenediamine disuccinate ([S,S] EDDS) genes. Environ. Microbiol., 18, 1249- 1263.

(7) Nishikiori, T., Okuyama, A., Naganawa, H., Takita, T., Hamada, M., Takeuchi, T., Aoyagi, T., Umezawa, H. (1984) Production by actinomycetes of (S,S)-N,N-ethylenediamine-disuccinic acid, an inhibitor of phospholipase c. J. Antibiot., 37, 426- 427

# Chapter 3

# Conclusions

In this chapter, I would like to further discuss the “resistance crisis” at hand, how the methods herein can aid the proposed solutions, what more can be added in the future, but also the misconceptions a researcher might have when using the pipelines described and finally conclude my Ph.D. work.

Highlighted throughout the thesis, humanity is facing a grave danger of returning to the dark ages where bacterial infections roam unchallenged. Before jumping to the conclusion of “we need new antibiotics” there are additional issues we must factor into a plan of action against the infectious bacteria. As a wise man once said: “Bacteria... Can’t live with them, can’t live without them.” This is especially right considering all the microbes we live with, forming mutualistic arrangements [135]. Although it is not possible (for now) to live without microbes and there is the fact that resistance will emerge as per adaptation of the target bacterium, we can still slow the whole process down and gain more time to fight back. There are multiple implementations, which also helped long before the golden age of antibiotics, such as improved sanitation measures in everyday life settings, increased availability of clean food and water [136] and the prevention of unnecessary use of antibiotics [137]. Another important part of the solution rests with the legislative and political decisions in order to spark more investments for the steady stream of new compounds [138, 139]. Approved by the FDA in 2018, Achaogen created the aminoglycoside antibiotic "plazomicin" that demonstrated potent bioactivity against carbapenem-resistant Enterobacteriaceae. However, just under a year after, they filed for bankruptcy in April 2019. This whole ordeal made it even clearer that investments in antibiotics require more support from governments and better technological pipelines to save money and decrease the risk of hitting a "dead-end" [140]. As mentioned in the introduction, a large portion of drugs is inspired by natural products. Plazomicin as well is a semi-synthetic drug derivative of "sisomicin" which is naturally produced by *Micromonospora inyoensis*. The BGC responsible for sisomicin production is extensively analyzed in terms of functional annotations of the included genes (e.g. conferring resistance, regulators, etc.), its biosynthetic pathway and other structurally similar products (such as gentamicin) [141]. All of these were done in order to help the future biosynthesis studies of sisomicin, e.g. finding more potent producer strains, guiding genetic engineering attempts or finding other metabolites with similar bioactivity. Now that is just one work, describing one natural product from one genome which is used to create a potent antibiotic that eventually led to the bankruptcy of a company. Considering all the work that has been done to generate hundreds of thousands of BGCs and describe how they work, a crucial thing to do now in natural product research is the effective prioritization of our primary objectives leading to a minimized risk. Concordantly, the main aim of the thesis is to create automated bioinformatics tools in order to guide researchers in their efforts of finding novel BGCs and producing them.

To aid the discovery efforts, the first three published works herein have been built to utilize TDGM methodologies. Since its first iteration in 2017, the ultimate goal of ARTS has been the prioritization of a BGC which in turn would lead to a discovery of a novel antibacterial natural product, ideally with a novel mode of action. Unfortunately (and understandably), this has not yet been accomplished since it can take quite some time and effort to find such a compound. However, ARTS has been widely used by the community (academic and private), analyzing more than a hundred thousand genomes both locally and through the web server. Several example publications since the past year show that the updated version of ARTS has been used to analyze genomes' potency to produce antibiotics by looking at putative self-resistance mechanisms within BGCs [142, 143, 144, 145, 146, 147], as well as comparative genome mining for the said potential of multiple strains and their BGC diversity [148, 149, 150, 151], larger pan-genome analyses to look for promising antibiotic encoding BGCs [152], and sometimes only to check known AMR genes [153]. So far, researchers who used and cited ARTS were mainly interested in the co-localization of their genes of interest with the predicted BGCs. A logical predicament, since co-localization of the resistance gene is the backbone of TDGM approaches [154], also used to locate the BGC itself [155, 156]. Other than its easy-to-use design and visualization, and a solid biological rationale, ARTS owes its current popularity to its speed. Especially for the "Duplication" and "HGT" criteria, leveraging pre-computed reference sets dramatically reduces the processing time, saving several to tens of hours depending on the genome size [157]. Consequently, ARTS makes it possible to analyze large amounts of genomes at once. At this time, except for the metagenomic sequences, multi-genome analysis is restricted to sequences from the same taxonomic clade (e.g., Actinobacteria). However, a metagenome analysis is bereft of duplication and phylogeny criteria, since they are specific to each reference set. One way to enable these criteria is to infer the taxonomy of the input metagenome sequences by using classification tools such as GTDB-Tk [158]. Additionally, such taxonomical placement can be applied as an optional first step in the ARTS pipeline, since this information can be untrustworthy and highly inconsistent amongst currently used databases [159]. Considering that many studies are on a set of genera such as *Streptomyces*, it could also prove useful to create reference sets describing core genes for specific and smaller clades of bacteria. Since identifying core genes are very much dependent on the phylum [130], creating more specific reference sets would also reduce the number of genes with multiple ARTS hits, which might also be considered false positive results. This low specificity issue of the ARTS pipeline is mostly due to the large BGC region boundaries defined by antiSMASH. Also pointed out by multiple publications, the issues with BGC boundary definitions are quite problematic to genome mining based applications [152, 160]. So far cited by one analysis (non-review) publication [161], we have shown that SYN-View can aid in dealing with this problem. The workflow of SYN-View is based on the fact that microorganisms can employ neighboured set of genes to work together as clusters (or operons) which can co-evolve together [162]. This co-evolution can give hints to the actual function since it is quite difficult to differentiate if the gene is conferring a resistance mechanism or it is playing a part in the biosynthesis of the metabolite or only there as a housekeeping gene just minding its own business. SYN-View leverages the fact that the comparison of gene orthologue neighbours can be used to illuminate this ambiguity [133], and would be a perfect addition to TDGM methodologies specifically as an optional analysis included in the ARTS pipeline. The ARTS database can also help track down the functionality of a gene through the "Target-Oriented Search" section since in ARTS-DB the user can easily investigate a gene's affinity to be proximate to certain BGC type(s). However, the main reason for creating ARTS-DB was to offer researchers a comprehensive and easy-to-use resource to aid their TDGM-based discovery efforts. As shown many times, studying the genomic expansion events by duplication or HGT can lead to the discovery of novel natural products [163, 164], or can be used to decipher the adaptation mechanisms of microorganisms [165, 166, 167]. The complete database is also available in SQLite's "db" format, which allows further mining possibilities for future applications, which is already in use by a few of our collaborators with their TDGM efforts.

Once we have an interesting BGC in mind, it's time to facilitate the production of its corresponding compound. With SeMa-Trap, we aimed to find promising target genes for genetic manipulation applications in order to increase the expression of a BGC, in turn increasing its production. Although the correlation of a genes and a BGCs expression from a relatively small amount of samples are not conclusive, it is clearly an advantageous start for an overexpression experiment. Especially with the rise of the number of experiments fed to the pipeline, comes better confidence in the target genes. There are many researchers working on the pipelines that conduct the analysis of RNA-Seq data [168, 169] or predict BGCs [170, 171]. However, there is no web server that can be used to combine these pipelines in order to mine the genome and transcriptome of an organism to prioritize genes which can be targeted in wet-lab for secondary metabolite overproduction. That is precisely why the aim of SeMa-Trap is not to improve the existing workflows or methods but to create an easy-to-use server to locate promising genes. Another important part of the analysis pipelines is the visualization of results [172, 173]. I myself can't take the credit for it apart from its constituent data however, one of the best things (in ARTS as well) about SeMa-Trap is its nice and informative way of visualization of results. As the output of SeMa-Trap can become quite extensive, data visualization is of utmost importance for converting results into insights. As shown in the example results, such visualization can also help with the identification of BGC boundaries based on the expression levels, which are shown to be much closer to the boundaries defined by MiBIG, rather than the large region predicted by antiSMASH.

Although the usefulness of the tools described in this work has been proven multiple times, there might be some misconceptions about their usage. The biggest of them is that sometimes users might consider the given results in their wet-lab experiments without further and closer inspection. As is the case in screening methodologies, pipelines described in this thesis were merely created to find promising targets based on the information we have from all the genetic research made in the past decades. For example, HGT detection is a highly complex issue with multiple proposed solutions [174, 175, 176]. Incorporated to ARTS, a way of detecting HGT events is based on the evidence provided by the incongruences between species and gene trees [177]. However, it must be noted that ARTS does not find an HGT event with 100% confidence. Furthermore, finding HGT events are highly dependent on the quality of branch placements in the generated trees, and the strength of the phylogenetic signal suggesting the occurrence of an HGT. Similarly, detection of the “duplication” event is based on the occurrence frequencies of a gene of interest amongst a set of genomes available from a database (e.g., NCBI’s RefSeq). This means that such numbers can change with each update on ARTS with the increased amount of available genomes representing a reference set. For SeMa-Trap as well, an apparent correlation between the expression of genetic elements does not definitively constitute a direct relationship of causality. Such deductions can only be made after analyzing a much larger amount of data [178].

In concluding remarks, investing in genome and transcriptome mining of microorganisms is a very fruitful venture for natural product discovery as it is proven many times that they can effectively help wet-lab applications. Also, as we need every organism on our side to strengthen our arsenal against infectious diseases, our main aim for future work is to enable the tools described for the analysis of fungal sequences. Obviously, the fungal kingdom has its characteristic differences from bacteria however, both the TDGM and RNA-Seq based prioritization techniques have been shown many times to be applicable for fungi [179, 133, 180, 181]. With the technological advancements, especially in the sequencing methodologies, the amount of available “omics” data has been skyrocketing. This increase also brings challenges in the creation of efficient algorithms. Indeed, if not the most, it is a very exciting time for computational biologists filled with endless possibilities. As for the natural product discovery and production efforts, this work is set to serve as a small piece in the puzzle, playing its role in the early stages of the “Renaissance” we are currently experiencing.

# Abbreviations

| | |
|---|---|
| ARTS | Antibiotic Resistant Target Seeker |
| ABC | Atlas of Biosynthetic gene Clusters |
| AMR | Antimicrobial Resistance |
| ANI | Average Nucleotide Identity |
| antiSMASH | The antibiotics & Secondary Metabolite Analysis Shell |
| BGC | Biosynthetic Gene Cluster |
| HGT | Horizontal Gene Transfer |
| HMM | Hidden Markov Model |
| IMG | Integrated Microbial Genomes |
| KS | ketosynthase |
| MDR | Multi-Drug Resistant |
| MRSA | Methicillin-resistant *Staphylococcus aureus* |
| MiBIG | Minimum Information about a Biosynthetic Gene Cluster |
| NaPDoS | Natural Product Domain Seeker |
| NCBI | National Center for Biotechnology Information |
| NP | Natural product |
| PKS | Polyketide Synthase |
| PRISM | Prediction Informatics for Secondary Metabolomes |
| SeMa-Trap | Secondary Metabolite Transcriptomic Pipeline |
| SM | Secondary metabolite |
| TDGM | Target-Directed Genome Mining |

# Bibliography

[1] Daniel A Dias, Sylvia Urban, and Ute Roessner. A historical overview of natural products in drug discovery. *Metabolites*, 2(2):303–336, 2012.

[2] Sam Hicks. Desert plants and people. 1966.

[3] A Douglas Kinghorn, Li Pan, Joshua N Fletcher, and Heebyung Chai. The relevance of higher plants in lead compound discovery programs. *Journal of natural products*, 74(6):1539–1555, 2011.

[4] Howard Gest. The discovery of microorganisms by robert hooke and antoni van leeuwenhoek, fellows of the royal society. *Notes and records of the Royal Society of London*, 58(2):187–201, 2004.

[5] Margaret M Lock and Vinh-Kim Nguyen. *An anthropology of biomedicine*. John Wiley & Sons, 2018.

[6] Gervase Vernon. Syphilis and salvarsan. *British Journal of General Practice*, 69(682):246–246, 2019.

[7] KJ Williams. The introduction of 'chemotherapy'using arsphenamine–the first magic bullet. *Journal of the Royal Society of Medicine*, 102(8):343–348, 2009.

[8] Robert Gaynes. The discovery of penicillin—new insights after more than 75 years of clinical use. *Emerging infectious diseases*, 23(5):849, 2017.

[9] Ernst Chain, Howard W Florey, Arthur D Gardner, Norman G Heatley, Margaret A Jennings, Jean Orr-Ewing, and A Gordon Sanders. Penicillin as a chemotherapeutic agent. *The lancet*, 236(6104):226–228, 1940.

[10] Julian Davies. Where have all the antibiotics gone? *Canadian Journal of Infectious Diseases and Medical Microbiology*, 17(5):287–290, 2006.

[11] Atanas G Atanasov, Sergey B Zotchev, Verena M Dirsch, and Claudiu T Supuran. Natural products in drug discovery: advances and opportunities. *Nature reviews Drug discovery*, 20(3):200–216, 2021.

[12] Julian Davies. Specialized microbial metabolites: functions and origins. *The Journal of antibiotics*, 66(7):361–364, 2013.

[13] Dão Pedro de Carvalho Neto, Xavier P Gonot-Schoupinsky, and Freda N Gonot-Schoupinsky. Coffee as a naturally beneficial and sustainable ingredient in personal care products: A systematic scoping review of the evidence. *Frontiers in Sustainability*, page 87, 2021.

[14] Rita Roque Bravo, Ana Carolina Faria, Andreia Machado Brito-da Costa, Helena Carmo, Přemysl Mladěnka, Diana Dias da Silva, Fernando Remião, and OEMONOM Researchers. Cocaine: An updated overview on chemistry, detection, biokinetics, and pharmacotoxicological aspects including abuse pattern. *Toxins*, 14(4):278, 2022.

[15] Lili Li, Jieyu Zhao, Yanni Zhao, Xin Lu, Zhihui Zhou, Chunxia Zhao, and Guowang Xu. Comprehensive investigation of tobacco leaves during natural early senescence via multi-platform metabolomics analyses. *Scientific reports*, 6(1):1–10, 2016.

[16] David J Newman and Gordon M Cragg. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of natural products*, 83(3):770–803, 2020.

[17] Shuo-yu Lin, Kyle Baumann, Chenxuan Zhou, Weiyu Zhou, Alison Evans Cuellar, and Hong Xue. Trends in use and expenditures for brand-name statins after introduction of generic statins in the us, 2002-2018. *JAMA network open*, 4(11):e2135371–e2135371, 2021.

[18] Bruno David, Jean-Luc Wolfender, and Daniel A Dias. The pharmaceutical industry and natural products: historical status and new trends. *Phytochemistry Reviews*, 14(2):299–315, 2015.

[19] David J Payne, Michael N Gwynn, David J Holmes, and David L Pompliano. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nature reviews Drug discovery*, 6(1):29–40, 2007.

[20] Christine Årdal, Enrico Baraldi, Ursula Theuretzbacher, Kevin Outterson, Jens Plahte, Francesco Ciabuschi, and John-Arne Røttingen. Insights into early stage of antibiotic development in small-and medium-sized enterprises: a survey of targets, costs, and durations. *Journal of pharmaceutical policy and practice*, 11(1):1–10, 2018.

[21] Christine Årdal, Manica Balasegaram, Ramanan Laxminarayan, David McAdams, Kevin Outterson, John H Rex, and Nithima Sumpradit. Antibiotic development—economic, regulatory and societal challenges. *Nature Reviews Microbiology*, 18(5):267–274, 2020.

[22] Bernardo Ribeiro da Cunha, Luís P Fonseca, and Cecília RC Calado. Antibiotic discovery: where have we come from, where do we go? *Antibiotics*, 8(2):45, 2019.

[23] Tanvir Mahtab Uddin, Arka Jyoti Chakraborty, Ameer Khusro, BM Redwan Matin Zidan, Saikat Mitra, Talha Bin Emran, Kuldeep Dhama, Md Kamal Hossain Ripon, Márió Gajdács, Muhammad Umar Khayam Sahibzada, et al. Antibiotic resistance in microbes: History, mechanisms,

therapeutic strategies and future prospects. *Journal of Infection and Public Health*, 14(12):1750–1766, 2021.

[24] Alison H Skalet, Vicky Cevallos, Berhan Ayele, Teshome Gebre, Zhaoxia Zhou, James H Jorgensen, Mulat Zerihun, Dereje Habte, Yared Assefa, Paul M Emerson, et al. Antibiotic selection pressure and macrolide resistance in nasopharyngeal streptococcus pneumoniae: a cluster-randomized clinical trial. *PLoS medicine*, 7(12):e1000377, 2010.

[25] C Lee Ventola. The antibiotic resistance crisis: part 1: causes and threats. *Pharmacy and therapeutics*, 40(4):277, 2015.

[26] Kristofer Wollein Waldetoft, James Gurney, Joseph Lachance, Paul A Hoskisson, and Sam P Brown. Evolving antibiotics against resistance: a potential platform for natural product development? *MBio*, 10(6):e02946–19, 2019.

[27] Barry G Hall, Stephen J Salipante, and Miriam Barlow. Independent origins of subgroup bl+ b2 and subgroup b3metallo-$\beta$-lactamases. *Journal of molecular evolution*, 59(1):133–141, 2004.

[28] Kelly C Peach, Walter M Bray, Dustin Winslow, Peter F Linington, and Roger G Linington. Mechanism of action-based classification of antibiotics using high-content bacterial image analysis. *Molecular BioSystems*, 9(7):1837–1848, 2013.

[29] I Nikolaidis, S Favini-Stabile, and Andréa Dessen. Resistance to antibiotics targeted to the bacterial cell wall. *Protein science*, 23(3):243–259, 2014.

[30] KOK-FAI KONG, Lisa Schneper, and Kalai Mathee. Beta-lactam antibiotics: from antibiosis to resistance and bacteriology. *Apmis*, 118(1):1–36, 2010.

[31] Kimberly G Blumenthal, Jonny G Peter, Jason A Trubiano, and Elizabeth J Phillips. Antibiotic allergy. *The Lancet*, 393(10167):183–198, 2019.

[32] Sojib Bin Zaman, Muhammed Awlad Hussain, Rachel Nye, Varshil Mehta, Kazi Taib Mamun, and Naznin Hossain. A review on antibiotic resistance: alarm bells are ringing. *Cureus*, 9(6), 2017.

[33] Stephen H Gillespie. Evolution of drug resistance in mycobacterium tuberculosis: clinical and molecular perspective. *Antimicrobial agents and chemotherapy*, 46(2):267–274, 2002.

[34] Julian Davies and Dorothy Davies. Origins and evolution of antibiotic resistance. *Microbiology and molecular biology reviews*, 74(3):417–433, 2010.

[35] Henrietta Venter, Rumana Mowla, Thelma Ohene-Agyei, and Shutao Ma. Rnd-type drug efflux pumps from gram-negative bacteria: molecular mechanism and inhibition. *Frontiers in microbiology*, 6:377, 2015.

[36] Bart N Green, Claire D Johnson, Jonathon Todd Egan, Michael Rosenthal, Erin A Griffith, and Marion Willard Evans. Methicillin-resistant staphylococcus aureus: an overview for manual therapists. *Journal of chiropractic medicine*, 11(1):64–76, 2012.

[37] Mansura S Mulani, Ekta E Kamble, Shital N Kumkar, Madhumita S Tawre, and Karishma R Pardesi. Emerging strategies to combat eskape pathogens in the era of antimicrobial resistance: a review. *Frontiers in microbiology*, 10:539, 2019.

[38] L Tsegaye, P Huston, R Milliken, K Hanniman, C Nesbeth, and L Noad. Antimicrobial resistance (amr): How is an international public health threat advanced in canada? the case of antimicrobial resistance. *Canada Communicable Disease Report*, 42(11):223, 2016.

[39] Giuseppe Mancuso, Angelina Midiri, Elisabetta Gerace, and Carmelo Biondo. Bacterial antibiotic resistance: the most critical pathogens. *Pathogens*, 10(10):1310, 2021.

[40] Marnix H Medema, Tristan de Rond, and Bradley S Moore. Mining genomes to illuminate the specialized chemistry of life. *Nature Reviews Genetics*, 22(9):553–571, 2021.

[41] David J Newman and Gordon M Cragg. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *Journal of natural products*, 75(3):311–335, 2012.

[42] Rene Niehus, Aurore Picot, Nuno M Oliveira, Sara Mitri, and Kevin R Foster. The evolution of siderophore production as a competitive trait. *Evolution*, 71(6):1443–1455, 2017.

[43] DG Larsson and Carl-Fredrik Flach. Antibiotic resistance in the environment. *Nature Reviews Microbiology*, 20(5):257–269, 2022.

[44] Yōko Takahashi and Takuji Nakashima. Actinomycetes, an inexhaustible source of naturally occurring antibiotics. *Antibiotics*, 7(2):45, 2018.

[45] Karan Khadayat, Dawa Dindu Sherpa, Krishna Prakash Malla, Sunil Shrestha, Nabin Rana, Bishnu P Marasini, Santosh Khanal, Binod Rayamajhee, Bibek Raj Bhattarai, and Niranjan Parajuli. Molecular identification and antimicrobial potential of streptomyces species from nepalese soil. *International journal of microbiology*, 2020, 2020.

[46] 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*, 7(5):726–735, 2022.

[47] Paul R Jensen, Krystle L Chavarria, William Fenical, Bradley S Moore, and Nadine Ziemert. Challenges and triumphs to genomics-based natural product discovery. *Journal of Industrial Microbiology and Biotechnology*, 41(2):203–209, 2014.

[48] Nadine Ziemert, Mohammad Alanjary, and Tilmann Weber. The evolution of genome mining in microbes–a review. *Natural product reports*, 33(8):988–1005, 2016.

[49] Mohammad R Seyedsayamdost, Matthew F Traxler, Jon Clardy, and Roberto Kolter. Old meets new: using interspecies interactions to detect secondary metabolite production in actinomycetes. In *Methods in enzymology*, volume 517, pages 89–109. Elsevier, 2012.

[50] Lin Du, Jarrod B King, Brian H Morrow, Jana K Shen, Andrew N Miller, and Robert H Cichewicz. Diarylcyclopentendione metabolite obtained from a preussia typharum isolate procured using an unconventional cultivation approach. *Journal of natural products*, 75(10):1819–1823, 2012.

[51] Yunzi Luo, Ryan E Cobb, and Huimin Zhao. Recent advances in natural product discovery. *Current opinion in biotechnology*, 30:230–237, 2014.

[52] Brian O Bachmann, Steven G Van Lanen, and Richard H Baltz. Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *Journal of industrial Microbiology and Biotechnology*, 41(2):175–184, 2014.

[53] Frederick Sanger and EOP Thompson. The amino-acid sequence in the glycyl chain of insulin. 1. the identification of lower peptides from partial hydrolysates. *Biochemical Journal*, 53(3):353, 1953.

[54] Glyn Moody. *Digital code of life: how bioinformatics is revolutionizing science, medicine, and business*. John Wiley & Sons, 2004.

[55] Margaret O Dayhoff. *Atlas of protein sequence and structure*. National Biomedical Research Foundation., 1972.

[56] Margaret Oakley Dayhoff and Robert S Ledley. Comprotein: a computer program to aid primary protein structure determination. In *Proceedings of the December 4-6, 1962, fall joint computer conference*, pages 262–274, 1962.

[57] Frederick Sanger, Gilian M Air, Bart G Barrell, Nigel L Brown, Alan R Coulson, John C Fiddes, CA Hutchison, Patrick M Slocombe, and Mo Smith. Nucleotide sequence of bacteriophage $\varphi$x174 dna. *nature*, 265(5596):687–695, 1977.

[58] John W Erickson and Gary G Altman. A search for patterns in the nucleotide sequence of the ms2 genome. *Journal of Mathematical Biology*, 7(3):219–230, 1979.

[59] Patrick Argos, Michael Hanei, and R Michael Garavito. The chou-fasman secondary structure prediction method with an extended data base. *FEBS letters*, 93(1):19–24, 1978.

[60] Emilia Palazzotto and Tilmann Weber. Omics and multi-omics approaches to study the biosynthesis of secondary metabolites in microorganisms. *Current opinion in microbiology*, 45:109–116, 2018.

[61] Gang Liu, Keith F Chater, Govind Chandra, Guoqing Niu, and Huarong Tan. Molecular regulation of antibiotic biosynthesis in streptomyces. *Microbiology and molecular biology reviews*, 77(1):112–143, 2013.

[62] Ulfert Hornemann and David A Hopwood. Biosynthesis of methylenomycin a: a plasmid-determined antibiotic. In *Biosynthesis*, pages 123–131. Springer, 1981.

[63] Brian AM Rudd and David A Hopwood. Genetics of actinorhodin biosynthesis by streptomyces coelicolor a3 (2). *Microbiology*, 114(1):35–43, 1979.

[64] Ff Malpartida and DA Hopwood. Molecular cloning of the whole biosynthetic pathway of a streptomyces antibiotic and its expression in a heterologous host. *Nature*, 309(5967):462–464, 1984.

[65] Marion Steffensky, Agnes Muhlenweg, Zhao-Xin Wang, Shu-Ming Li, and Lutz Heide. Identification of the novobiocin biosynthetic gene cluster of streptomyces spheroides ncib 11891. *Antimicrobial agents and chemotherapy*, 44(5):1214–1222, 2000.

[66] Mervyn J Bibb. Regulation of secondary metabolism in streptomycetes. *Current opinion in microbiology*, 8(2):208–215, 2005.

[67] Carmen Méndez and José A Salas. The role of abc transporters in antibiotic-producing organisms: drug secretion and resistance mechanisms. *Research in microbiology*, 152(3-4):341–350, 2001.

[68] Juan F Martín, Javier Casqueiro, and Paloma Liras. Secretion systems for secondary metabolites: how producer cells send out messages of intercellular communication. *Current opinion in microbiology*, 8(3):282–293, 2005.

[69] Aleksandra Nivina, Kai P Yuet, Jake Hsu, and Chaitan Khosla. Evolution and diversity of assembly-line polyketide synthases: Focus review. *Chemical reviews*, 119(24):12524–12547, 2019.

[70] Bradley S Moore and Christian Hertweck. Biosynthesis and attachment of novel bacterial polyketide synthase starter units. *Natural product reports*, 19(1):70–99, 2002.

[71] Michael A Fischbach and Christopher T Walsh. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chemical reviews*, 106(8):3468–3496, 2006.

[72] Manmeet Ahuja, Yi-Ming Chiang, Shu-Lin Chang, Mike B Praseuth, Ruth Entwistle, James F Sanchez, Hsien-Chun Lo, Hsu-Hua Yeh, Berl R Oakley, and Clay CC Wang. Illuminating the diversity of aromatic polyketide synthases in aspergillus nidulans. *Journal of the American Chemical Society*, 134(19):8212–8221, 2012.

[73] Emma Kenshole, Marion Herisse, Michael Michael, and Sacha J Pidot. Natural product discovery through microbial genome mining. *Current Opinion in Chemical Biology*, 60:47–54, 2021.

[74] Jessie James Limlingan Malit, Hiu Yu Cherie Leung, and Pei-Yuan Qian. Targeted large-scale genome mining and candidate prioritization for natural product discovery. *Marine Drugs*, 20(6):398, 2022.

[75] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[76] Emmanuel Zazopoulos, Kexue Huang, Alfredo Staffa, Wen Liu, Brian O Bachmann, Koichi Nonaka, Joachim Ahlert, Jon S Thorson, Ben Shen, and Chris M Farnet. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. *Nature biotechnology*, 21(2):187–190, 2003.

[77] Mohd Zeeshan Ansari, Gitanjali Yadav, Rajesh S Gokhale, and Debasisa Mohanty. Nrps-pks: a knowledge-based resource for analysis of nrps/pks megasynthases. *Nucleic acids research*, 32(suppl_2):W405–W413, 2004.

[78] Sean R Eddy. Accelerated profile hmm searches. *PLoS computational biology*, 7(10):e1002195, 2011.

[79] Hengqian Ren, Chengyou Shi, and Huimin Zhao. Computational tools for discovering and engineering natural product biosynthetic pathways. *Iscience*, 23(1):100795, 2020.

[80] Marnix H Medema, Kai Blin, Peter Cimermancic, Victor De Jager, Piotr Zakrzewski, Michael A Fischbach, Tilmann Weber, Eriko Takano, and Rainer Breitling. antismash: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research*, 39(suppl_2):W339–W346, 2011.

[81] Kai Blin, Simon Shaw, Alexander M Kloosterman, Zach Charlop-Powers, Gilles P Van Wezel, Marnix H Medema, and Tilmann Weber. antismash 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research*, 49(W1):W29–W35, 2021.

[82] Martina Adamek, Mohammad Alanjary, and Nadine Ziemert. Applied evolution: phylogeny-based approaches in natural products research. *Natural Product Reports*, 36(9):1295–1312, 2019.

[83] Nelly Sélem-Mojica, César Aguilar, Karina Gutiérrez-García, Christian E Martínez-Guerrero, and Fancisco Barona-Gómez. Evomining reveals the origin and fate of natural product biosynthetic enzymes. *Microbial genomics*, 5(12), 2019.

[84] Jacob J Banik and Sean F Brady. Recent application of metagenomic approaches toward the discovery of antimicrobials and other bioactive small molecules. *Current opinion in microbiology*, 13(5):603–609, 2010.

[85] Zhiyang Feng, Dimitris Kallifidas, and Sean F Brady. Functional analysis of environmental dna-derived type ii polyketide synthases reveals structurally diverse secondary metabolites. *Proceedings of the National Academy of Sciences*, 108(31):12629–12634, 2011.

[86] Hahk-Soo Kang and Sean F Brady. Arimetamycin a: improving clinically relevant families of natural products through sequence-guided screening of soil metagenomes. *Angewandte Chemie International Edition*, 52(42):11063–11067, 2013.

[87] Nadine Ziemert, Sheila Podell, Kevin Penn, Jonathan H Badger, Eric Allen, and Paul R Jensen. The natural product domain seeker napdos: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PloS one*, 7(3):e34064, 2012.

[88] Gabriel A Vignolle, Denise Schaffer, Leopold Zehetner, Robert L Mach, Astrid R Mach-Aigner, and Christian Derntl. Funorder: A robust and semi-automated method for the identification of essential biosynthetic genes through computational molecular co-evolution. *PLoS computational biology*, 17(9):e1009372, 2021.

[89] Boojala Vijay B Reddy, Aleksandr Milshteyn, Zachary Charlop-Powers, and Sean F Brady. esnapd: a versatile, web-based bioinformatics platform for surveying and mining natural product biosynthetic diversity from metagenomes. *Chemistry & biology*, 21(8):1023–1033, 2014.

[90] Jorge C Navarro-Muñoz, Nelly Selem-Mojica, Michael W Mullowney, Satria A Kautsar, James H Tryon, Elizabeth I Parkinson, Emmanuel LC De Los Santos, Marley Yeong, Pablo Cruz-Morales, Sahar Abubucker, et al. A computational framework to explore large-scale biosynthetic diversity. *Nature chemical biology*, 16(1):60–68, 2020.

[91] Zhenyu Liu, Yatong Zhao, Chaoqun Huang, and Yunzi Luo. Recent advances in silent gene cluster activation in streptomyces. *Frontiers in Bioengineering and Biotechnology*, 9:632230, 2021.

[92] Min Xu and Gerard D Wright. Heterologous expression-facilitated natu-

ral products' discovery in actinomycetes. *Journal of Industrial Microbiology and Biotechnology*, 46(3-4):415–431, 2019.

[93] Bikash Baral, Amir Akhgari, and Mikko Metsä-Ketelä. Activation of microbial secondary metabolic pathways: Avenues and challenges. *Synthetic and Systems Biotechnology*, 3(3):163–178, 2018.

[94] Jeremy G Owen, Boojala Vijay B Reddy, Melinda A Ternei, Zachary Charlop-Powers, Paula Y Calle, Jeffrey H Kim, and Sean F Brady. Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proceedings of the National Academy of Sciences*, 110(29):11797–11802, 2013.

[95] Fabrizio Alberti, Daniel J Leng, Ina Wilkening, Lijiang Song, Manuela Tosin, and Christophe Corre. Triggering the expression of a silent gene cluster from genetically intractable bacteria results in scleric acid discovery. *Chemical science*, 10(2):453–463, 2019.

[96] Bradley M Hover, Seong-Hwan Kim, Micah Katz, Zachary Charlop-Powers, Jeremy G Owen, Melinda A Ternei, Jeffrey Maniko, Andreia B Estrela, Henrik Molina, Steven Park, et al. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant gram-positive pathogens. *Nature microbiology*, 3(4):415–422, 2018.

[97] Fang-Yuan Chang and Sean F Brady. Discovery of indolotryptoline antiproliferative agents by homology-guided metagenomic screening. *Proceedings of the National Academy of Sciences*, 110(7):2478–2483, 2013.

[98] Leonard Katz and Richard H Baltz. Natural product discovery: past, present, and future. *Journal of Industrial Microbiology and Biotechnology*, 43(2-3):155–176, 2016.

[99] Thorger Lincke, Swantje Behnken, Keishi Ishida, Martin Roth, and Christian Hertweck. Closthioamide: an unprecedented polythioamide antibiotic from the strictly anaerobic bacterium clostridium cellulolyticum. *Angewandte Chemie*, 122(11):2055–2057, 2010.

[100] Takeshi Hosaka, Mayumi Ohnishi-Kameyama, Hideyuki Muramatsu, Kana Murakami, Yasuhisa Tsurumi, Shinya Kodani, Mitsuru Yoshida, Akihiko Fujie, and Kozo Ochi. Antibacterial discovery in actinomycetes strains with mutations in rna polymerase or ribosomal protein s12. *Nature biotechnology*, 27(5):462–464, 2009.

[101] Amy M Lum, Jianqiang Huang, C Richard Hutchinson, and Camilla M Kao. Reverse engineering of industrial pharmaceutical-producing actinomycete strains using dna microarrays. *Metabolic engineering*, 6(3):186–196, 2004.

[102] Ying Zhuo, Wenquan Zhang, Difei Chen, Hong Gao, Jun Tao, Mei Liu, Zhongxuan Gou, Xianlong Zhou, Bang-Ce Ye, Qing Zhang, et al. Reverse biological engineering of hrdb to enhance the production of avermectins in an industrial strain of streptomyces avermitilis. *Proceedings of the National Academy of Sciences*, 107(25):11250–11254, 2010.

[103] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[104] Gregory CA Amos, Takayoshi Awakawa, Robert N Tuttle, Anne-Catrin Letzel, Min Cheol Kim, Yuta Kudo, William Fenical, Bradley S. Moore, and Paul R Jensen. Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. *Proceedings of the National Academy of Sciences*, 114(52):E11121–E11130, 2017.

[105] Xiaobo Li, Xiang Ke, Lijia Qiao, Yufei Sui, and Ju Chu. Comparative genomic and transcriptomic analysis guides to further enhance the biosynthesis of erythromycin by an overproducer. *Biotechnology and Bioengineering*, 119(6):1624–1640, 2022.

[106] Maria Sorokina and Christoph Steinbeck. Review on natural products databases: where to find data in 2020. *Journal of cheminformatics*, 12(1):1–51, 2020.

[107] Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 44(D1):D67–D72, 2016.

[108] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016.

[109] Rasko Leinonen, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Neil Goodgame, Richard Gibson, et al. The european nucleotide archive. *Nucleic acids research*, 39(suppl_1):D28–D31, 2010.

[110] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research*, 39(suppl_1):D19–D21, 2010.

[111] Thomas W McCarthy, Hsien-chao Chou, and Volker P Brendel. Srassembler: Selective recursive local assembly of homologous genomic regions. *BMC bioinformatics*, 20(1):1–13, 2019.

[112] Dmitry Meleshko, Hosein Mohimani, Vittorio Tracanna, Iman Hajirasouliha, Marnix H Medema, Anton Korobeynikov, and Pavel A Pevzner.

Biosyntheticspades: reconstructing biosynthetic gene clusters from assembly graphs. *Genome research*, 29(8):1352–1362, 2019.

[113] Natsuko Ichikawa, Machi Sasagawa, Mika Yamamoto, Hisayuki Komaki, Yumi Yoshida, Shuji Yamazaki, and Nobuyuki Fujita. Dobiscuit: a database of secondary metabolite biosynthetic gene clusters. *Nucleic acids research*, 41(D1):D408–D414, 2012.

[114] Kyle R Conway and Christopher N Boddy. Clustermine360: a database of microbial pks/nrps biosynthesis. *Nucleic acids research*, 41(D1):D402–D407, 2012.

[115] Xavier Lucas, Christian Senger, Anika Erxleben, Björn A Grüning, Kersten Döring, Johannes Mosch, Stephan Flemming, and Stefan Günther. Streptomedb: a resource for natural compounds isolated from streptomyces species. *Nucleic acids research*, 41(D1):D1130–D1136, 2012.

[116] Kai Blin, Marnix H Medema, Daniyal Kazempour, Michael A Fischbach, Rainer Breitling, Eriko Takano, and Tilmann Weber. antismash 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic acids research*, 41(W1):W204–W212, 2013.

[117] Peter Cimermancic, Marnix H Medema, Jan Claesen, Kenji Kurita, Laura C Wieland Brown, Konstantinos Mavrommatis, Amrita Pati, Paul A Godfrey, Michael Koehrsen, Jon Clardy, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, 158(2):412–421, 2014.

[118] Michalis Hadjithomas, I-Min Amy Chen, Ken Chu, Anna Ratner, Krishna Palaniappan, Ernest Szeto, Jinghua Huang, TBK Reddy, Peter Cimermančič, Michael A Fischbach, et al. Img-abc: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *MBio*, 6(4):e00932–15, 2015.

[119] Krishnaveni Palaniappan, I-Min A Chen, Ken Chu, Anna Ratner, Rekha Seshadri, Nikos C Kyrpides, Natalia N Ivanova, and Nigel J Mouncey. Img-abc v. 5.0: an update to the img/atlas of biosynthetic gene clusters knowledgebase. *Nucleic acids research*, 48(D1):D422–D430, 2020.

[120] Marnix H Medema, Renzo Kottmann, Pelin Yilmaz, Matthew Cummings, John B Biggins, Kai Blin, Irene De Bruijn, Yit Heng Chooi, Jan Claesen, R Cameron Coates, et al. Minimum information about a biosynthetic gene cluster. *Nature chemical biology*, 11(9):625–631, 2015.

[121] Tilmann Weber and Hyun Uk Kim. The secondary metabolite bioinformatics portal: computational tools to facilitate synthetic biology of secondary metabolite production. *Synthetic and Systems Biotechnology*, 1(2):69–79, 2016.

[122] Nadja B Cech, Marnix H Medema, and Jon Clardy. Benefiting from big data in natural products: importance of preserving foundational skills and prioritizing data quality. *Natural Product Reports*, 38(11):1947–1953, 2021.

[123] Emma Ricart, Valérie Leclère, Areski Flissi, Markus Mueller, Maude Pupin, and Frederique Lisacek. rban: retro-biosynthetic analysis of nonribosomal peptides. *Journal of cheminformatics*, 11(1):1–14, 2019.

[124] Hosein Mohimani, Wei-Ting Liu, Roland D Kersten, Bradley S Moore, Pieter C Dorrestein, and Pavel A Pevzner. Nrpquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. *Journal of natural products*, 77(8):1902–1909, 2014.

[125] James R Doroghazi, 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. A roadmap for natural product discov-

ery based on large-scale genomics and metabolomics. *Nature chemical biology*, 10(11):963–968, 2014.

[126] Marnix H Medema, Yared Paalvast, Don D Nguyen, Alexey Melnik, Pieter C Dorrestein, Eriko Takano, and Rainer Breitling. Pep2path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS computational biology*, 10(9):e1003822, 2014.

[127] Katherine R Duncan, Max Crüsemann, Anna Lechner, Anindita Sarkar, Jie Li, Nadine Ziemert, Mingxun Wang, Nuno Bandeira, Bradley S Moore, Pieter C Dorrestein, et al. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from salinispora species. *Chemistry & biology*, 22(4):460–471, 2015.

[128] Jeffrey D Rudolf, Xiaohui Yan, and Ben Shen. Genome neighborhood network reveals insights into enediyne biosynthesis and facilitates prediction and prioritization for discovery. *Journal of Industrial Microbiology and Biotechnology*, 43(2-3):261–276, 2016.

[129] Gerard D Wright. Molecular mechanisms of antibiotic resistance. *Chemical communications*, 47(14):4055–4061, 2011.

[130] Mehmet Direnç Mungan, Mohammad Alanjary, Kai Blin, Tilmann Weber, Marnix H Medema, and Nadine Ziemert. Arts 2.0: feature updates and expansion of the antibiotic resistant target seeker for comparative genome mining. *Nucleic acids research*, 48(W1):W546–W552, 2020.

[131] Jason Stahlecker, Erik Mingyar, Nadine Ziemert, and Mehmet Direnç Mungan. Syn-view: a phylogeny-based synteny exploration tool for the identification of gene clusters linked to antibiotic resistance. *Molecules*, 26(1):144, 2020.

[132] Mehmet Direnç Mungan, Kai Blin, and Nadine Ziemert. Arts-db: a database for antibiotic resistant targets. *Nucleic acids research*, 50(D1):D736–D740, 2022.

[133] Ellis C O'Neill, Michelle Schorn, Charles B Larson, and Natalie Millán-Aguiñaga. Targeted antibiotic discovery through biosynthesis-associated resistance determinants: Target directed genome mining. *Critical reviews in microbiology*, 45(3):255–277, 2019.

[134] Mehmet Direnç Mungan, Theresa Anisja Harbig, Naybel Hernandez Perez, Simone Edenhart, Evi Stegmann, Kay Nieselt, and Nadine Ziemert. Secondary metabolite transcriptomic pipeline (sema-trap), an expression-based exploration tool for increased secondary metabolite production in bacteria. *Nucleic Acids Research*, 2022.

[135] Fredrik Backhed, Ruth E Ley, Justin L Sonnenburg, Daniel A Peterson, and Jeffrey I Gordon. Host-bacterial mutualism in the human intestine. *science*, 307(5717):1915–1920, 2005.

[136] Jean-Yves Maillard, Sally F Bloomfield, Patrice Courvalin, Sabiha Y Essack, Sumanth Gandra, Charles P Gerba, Joseph R Rubino, and Elizabeth A Scott. Reducing antibiotic prescribing and addressing the global problem of antibiotic resistance by targeted hygiene in the home and everyday life settings: A position paper. *American journal of infection control*, 48(9):1090–1099, 2020.

[137] Laura J Shallcross and Dame Sally C Davies. Antibiotic overuse: a key driver of antimicrobial resistance, 2014.

[138] Seema Verma. Aligning Payment And Prevention To Drive Antibiotic Innovation For Medicare Beneficiaries. `https://www.healthaffairs.org/do/10.1377/forefront.20190802.505113/full/`, 2019.

[139] Christine Oline Årdal, David Findlay, Miloje Savic, Yehuda Carmeli, Inge Gyssens, Ramanan Laxminarayan, Kevin Outterson, and John H Rex. Revitalizing the antibiotic pipeline: Stimulating innovation while driving sustainable use and global access. 2018.

[140] Asher Mullard. Achaogen bankruptcy highlights antibacterial development woes. *Nature Reviews Drug Discovery*, 18(6):411–412, 2019.

[141] Wun-Rong Hong, Mei Ge, Zhi-Hong Zeng, Li Zhu, Min-Yu Luo, Lei Shao, and Dai-Jie Chen. Molecular cloning and sequence analysis of the sisomicin biosynthetic gene cluster from micromonospora inyoensis. *Biotechnology letters*, 31(3):449–455, 2009.

[142] Lesley-Ann Giddings, Kevin Kunstman, Bouziane Moumen, Laurent Asiama, Stefan Green, Vincent Delafont, Matthew Brockley, and Ascel Samba-Louaka. Isolation and genome analysis of an amoeba-associated bacterium dyella terrae strain ely copper mine from acid rock drainage in vermont, united states. *Frontiers in microbiology*, 13, 2022.

[143] Hilal Ay. Genomic insight into a novel actinobacterium, actinomadura rubrisoli sp. nov., reveals high potential for bioactive metabolites. *Antonie van Leeuwenhoek*, 114(2):195–208, 2021.

[144] Ana Catalina Lara, Erika Corretto, Lucie Kotrbová, František Lorenc, Kateřina Petříčková, Roman Grabic, and Alica Chroňáková. The genome analysis of the human lung-associated streptomyces sp. tr1341 revealed the presence of beneficial genes for opportunistic colonization of human tissues. *Microorganisms*, 9(8):1547, 2021.

[145] Izzet Burcin Saticioglu. Flavobacterium erciyesense sp. nov., a putative non-pathogenic fish symbiont. *Archives of Microbiology*, 203(9):5783–5792, 2021.

[146] Hayrettin Saygin, Hilal Ay, Kiymet Guven, Demet Cetin, and Nevzat Sahin. Comprehensive genome analysis of a novel actinobacterium with high potential for biotechnological applications, nonomuraea aridisoli sp. nov., isolated from desert soil. *Antonie van Leeuwenhoek*, 114(12):1963–1975, 2021.

[147] Xiaohe Jin, Eric S Miller, and Jonathan S Lindsey. Natural product gene clusters in the filamentous nostocales cyanobacterium ht-58-2. *Life*, 11(4):356, 2021.

[148] Hayrettin Saygin, Hilal Ay, Kiymet Guven, Kadriye Inan-Bektas, Demet Cetin, and Nevzat Sahin. Saccharopolyspora karakumensis sp. nov., saccharopolyspora elongata sp. nov., saccharopolyspora aridisoli sp. nov., saccharopolyspora terrae sp. nov. and their biotechnological potential revealed by genome analysis. *Systematic and Applied Microbiology*, 44(6):126270, 2021.

[149] Ali Budhi Kusuma. *Microbiology of Indonesian extremobiospheres: from unexplored actinobacteria diversity to novel antimicrobial discovery*. PhD thesis, Newcastle University, 2021.

[150] Ashley Isaac, Ben Francis, Rudolf I Amann, and Shady A Amin. Tight adherence (tad) pilus genes indicate putative niche differentiation in phytoplankton bloom associated rhodobacterales. *Frontiers in microbiology*, 12, 2021.

[151] Robert Murphy, René Benndorf, Z Wilhelm De Beer, John Vollmers, Anne-Kristin Kaster, Christine Beemelmanns, and Michael Poulsen. Comparative genomics reveals prophylactic and catabolic capabilities of actinobacteria within the fungus-farming termite symbiosis. *MSphere*, 6(2):e01233–20, 2021.

[152] Carlos Caicedo-Montoya, Monserrat Manzo-Ruiz, and Rigoberto Ríos-

Estepa. Pan-genome of the genus streptomyces and prioritization of biosynthetic gene clusters with potential to produce antibiotic compounds. *Frontiers in microbiology*, 12, 2021.

[153] Sylvia Valdezate, Fernando Cobo, Sara Monzón, María J Medina-Pascual, Ángel Zaballos, Isabel Cuesta, Silvia Pino-Rosa, and Pilar Villalón. Genomic background and phylogeny of cfi a-positive bacteroides fragilis strains resistant to meropenem-edta. *Antibiotics*, 10(3):304, 2021.

[154] Yan Yan, Nicholas Liu, and Yi Tang. Recent developments in self-resistance gene directed natural product discovery. *Natural Product Reports*, 37(7):879–892, 2020.

[155] Zhao-Xin Wang, Shu-Ming Li, and Lutz Heide. Identification of the coumermycin a1biosynthetic gene cluster of streptomyces rishiriensis dsm 40489. *Antimicrobial agents and chemotherapy*, 44(11):3040–3048, 2000.

[156] Torsten Bak Regueira, Kanchana Rueksomtawin Kildegaard, Bjarne Gram Hansen, Uffe H Mortensen, Christian Hertweck, and Jens Nielsen. Molecular basis for mycophenolic acid biosynthesis in penicillium brevicompactum. *Applied and environmental microbiology*, 77(9):3035–3043, 2011.

[157] Mohammad Alanjary, Brent Kronmiller, Martina Adamek, Kai Blin, Tilmann Weber, Daniel Huson, Benjamin Philmus, and Nadine Ziemert. The antibiotic resistant target seeker (arts), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic acids research*, 45(W1):W42–W48, 2017.

[158] Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. Gtdb-tk: a toolkit to classify genomes with the genome taxonomy database, 2020.

[159] Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology*, 36(10):996–1004, 2018.

[160] Hye-Seon Kim, Jessica M Lohmar, Mark Busman, Daren W Brown, Todd A Naumann, Hege H Divon, Erik Lysøe, Silvio Uhlig, and Robert H Proctor. Identification and distribution of gene clusters required for synthesis of sphingolipid metabolism inhibitors in diverse species of the filamentous fungus fusarium. *BMC genomics*, 21(1):1–24, 2020.

[161] Nitish Sharma, Reena Kumari, Monika Thakur, Amit K Rai, and Sudhir P Singh. Molecular dissemination of emerging antibiotic, biocide, and metal co-resistomes in the himalayan hot springs. *Journal of Environmental Management*, 307:114569, 2022.

[162] Francesco Del Carratore, Konrad Zych, Matthew Cummings, Eriko Takano, Marnix H Medema, and Rainer Breitling. Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters. *Communications biology*, 2(1):1–10, 2019.

[163] Pablo Cruz-Morales, Johannes Florian Kopp, Christian Martínez-Guerrero, Luis Alfonso Yáñez-Guerra, Nelly Selem-Mojica, Hilda Ramos-Aboites, Jörg Feldmann, and Francisco Barona-Gómez. Phylogenomic analysis of natural products biosynthetic gene clusters allows discovery of arseno-organic metabolites in model streptomycetes. *Genome biology and evolution*, 8(6):1906–1916, 2016.

[164] Yan Yan, Qikun Liu, Xin Zang, Shuguang Yuan, Undramaa Bat-Erdene, Calvin Nguyen, Jianhua Gan, Jiahai Zhou, Steven E Jacobsen, and Yi Tang. Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action. *Nature*, 559(7714):415–418, 2018.

[165] Jana K Schniete, Pablo Cruz-Morales, Nelly Selem-Mojica, Lorena T Fernández-Martínez, Iain S Hunter, Francisco Barona-Gómez, and Paul A Hoskisson. Expanding primary metabolism helps generate the metabolic robustness to facilitate antibiotic biosynthesis in streptomyces. *MBio*, 9(1):e02283–17, 2018.

[166] Hervé Isambert and Richard R Stein. On the need for widespread horizontal gene transfers under genome size constraint. *Biology direct*, 4(1):1–10, 2009.

[167] Marit S Bratlie, Jostein Johansen, Brad T Sherman, Da Wei Huang, Richard A Lempicki, and Finn Drabløs. Gene duplications in prokaryotes can be associated with environmental adaptation. *BMC genomics*, 11(1):1–17, 2010.

[168] AKM Firoj Mahmud, Nicolas Delhomme, Soumyadeep Nandi, and Maria Fällman. Prokseq for complete analysis of rna-seq data from prokaryotes. *Bioinformatics*, 37(1):126–128, 2021.

[169] Dechao Bu, Haitao Luo, Peipei Huo, Zhihao Wang, Shan Zhang, Zihao He, Yang Wu, Lianhe Zhao, Jingjia Liu, Jincheng Guo, et al. Kobas-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic acids research*, 49(W1):W317–W325, 2021.

[170] Michael A Skinnider, Chad W Johnston, Mathusan Gunabalasingam, Nishanth J Merwin, Agata M Kieliszek, Robyn J MacLellan, Haoxin Li, Michael RM Ranieri, Andrew LH Webster, My Cao, et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nature communications*, 11(1):1–9, 2020.

[171] Thomas Wolf, Vladimir Shelest, Neetika Nath, and Ekaterina Shelest.

Cassis and smips: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. *Bioinformatics*, 32(8):1138–1143, 2016.

[172] Mackinlay Card. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[173] Seán I O'Donoghue. Grand challenges in bioinformatics data visualization. *Frontiers in Bioinformatics*, page 13, 2021.

[174] Gur Sevillya, Orit Adato, and Sagi Snir. Detecting horizontal gene transfer: a probabilistic approach. *BMC genomics*, 21(1):1–11, 2020.

[175] Luay Nakhleh, Derek Ruths, and Li-San Wang. Riata-hgt: a fast and accurate heuristic for reconstructing horizontal gene transfer. In *International Computing and Combinatorics Conference*, pages 84–93. Springer, 2005.

[176] J Peter Gogarten and Jeffrey P Townsend. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9):679–687, 2005.

[177] Deise JP Goncalves, Beryl B Simpson, Edgardo M Ortiz, Gustavo H Shimizu, and Robert K Jansen. Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. *Molecular phylogenetics and evolution*, 138:219–232, 2019.

[178] Franziska Liesecke, Dimitri Daudu, Rodolphe Dugé de Bernonville, Sébastien Besseau, Marc Clastre, Vincent Courdavault, Johan-Owen De Craene, Joel Crèche, Nathalie Giglioli-Guivarc'h, Gaëlle Glévarec, et al. Ranking genome-wide correlation measurements improves microarray and rna-seq based global and targeted co-expression networks. *Scientific reports*, 8(1):1–16, 2018.

[179] Nicholas Liu, Elizabeth D Abramyan, Wei Cheng, Bruno Perlatti, Colin JB Harvey, Gerald F Bills, and Yi Tang. Targeted genome mining reveals the biosynthetic gene clusters of natural product cyp51 inhibitors. *Journal of the American Chemical Society*, 143(16):6043–6047, 2021.

[180] Philipp Wiemann and Nancy P Keller. Strategies for mining fungal natural products. *Journal of Industrial Microbiology and Biotechnology*, 41(2):301–313, 2014.

[181] Zhuang Ding, Xiao Wang, Fan-Dong Kong, Hui-Ming Huang, Yan-Na Zhao, Min Liu, Zheng-Ping Wang, and Jun Han. Overexpression of global regulator talae1 leads to the discovery of new antifungal polyketides from endophytic fungus trichoderma afroharzianum. *Frontiers in microbiology*, 11:622785, 2020.