MICROBIAL GENOMICS

RESEARCH ARTICLE

Undabarrena *et al.*, *Microbial Genomics* 2021;7:000621

DOI 10.1099/mgen.0.000621

Rhodococcus comparative genomics reveals a phylogenomic-dependent non-ribosomal peptide synthetase distribution: insights into biosynthetic gene cluster connection to an orphan metabolite

Agustina Undabarrena[1,*], Ricardo Valencia[1]†, Andrés Cumsille[1], Leonardo Zamora-Leiva[1], Eduardo Castro-Nallar[2], Francisco Barona-Gomez[3] and Beatriz Cámara[1,*]

Abstract

Natural products (NPs) are synthesized by biosynthetic gene clusters (BGCs), whose genes are involved in producing one or a family of chemically related metabolites. Advances in comparative genomics have been favourable for exploiting huge amounts of data and discovering previously unknown BGCs. Nonetheless, studying distribution patterns of novel BGCs and elucidating the biosynthesis of orphan metabolites remains a challenge. To fill this knowledge gap, our study developed a pipeline for high-quality comparative genomics for the actinomycete genus *Rhodococcus*, which is metabolically versatile, yet understudied in terms of NPs, leading to a total of 110 genomes, 1891 BGCs and 717 non-ribosomal peptide synthetases (NRPSs). Phylogenomic inferences showed four major clades retrieved from strains of several ecological habitats. BiG-SCAPE sequence similarity BGC networking revealed 44 unidentified gene cluster families (GCFs) for NRPS, which presented a phylogenomic-dependent evolution pattern, supporting the hypothesis of vertical gene transfer. As a proof of concept, we analysed in-depth one of our marine strains, *Rhodococcus* sp. H-CA8f, which revealed a unique BGC distribution within its phylogenomic clade, involved in producing a chloramphenicol-related compound. While this BGC is part of the most abundant and widely distributed NRPS GCF, CORASON analysis unveiled major differences regarding its genetic context, co-occurrence patterns and modularity. This BGC is composed of three sections, two well-conserved right/left arms flanking a very variable middle section, composed of *nrps* genes. The presence of two non-canonical domains in H-CA8f's BGC may contribute to adding chemical diversity to this family of NPs. Liquid chromatography-high resolution MS and dereplication efforts retrieved a set of related orphan metabolites, the corynecins, which to our knowledge are reported here for the first time in *Rhodococcus*. Overall, our data provide insights to connect BGC uniqueness with orphan metabolites, by revealing key comparative genomic features supported by models of BGC distribution along phylogeny.

DATA SUMMARY

All supporting data and protocols have been provided within the article or through supplementary data files or Figshare repositories (https://doi.org/10.6084/m9.figshare.13158086.v2). Public genome data were retrieved from the National Center for Biotechnology Information GenBank (Table S1, available in the online version of this article). Code scripts are available as Jupyter notebooks in a GitHub repository (https://github.com/

Received 16 February 2021; Accepted 04 June 2021; Published 09 July 2021

Author affiliations: [1]Laboratorio de Microbiología Molecular y Biotecnología Ambiental, Departamento de Química y Centro de Biotecnología Daniel Alkalay Lowitt, Universidad Técnica Federico Santa María, Valparaíso 2340000, Chile; [2]Center for Bioinformatics and Integrative Biology, Facultad de Ciencias de la Vida, Universidad Andres Bello, Santiago, Chile; [3]Evolution of Metabolic Diversity Laboratory, Unidad de Genómica Avanzada (Langebio), Cinvestav, Irapuato, Guanajuato, Mexico.

***Correspondence:** Beatriz Cámara, beatriz.camara@usm.cl; Agustina Undabarrena, agustina.undabarrena@usm.cl

Keywords: biosynthetic gene clusters; comparative genomics; non-ribosomal peptide synthetase evolution; orphan metabolites; *Rhodococcus*.

Abbreviations: ASW, artificial sea water; BGC, biosynthetic gene cluster; BY, Bayesian multilocus phylogeny; GCF, gene cluster family; LC-HRMS, liquid chromatography-high resolution; ML, maximum likelihood; MS, mass spectrometry; NCBI, National Center for Biotechnology Information; NP, natural product; NRPS, non-ribosomal peptide synthetase; PERMANOVA, permutational multivariate analysis of variance.

†Present address: Institute of Quantitative Biology, Biochemistry and Biotechnology, School of Biological Sciences, University of Edinburgh, King's Buildings, Edinburgh, UK.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Five supplementary tables and five supplementary figures are available with the online version of this article.

rvalenciaaz/rhodococcus-bgc). All supplementary material can be found on Figshare (https://doi.org/10.6084/m9.figshare.13158086.v2).

Impact Statement

Biosynthetic gene clusters (BGCs) harbour genetic information to build a myriad of natural products (NPs). Actinomycete NPs provide an unsurpassed resource in drug discovery to face multi-resistant pathogenic bacteria. Although researchers have been describing how BGCs play a role in their biosynthesis, little is known regarding the patterns modelling BGC structure and distribution. Understanding these has an important effect in linking the vast amount of genomic information with the production of NPs, especially to orphan metabolites. This study performed a comparative genomics analysis of the underexplored genus *Rhodococcus*, using *Rhodococcus* sp. H-CA8f, one of our Chilean fjord-derived marine strains, as a model to perform an in-depth analysis of BGC distribution patterns. A BGC network revealed that the main category was encompassed by non-ribosomal peptide synthetase (NRPS) pathways, retrieving 44 gene cluster families (GCFs). Our results support a strong correlation with phylogeny, revealing clade-specific GCFs. Deeper understanding of a NRPS in *Rhodococcus* sp. H-CA8f, likely to be producing an orphan chloramphenicol-related compound, revealed that its BGC distribution is unique among its phylogenomic clade. This study contributes to unveiling unique BGCs, understanding their distribution among clades and the proposal of the involvement of the production of an orphan metabolite, never described before in *Rhodococcus*.

INTRODUCTION

Natural products (NPs) are commonly synthesized by complex specialized metabolic pathways, whose genes are physically grouped together in biosynthetic gene clusters (BGCs) [1, 2]. The advances in sequencing technologies and bioinformatics tools of the genomic era have played an essential role in the discovery of BGCs through genome mining [3, 4]. Thousands of sequences have become available, containing an even larger number of BGCs with overwhelming diversity, making a roadmap for their characterization necessary [5]. For instance, classifying BGCs into gene cluster families (GCFs) allows further prioritization based on the similarities shared between NP structural scaffolds [6, 7]. However, there are knowledge gaps regarding BGC linkage with NPs, leaving a vast abundance of orphan metabolites. Moreover, understanding BGC diversity, maintenance and distribution patterns, to ultimately decipher how these contribute to environmental adaptations, remains a challenge [8]. In this sense, comparative genomics allows a comprehensive exploration of BGCs based on high-throughput mining, providing the much-needed evidence to target certain BGCs, augmenting the knowledge to empower the genomic-guided bioprospection for NPs.

Actinomycetes have been in the spotlight as a renowned source of NPs, due to their ability to produce a myriad of structurally rich bioactive compounds [9–11]. Although focus has been historically placed on the soil-derived genus *Streptomyces* [12, 13], bioprospecting underexploited environments with strong selective pressures such as the ocean [14], along with the study of other genera – rather than *Streptomyces* [15] – has proven to be a successful strategy to enrich screening collections [16]. As actinomycete genome sequencing increases, a correlation between genome size and BGC abundance was evidenced for the genera *Actinomadura*, *Gordonia*, *Micromonospora*, *Nocardia*, *Nocardiopsis* and *Rhodococcus*, which were demonstrated to harbour an unexplored reservoir for unique BGCs [17]. Historically, the genus *Rhodococcus* has been largely explored for its extensive catabolic versatility, including bioremediation, biotransformation and biocatalysis applications [18–20]. In contrast, scarce knowledge is available regarding comparative BGC analysis, although *Rhodococcus* genomes currently add up to ~500 in the National Center for Biotechnology Information (NCBI) database.

Comparative studies of the genus *Rhodococcus* have been mainly focused on defining phylogeny, determining the core genome and to functionally analyse their catabolic potential and stress responses [21, 22]. Notably, a prior study contemplating 20 *Rhodococcus* genomes showed a mostly uncharacterized BGC repertoire, revealing certain strain-specific GCFs [23]. However, NP BGCs and the connection with the roles of their metabolites are mosly unknown [24]. A few studies connect non-ribosomal peptide synthetase (NRPS) pathways to their products, mostly with siderophores, such as heterobactin [25], rhequichelin [26] and rhodochelin [27], and also to a lipopeptide surfactant [28]. Other efforts have yielded humimycins, a synthetic NRPS-inspired NP [29], which no doubt validates the use of genome mining of BGCs. Still, little is known regarding NPs with antibiotic activity in the genus *Rhodococcus*. The main compounds known to date are the cyclic tetrapeptide rhodopeptins [30], the cyclic lasso peptides lariatins [31] and quinoline aurachins [32, 33], although none are reported from marine-derived *Rhodococcus*. Thus, there are still open questions about the main mechanisms underlying BGC distribution in rhodococci, and insights into their connection to specialized metabolites.

In this work, we aim to augment the knowledge of NRPS distribution across phylogeny, by performing an in-depth BGC comparative genomics analysis of the genus *Rhodococcus*. We developed a bioinformatics pipeline to address the selection of high-quality data, phylogenomics (CORASON) and sequence similarity BGC networking analyses to reveal patterns that model BGC diversity and structure. Moreover, the bioprospection of orphan metabolites was unveiled by using one of our bioactive strains as a proof of concept, the marine-derived *Rhodococcus* sp. H-CA8f [34]. Complementing high-throughput comparative genomics

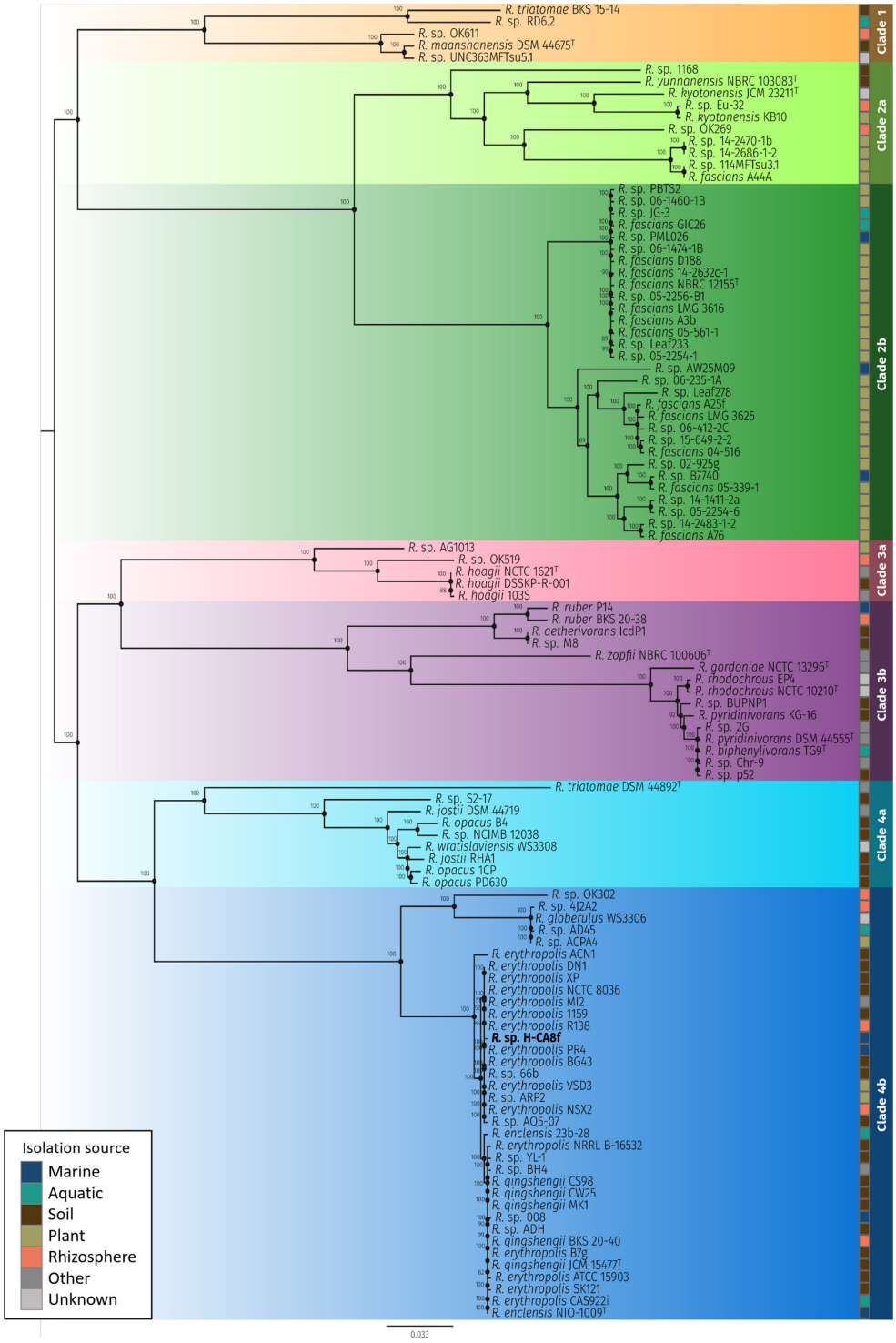

Fig. 1. Phylogenomic inference of *Rhodococcus* evolutionary relationships. Phylogeny of selected *Rhodococcus* genomes (filtered dataset of *n*=110, see Fig. S1) inferred using Orthofinder v2.2.7, identifying 613931 genes assigned to orthogroups (orthologous genes translated to protein sequences). FastTree was used for approximate ML tree inference. Bar, evolutionary distance, considering 0.033 substitutions per amino acid position. Clades (1 to 4) are represented in colours: clade 1, orange; subclade 2a, light green; subclade 2b, dark green; subclade 3a, light magenta; subclade 3b, dark magenta; subclade 4a, light blue; subclade 4b, dark blue. Coloured squares represent the isolation source of each strain, depicted as: blue, marine environment; emerald green, aquatic; brown, soil; olive green, plant; coral, rhizosphere; grey, other source; and light-grey, unknown. *Rhodococcus* sp. H-CA8f is depicted in black bold font, and is located within subclade 4b.

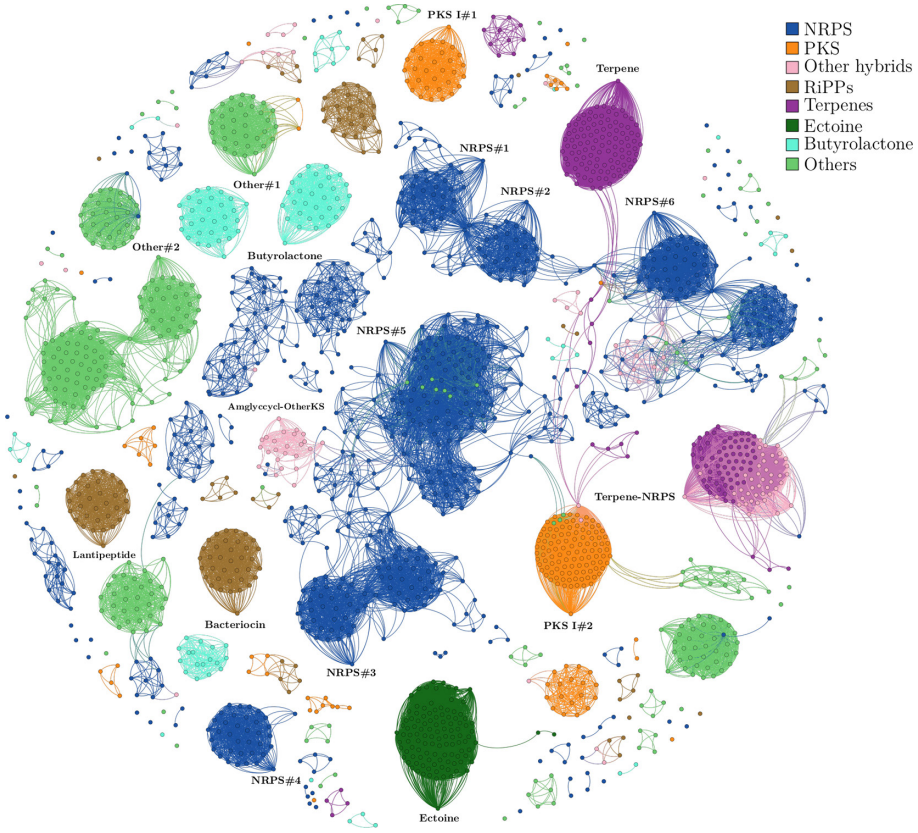

Fig. 2. *Rhodococcus* BGC networking. The distance network was constructed using BiG-SCAPE based on the *Rhodococcus* genomes filtered dataset, leading to a total of 1891 BGCs grouped by different categories. Each node represents one BGC, connected by edges when sharing a raw distance ≤0.6. *Rhodococcus* sp. H-CA8f BGCs, shown in black bold font, were apart from the main group of nodes but maintaining their connections. Colours represent BGC categories used in this study (slightly modified, see Table S3) depicted as follows: blue, NRPS; orange, polyketide synthase (PKS); pink, other hybrids; brown, (ribosomally synthesized and post-translationally modified peptide) RiPPs; purple, terpenes; dark green, ectoine; turquoise, butyrolactone; and green, other.

with phylogenomic and GCF network analysis sustains BGC correlations; thus, enhancing genome mining predictions. Our results ultimately bear potential connections through biosynthesis, evolution and ecological implications of the genus *Rhodococcus*.

METHODS

Comparative genomics pipeline

Rhodococcus genomes were downloaded from the NCBI RefSeq FTP server (306 entries as of 12th September 2018). Additionally, *Rhodococcus* sp. H-CA8f was selected from our culture collection, since it bears unique genomic features [34] and displayed antibacterial activity against both Gram-negative and Gram-positive target pathogens [35]. A comparative genomics pipeline was developed to comprehensively analyse high-throughput genome datasets. A schematic representation of the bioinformatic and biological criteria used to filter non-informative data is presented in Fig. S1.

Multiple data filtering criteria were performed in the pipeline on three levels: genomes (Fig. S1, blue box); BGCs (Fig. S1, red box); and NRPSs (Fig. S1, green box). Briefly, 'Green Yes boxes' indicate that data fulfil defined criteria and, thus, can be downstream analysed. 'Yellow No boxes' indicate that data conditionally fulfilled criteria and, hence, another filter was applied. 'Red No boxes' indicate that data did not fulfil the criteria and, thus, was subsequently discarded for further analyses. In the first level (Fig. S1, blue box), genomes with <200 contigs were selected, and analysed for completeness (>98%) and contamination (<5%), as implemented in CheckM v1.012 [36]. Although a rigorous completeness filter was applied and excessive fragmentation was avoided, some BGCs were predicted on contig edges, and those were still maintained for further analyses. Additionally, a manual bibliographical filter was performed to remove redundant genomes, checking for: (i) synonym strains – the same strains with different culture collection numbers; (ii) synonym genomes – with different entry names due to genome assembly improvements; or (iii) mutant strains – checked using culture collection database and bibliography [37]. ANIb (average nucleotide identity by BLAST alignments) between genomes was calculated using the pyANI package [38], to identify and discard highly similar

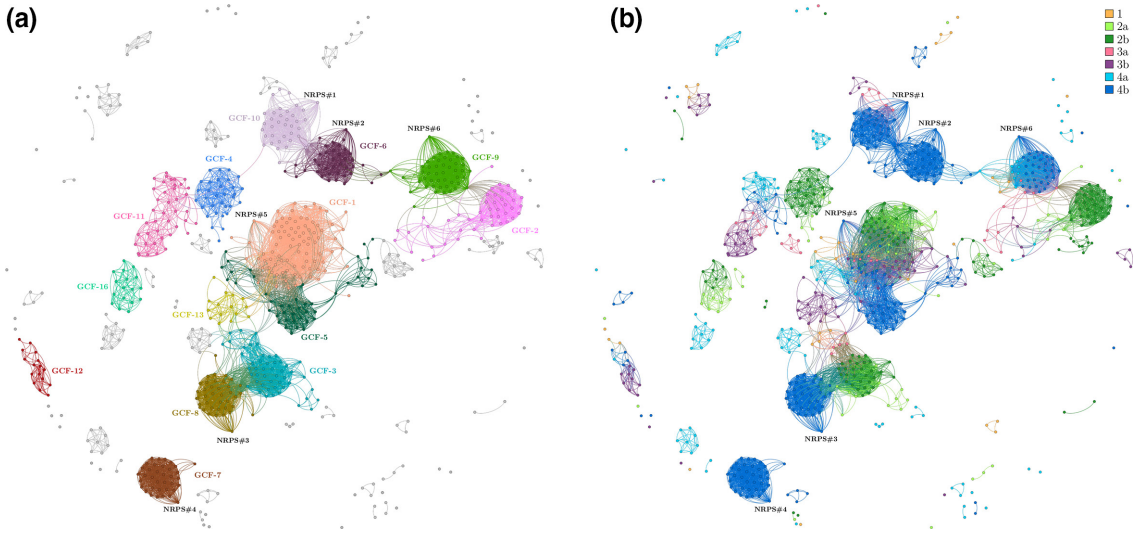

Fig. 3. NRPS BGC network. NRPS nodes (*n*=717) were retrieved from the full BGC network (see Fig. 2). *Rhodococcus* sp. H-CA8f BGCs are depicted as NRPS 1–6 (for details, see Table S4) with black labels. (a) Colours depict the GCFs' pattern of distribution, formed by ≥10 BGCs. The remaining GCFs are shown in grey. (b) Colours depict the phylogenomic distribution, correlated with the subclade colours from the phylogenomic tree from Fig. 1.

genomes (>98%). This threshold has been used for the dereplication of genomes and metagenome-assembly genomes in BGC biodiversity studies [39] and environmental microbial genomics [40]. Finally, if two or more *Rhodococcus* entries were redundant, only one genome was selected considering the following criteria: (i) fewer contigs; (ii) total assembly length in base pairs; and (iii) a recent year of entry publication at the NCBI database.

In the second level (Fig. S1, green box), selected genomes were submitted to standalone antiSMASH v4.1.0 [41] for BGC prediction, and BiG-SCAPE v.20181005 [42] was used to obtain cluster similarities. Similarly, redundant clusters were filtered using genomic ANIb (≥98%) and BiG-SCAPE raw distance (≤10^{-3}). A Python workflow was constructed to select BGCs that were composed of at least one biosynthetic plus one non-biosynthetic gene (https://github.com/rvalenciaaz/rhodococcus-bgc). Finally, at the third level (Fig. S1, red box), NRPS BGCs were manually corroborated for presenting two or more adenylation domains by using antiSMASH v4.1.0 [41].

Phylogenomic analysis

A phylogenomic tree (Fig. 1) was inferred with Orthofinder v2.2.7 [43] using the selected *Rhodococcus* genomes (Fig. S1). DIAMOND aligner was used for orthogroup retrieval [44], MAFT [45] for multiple sequence alignment and FastTree [46] for approximate maximum-likelihood (ML) tree inference. Additionally, a phylogenomic method involving multilocus sequence analysis (MLSA) based on 100 highly conserved single copy genes using Automated Multi-locus Species Tree (AutoMLST) (http://automlst.ziemertlab.com/) was performed for *Rhodococcus* strains comprising subclade 4b, considering AutoMLST strain upload limitations [47]. For phylogeny inference, *de novo mode* was used with the option of concatenated alignment under the following configuration parameters: (i) strains from subclade 4b were manually selected from the AutoMLST in-house database, with the addition of three strains: *Rhodococcus* sp. NACPA4, *Rhodococcus* sp. H-CA8f and *Rhodococcus* sp. AQ5-07; (ii) IQ-TREE Ultrafast Bootstrap analysis was performed with 1000 replicates [48]; (iii) ModelFinder was used to find the best algorithm for tree reconstruction; (iv) inconsistent MLST genes were filtered (i.e. genes with greatest topology differences), and (v) fast alignment mode was activated. The final tree was modified with Dendroscope 3.6.2 [49] and MEGAX [50] (Fig. S2). Table S2 lists the 100 conserved single-copy genes from which 88 were selected based on neutral *dN*/*dS* values, applying software default parameters [47]. To complement tree topology, a Bayesian multilocus phylogeny (BY) was inferred (Fig. S3) using MAFFT [45] and the concatenated nucleotide alignment of the genes *gyrB*, *rpoB*, *rpoC*, *secY* and *recA* [21]. Tree inference was accomplished with MrBayes v.3.2.7 [51, 52] using one million generations and two runs, while PartitionFinder2 [53] was used for fitting substitutions models. Orthofinder and Bayesian trees were compared using Robinson–Foulds [54] and SPR metrics in R, using the phangorn package [55]. The quartet distance [56], which considers tree similarity using small taxa groups, between the ML and BY trees was calculated using the TreeCmp webserver [57]. The *prunes trees* option was used to compare common taxa. The metric was normalized with respect to the mean value for random trees generated with the Yule and uniform model, respectively. To investigate putative ecological relationships, the isolation source of each strain was obtained from the NCBI and Joint Genome Institute (JGI) online servers and depicted in both trees with a colour legend next to each strain.

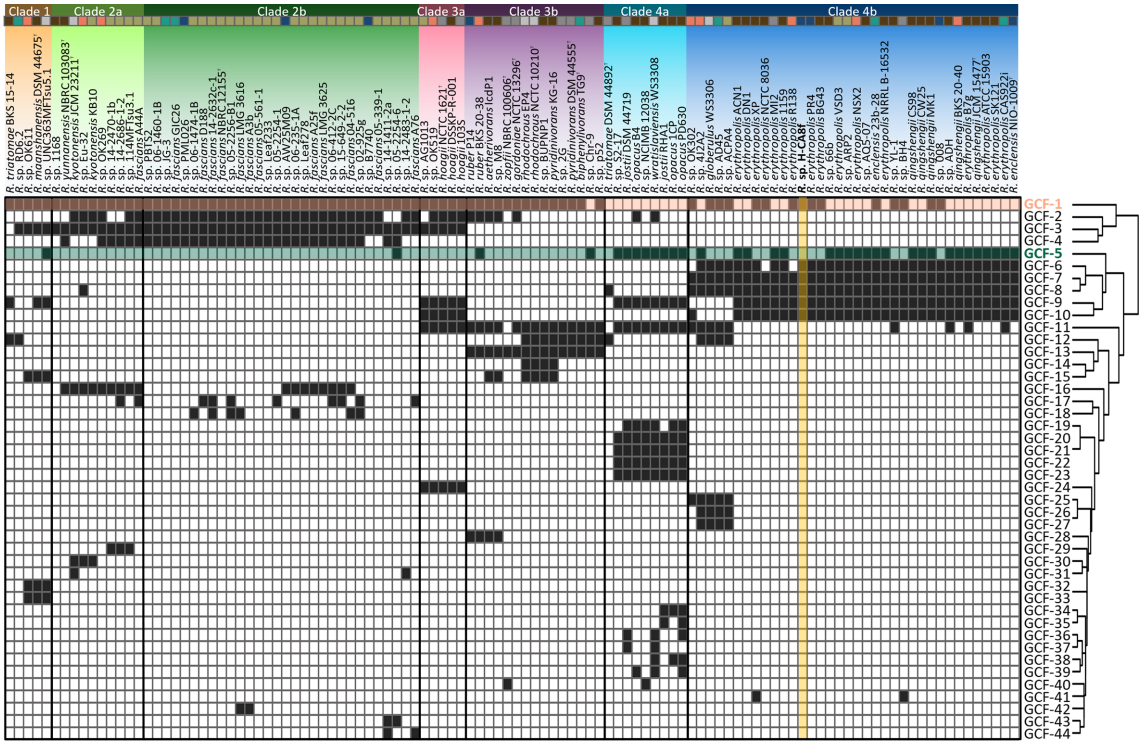

Fig. 4. Hierarchical clustering of NRPS GCFs. NRPS GCFs (*n*=44, right side) considering the presence/absence barcoding depicted in *Rhodococcus* genomes represented according to phylogenomic inference and including isolation source, according to Fig. 1. Presence of a GCF in a *Rhodococcus* genome is represented by a filled square, while its absence is represented by an empty square. Related GCF-1 (light orange) and GCF-5 (green) are highlighted for better visualization. *Rhodococcus* sp. H-CA8f is shown in bold font within subclade 4b, and its GCF representatives are highlighted in yellow.

BGC networking and GCF analysis of NRPS

Selected *Rhodococcus* genomes (see Table S2) were uploaded to the antiSMASH v4.1.0 tool [41] to identify BGCs (Table S3) and into BiG-SCAPE v.20181005 [42] to calculate raw distances between clusters, by which a BGC network was constructed (Fig. 2). For this analysis, only the sequence of *Rhodococcus* sp. H-CA8f's chromosome (GenBank accession no. CP023720) was used [34], and detailed genome mining is presented in Table S4. For network construction, several raw distance cut-offs were tested, ranging from 0 to 1 (0 being the most restrictive scenario) with a step of 0.1, where 0.6 was finally selected, aiming for a balanced connectivity of the overall network. Final graph layout was obtained using a combination of Fruchterman–Reingold [58] and Yifan Hu [59] algorithms, adjusting balance between node sparsity and agglomeration. Visualization of the networks were performed in Gephi v0.9.2 [60]. A reduced classification of BGC categories is presented, based on the following modifications: (i) 'PKS I' was grouped together with 'PKS'; (ii) 'Other hybrids' was created to contain any hybrid combination; (iii) 'ectoine' and 'butyrolactone' were dropped from 'Others' and annotated as individual separated categories. Furthermore, a NRPS network was generated as a subgraph of the whole BGC network (Fig. 3), coloured by GCFs (Fig. 3a) and the phylogenomic clades (Fig. 3b). To group NRPSs into GCFs, the Louvain algorithm for community detection was applied with a default resolution parameter value of 1 [61, 62]. Unconnected nodes were excluded from the GCF definition. Manual inspection of selected GCFs was performed by uploading into antiSMASH v.4.1.0 all its BGCs.

Phylogenomic-dependent patterns of NRPS GCFs

Presence/absence matrix patterns of each NRPS GCF were determined with a binary set in R, using the pheatmap v1.0.10 package [63]. Filled squares denote the presence of a certain NRPS GCF in a *Rhodococcus* genome (Fig. 4). A hierarchical clustering of the presence/absence map of the NRPS GCFs is shown as a dendrogram alongside the vertical axis. The horizontal axis considers the clades from the phylogenomic tree, maintaining the respective clade colour as depicted in Fig. 1. GCF-1/GCF-5 are highlighted in their respective colours for better visualization. NRPS GCF rarefaction curves (Fig. S4) were generated using the GCF presence/absence matrix plotted against the surveyed genomes. Richness calculations were performed using the iNEXT package in R [64]. NRPS GCF richness was considered for the diversity index, and default bootstrap iterations (*n*=50) with 95% confidence intervals were used in the run. Interpolation and extrapolation data were inferred by iNEXT. GCF presence/absence pattern similarity within and between clades was assessed

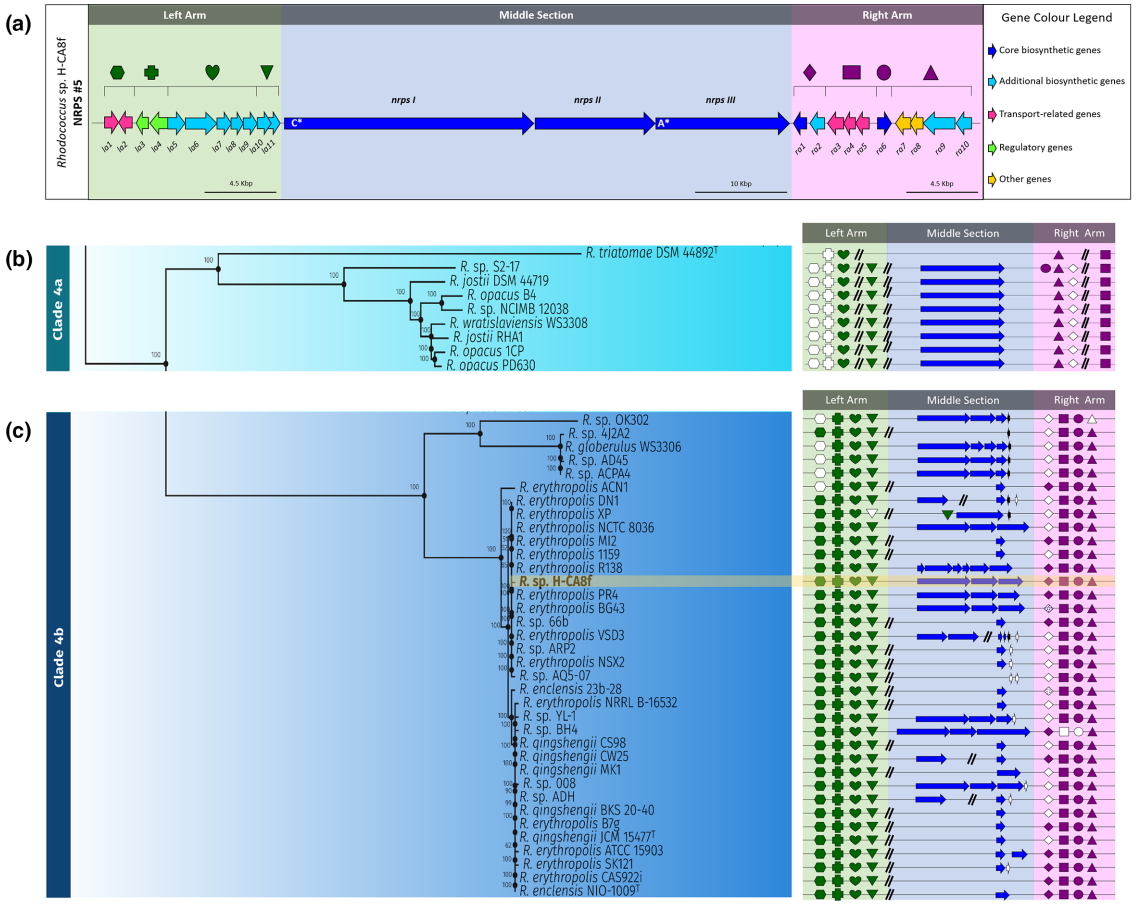

Fig. 5. Gene distribution patterns of GCF-1/GCF-5. (a) Genetic representation of NRPS #5 BGC belonging to GCF-1 from *Rhodococcus* sp. H-CA8f. NRPS #5 is grouped into three regions: left arm (green); middle section (blue); and right arm (purple). Genes grouped into blocks are represented with the following symbols: (i) left arm region – hexagon, *la1–la2*; cross, *la3–la4*; heart, *la5–la9*; and inverted triangle, *la10–la11;* (ii) right arm region – diamond, *ra1–ra2*; rectangle; *ra3–ra5*; circle, *ra6*; and triangle, *ra7–ra10*. For detailed predicted functions of genes, see Table 1. In each section, genes are drawn according to the size bar. In the middle section, letters within genes represent special domains: C*, starter condensation domain in *nrps I*; A*, non-classical adenylation domain in *nrps III*. (b) and (c) Genomic context comparison using CORASON of the GCF-1/GCF-5 BGC distribution from phylogenomic subclades 4a and 4b, respectively. Every gene comprising NRPS #5 of strain H-CA8f (shown in black bold font within subclade 4b and highlighted in yellow) was used as a query. Gene orientation and genetic organization are depicted similarly to NRPS #5 of *Rhodococcus* sp. H-CA8f, unless otherwise indicated. Filled symbols represent the presence of all genes constituting a block, whereas empty symbols indicate that at least one gene of that block is missing. Symbol size is not representative of gene size, and intergenic spaces are not to scale. Parallel lines indicate that genes are present elsewhere in the genome. Other genes – different from those previously mentioned – are represented as follows: black arrows, tRNAs; white arrows, hypothetical proteins.

with a non-parametric multivariate statistical test [65]. Since GCF presence/absence is a binary trait, the Jaccard distance was employed to generate a distance matrix. Then, we performed a permutational multivariate analysis of variance (PERMANOVA) [66] for 999 permutations, considering the clades as groups, in the vegan package of R (https://CRAN.R-project.org/package=vegan).

Evolutionary relationships of GCF-1/GCF-5 NRPS

GCF-1/GCF-5 genetic context across clade four was evaluated with CORe Analysis of Syntenic Orthologs to prioritize NP-biosynthetic gene clusters (CORASON) [42] (Fig. 5), using every gene from H-CA8f's NRPS #5 as the query (Fig. 5a). The level of gene conservation was analysed by three criteria: (i) gene co-occurrence pattern across the phylogenomic clade, (ii) putative function based on BLASTP annotation and (iii) genetic organization (i.e. whether genes are in the same position and codified in the same direction). According to this, genes were grouped into blocks and represented with symbols when presenting a co-occurrence pattern. If at least one gene from the block is missing, the symbol is depicted empty. Otherwise, filled symbols represent the presence of the same genes shown for NRPS #5 (Fig. 5a). Final construction was manually edited to maintain schematic representation according to the phylogenomic subclades 4a (Fig. 5b) and 4b (Fig. 5c). For NRPS modularity analysis (Fig. S5), domain

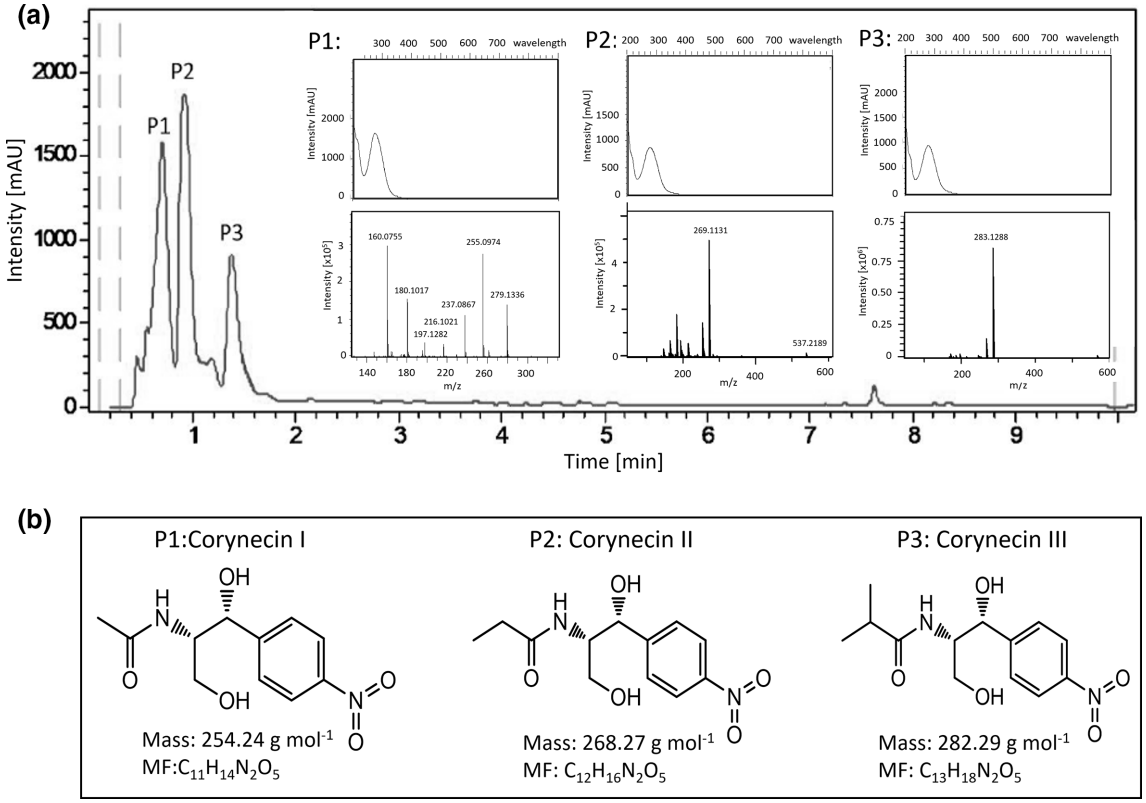

Fig. 6. Chromatographic profile based on LC-HRMS analysis of *Rhodococcus* sp. H-CA8f bioactive crude extract. (a) Dereplication of *Rhodococcus* sp. H-CA8f's ISP2-ASW derived crude extract depicting three major peaks: P1, P2 and P3. Inset images show putative compound identification based on UV spectra (top) and mass spectra (bottom). (b) Chemical structures for the identified corynecin I (P1), II (P2) and III (P3), along with their respective masses and molecular formulae.

prediction was conducted using antiSMASH v5.2.0 [67], which incorporates NRPSpredictor3 [68], latent semantic indexing (LSI) based A-domain function predictor [69] and NRPSsp [70]. Additionally, Prediction Informatics for Secondary Metabolomes (PRISM 3) [71] was used for the detection of non-canonical domains.

NP extraction and assessment of antibacterial activity

Five culture media were used to test varying culture conditions for *Rhodococcus* sp. H-CA8f: ISP1 (5 g tryptone l^{-1}, 3 g yeast extract l^{-1}), ISP2 (10 g malt extract l^{-1}, 4 g yeast extract l^{-1}, 4 g glucose l^{-1}), R5A [72], SM19 [73] and a modified-SG medium, with soytone peptone –instead of soytone – and with no added glucose. ISP1 and ISP2 media were prepared with artificial sea water (ASW) (i.e. ISP1-ASW and ISP2-ASW). Fermentations were performed in a 250 ml Erlenmeyer flask containing 50 ml culture media, in a rotary shaker at 200 r.p.m. at 30 °C for 10 days. Afterwards, cells were separated from the supernatant by centrifugation at 5000 r.p.m. for 10 min. Supernatants were extracted twice in a decantation funnel using ethylacetate (EtOAc) in a 1:1 (v/v) ratio. The recovered organic phase was almost completely evaporated with a speed vacuum. Crude extracts were dissolved in methanol:water (HPLC-grade MeOH:MQ-H_2O, 1:1) to a final concentration of 5 mg ml^{-1}, and subsequently stored at −20 °C until further use.

The antibacterial activity of crude extracts was assessed as previously described [74], with minor modifications. In this study, seven model bacteria were used to test susceptibility: *Staphylococcus aureus* NBRC 227 100910^T (STAU), *Listeria monocytogenes* 07PF0776 (LIMO), *Salmonella enterica* subsp. *enterica* LT2^T 228 (SAEN), *Escherichia coli* FAP1 (ESCO), *Pseudomonas aeruginosa* DSM 50071^T (PSAU), *Clavibacter michiganensis* subsp. *michiganensis* VL493 (CLMI), a phytopathogenic strain isolated from an infected tomato plant obtained from Limache, Chile [75], and *Micrococcus luteus* H-CD9b (MILU), an actinomycete previously isolated by our group from the Northern Chilean Patagonia [35]. Model bacteria were grown overnight in a 5 ml LB culture at either 37 °C (PSAU, SAEN, ESCO and STAU) or 30 °C (MILU, LIMO and CLMI). The inoculum was adjusted to a final OD_{600} of 0.2. Subsequently, model bacteria were streaked as a fine lawn on LB agar plates and 10 µl extract was placed on top. Inhibition zones were observed after overnight incubation. Results are shown in Table S5. Extractions of the media were also tested for antibacterial activity, and methanol:water was

used as negative control. Extracts with antibacterial activity were selected for further chemical dereplication.

Chemical dereplication of NPs

Chemical dereplication was accomplished using liquid chromatography-high resolution MS (LC-HRMS) performed by Fundación MEDINA (Fig. 6). Experiments were carried out using an HPLC 1200 Rapid Resolution system (Agilent) coupled to a high-resolution mass spectrometer, MaXiz (Bruker). For separation, a SB-C8 Zorbax column (2.1×30 mm, 3.5 µm) was used with a flow rate of 0.3 ml min^{-1}. The mobile phase consisted of solvent A, H_2O:acetonitrile (AcN) (90:10), and solvent B, H_2O:AcN (10:90), both with ammonium formate 13 mM and 0.01% trifluoroacetic acid (TFA). Gradient composition started with a linear decrease of solvent A from 90–0%, and a linear increase of solvent B from 10–100%, in 8 min. Then, the following 2 min were as for the initial maintaining conditions with 90% of solvent A and 10% of solvent B. MS was operated in positive mode (ESI+) with a spray voltage at 4kV, 11l N_2 min^{-1} at 200 °C capillary temperature and 280 KPa of nebulizer pressure. Absorbance was measured at 210 nm wavelength. Data analysis for NP identification was performed concerning: (i) retention time; (ii) UV absorbance spectrum; and (iii) accurate masses, obtained for every peak from the crude extract chromatogram profile (Fig. 6a). These criteria were used for comparison with MEDINA's in-house database along with the Dictionary of Natural Products (DNP) database of Chapman and Hall, where molecules were searched for their identification (Fig. 6b).

RESULTS AND DISCUSSION

Comparative genomics pipeline

At the time of writing, up to ~300 *Rhodococcus* genomes were available from the NCBI database. This number is rising rapidly, currently being around 500 assemblies, although most of them stand as draft versions. Highly fragmented genomes can represent a potential drawback for genome mining, especially when BGC comparative inferences are addressed. The usually long, repetitive organizations of the BGC assembly lines can end up being split on multiple contigs [76, 77]. A recent study showed that 25% of publicly available genomes were fragmented in more than 200 contigs [76]. Thus, application of thresholds on contig numbers for comparative genome studies have been discussed extensively [78–81]. In this study, to create a rigorous high-throughput comparative genomics pipeline, several informatics filters with subsequential biological criteria were applied, including: (i) genome quality, (ii) phylogenetic relatedness, (iii) average nucleotide identity (ANI) and (iv) BGC dereplication (Fig. S1). Filtering criteria were applied on three levels – genomes (blue box, Fig. S1), BGCs (green box, Fig. S1) and NRPS (red box, Fig. S1) – to select high-quality data for robust downstream analysis. However, due to the lack of complete *Rhodococcus* genomes, several retrieved BGCs were on contig edges. Nevertheless, these BGCs were manually inspected and carefully considered. The outcome was that our pipeline led to the selection of 110 rhodococci genomes (<200 contigs with CheckM completeness >98% and contamination <5%; metrics detailed in Table S1) harbouring 1891 BGCs, from which we specially focused on the 717 NRPS BGCs (Fig. S1).

Phylogenomic analysis

A phylogenomic inference was carried out using orthologue analysis and a ML approach, supporting *Rhodococcus* genus evolutionary relationships falling into four major clades (1 to 4) and respective subclades (a to b) (Fig. 1). *Rhodococcus* species' distribution among clades is scattered, with some representatives placed in two distinct clades (e.g. *Rhodococcus triatomae* BKS 15–14 in clade 1 versus *R. triatomae* DSM 44892^T in subclade 4a) (Fig. 1). However, this phylogenomic tree is consistent with the species distribution found in other systematics *Rhodococcus* studies [82–84], where the position of *R. triatomae* strains is also unclear and they are often classified as part of a new clade [82]. Regarding the isolation sources, strains exhibit widespread variation across the four phylogenomic clades. However, correlation could be observed at a species level, where niche partitioning is reflected (i.e. all *Rhodococcus fascians* are plant-retrieved, except for one; all *Rhodococcus opacus* are soil-derived) (Fig. 1). The phylogenomic tree shows that clade 1 (orange) has the least number of representatives, clade 2 (green) groups mostly with plant-associated strains (75%) related to *R. fascians*, and clade 3 (magenta) is composed of diverse representatives mostly known by their pathogenicity (subclade 3a) and their ability to degrade a wide range of aromatic and recalcitrant compounds (subclade 3b) (Fig. 1). Clade 4 (blue) comprises the greatest number of representatives, grouping model strains known for their capabilities to degrade a variety of aromatic compounds (subclade 4a); and subclade 4b where *Rhodococcus erythropolis*, *Rhodococcus qingshengii* and *Rhodococcus enclensis* species are grouped, including our marine *Rhodococcus* sp. H-CA8f (Fig. 1). Strains from subclade 4b are retrieved from several isolation sources, although most are soil-derived (50%). Notably, strain H-CA8f groups with another marine strain, *R. erythropolis* PR4, isolated from Japan [85]. An additional phylogenetic analysis was performed for subclade 4b, using 88 highly conserved housekeeping genes (Table S2) through the automated multilocus species tree (AutoMLST) tool [47], where an improved separation between the closely related strains to the *R. erythropolis*, *R. qingshengii* and *R. enclensis* species is observed (Fig. S2).

Furthermore, ML phylogenomic inference was further compared by BY (Fig. S3). Changes in topology make it difficult to obtain taxonomic resolution, but both trees suggest that subclade 4b harbours closely related strains that probably share a high percentage of the core genome. Although the ancestral placement of subclades in BY differs from the ML tree, which can be especially appreciated for clade 2 that places between subclade 4a and 4b, this does not alter the overall subclade topology and similar branching patterns can be validated for all subclades (accounted by tree comparison metrics; normalized Robinson–Foulds distance 0.51; SPR

metric 20). In addition, we calculated a distance metric based on small groups of taxa, named quartet distance [56], where 0 is assigned to identical trees. The resulting normalized scores with respect to the mean value for random trees generated with the Yule and uniform model are 0.1917 and 0.1916, respectively. This indicates that intra-clade topology between the two trees is similar. This is further observed with model strains such as *Rhodococcus jostii* RHA1 [86], known to degrade a variety of aromatic compounds including polychlorinated biphenyls, and *R. opacus* 1CP, a well-known chlorophenol degrader [87], where their placement is maintained within the same subclade 4a. Despite the incongruence found with *R. triatomae* placement, also previously observed [82], our overall results support the ML-based tree clade definition to be the basis for subsequent analyses.

BGC networking and GCF analysis of NRPS

Previously, genome mining of *Rhodococcus* sp. H-CA8f using antiSMASH v4.0.2 retrieved 17 BGCs [34]. In this study, we updated genome mining predictions using antiSMASH v4.1.0 and, remarkably, most *Rhodococcus* sp. H-CA8f putative pathways (65%) harbour low similarity scores (≤ 50%) to the Minimum Information about a Biosynthetic Gene Cluster (MIBiG) repository, while another four are completely unknown. From the total of 17 BGCs, only 2 presented high similarity (≥75%) to MIBiG hits: NRPS 4 with 100% similarity to heterobactin, and the ectoine BGC (Table S4). This result supports that our marine strain H-CA8f harbours a mostly underexplored repertoire of BGCs, with a biosynthetic potential that remains largely unknown.

To obtain an overview about *Rhodococcus* BGC diversity, a BiG-SCAPE sequence similarity BGC network was constructed with a total of 1891 BGCs (Table S3). *Rhodococcus* sp. H-CA8f are highlighted in the network in black bold labels for better visualization (Fig. 2). Nodes are coloured by BGC categories and connected when sharing a BiG-SCAPE raw distance cut-off of ≤0.6 (Fig. 2).

Network similarities showed that the rhodococci repertoire of BGCs form defined connecting groups mostly represented by nodes within a specific category, and showing a balloon-shape structure (Fig. 2). From these, ectoine (dark green) and butyrolactone (calypso) are the most conserved categories, being present in 100 and 78% of *Rhodococcus* genomes, respectively (Fig. 2), which is in line with previous studies [23, 88, 89]. BGC conservation seems to be correlated with an essential function of their respective metabolites; while ectoines are protective osmolytes that help bacteria thrive under osmotic stress [90], butyrolactones are hormone-like signalling molecules synthesized to coordinate communication [91]. These BGCs stand as an example of how a highly conserved clustering, represented as a close pattern of distribution in the network (Fig. 2), could indicate a correlation of BGC-associated functional traits with ecological importance for *Rhodococcus* lifestyle.

Conversely, the NRPS category presented a dispersed pattern of distribution. It is the most prevalent category of the network comprising 38% of the total BGCs, followed by 10.5% PKS and 9.5% terpenoid BGCs (Fig. 2). This result shows that the genus *Rhodococcus* is rich in these pathways as previously observed [23], similar to what is reported for their phylogenetically close genus *Nocardia* [89]. To visualize correlations within this category, a zoom-in into the NRPSs network (n=717) was constructed, coloured by GCFs (Fig. 3a) and phylogenomic inferences (Fig. 3b). Using connectivity and edge weights data reported by BiG-SCAPE, GCF relatedness were tested against the MIBiG repository [92]. Our data showed that only seven GCFs presented to some extent similarity to a MIBiG hit. However, scores were too low to assign the corresponding product pathway name and, thus, they were number-labelled from GCF-1 to GCF-44 (Fig. 3a). This result suggests that NRPS pathways represent a big family of BGCs widely distributed over several *Rhodococcus* genomes. However, they are poorly characterized, as all GCFs remained mostly unknown.

Furthermore, NRPSs were coloured according to phylogenomic clades, and a clear correlation between GCF distribution and phylogeny was evidenced, presenting a pronounced clade-specific distribution pattern (Fig. 3b). However, this is not observed for its sister genus *Nocardia*, where slight correlations between BGCs and phylogenetic clades limited to only one NRPS GCF have been described [89]. In our study, these phylogenomic-dependent patterns correlate with specific GCFs, mainly for GCF-5, GCF-6, GCF-7, GCF-8, GCF-9 and GCF-10 (Fig. 3a), which are essentially distributed in subclade 4b (Fig. 3b). For instance, GCF-7 nodes (n=36) are found exclusively in subclade 4b and present a closed balloon-shape structure with no connections to other GCFs (Fig. 3), suggesting a correlation with a relevant functional trait. To test this hypothesis, all nodes were submitted to antiSMASH and a high similarity (>63%) to the heterobactin MIBiG BGC was revealed. Heterobactins are siderophores, low-molecular-mass organic compounds that scavenge iron with high affinity and specificity [93]. These results demonstrate how network structures can provide insights into highly conserved BGCs, and they seem to be related with essential roles in nature of their derived NPs. Similarly, this network pattern was observed for other NRPS siderophore pathways in a previous study, such as rhodochelin/rhequichelin and rhequibactin GCFs [23]. If these BGCs are maintained throughout the same evolutionary lineage, then a strong phylogenomic signal of that GCF can be assumed. As observed previously in a study limited to 20 genomes, most predicted GCFs in *Rhodococcus* are clade-specific and lack sequence similarity with MIBiG hits [23]. Our analysis, covering a much larger dataset, supports these observations, broadening the information regarding BGC comparative genomics in the genus *Rhodococcus*. Their vast uncharacterized repertoire of GCFs provides interesting opportunities for exploiting NPs, while their phylogenomic-specific patterns unveil unprecedented insights into the rhodococcal distribution of NRPSs.

Table 1. Genome-based prediction of NRPS #5 BGC of *Rhodococcus* sp. H-CA8f

Gene ID	Gene name[a]	Geometric symbol[a]	Location[a]	Predicted function[b]
CPI83_20045	*la1*	hexagon	Left arm	C_4-Dicarboxylate Transporter
CPI83_20040	*la2*			Arginine/Ornithine Antiporter
CPI83_20035	*la3*	cross		GTP Pyrophosphokinase
CPI83_20030	*la4*			Transcriptional Regulator (TetR/AcrR Family)
CPI83_20025	*la5*	heart		CoA Carboxylase (subunit *β*)
CPI83_20020	*la6*			CoA Carboxylase (subunit α)
CPI83_20015	*la7*			Acyl-CoA Dehydrogenase
CPI83_20010	*la8*			Dehydratase
CPI83_20005	*la9*			CoA Ester Lyase
CPI83_20000	*la10*	inverted triangle		CoA Transferase (subunit α)
CPI83_19995	*la11*			CoA Transferase (subunit *β*)
CPI83_19990	*nrps I*	–	Middle section	NRPS (C*-starter domain)
CPI83_19985	*nrps II*	–		NRPS
CPI83_19980	*nrps III*	–		NRPS (A*-domain)
CPI83_19975	*ra1*	diamond	Right arm	Lipase (α/*β* fold hydrolase)
CPI83_19970	*ra2*			Esterase
CPI83_19965	*ra3*			Peptide Antibiotic ABC Transporter
CPI83_19960	*ra4*	rectangle		
CPI83_19955	*ra5*			
CPI83_19950	*ra6*	circle		Glycosyltransferase
CPI83_19945	*ra7*			*p*-Aminobenzoate *N*-Oxygenase (AurF)
CPI83_19940	*ra8*	triangle		Prephenate dehydrogenase (TyrA)
CPI83_19935	*ra9*			Aminodeoxy Chorismate Synthase
CPI83_19930	*ra10*			Chorismate Mutase

[a]Based on Fig. 4(a).
[b]Based on BLASTP results.

Phylogenomic-dependent patterns of NRPS GCFs

To unveil BGC distribution patterns along the genus *Rhodococcus*, a NRPSs hierarchical clustering was performed (Fig. 4) integrating the GCF network (Fig. 3a) with phylogenomic correlations (Fig. 3b). NRPS hierarchical clustering displayed that GCFs are exclusively distributed along certain clades (Fig. 4). An interesting correlation between subclade 2b and 4b is evidenced, where GCF-1 to GCF-4 are overrepresented in the former, while GCF-5 to GCF-10 are in the latter (Fig. 4). This is further supported by BY, where these two clades appear as closely related (Fig. S3). Additionally, correlations between subclade 4a and GCF-19 to GCF-23, or in subclade 3b which is enriched in GCF-11 to GCF-15, can be observed (Fig. 4). These correlations further support the previously observed *Rhodococcus* clade-specific GCF distribution and suggest that different species bear a specific repertoire of BGCs possibly associated with an essential function, most likely to be maintained across lineages.

Rarefaction curves from diverse biomes (i.e. plant-, soil- and water-associated) were performed to investigate NRPS GCF richness (Fig. S4). Despite the number of strains per biome being imbalanced, a trend was observed. Most rhodococci have been isolated from terrestrial sources (i.e. soil- and plant-derived, n=34), and the beginning of a plateau can be observed bordering 30–35 GCFs, meaning those niches are starting to saturate and the maximum diversity has almost been achieved (Fig. S4). In contrast, fewer strains have been cultured from water-associated environments, represented by a steeper slope (i.e. n=7 and 8, for aquatic and marine sources, respectively). This indicates that it is likely that more NRPS BGCs will be discovered if more isolates from these habitats

are sampled (Fig. S4). These results encourage the sampling of underexplored habitats, rather than soil, to retrieve currently unknown NRPS pathways.

Overall, our results demonstrate that in the genus *Rhodococcus*, phylogeny correlates with NRPS GCFs (PERMANOVA, *P* value 0.001). Vertical gene transfer might be the most important driver for phylogenomically dependent BGCs, having arisen from the same ancestral origin [94, 95]. Evidence of clade-specific BGC distribution has been previously reported in other actinomycete genera. For instance, no correlation between BGCs and geographical distribution was found in *Amycolaptosis*, although it was indeed observed between species' phylogeny [96]. In *Streptomyces*, an in-depth analysis revealed that closely related species have genetic and metabolic overlap, although environmental selective pressures shape metabolic traits giving rise to unique evolutionary histories [97]. Gene acquisition by lateral gene transfer is surprisingly rare and, instead, *Streptomyces* tend to accumulate point mutations as drivers of evolution [97]. Recently, a sequence similarity network with the *Rhodococcus* sister genus *Nocardia* showed minor correlation between phylogenetic clades and BGC distribution [89]. On the contrary, ecology could especially influence biosynthetic potential in close interactions such as symbiotic relationships, as reported for insect-associated *Streptomyces* [98] and *Pseudonocardia* [99], or in *Salinispora*, a marine genus that displays strong environmental adaptations, in which closely related species have faced ecological differentiation as drivers for speciation [100, 101]. To our knowledge, such a strong phylogenomic-dependent GCF pattern of BGC distribution as demonstrated in our study has not been previously reported. Efforts in elucidating BGC diversity had been described through sparse scenarios [6, 102–104], while evolutionary interconnections between BGCs and the forces shaping their specialised metabolites are just beginning to be recognized [105]. The postgenomic era has revealed some hypotheses, such as the dynamic chemical matrix evolutionary hypothesis, which reconciles chemical, functional and genetic data [106]. Overall, these studies reflect the importance of broadening the analysis of BGC dynamics, taking into consideration phylogeny, isolation source and evolutionary history of the studied genus, which was our aim for *Rhodococcus* after our comparative genomics analysis.

Evolutionary relationships of GCF-1/GCF-5 NRPS

Our previous results showed that GCF-1 (light orange) and GCF-5 (green) are highly interconnected (Fig. 3a) and widely distributed along the different phylogenomic clades (Fig. 3b). Barcode analysis showed a complementary dynamic of these GCFs, meaning that when one is present, the other is absent, this becoming more evident in subclade 4b (Fig. 4). A manual inspection of the conforming BGCs showed that for GCF-5, the majority (80%) were present on contig edges, only those from subclade 4a being complete. If more complete rhodococci genomes were available, the formation of an entire GCF could be a possibility, as they share similar features. All these observations prompted us to perform a deeper analysis on GCF-1 and GCF-5 dynamics, for which BGCs were submitted to antiSMASH and manually curated (n=134). For GCF-1, 42% of BGCs presented no similarity hit at all with MIBiG repository BGCs, whereas the rest presented a gene similarity score not higher than 23%, attributed to accessory genes. Within these, the chloramphenicol-BGC (MIBiG: BGC0000893) was noticed with up to 17% of gene similarity, and present in subclade 2a and 4b strains (data not shown). This similarity is explained by four genes encoding chorismate mutase, aminodeoxy chorismate synthase, deoxyprephenate dehydrogenase and *p*-aminobenzoate synthase, respectively. However, the GCF-1 BGCs differ substantially with the biosynthetic genes of the chloramphenicol pathway: while the latter harbours only one monomodular *nrps* gene [107], BGCs of subclade 4b present a variety of mono- or multi-modular *nrps* genes, suggesting their involvement in the production of a more diverse molecular family. Furthermore, similarities to MIBiG BGCs in GCF-5 were exceptionally low, presenting a ≤20% score for the 78% of the BGCs. These results support the importance of pursuing high-quality assemblies for genome-based BGCs prediction, even more when further comparative genome mining conclusions are desired. As mentioned before, they represent the basis for the appropriate analysis, and major effects in the outcome can be obtained when using fragmented assemblies, as BGCs are likely to be broken up into many contigs [77].

To better understand GCF-1/GCF-5 interconnectedness, we used *Rhodococcus* sp. H-CA8f's NRPS #5 from GCF-1 as a model to comprehend patterns of gene distribution. Exploration of this BGC along phylogeny was achieved by applying CORASON tool [42]. Every gene of NRPS #5 was used as a query and compared within strains from subclade 4a and 4b (Fig. 5). *Rhodococcus* sp. H-CA8f NRPS #5 is composed of 24 genes segmented into three regions: left arm (11 genes, green); middle section (3 genes, blue); and right arm (10 genes, purple) (Fig. 5a). Predicted gene functions are summarized in Table 1, along with their respective grouping into blocks represented with geometric symbols accounting for their co-occurrence pattern of distribution (Table 1, Fig. 5a).

CORASON results revealed the BGC genetic distribution pattern observed for both left and right arms along subclade 4a (Fig. 5b) and 4b (Fig. 5c). In general, most genes are found within clade 4, although differences between the co-occurrence pattern of gene blocks can be appreciated. Both left and right arms present four geometric symbols each, whose genes are grouped as follows: *la1–la2* (hexagon); *la3–la4* (cross); *la5–la9* (heart); and *la10–la11* (inverted triangle); and *ra1–ra2* (diamond); *ra3–ra5* (rectangle); *ra6* (circle); and *ra7–ra10* (triangle), respectively (Fig. 5a, Table 1). Genetic organization unveiled that across subclade 4a, genes from the left arm are usually observed in other genome locations (i.e. depicted with parallel lines), while some genes from the right arm (triangle, *ra7–ra10*) are co-located contiguous to the middle section, and composed solely of one *nrps* gene (Fig. 5b). This genetic organization is broadly maintained across subclade 4b, although in some strains, the left arm is also co-located with the middle section. In addition,

transporter-related genes (rectangle, *ra3–ra5*) are always co-located within the right arm (Fig. 5c), conversely to what is observed in subclade 4a (Fig. 5b).

Next, we focused on gene co-occurrence pattern for GCF-1/GCF-5 BGCs in clade 4 (Fig. 5b, c). Coloured symbols represent that all the genes conforming that block are present. On the contrary, an unfilled symbol denotes that at least one gene is missing. For subclade 4a, genes conforming to hexagon, cross and diamond are never found within the BGC (Fig. 5b). However, for subclade 4b, diversity is mainly observed in the left arm associated with *la1–la2* genes (hexagon), which encode transporters, which are absent in the more distant strains (Fig. 5c). Conversely, a broader pattern of gene co-occurrence is appreciated in the right arm, mainly given by *ra1–ra2* genes (diamond) encoding a lipase and an esterase, respectively, which seem to be present or absent in the different BGC organizations (Fig. 5c). The right arm was found to be associated with at least one monomodular *nrps* gene in almost every rhodococcal strain analysed (92%), suggesting a strong dependence for the assembly line of their putative product (Fig. 5c). In contrast, the left arm in most strains was not observed contiguous to a *nrps* gene although it was physically present elsewhere (Fig. 5b). These observations suggest that evolution of the left arm seems to be mainly derived through duplication events, whereas the right arm could be more associated with gene insertions/deletions [107]. Nevertheless, our results demonstrated that although some differences in co-occurrence patterns in BGC genetic structure can be observed between subclades, overall, the genes grouping into blocks showed a high conservation within left/right arms.

Regarding the middle section, which represents the most variable section of the cluster, CORASON revealed a BGC configuration consisting of one large *nrps* gene for subclade 4a (Fig. 5b), in contrast to a wider diversity of *nrps* genes within subclade 4b (Fig. 5c). Among this vast BGC diversity, NRPS #5 from *Rhodococcus* sp. H-CA8f was demonstrated to have a unique BGC configuration (Fig. 5c), encompassing three main biosynthetic genes, *nrps I*, *nrps II* and *nrps III* (Fig. 5a), responsible for the assembly of a peptidic core predicted to have a total of 18 monomers (Fig. S5). Interestingly, the two strains within subclade 4b that present a similar BGC structure to strain H-CA8f are the phylogenomically related *R. erythropolis* R138 and *R. erythropolis* PR4 (Fig. 1), and differences are observed only within the middle section. Even when an identical BGC predisposition between two strains is found, it does not necessarily imply chemical uniformity [108]. Thus, we aimed to expand the middle section analysis of most similar BGCs related to strain H-CA8f comprising strains NCTC 8036, R138, PR4, BG43, YL-1 BH4 and 008 (Fig. S5). Domain structure comparative analysis revealed that even though their numbers of *nrps* genes are similar, their modularity is somewhat different, presenting differences in the *nrps III* gene length and, thus, the respective assembling monomers, ranging from 17 to 21 amino acids (i.e. a module is represented with the same colour domains; Fig. S5).

Moreover, we observed additional unique features of the *nrps* genes that add chemical diversification, supporting H-CA8f's NRPS #5 uniqueness. On one hand, bioinformatic antiSMASH- and PRISM-based predictions showed that *nrps I* encompass a C*-starter domain, predicted to accept a *β*-hydroxy acid as a monomer (Fig. 5a, Table 1); a feature also conserved in the eight strains mentioned above (Fig. S5). On the other hand, *nrps III* harbours a non-classical adenylation domain (i.e. A*-domain) (Fig. 5a, Table 1), which is not associated with a condensation domain (i.e. A-PCP); thus, forming an incomplete module, depicted colourless in Fig. S5. This feature is also observed for subclade 4b strains NCTC 8036, R138, PR4 and BG43; although absent for the more distant strains YL-1, BH4 and 008 (Fig. S5). The observed differences could be the result of an ongoing evolutionary process, for which two ways can be equally possible. (i) Domain loss – the last module of *nrps II* was sometime complete and the loss of its domains is ongoing. In this scenario, a PCP could have been lost (a configuration observed for the upper five strains), while the last three strains YL-1, BH4 and 008 have the additional loss of an A-domain. (ii) Domain gain – a duplication of the A-domain occurred, leading to extra copies in *nrps II* and in *nrps III*, explaining the configuration observed for strains NCTC 8036, R138, H-CA8f, PR4 and BG43. Such duplication events have been reported in fungi [109], where the cost of its maintenance can be explained through the promiscuous incorporation of monomer(s), adding chemical diversity. Overall, our results support that although this BGC is widely distributed along clade 4, *Rhodococcus* sp. H-CA8f NRPS #5 bears a unique BGC in terms of genetic configuration, co-occurrence patterns and modularity. The ubiquity of these related BGCs suggests that a parental BGC was probably fixed early in the evolutionary history of the genus and their derivatives are under a dynamic evolutionary process, deriving into the vast BGC distribution observed along the clade 4 (Fig. 5).

Chemical dereplication of NPs

To link this unique BGC to its NP(s), *Rhodococcus* sp. H-CA8f was further explored for its ability to produce bioactive metabolites. Fermentations were performed in five different culture media, and crude extracts' antimicrobial activities against seven model Gram-positive and Gram-negative bacteria of clinical interest were tested. Notably, all model bacteria were at least inhibited once in the different media (Table S5). An important activity was observed against *L. monocytogenes*, *C. michiganensis* and *M. luteus*, specifically in ISP2-ASW, R5A and SG media (Table S5). Furthermore, bioactive extracts were submitted for chemical dereplication, by high-performance LC-HRMS analysis (Fig. 6). Data concerning UV spectra (Fig. 6a), molecular formulae (Fig. 6b) and accurate masses (Fig. 6a) were compared against Fundación MEDINA's in-house database to achieve possible compound identification. The ISP2-ASW chromatographic profile showed the presence of peaks P1, P2 and P3, which have similar chemical features to a previously isolated orphan metabolite, namely corynecin I, II and III, respectively (Fig. 6). Corynecins were first discovered from

Corynebacterium hydrocarboclastus culture broth, when using *n*-alkanes as the sole carbon source [110]. To our knowledge, this is the first report describing the detection of the orphan metabolite corynecins in *Rhodococcus*.

Corynecins are a family of *N*-acyl derivatives of D-threo-*p*-nitrophenylserinol compounds, structurally related to the antibiotic chloramphenicol [111]. Since NRPS #5 of *Rhodococcus* sp. H-CA8f presents a 17% similarity to the chloramphenicol BGC from *Streptomyces venezuelae* ATCC 10712 (MIBiG entry: BGC0000893 [107]) (Table S4), relatedness between them was explored. Their resemblance is mostly explained by four genes, the *ra7–ra10* genes grouped as triangles in Fig. 5a, which bears the following BLAST with BGC0000893: *p*-aminobenzoate *N*-oxygenase (SVEN_0924:43.2% identity, 94% coverage); prephenate dehydrogenase (SVEN_0919: 48% identity, 72% coverage); aminodeoxy chorismate synthase (SVEN_0920:54.4% identity, 98% coverage) and chorismate mutase (SVEN_0918: 48.4% identity, 68% coverage).

Considering our observations, we propose that the right arm of NRPS #5 from *Rhodococcus* sp. H-CA8f is involved in the production of the orphan corynecins, and the overall NRPS #5 may be involved in the synthesis of at least four different molecules (*m1–m4*). First, the left arm presents all the genes necessary for the synthesis and modification of a 4-carbon (C_4) molecule (*m1*). In the middle section, the presence of a C*-starter domain within *nrps I* suggests that *m1* could be incorporated into the assembly line as the initial monomer, with subsequent elongation by incorporating the amino acid core (Fig. S5); thus, creating *m2* (i.e. *m1*+peptide core). At the other end, the right arm is probably involved in the synthesis of corynecins (*m3*), which was functionally observed in H-CA8f's bioactive extract (Fig. 6). The non-classical A*-domain within *nrps III* may have the ability to interact with corynecins (*m3*) and modify them for incorporation into the assembly line as monomers, resembling the stand-alone A-domain interaction observed in chloramphenicol biosynthesis using a monomodular NRPS that solely harbours A-PCP domains [107]. Furthermore, the A*-domain of *nrps III* could be acting as a starting domain [112], as the result of the recombination of two clusters, left arm and the two *nrps* genes from the middle section, and the *nrps III* of the middle section together with the right arm, meaning that there could be more than one cluster on this genomic space [113]. Some studies support NRPS non-canonical function, reporting that some modules do not necessarily work in a sequential fashion [114]. In this regard, the last C-domain of *nrps II* may be acting on the condensation of either the A-domain next to it (*nrps II*) or the A-domain from *nrps III* (A*-domain). Alternatively, it can act in an iterative fashion by condensing both monomers, providing further chemical diversification [112, 115].

Therefore, we hypothesize that a potential final product of the biosynthetic assembly (*m4*) will result from the synergistic action of all the NRPS #5 regions, addressed by the C_4-molecule (*m1*, left arm product) attached to the peptide core (*m2*, middle section product), and with the possibility of the incorporation of corynecins (*m3*, right arm product). However, so far, we acknowledge that out of all the products proposed above, only corynecins (*m3*) were chemically detected in strain H-CA8f bioactive crude extracts. The LC-HRMS detection limit, using only positive mode (ESI+) ionization, or the need of a higher ionization voltage, may be among the chemical variables that could explain our inability to detect the proposed metabolites [116, 117].

Based on high-quality comparative genomic analyses, our results broaden our comprehension of the distribution and diversity of BGCs in *Rhodococcus*, supporting a phylogenomic signal underlying NRPS pathways across the genus. Deeper understanding of a NRPS GCF widely distributed along clade 4 unveiled that *Rhodococcus* sp. H-CA8f, our marine strain, harbours a unique BGC configuration in terms of gene context, co-occurrence patterns and modularity, which may be possibly connected to chloramphenicol-related orphan metabolites, namely, the corynecins. Overall, these findings enrich comparative BGC analyses that attempt to link orphan metabolites and will help future genome-guided efforts for NP discovery in *Rhodococcus*.

Funding information
This work was supported by funds from the Comisión Nacional de Investigación Científica y Tecnológica (FONDECYT regular no. 1171555 to B. C.; FONDECYT postdoctorado no. 3180399 to A. U.), and CONICYT PIA GAMBIO project no. ACT172128 to B. C. In addition, E. C.-N. was funded by FONDECYT regular no. 1200834 and by PIA-Anillo ACT192057. A. C. and L. Z.-L. acknowledge Beca de Doctorado ANID no. 21191625 and Beca CONICYT de Doctorado Nacional no. 21180908, respectively, and together they acknowledge Programa de Incentivos a la Iniciación Científica de la Dirección de Postgrado y Programas de la UTFSM. R. V. acknowledges the Darwin Trust of Edinburgh for a PhD scholarship. F. B.-G. is the recipient of a Newton Advanced Fellowship, Royal Society, UK (NAF\R2\180631).

Acknowledgements
We thank Dr Fernando Reyes for providing services with Fundación Medina (Granada, Spain), Dr Miryam Valenzuela for kindly sharing the CLMI strain and Dr Danilo Pérez-Pantoja for facilitating computational access. We also thank Matt Woolery and Dr William Fenical for kindly and critically reading our manuscript.

Author contributions
A. U. and B. C. conceived the idea for the project. A. U., R. V., A. C. and L. Z.-L. performed experiments and analysed results. A. U. and R. V. performed formal analysis on curated data. E. C.-N., F. B.-G. and B. C. helped with data analysis/interpretation. A. U. wrote the paper; and B. C. reviewed and edited the paper, using input from all authors.

Conflicts of interest
The authors declare there are no conflicts of interest.

References

1. Zotchev SB. Genomics-based insights into the evolution of secondary metabolite biosynthesis in actinomycete bacteria. In: Pontarotti P (eds). *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life*. Cham: Springer International Publishing; 2014. pp. 35–45.
2. Jensen PR. Natural products and the gene cluster revolution. *Trends Microbiol* 2016;24:968–977.
3. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;17:333–351.

4. Miller IJ, Chevrette MG, Kwan JC. Interpreting microbial biosynthesis in the genomic age: Biological and practical considerations. *Mar Drugs* 2017;15:165.
5. Ziemert N, Alanjary M, Weber T. The evolution of genome mining in microbes – a review. *Nat Prod Rep* 2016;33:988–1005.
6. Cruz-Morales P, Kopp JF, Martínez-Guerrero C, Yáñez-Guerra LA, Selem-Mojica N, *et al*. Phylogenomic analysis of natural products biosynthetic gene clusters allows discovery of arseno-organic metabolites in model streptomycetes. *Genome Biol Evol* 2016;8:1906–1916.
7. Sélem-Mojica N, Aguilar C, Gutiérrez-García K, Martínez-Guerrero CE, Barona-Gómez F. EvoMining reveals the origin and fate of natural product biosynthetic enzymes. *Microb Genomics* 2019;5:e000260.
8. O'Brien J, Wright GD. An ecological perspective of microbial secondary metabolism. *Curr Opin Biotechnol* 2011;22:552–558.
9. Genilloud O. Actinomycetes: still a source of novel antibiotics. *Nat Prod Rep* 2017;34:1203–1232.
10. Zotchev SB. Marine actinomycetes as an emerging resource for the drug development pipelines. *J Biotechnol* 2012;158:168–175.
11. Monciardini P, Iorio M, Maffioli S, Sosio M, Donadio S. Discovering new bioactive molecules from microbial sources. *Microb Biotechnol* 2014;7:209–220.
12. Bérdy J. Thoughts and facts about antibiotics: Where we are now and where we are heading. *J Antibiot* 2012;65:385–395.
13. de Lima Procópio RE, da Silva IR, Martins MK, de Azevedo JL, de Araújo JM. Antibiotics produced by *Streptomyces*. *Braz J Infect Dis* 2012;16:466–471.
14. Rocha-Martin J, Harrington C, Dobson ADW, O'Gara F. Emerging strategies and integrated systems microbiology technologies for biodiscovery of marine bioactive compounds. *Mar Drugs* 2014;12:3516–3559.
15. Tiwari K, Gupta RK. Rare actinomycetes: a potential storehouse for novel antibiotics. *Crit Rev Biotechnol* 2012;32:108–132.
16. Subramani R, Aalbersberg W. Culturable rare Actinomycetes: diversity, isolation and marine natural product discovery. *Appl Microbiol Biotechnol* 2013;97:9291–9321.
17. Schorn MA, Alanjary MM, Aguinaldo K, Korobeynikov A, Podell S, *et al*. Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. *Microbiology (Reading)* 2016;162:2075–2086.
18. Larkin MJ, Kulakov LA, Allen CC. Biodegradation and *Rhodococcus* – masters of catabolic versatility. *Curr Opin Biotechnol* 2005;16:282–290.
19. Larkin MJ, Kulakov LA, Allen CCR. Biodegradation by members of the genus *Rhodococcus*: biochemistry, physiology, and genetic adaptation. *Adv Appl Microbiol* 2006;59:1–29.
20. Cappelletti M, Presentato A, Piacenza E, Firrincieli A, Turner RJ, *et al*. Biotechnology of *Rhodococcus* for the production of valuable compounds. *Appl Microbiol Biotechnol* 2020;104:8567–8594.
21. Orro A, Cappelletti M, D'Ursi P, Milanesi L, Di Canito A, *et al*. Genome and phenotype microarray analyses of *Rhodococcus* sp. Bcp1 and *Rhodococcus opacus* R7: genetic determinants and metabolic abilities with environmental relevance. *PLoS One* 2015;10:e0139467.
22. Cappelletti M, Fedi S, Zampolli J, Di Canito A, D'Ursi P, *et al*. Phenotype microarray analysis may unravel genetic determinants of the stress response by *Rhodococcus aetherivorans* BCP1 and *Rhodococcus opacus* R7. *Res Microbiol* 2016;167:766–773.
23. Ceniceros A, Dijkhuizen L, Petrusma M, Medema MH. Genome-based exploration of the specialized metabolic capacities of the genus *Rhodococcus*. *BMC Genomics* 2017;18:593.
24. Rigali S, Anderssen S, Naômé A, van Wezel GP. Cracking the regulatory code of biosynthetic gene clusters as a strategy for natural product discovery. *Biochem Pharmacol* 2018;153:24–34.
25. Bosello M, Zeyadi M, Kraas FI, Linne U, Xie X, *et al*. Structural characterization of the heterobactin siderophores from *Rhodococcus erythropolis* PR4 and elucidation of their biosynthetic machinery. *J Nat Prod* 2013;76:2282–2290.
26. Miranda-CasoLuengo R, Coulson GB, Miranda-Casoluengo A, Vázquez-Boland JA, Hondalus MK, *et al*. The hydroxamate siderophore rhequichelin is required for virulence of the pathogenic actinomycete *Rhodococcus equi*. *Infect Immun* 2012;80:4106–4114.
27. Bosello M, Robbel L, Linne U, Xie X, Marahiel MA. Biosynthesis of the siderophore rhodochelin requires the coordinated expression of three independent gene clusters in *Rhodococcus jostii* RHA1. *J Am Chem Soc* 2011;133:4587–4595.
28. Habib S, Ahmad SA, Wan Johari WL, Abd Shukor MY, Alias SA, *et al*. Production of lipopeptide biosurfactant by a hydrocarbon-degrading Antarctic *Rhodococcus*. *Int J Mol Sci* 2020;21:6138.
29. Chu J, Vila-Farres X, Inoyama D, Ternei M, Cohen LJ, *et al*. Discovery of MRSA active antibiotics using primary sequence from the human microbiome. *Nat Chem Biol* 2016;12:1004–1006.
30. Chiba H, Agematu H, Kaneto R, Terasawa T, Sakai K, *et al*. *Rhodopeptins* (Mer-N1033), novel cyclic tetrapeptides with antifungal activity from *Rhodococcus* sp. *J Antibiot* 1999;52:695–699.
31. Iwatsuki M, Tomoda H, Uchida R, Gouda H, Hirono S, *et al*. Lariatins, antimycobacterial peptides produced by *Rhodococcus* sp. K01-B0171, have a lasso structure. *J Am Chem Soc* 2006;128:7486–7491.
32. Kitagawa W, Tamura T. A quinoline antibiotic from *Rhodococcus erythropolis* JCM 6824. *J Antibiot* 2008;61:680–682.
33. Nachtigall J, Schneider K, Nicholson G, Goodfellow M, Zinecker H, *et al*. Two new aurachins from *Rhodococcus* sp. Acta 2259. *J Antibiot* 2010;63:567–569.
34. Undabarrena A, Salvà-Serra F, Jaén-Luchoro D, Castro-Nallar E, Mendez KN, *et al*. Complete genome sequence of the marine *Rhodococcus* sp. H-CA8f isolated from Comau fjord in Northern Patagonia, Chile. *Mar Genomics* 2018;40:13–17.
35. Undabarrena A, Beltrametti F, Claverías FP, González M, Moore ERB, *et al*. Exploring the diversity and antimicrobial potential of marine actinobacteria from the Comau fjord in Northern Patagonia, Chile. *Front Microbiol* 2016;7:1135.
36. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–1055.
37. Savory EA, Fuller SL, Weisberg AJ, Thomas WJ, Gordon MI, *et al*. Evolutionary transitions between beneficial and phytopathogenic *Rhodococcus* challenge disease management. *eLife* 2017;6:e30925.
38. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* 2015;8:12–24.
39. Sharrar AM, Crits-Christoph A, Méheust R, Diamond S, Starr EP, *et al*. Bacterial secondary metabolite biosynthetic potential in soil varies with phylum, depth, and vegetation type. *mBio* 2020;11:e00416-20.
40. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, *et al*. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* 2016;7:13219.
41. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, *et al*. antiSMASH 4.0–improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* 2017;45:W36–W41.
42. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, *et al*. A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* 2020;16:60–68.
43. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015;16:157.
44. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12:59–60.

45. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;30:772–780.
46. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490.
47. 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.
48. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;32:268–274.
49. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, *et al*. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 2007;8:460.
50. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 2018;35:1547–1549.
51. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 2001;17:754–755.
52. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 2003;19:1572–1574.
53. Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol Biol Evol* 2017;34:772–773.
54. Robinson DF, Foulds LR. Comparison of phylogenetic trees. *Math Biosci* 1981;53:131–147.
55. Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics* 2011;27:592–593.
56. Estabrook GF, McMorris FR, Meacham CA. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Syst Zool* 1985;34:193–200.
57. Goluch T, Bogdanowicz D, Giaro K, Price S. Visual TreeCmp: comprehensive comparison of phylogenetic trees on the web. *Methods Ecol Evol* 2020;11:494–499.
58. Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. *Softw Pract Exp* 1991;21:1129–1164.
59. Hu Y. Efficient high-quality force-directed graph drawing. *Math J* 2006;10:35.
60. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. 2009. https://gephi.org/publications/gephi-bastian-feb09.pdf
61. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* 2008:2008:P10008.
62. Needham M, Hodler AE. *Graph Algorithms: Practical Examples in Apache Spark and Neo4j*. Sebastopol, CA: O'Reilly Media; 2019.
63. Kolde R. Pheatmap: Pretty heatmaps. 2019. https://cran.r-project.org/web/packages/pheatmap/index.html [accessed 04 Jan 2019].
64. Hsieh TC, Ma KH, Chao A. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods Ecol Evol* 2016;7:1451–1456.
65. Chase AB, Sweeney D, Muskat MN, Guillén-Matus D, Jensen PR. Vertical inheritance governs biosynthetic gene cluster evolution and chemical diversification. *bioRxiv* 2021.
66. Clarke KR. Non-parametric multivariate analyses of changes in community structure. *Austral Ecol* 1993;18:117–143.
67. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, *et al*. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* 2019;47:W81–W87.
68. Röttig M, Medema MH, Blin K, Weber T, Rausch C, *et al*. NRPSpredictor2 – a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res* 2011;39:W362-7.
69. Baranašić D, Zucko J, Diminic J, Gacesa R, Long PF, *et al*. Predicting substrate specificity of adenylation domains of nonribosomal peptide synthetases and other protein properties by latent semantic indexing. *J Ind Microbiol Biotechnol* 2014;41:461–467.
70. Prieto C, García-Estrada C, Lorenzana D, Martín JF. NRPSsp: non-ribosomal peptide synthase substrate predictor. *Bioinformatics* 2012;28:426–427.
71. Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res* 2017;45:W49–W54.
72. Fernández E, Weißbach U, Reillo CS, Braña AF, Méndez C, *et al*. Identification of two genes from *Streptomyces argillaceus* encoding glycosyltransferases involved in transfer of a disaccharide during biosynthesis of the antitumor drug mithramycin. *J Bacteriol* 1998;180:4929–4937.
73. Malmierca MG, González-Montes L, Pérez-Victoria I, Sialer C, Braña AF, *et al*. Searching for glycosylated natural products in actinomycetes and identification of novel *Macrolactams* and *Angucyclines*. *Front Microbiol* 2018;9:39.
74. Cumsille A, Undabarrena A, González V, Claverías F, Rojas C, *et al*. Biodiversity of actinobacteria from the South Pacific and the assessment of *Streptomyces* chemical diversity with metabolic profiling. *Mar Drugs* 2017;15:286.
75. Valenzuela M, Besoain X, Durand K, Cesbron S, Fuentes S, *et al*. *Clavibacter michiganensis* subsp. michiganensis strains from central Chile exhibit low genetic diversity and sequence types match strains in other parts of the world. *Plant Pathol* 2018;67:1944–1954.
76. Busch J, Agarwal V, Schorn M, Machado H, Moore BS, *et al*. Diversity and distribution of the bmp gene cluster and its polybrominated products in the genus *Pseudoalteromonas*. *Environ Microbiol* 2019;21:1575–1585.
77. van der Hooft JJJ, Mohimani H, Bauermeister A, Dorrestein PC, Duncan KR, *et al*. Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem Soc Rev* 2020;49:3297–3314.
78. Blin K, Medema MH, Kottmann R, Lee SY, Weber T. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res* 2017;45:D555–D559.
79. Blin K, Andreu P, de los Santos ELC, Del Carratore F, Lee SY, *et al*. The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res* 2019;47:D625–D630.
80. Blin K, Kim HU, Medema MH, Weber T. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief Bioinform* 2019;20:1103–1113.
81. Blin K, Shaw S, Kautsar SA, Medema MH, Weber T. The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Res* 2021;49:D639–D643.
82. Sangal V, Goodfellow M, Jones AL, Seviour RJ, Sutcliffe IC. Refined systematics of the genus *Rhodococcus* based on whole genome analyses. In: Alvarez H (ed). *Biology of Rhodococcus*, vol. 16. Cham: Springer International Publishing; 2019. pp. 1–21.
83. Anastasi E, MacArthur I, Scortti M, Alvarez S, Giguère S, *et al*. Pangenome and phylogenomic analysis of the pathogenic actinobacterium *Rhodococcus equi*. *Genome Biol Evol* 2016;8:3140–3148.
84. Creason AL, Davis EWI, Putnam ML, Vandeputte OM, Chang JH. Use of whole genome sequences to develop a molecular phylogenetic framework for *Rhodococcus fascians* and the *Rhodococcus* genus. *Front Plant Sci* 2014;5:00406.
85. Komukai-Nakamura S, Sugiura K, Yamauchi-Inomata Y, Toki H, Venkateswaran K, *et al*. Construction of bacterial consortia that degrade Arabian light crude oil. *J Ferment Bioeng* 1996;82:570–574.
86. McLeod MP, Warren RL, Hsiao WWL, Araki N, Myhre M, *et al*. The complete genome of Rhodococcus sp. RHA1 provides

insights into a catabolic powerhouse. *Proc Natl Acad Sci U S A* 2006;103:15582–15587.

87. Moiseeva OV, Solyanikova IP, Kaschabek SR, Gröning J, Thiel M, *et al*. A new modified ortho cleavage pathway of 3-chlorocatechol degradation by *Rhodococcus opacus* 1CP: genetic and biochemical evidence. *J Bacteriol* 2002;184:5282–5292.
88. Doroghazi JR, Metcalf WW. Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics* 2013;14:611.
89. Männle D, McKinnie SMK, Mantri SS, Steinke K, Lu Z, *et al*. Comparative genomics and metabolomics in the genus *Nocardia*. *mSystems* 2020;5:e00120-25.
90. Graf R, Anzali S, Buenger J, Pfluecker F, Driller H. The multifunctional role of ectoine as a natural cell protectant. *Clin Dermatol* 2008;26:326–333.
91. Du Y-L, Shen X-L, Yu P, Bai L-Q, Li Y-Q. Gamma-butyrolactone regulatory system of *Streptomyces chattanoogensis* links nutrient utilization, metabolism, and development. *Appl Environ Microbiol* 2011;77:8415–8426.
92. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, *et al*. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res* 2020;48:D454–D458.
93. Kraemer SM. Iron oxide dissolution and solubility in the presence of siderophores. *Aquat Sci* 2004;66:3–18.
94. Doroghazi JR, Albright JC, Goering AW, Ju K-S, Haines RR, *et al*. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol* 2014;10:963–968.
95. Reitz ZL, Hardy CD, Suk J, Bouvet J, Butler A. Genomic analysis of siderophore *β*-hydroxylases reveals divergent stereocontrol and expands the condensation domain family. *Proc Natl Acad Sci U S A* 2019;116:19805–19814.
96. Adamek M, Alanjary M, Sales-Ortells H, Goodfellow M, Bull AT, *et al*. Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in *Amycolatopsis* species. *BMC Genomics* 2018;19:426.
97. Chevrette MG, Carlos-Shanley C, Louie KB, Bowen BP, Northen TR, *et al*. Taxonomic and metabolic incongruence in the ancient genus *Streptomyces*. *Front Microbiol* 2019;10:2170.
98. Chevrette MG, Carlson CM, Ortega HE, Thomas C, Ananiev GE, *et al*. The antimicrobial potential of *Streptomyces* from insect microbiomes. *Nat Commun* 2019;10:516.
99. Caldera EJ, Chevrette MG, McDonald BR, Currie CR. Local adaptation of bacterial symbionts within a geographic mosaic of antibiotic coevolution. *Appl Environ Microbiol* 2019;85:e01580-19.
100. Jensen PR, Mafnas C. Biogeography of the marine actinomycete *Salinispora*. *Environ Microbiol* 2006;8:1881–1888.
101. Letzel AC, Li J, Amos GCA, Millán-Aguiñaga N, Ginigini J, *et al*. Genomic insights into specialized metabolism in the marine actinomycete *Salinispora*. *Environ Microbiol* 2017;19:3660–3673.
102. Cruz-Morales P, Ramos-Aboites HE, Licona-Cassani C, Selem-Mójica N, Mejía-Ponce PM, *et al*. Actinobacteria phylogenomics, selective isolation from an iron oligotrophic environment and siderophore functional characterization, unveil new desferrioxamine traits. *FEMS Microbiol Ecol* 2017;93:fix086.
103. Gutiérrez-García K, Neira-González A, Pérez-Gutiérrez RM, Granados-Ramírez G, Zarraga R, *et al*. Phylogenomics of 2,4-diac etylphloroglucinol-producing *Pseudomonas* and novel antiglycation endophytes from *Piper auritum*. *J Nat Prod* 2017;80:1955–1963.
104. Juárez-Vázquez AL, Edirisinghe JN, Verduzco-Castro EA, Michalska K, Wu C, *et al*. Evolution of substrate specificity in a retained enzyme driven by gene loss. *eLife* 2017;6:e22679.
105. Adamek M, Alanjary M, Ziemert N. Applied evolution: phylogeny-based approaches in natural products research. *Nat Prod Rep* 2019;36:1295–1312.
106. Chevrette MG, Gutiérrez-García K, Selem-Mojica N, Aguilar-Martínez C, Yañez-Olvera A, *et al*. Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat Prod Rep* 2020;37:566–599.
107. He J, Magarvey N, Piraee M, Vining LC. The gene cluster for chloramphenicol biosynthesis in *Streptomyces venezuelae* ISP5230 includes novel shikimate pathway homologues and a monomodular non-ribosomal peptide synthetase gene. *Microbiology (Reading)* 2001;147:2817–2829.
108. Martinet L, Naômé A, Baiwir D, De Pauw E, Mazzucchelli G, *et al*. On the risks of phylogeny-based strain prioritization for drug discovery: *Streptomyces lunaelactis* as a case study. *Biomolecules* 2020;10:1027.
109. Bushley KE, Turgeon BG. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. *BMC Evol Biol* 2010;10:26.
110. Suzuki T, Honda H, Katsumata R. Production of antibacterial compounds analogous to chloramphenicol by a n-paraffin-grown bacterium. *Agricultural and Biological Chemistry* 2014;36:2223–2228.
111. Nakano H, Tomita F, Yamaguchi K, Nagashima M, Suzuki T. Corynecin (chloramphenicol analogs) fermentation studies: selective production of corynecin I by *Corynebacterium hydrocarboclastus* grown on acetate. *Biotechnol Bioeng* 1977;19:1009–1018.
112. Calcott MJ, Ackerley DF. Portability of the thiolation domain in recombinant pyoverdine non-ribosomal peptide synthetases. *BMC Microbiol* 2015;15:162.
113. Baunach M, Chowdhury S, Stallforth P, Dittmann E. The landscape of recombination events that create nonribosomal peptide diversity. *Mol Biol Evol* 2021;38:2116–2130.
114. Ali H, Ries MI, Lankhorst PP, van der Hoeven RAM, Schouten OL, *et al*. A non-canonical NRPS is involved in the synthesis of fungisporin and related hydrophobic cyclic tetrapeptides in *Penicillium chrysogenum*. *PLoS One* 2014;9:e98212.
115. Motz HD, Schwarzer D, Marahiel MA. Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *ChemBioChem* 2002;3:490–504.
116. Mirzaei H, Regnier F. Enhancing electrospray ionization efficiency of peptides by derivatization. *Anal Chem* 2006;78:4175–4183.
117. Brodbelt JS. Ion activation methods for peptides and proteins. *Anal Chem* 2016;88:30–51.