Exploring Hidden Diversity: Genome Mining and Evolutionary Insights into Noncanonical Fungal Biosynthetic Gene Clusters and Understudied Fungal Natural Products

By

Grant R. Nickles

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Cellular and Molecular Biology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2025

Date of final oral examination: 05/28/2025

The dissertation is approved by the following members of the Final Oral Committee:

Nancy P. Keller, Professor, Medical Microbiology & Immunology and Plant Pathology

Mehdi Kabbage, Professor, Plant Pathology

Chris T. Hittinger, Professor, Genetics

Jason C. Kwan, Associate Professor, Pharmaceutical Sciences

Claudia Solis Lemus, Assistant Professor, Plant Pathology

DISSERTATION ABSTRACT

Fungi and bacteria have evolved specialized chemical pathways that convert primary metabolites in specialized/secondary metabolites (SMs) that provide a variety of beneficial functions. For the producing organism, SMs may help in microbial niche adaptation development, signaling, symbiosis, defense, and UV protection, among other functions. Humanity has leveraged these compounds for their own uses across medicinal, agricultural, and food safety sectors to combat pathogens and contaminants with antimicrobials, lower cholesterol with statins, and immune system suppression for life-saving operations. Astonishingly, even though a small fraction of microbial SMs have been characterized, their derivatives represent one-third of all FDA-approved drugs and two-thirds of FDA-approved antibiotics.

During the golden age of drug discovery (1940s–1960s), natural product scientists identified thousands of bioactive compounds with antibacterial and antifungal properties, many of which became clinically approved medicines still commonly used today (e.g., penicillins, bacitracin, macrolides). However, modern drug discovery faces a major bottleneck: the persistent rediscovery of known compounds, particularly in antibiotic development, where no truly novel class has been discovered since the early 1980s.

To overcome this rediscovery bottleneck, our work focused on uncovering SM diversity by targeting noncanonical biosynthetic pathways and microbial lineages that have been largely excluded from traditional genome mining efforts. First, we created the first ever genome mining pipeline capable of finding the genes that produce a noncanonical class of SMs called isocyanides (Chapter 2). We then explored Lecanoromycetes isocyanide SM diversity and unveil the stepwise evolution of a widespread hybrid megasynthase (Chapter 3). Lastly, we looked at SM biodiversity within invasive South African *Amanita muscaria* as a case study for tracking SM evolution between native vs invasive populations (Chapter 4).

ACKNOWLEDGMENTS

The last five years have been the more formative of my life. I think back on my first week of graduate school. A name I had never heard of, Dr. Nancy Keller, emailed me asking if I was interested in doing a rotation with her group. It is almost impossible for me to entertain what kind of scientist I would be had I not joined Dr. Kellers group as I can't imagine myself doing a PhD with anyone else. Nancy has been one of the most formative mentors of my life. While we sometimes butt heads, there has always been a mutual respect between us. Nancy is an almost endless reservoir of knowledge of secondary metabolites, fungal biology, and natural product biosynthesis. She is a mentor, a friend, and someone I am so thankful to have in my network.

I'd also like to specifically thank Dr. Milton Drott (Mickey). Mickey challenged me to think in ways no one ever had. He pushed me to ground my ideas in a scientific purpose. In many ways, I feel I strive to emulate him in my current work. I learned from Mickey a framework for identifying and approaching computational problems and I am so thankful for the years we overlapped within the Keller lab. Without Mickey's mentorship, friendship, and professional support I would not be the scientist I am today.

I thank all my Keller lab members that I have overlapped with. A special mention to Sung Chul, Dianiris, Wenjie, Nandhitha, Nick, Natalia, Justin, Harrison, and Jin Woo for your support on some aspect of my work during my PhD. I wouldn't have accomplished everything that I did without the endless support from everyone mentioned and in addition to so many other names not listed.

To my CMB cohort and best friends Dylan, Dom, Alex, and Owen, love you guys! Our annual camping trips to the boundary waters have marked the passage of years for me as I have progressed through graduate school and I'm so grateful to have gone through the PhD process with you all.

To my close family, Pat, Dawn, Jason, Megan, Alec, Kylie, Luke, Jenna, and Connor. I love you all so much. Both my grandfather, Bob McQueen, and uncle, Sean McQueen, passed while I was a PhD

dissertator. The passing of these two amazing men were some of the hardest moments in my life, and in many ways are still events I'm processing today. You all were my rock in more ways than I can describe.

Lastly but most importantly I would like to thank Brandon. Dating and supporting someone through a PhD can be very difficult. My work occasionally came before our relationship, and you have always been supportive and patient during my personal and professional highs and lows.

TABLE OF CONTENTS

CHAPTER 1: Introduction

We are in an antibiotic and chemical therapeutics crisis. The rate of antibiotic development has plummeted since the end of the 20th century, in large part because existing methods keep re-discovering the same types of chemicals. For example, no truly novel class of antibiotics has been discovered since the early 1980s,[1] and fewer and fewer antibiotics and antifungals achieve regulatory approval each year[2]. Meanwhile, the demand for new therapeutics is increasing rapidly as bacterial and fungal pathogens rapidly develop resistance to front line antibiotics and antifungals. A recent study estimated 1.14 million deaths in 2021 were directly due to antibiotic resistant bacterial infections, with certain bacterial infections like methicillin-resistant *Staphylococcus aureus* and carbapenem-resistant gram-negative bacteria driving a large amount of mortality increases[2,3]. Similarly, antifungal resistance poses a growing threat as more fungi evolve resistance to key antifungal agents like azoles and echinocandins[4]. It is estimated fungal infections caused approximately 3.8 million deaths in 2023[5] which is likely to worsen if new antifungals are not developed.

Microbes like fungi present the best opportunity to overcome the discovery crisis because they have evolved a functionally diverse natural pharmacopeia whose full potential remains untapped. For hundreds of millions of years, fungi have evolved chemical pathways that convert primary metabolites into specialized (secondary) metabolites (SMs)[6]. For the producing organism, SMs may aid in microbial competition, signaling, symbiosis and defense, among other functions. For humans, these microbial SMs are a major source of antimicrobial (e.g. penicillin, griseofulvin) and therapeutic (e.g. cyclosporine, mycophenolate) compounds. Astonishingly, even though a small fraction of microbial SMs have been characterized, their derivatives represent one-third of all FDA-approved drugs and two-thirds of FDA-approved antibiotics[7].

Fortunately for the pursuit of discovering novel compounds to expand our chemical repertoire, the biosynthetic genes responsible for SM synthesis and transport are often arranged in contiguous clusters,

known as biosynthetic gene clusters (BGCs), which enables their *de novo* prediction. Bioinformatic pipelines have been developed to take advantage of this arrangement to identify new BGCs in genomes, a process known as genome mining.

It has been estimated the scientific community has only characterized 1-3% of the millions of detectable microbial BGCs[8–10]. The sheer volume of uncharacterized BGCs presents the natural product community with a critical prioritization problem. Selecting promising candidates for study is further complicated because many SMs are poorly expressed, or not produced at all, under conventional laboratory conditions. Coupled with the lengthy timeline required to genetically and chemically characterize just a single BGC, the average graduate student is consequently limited to investigating only a handful of SM pathways during their PhD. The work in this dissertation is unified by the framework of prioritizing the study of SMs from poorly studied fungal taxa or SMs made by noncanonical BGCs (i.e., BGCs not predictable by conventional genome mining).

One such example of a noncanonical BGC is the isocyanide metabolite producing BGCs. Isocyanide metabolites have been found almost exclusively in fungi, bacteria, and marine sponges[12]. Extracted fungal isocyanide metabolites exhibit tremendous bioactivity. Among other things, isocyanides have been shown to 1) exhibit antimicrobial properties[12], 2) are the causative agent in ovine ill thrift disease[13], 3) can sequester external copper[14], and 4) have inhibited tumor cells *in vitro*[12]. While dozens of fungal isocyanide metabolites have been characterized, their biosynthesis remains almost completely unknown. This has drastically hindered our ability to understand isocyanide's ecological function and pharmacological potential[15]. Our lab discovered and characterized the first example of an isocyanide producing BGC in fungi, which featured an isocyanide synthase backbone (ICS). Prior to this discovery, it remained unknown that ICSs could serve as the backbone genes within fungal BGCs[16]. As existing genome mining algorithms

were unable to detect the presence of ICS BGCs within sequencing data, the extent of ICS BGC diversity and conservation remained unknown was I started my dissertation work in 2021.

Chapter 2 presents the first genome mining pipeline capable of predicting ICS BGCs within fungi. Using a dataset of 3,300 fungal genomes, we were able to locate 3,800 ICS BGCs using promoter motif genome mining[11], making ICS BGCs the fifth largest class of predictable fungal BGCs relative to canonical classes found by antiSMASH. We then characterized the ten largest ICS gene cluster families (GCF; i.e., ICS BGCs that share a set of co-localized genes) and showed that there is strong evidence that selection is maintaining the structure of these ICS GCFs. Finally, we explored the evolutionary origins of the largest ICS GCF, *dit*1/2, which has been previously characterized in yeast. Taken together, our results create a roadmap for future research into fungal ICS BGCs.

Chapter 3 explores the biodiversity of ICS BGCs in a poorly studied fungal taxon, the Lecanoromycetes, which is comprised of almost entirely lichen-forming fungi (LFF). LFF are renowned for their ability to thrive in metal-contaminated environments through multiple evolved strategies, such as sequestering metal in their thalli[12,13]. Considering this, and the established association of ICS-produced isocyanide SMs in metal-associated ecology[14,15], we hypothesized that LFF could provide unique insights into the diversity and evolutionary trajectories of ICSs. Our subsequent genomic and phylogenetic analyses revealed previously uncharacterized lichen-associated ICS BGCs and, significantly, led to the discovery and evolutionary characterization of a highly unusual ICS-NRPS hybrid enzyme that is found in 28% of Ascomycetes.

Chapter 4 uncovers the geographic origins of introduced *Amanita muscaria*, the world's most (in)famous mushroom, in South Africa, establishing a European origin. This work further investigates the specialized metabolomes of mushrooms from the invasive (South African) and native (European) ranges through

characterization of full genomes, biosynthetic gene clusters (BGCs), and metabolomes. Our findings reveal a remarkable conservation of the *A. muscaria* BGCs and mushroom cap metabolites between native and invasive ranges, despite geographic isolation and hundreds of years of evolution on a new continent[16] . Furthermore, we assessed the bioactivities of SMs produced by *A. muscaria* on bacteria, fungi, nematodes, and insects to gain insight into the potential ecological role that these SMs could have played in the success of *A. muscaria* as an invasive species.

References

1. Silver, L.L. (2011). Challenges of Antibacterial Discovery. Clin. Microbiol. Rev. *24*, 71–109. https://doi.org/10.1128/cmr.00030-10.

2. Rodriguez-Monguio, R., Seoane-Vazquez, E., and Powers, J.H. (2023). A Comparative Assessment of Approvals and Discontinuations of Systemic Antibiotics and Other Therapeutic Areas. Healthcare *11*, 1759. https://doi.org/10.3390/healthcare11121759.

3. Collaborators, G. 2021 A.R., Naghavi, M., Vollset, S.E., Ikuta, K.S., Swetschinski, L.R., Gray, A.P., Wool, E.E., Aguilar, G.R., Mestrovic, T., Smith, G., et al. (2024). Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050. Lancet *404*, 1199–1226. https://doi.org/10.1016/s0140-6736(24)01867-1.

4. Vitiello, A., Ferrara, F., Boccellino, M., Ponzo, A., Cimmino, C., Comberiati, E., Zovi, A., Clemente, S., and Sabbatucci, M. (2023). Antifungal Drug Resistance: An Emergent Health Threat. Biomedicines *11*, 1063. https://doi.org/10.3390/biomedicines11041063.

5. Steward, M. (2024). Global Burden of Fungal Infections and Antifungal Resistance from 1961 to 2024: Findings and Future Implications. Pharmacol. Pharm. *15*, 81–112. https://doi.org/10.4236/pp.2024.154007.

6. Keller, N.P. (2019). Fungal secondary metabolism: regulation, function and drug discovery. Nat Rev Microbiol *17*, 167–180. https://doi.org/10.1038/s41579-018-0121-1.

7. Patridge, E., Gareiss, P., Kinch, M.S., and Hoyer, D. (2016). An analysis of FDA-approved drugs: natural products and their derivatives. Drug Discov. Today *21*, 204–207. https://doi.org/10.1016/j.drudis.2015.01.009.

8. Bağcı, C., Nuhamunada, M., Goyat, H., Ladanyi, C., Sehnal, L., Blin, K., Kautsar, S.A., Tagirdzhanov, A., Gurevich, A., Mantri, S., et al. (2024). BGC Atlas: a web resource for exploring the global chemical diversity encoded in bacterial genomes. Nucleic Acids Res. *53*, D618–D624. https://doi.org/10.1093/nar/gkae953.

9. Medema, M.H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J.B., Blin, K., Bruijn, I.D., Chooi, Y.H., Claesen, J., Coates, R.C., et al. (2015). Minimum Information about a Biosynthetic Gene cluster. Nature Chemical Biology *11*, 625–631. https://doi.org/10.1038/nchembio.1890.

10. Robey, M.T., Caesar, L.K., Drott, M.T., Keller, N.P., and Kelleher, N.L. (2021). An Interpreted Atlas of Biosynthetic Gene Clusters from 1000 Fungal Genomes. PNAS *1*, https://doi.org/10.1073/pnas.2020230118.

11. Wolf, T., Shelest, V., and Shelest, E. (2013). Motif-based method for the genome-wide prediction of eukaryotic gene clusters. Lecture Notes in Computer Science *8158 LNCS*, 389–398. https://doi.org/10.1007/978-3-642-41190-8_42.

12. Rola, K., Osyczka, P., and Kafel, A. (2016). Different Heavy Metal Accumulation Strategies of Epilithic Lichens Colonising Artificial Post-Smelting Wastes. Arch. Environ. Contam. Toxicol. *70*, 418–428. https://doi.org/10.1007/s00244-015-0180-5.

13. Sarret, G., Manceau, A., Cuny, D., Haluwyn, C.V., Déruelle, S., Hazemann, J.-L., Soldo, Y., Eybert-Bérard, L., and Menthonnex, J.-J. (1998). Mechanisms of Lichen Resistance to Metallic Pollution. Environ. Sci. Technol. *32*, 3325–3330. https://doi.org/10.1021/es970718n.

14. Won, T.H., Bok, J.W., Nadig, N., Venkatesh, N., Nickles, G., Greco, C., Lim, F.Y., González, J.B., Turgeon, B.G., Keller, N.P., et al. (2022). Copper starvation induces antimicrobial isocyanide integrated into two distinct biosynthetic pathways in fungi. Nat Commun *13*, 4828. https://doi.org/10.1038/s41467-022-32394-x.

15. Lim, F.Y., Won, T.H., Raffa, N., Baccile, J.A., Wisecaver, J., Keller, N.P., Rokas, A., and Schroeder, F.C. (2018). Fungal isocyanide synthases and xanthocillin biosynthesis in Aspergillus fumigatus. Mbio *9*. https://doi.org/10.1128/mbio.00785-18.

16. Reid, D.A., and Eicker, A. (1991). South African fungi: the genus Amanita. Mycol. Res. *95*, 80–95. https://doi.org/10.1016/s0953-7562(09)81364-6.

CHAPTER 2: Mining for a new class of fungal natural products: the evolution, diversity, and distribution of isocyanide synthase biosynthetic gene clusters.

Grant R. Nickles[1], Brandon Oestereicher[2], Nancy P. Keller[1,3*], Milton T. Drott[4*]

[1] Department of Medical Microbiology and Immunology, University of Wisconsin—Madison, Madison, WI 53706, USA

[2] Madison, WI, 53703, USA

[3] Department of Plant Pathology, University of Wisconsin—Madison, Madison, WI 53706, USA

[4] USDA-ARS Cereal Disease Lab (CDL), St. Paul, MN 55108, USA

* To whom correspondence should be addressed. Email: Milton.Drott@usda.gov

Correspondence may also be addressed to Nancy P. Keller. Email: npkeller@wisc.edu

This work has been published as:

Nickles, G. R., Oestereicher, B., Keller, N. P. and Drott, M. T. (2023) Mining for a new class of fungal natural products: the evolution, diversity, and distribution of isocyanide synthase biosynthetic gene clusters. *Nucleic Acids Res.*, 10.1093/nar/gkad573.

Author Contributor Role Taxonomy (CRediT):

Conceptualization: GRN, NPK, MTD; Data curation: GRN; Formal analysis: GRN, MTD; Resources: NPK, MTD; Funding acquisition: GRN, NPK, MTD; Investigation: GRN; Methodology: GRN; Project administration: GRN, MTD; Software: GRN, BO, MTD; Validation: GRN; Visualization: GRN; Writing – Original Draft: GRN, MTD; Writing –Review & Editing: GRN, NPK, MTD; Project administration: NPK, MTD

2.1 Abstract

The products of non-canonical isocyanide synthase (ICS) biosynthetic gene clusters (BGCs) mediate pathogenesis, microbial competition, and metal-homeostasis through metal-associated chemistry. We sought to enable research into this class of compounds by characterizing the biosynthetic potential and evolutionary history of these BGCs across the Fungal Kingdom. We amalgamated a pipeline of tools to predict BGCs based on shared promoter motifs and located 3,800 ICS BGCs in 3,300 genomes, making ICS BGCs the fifth largest class of specialized metabolites compared to canonical classes found by antiSMASH. ICS BGCs are not evenly distributed across fungi, with evidence of gene-family expansions in several Ascomycete families. We show that the ICS *dit1/2* gene cluster family (GCF), which was prior only studied in yeast, is present in ~30% of all Ascomycetes. The *dit* variety ICS exhibits greater similarity to bacterial ICS than other fungal ICS, suggesting a potential convergence of the ICS backbone domain. The evolutionary origins of the *dit* GCF in Ascomycota are ancient and these genes are diversifying in some lineages. Our results create a roadmap for future research into ICS BGCs. We developed a website (https://isocyanides.fungi.wisc.edu/) that facilitates the exploration and downloading of all identified fungal ICS BGCs and GCFs.

2.2 Introduction

Fungal specialized metabolites (SMs; also called secondary metabolites or natural products) are essential sources of antimicrobial (e.g., penicillin, griseofulvin) and therapeutic (e.g., cyclosporine, mycophenolate) compounds[1–3]. In nature, SMs enable niche adaptation by conferring protection from abiotic and biotic stressors, as well as aiding in nutrient acquisition and competitive interactions[4]. The bioactive properties of these molecules have propelled the research community to identify strategies to find new fungal SMs. Current genome-mining algorithms that predict SM-producing biosynthetic gene clusters (BGCs) in fungal genomes utilize the contiguous physical arrangement of genes to enable BGC predictions[5–7]. This approach

has successfully revealed millions of putative BGCs, many of which remain uncharacterized[8,9]. However, these algorithms rely on prior knowledge of chemical-class defining "backbone" synthases/synthetases (e.g., nonribosomal synthetases, polyketide synthases, terpene synthases/cyclases), biasing results towards well-characterized classes of BGCs. This approach cannot find BGC architectures with uncharacterized backbone genes (i.e., non-canonical). While there have been several recent efforts to predict non-canonical BGC architectures[10,11], the uncharacterized nature of non-canonical BGCs has made them challenging to incorporate into genome-mining software, obfuscating our broader understanding of their prevalence and function. Our inability to mine “unknown” BGCs[12] remains one of the greatest challenges hindering the discovery of novel compounds with bioactive properties[13].

Isocyanides (also called isonitriles) are a chemical class of SM produced by bacteria and fungi[14]. These compounds are produced by non-canonical BGCs that are not detected by current genome-mining software[15]. They are characterized by the presence of the highly reactive isocyanide functional group ($R{\equiv}N^{+}\text{-}C^{-}$), which is formed by the conversion of the amino group on select amino acids. Isocyanide SMs can participate in unique chemical reactions[16,17] and possess potent antifungal, antibacterial, antitumor, and antiprotozoal bioactivity[18–20]. The isocyanide functional group is renowned for its ability to chelate to metals[21,22], and compounds with this group can preferentially bind to specific transition metals and metalloproteins[16,19,23].

The first isocyanide backbone enzyme, isocyanide synthase (ICS), was discovered by[24] while investigating the metabolome of an unculturable soil bacterium. Since this initial finding, a small number of ICSs have been shown to serve as backbone synthases in bacterial and fungal BGCs[23,25–27]. The only three characterized fungal ICS BGCS are: (i) *xan*[19], (ii) *crm*[16], and (iii) *dit*[28–30]. While the *dit1* (ICS) *dit2* (p450) cluster produces dityrosine, a core component in the cell wall of yeast ascospores that has no known metal association[28–30], the *xan* and *crm* BGCs are both strongly up-regulated under copper starvation[31]. Despite

the mounting evidence that ICS BGCs are a rich source of bioactive microbial natural products, many of which may have metal-associated properties[17,19,23], their exclusion from existing genome mining software has limited research on them relative to other canonical classes of SMs. Lim et al. (2018) reported a diversification of ICS proteins across both the fungal and bacterial kingdoms. However, the chemical diversity of compounds that result from these synthases depends on the extent to which ICSs function as backbones in BGCs, as opposed to existing as standalone genes[32], and has yet to be fully elucidated.

In the present study, we addressed the hypothesis that ICS genes are core members in diverse fungal BGCs. We present the first genome-mining pipeline, assembled from preexisting tools to identify fungal ICS BGCs. We discovered 3,800 ICS BGCs in 3,300 fungal genomes, making ICSs the fifth-largest class of SM compared to canonical classes found by antiSMASH[6]. In exploring this dataset, we focused on two unanswered questions: (i) What is the prevalence of ICS BGCs across different taxonomic groups? And (ii) what are the structural variations and commonalities among ICS BGCs in fungi (e.g., size, protein domain content)? Finally, we conducted additional analysis to examine the distinctive genomic and evolutionary signatures of the *dit1*/*2* gene cluster family (GCF) that was previously only described in yeast (referred to as the *dit* superfamily in this study).

2.3 Materials and Methods

2.3.1 Dataset and genomic-database annotation

All publicly available annotated fungal genomes were downloaded from the NCBI database on 09/09/2021 using the NCBI's Dataset tool, version 11.32.1. We generated protein domain predictions using HMMER v3.1b2 (e-value <= 1e-4)[33] with the Pfam database v34[34]. ICS proteins were identified based on the presence of the ICS-specific domain PF05141.1 (named 'DIT1_PvcA' in the Pfam database in reference to the *dit* cluster)[31]. All canonical BGCs were predicted using the default settings within the fungal version of

antiSMASH v5[6]. See ‘Step 1’ in the Reproducible Script for details on the code and specific parameters. See Supplemental Table S1 for information on the taxonomy and metadata of each genome.

2.3.2 Locating putative ICS BGCs from regulatory-motif conservation

Genes within a BGC often share cluster-specific promoter motifs (i.e., genetic regulatory elements) that allow for the co-regulation of a cluster by a single transcription factor[35]. To generate our BGC predictions, we examined the genes surrounding ICSs for shared promoter motifs using the eukaryotic BGC-detection algorithm, CASSIS v1[36]. All CASSIS predictions that contained one or more neighboring genes next to the ICS gene were considered BGCs and included in the subsequent analysis in accordance with past definitions of BGCs[7]. Scripts detailing the implementation of CASSIS can be found in ‘Step 2’ of the Reproducible Script.

2.3.3 Guidance for interpreting promoter-motif based BGC predictions

Resulting predictions from CASSIS assume co-regulation of all genes within a BGC and that co-regulation is mediated by a single, consistent regulatory motif[37]. A major strength of this approach is that differences in regulation may be biologically relevant (e.g., divergence of regulatory motifs might suggest a change in or loss of a cluster's functionality). However, we cannot be certain that all co-regulatory motifs across all genomes and/or genes are called correctly. We urge caution when interpreting individual predictions (or lack thereof) in individual species of interest. The focus of this study is to identify kingdom-wide patterns in ICS BGCs. At this scale, BGCs can be grouped into GCFs. This amalgamation offers additional data on the gene content of specific clusters. We have implemented strategies to harness patterns across GCF to strengthen our inferences and offer fundamental insights into ICS GCFs across the fungal kingdom.

2.3.4 Converting BGC predictions into standard file formats

Raw CASSIS predictions were run through a custom Python pipeline that converted each prediction into a GenBank (gbk), FASTA (fna), and General Feature Format (gff) file, enabling us to leverage existing BGC-processing software such as BiG-SCAPE[38], and cblaster[39]. The raw gbk, fna, and gff files for all 3,800 ICS BGCs can be downloaded from the Supplementary Repository. Coding regions (CDSes) encoding SM-related biosynthetic protein domains[6] were flagged as 'biosynthetic' because BiG-SCAPE places more emphasis on these proteins when determining the relatedness of BGCs. The scripts used in this step can be found in 'Step 3' of the Reproducible Script. Refer to Supplemental Tables 2-4 for detailed summaries of the number of ICS BGCs in each species/genome and the protein domains present in each BGC.

2.3.5 Grouping the ICS BGCs into GCFs

BGCs that belong to the same GCF are thought to produce identical or closely related natural products based on gene content and protein identity thresholds. ICS BGC gbk files were clustered into GCFs using BiG-SCAPE v1.0.1 ('Step 4' in Reproducible Script)[38]. We created a modified 'anchor file' to ensure BiG-SCAPE treated each ICS gene as a chemical-class defining synthase (i.e., a backbone gene). Parameters for ICS GCF classification were tested using 13 different cutoffs ranging from 0.2 to 0.8 in increments of 0.05. Cutoff values above 0.45 were found to be too relaxed, leading to the merging of major ICS GCFs that were separate at lower cutoffs. Conversely, cutoffs below 0.25 were too strict, forming only small GCFs comprising BGCs from closely related species. A value of 0.3 (which is also the default value) was confirmed as the optimal cutoff. A network visualization of the BiG-SCAPE output was created with Cytoscape v3.9.1[40] and can be found in Supplementary Figure S1. For more information on the dereplication pipeline run on the BiG-SCAPE output, refer to the Supplementary Methods.