Computational Approaches to Bacterial Phylogeny and Evolution of Secondary Metabolites

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M.Sc. Bita Pourmohsenin

aus Teheran / Iran

Tübingen

2026

Computational Approaches to Bacterial Phylogeny and Evolution of Secondary Metabolites

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M.Sc. Bita Pourmohsenin

aus Teheran / Iran

Tübingen

2026

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

Tag der mündlichen Qualifikation:	02.07.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatterin:	Prof. Dr. Nadine Ziemert
2. Berichterstatter:	Prof. Dr. Daniel H. Huson

Abstract

Specialized metabolites play central roles in microbial ecology, biotechnology, and medicine. Advances in sequencing technologies now make it possible to explore the biosynthetic potential of microbial genomes at unprecedented scale, offering new opportunities to discover bioactive compounds and understand their ecological functions. However, translating this wealth of genomic information into meaningful biological insight remains a major challenge. Biosynthetic pathways are often modular, frequently reshaped by horizontal gene transfer, and not always directly linked to associated functions such as transport or regulation. As a result, it is difficult to determine how these systems originate, diversify, and persist across bacterial lineages.

In this dissertation, siderophore and metallophore systems, which mediate microbial metal acquisition and interspecies interactions, were investigated as a model for complex secondary metabolite traits. By combining phylogeny-aware genome mining with large-scale comparative analyses, the distribution and evolutionary histories of biosynthetic pathways and transport systems were examined in the context of ecological settings and cell-envelope architecture.

To support these analyses, AutoMLST2 was developed as a scalable Genome Taxonomy Database (GTDB)-integrated platform for species-tree reconstruction and taxonomic placement. In another study, chelator-centric detection rules for nonribosomal peptide synthase (NRPS)-derived metallophores were implemented, extending functional annotation within existing biosynthetic gene clusters and revealing that metal-binding secondary metabolites are more widespread and structurally diverse than previously recognized.

Using this framework, siderophore biosynthetic pathways and associated transport systems were analyzed at large scale, revealing distinct evolutionary strategies among siderophore types and highlighting transport systems as key constraints shaped by cell-envelope architecture.

Beyond siderophore biology, this work demonstrates how combining refined annotations with a robust phylogenomic framework enables large-scale inference of evolutionary patterns across bacterial genomes. While siderophores serve as a focused case study, the approaches developed here are broadly applicable to other secondary metabolites and complex genomic traits, providing a general strategy for investigating how microbial metabolic systems diversify, persist, and adapt under structural and evolutionary constraints.

Kurzfassung

Spezialisierte Metabolite spielen eine zentrale Rolle in mikrobiellen Ökosystemen sowie in biotechnologischen und medizinischen Anwendungen. Fortschritte in Sequenzierungstechnologien ermöglichen es heute, das biosynthetische Potenzial mikrobieller Genome in bislang unerreichter Breite zu erfassen und eröffnen neue Möglichkeiten zur Entdeckung bioaktiver Verbindungen sowie zum Verständnis ihrer ökologischen Funktionen. Die Interpretation dieser umfangreichen genomischen Daten bleibt jedoch eine zentrale Herausforderung. Biosynthetische Signalwege sind häufig modular aufgebaut, werden durch horizontalen Gentransfer geprägt und sind nicht immer direkt mit Funktionen wie Transport oder Regulation verknüpft. Dies erschwert das Verständnis ihrer Entstehung, Diversifizierung und langfristigen Etablierung in bakteriellen Linien.

In der vorliegenden Arbeit werden Siderophor- und Metallophor-Systeme, die an der mikrobiellen Metallakquisition und an interspezifischen Interaktionen beteiligt sind, als Modellsystem für komplexe sekundärmetabolische Merkmale untersucht. Durch die Kombination phylogenie-sensitiver Genom-Mining-Ansätze mit großskaligen vergleichenden Analysen wurden die Verteilung sowie die evolutionären Muster biosynthetischer Signalwege und zugehöriger Transportsysteme im Kontext ökologischer Rahmenbedingungen und der Zellhüllenarchitektur analysiert.

Zur methodischen Unterstützung dieser Analysen wurde AutoMLST2 als skalierbare, in die Genome Taxonomy Database (GTDB) integrierte Plattform zur Rekonstruktion von Artenbäumen und zur taxonomischen Einordnung von Genomen entwickelt. Parallel dazu wurden chelator-zentrierte Erkennungsregeln für von nicht-ribosomalen Peptidsynthetasen (NRPS) abgeleitete Metallophor-Biosynthesewege implementiert. Diese erweitern die funktionelle Annotation bestehender biosynthetischer Gencluster und zeigen, dass metallbindende sekundäre Metabolite weiter verbreitet und strukturell vielfältiger sind als bislang angenommen.

Auf Grundlage dieses Rahmens wurden Siderophor-Biosynthesewege und zugehörige Transportsysteme in großem Maßstab analysiert. Dabei zeigen sich unterschiedliche evolutionäre Strategien zwischen verschiedenen Siderophortypen, wobei Transportsysteme als zentrale Einschränkungen wirken, die durch die Architektur der bakteriellen Zellhülle bestimmt werden.

Über die Siderophorbiologie hinaus zeigt diese Arbeit, wie die Kombination verfeinerter Annotationen mit einem robusten phylogenomischen Rahmenwerk großskalige Rückschlüsse auf evolutionäre Muster in bakteriellen Genomen ermöglicht. Obwohl Siderophore hier als fokussiertes Modellsystem dienen, sind die entwickelten Ansätze auf andere Klassen sekundärer Metabolite sowie auf weitere komplexe genomische Merkmale übertragbar

und bieten eine allgemeine Strategie zur Untersuchung der Diversifizierung, Persistenz und Anpassung mikrobieller Stoffwechselsysteme unter strukturellen und evolutionären Einschränkungen.

Acknowledgments

This dissertation was shaped not only by research, but by the support and kindness of many people along the way, and I am deeply grateful to all of them.

First and foremost, I would like to thank my supervisor, Nadine, who has been more than an academic advisor over these years. Her support, trust, and mentorship shaped not only this thesis but also my confidence as a scientist. Through exciting successes and frustrating dead ends, and everything in between, she always encouraged independence while making sure I was never truly on my own. I am incredibly grateful for the supportive environment she has created, and for the example she has set of what good mentorship looks like.

I am very grateful to Athina and Monica for welcoming me so quickly into the Council of Hobbits and making the office feel like a safe space rather than just a workplace. Athina, thank you for being my "tea" buddy, office-mate, and steady presence throughout this journey. I would also like to thank Direnc and Martina, who guided me patiently in my early steps in research, and Caner, our terminal wizard, who somehow solves problems faster than I can finish reading the error message. Thank you for always being willing to help, often saving entire days with a single command. I would like to thank my friend Claude, who was my copilot throughout all the coding, and Libera, who is a constant reminder that AI cannot replace human curiosity, intuition, and care. To my later office companions, Semih, Sasha, and Noel, thank you for the shared coffees and cookies, conversations, and quiet moments of parallel work. You made the long days lighter and the office a genuinely pleasant place to be. I am also grateful to everyone whose names are not mentioned here but who contributed through discussions, encouragement, and support along the way.

My deepest gratitude goes to my parents. My mother, Monir, taught me to stay curious, to ask questions, and to never stop pursuing what I want while staying kind

and open-minded. My father, Fariborz, always placed my education above everything else, offering unwavering support and encouragement, even when it meant sacrifice and distance.

My heartfelt thanks go to my friends who made Germany feel like home. Mary, who welcomed me when I first arrived, opened her home to me, and has been a best friend I can always rely on. Kimia, Majid, and Narjes, who later became my second family, bonded through shared experiences and resilience. Kimia, together with our feline flatmates Goofy and Rara, contributed through keyboard interference, moral support during late nights, and constant reminders that work-life balance matters. And finally, Dani, who somehow matches my level of anxiety and still manages to calm me down, reminding me that even the most stressful moments are easier when shared.

"So once you do know what the question actually is, you'll know what the answer means."

Contents

List of Abbreviations

ABC ATP-binding cassette

AleRax Algorithm for reconciliation analysis

AMR Antimicrobial resistance

antiSMASH Antibiotics & Secondary Metabolite Analysis Shell

BGC Biosynthetic gene cluster

DTL Duplication–transfer–loss

eMPRess Efficient Maximum Parsimony Reconciliation

Fe^{3+} Ferric iron

GTDB Genome Taxonomy Database

GTDB-Tk Genome Taxonomy Database Toolkit

HGT Horizontal gene transfer

HMM Hidden Markov model

IucA/IucC NRPS-independent siderophore biosynthetic enzymes

MFS Major Facilitator Superfamily

MIBiG Minimum Information about a Biosynthetic Gene cluster

MLST Multilocus sequence typing

NIS NRPS-independent siderophore

NRP Nonribosomal peptide

NRPS Nonribosomal peptide synthetase

RED Relative evolutionary divergence

REDgroup Taxonomic grouping based on relative evolutionary divergence

TBDR TonB-dependent receptor

TonB–ExbB–ExbD TonB energy-transduction complex

Zn^{2+} Zinc ion

Cu^{2+} Copper ion

Mn^{2+} Manganese ion

Introduction

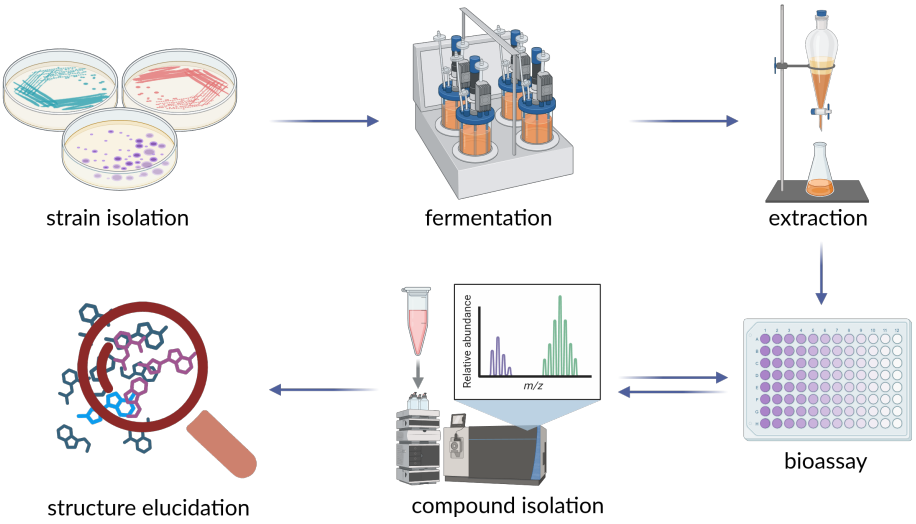
Background and Motivation

Natural products, also called specialized or secondary metabolites, have long served as a rich source of bioactive compounds. These molecules mediate microbial defense, signaling, nutrient acquisition, and symbiosis, and they have given rise to many of our most important drugs, especially in the fields of infectious disease and cancer therapy[1].

Despite their importance, only a small fraction of the microbial natural-product repertoire has been experimentally characterized. Recent large-scale analyses estimate that less than 3% of the potential encoded in bacterial genomes is known[2]. The vast remaining "dark matter" of chemical diversity, therefore, represents a major opportunity for discovery.

Historically, discovery relied on **bioactivity-guided isolation**, a stepwise process involving environmental sampling, microbial cultivation, extract screening, active-compound purification, and structure elucidation (Figure 1a). While transformative during the antibiotic “golden age,” this workflow is increasingly slow, resource-intensive, and prone to rediscovery of known scaffolds[3]. Concurrently, the global crisis of antimicrobial resistance (AMR) underscores the urgent need for novel compounds and delivery strategies, as multidrug-resistant pathogens threaten to reverse decades of progress in infectious disease treatment[4].

In this context, **genome mining** provides a transformative alternative. Natural products are typically encoded in co-localized sets of genes responsible for core biosynthesis, tailoring, regulation, and transport, collectively referred to as **biosynthetic gene clusters (BGCs)**. Bioinformatics tools can detect candidate pathways directly from sequence data, prioritize likely novel producers, and circumvent several limitations of classical discovery



(a)

(b)

Figure 1: **Comparison of classical natural-product discovery and genome-mining-based workflows. (a)** Classical bioactivity-guided discovery pipeline, in which environmental samples are cultivated, screened for biological activity, and subjected to iterative purification and structure elucidation. **(b)** Genome-mining-based discovery workflow, where sequencing data are used to identify biosynthetic gene clusters (BGCs), prioritize novelty in silico, and guide targeted experimental validation.

workflows[1, 5]. Figure 1b contrasts these two discovery paradigms, highlighting the shift from activity-driven screening to sequence-driven prioritization.

From Genome Mining to Function

The advent of high-throughput sequencing has transformed natural-product discovery by shifting the focus from phenotypic screens to genome-guided exploration. Rather than relying solely on bioassays, researchers can now analyze genomic data to predict and prioritize biosynthetic potential. Dedicated algorithms detect characteristic domain architectures, compare them to reference databases, and infer the likely product class, novelty, and taxonomic distribution[6, 7].

A key platform is **antiSMASH**, which scans microbial genomes for signature enzymes and predicts cluster types. Version 5 introduced faster algorithms and improved cluster-type recognition[8], while version 7 expanded support to 81 cluster classes and enhanced chemical-structure prediction and visualization[5]. Examples of such biosynthetic gene clusters and their corresponding chemical products are shown in Figure 22, illustrating how conserved enzymatic functions within BGCs can give rise to shared metal-binding motifs across structurally distinct siderophores.

These computational approaches have revealed the immense scale of biosynthetic diversity and created a need for rational **prioritization**. Strategies such as resistance-gene detection, novelty scoring, and phylogenetic filtering help select clusters most likely to encode new chemistry[9, 10]. Genome mining, therefore, serves not only as an annotation tool but as a functional bridge between sequence, chemistry, and ecology.

Interpreting this diversity, however, requires a reliable phylogenomic context. The evolutionary histories of biosynthetic systems, marked by horizontal gene transfer, duplication, and lineage-specific loss, can only be understood in the framework of high-quality species trees. Comparative analyses that integrate genome-mining with robust phylogenies reveal how chemical traits correlate with taxonomy and environment[11].

In this dissertation, genome mining is applied in two complementary ways. First, new detection rules for **metallophore-type biosynthetic clusters** were developed to improve sensitivity and functional coverage (Manuscript II, Chapter 2). Second, a scalable **phylogenomic framework** was developed in Manuscript I (Chapter 1) to contextualize BGC distributions across the bacterial tree of life. Together, these resources enabled the comparative evolutionary study of **siderophore biosynthesis and transport systems** presented in Manuscript III (Chapter 3).

Siderophores as a Model System

Among the diverse families of natural products, **siderophores** occupy a unique ecological and biomedical niche. Iron is essential for almost all forms of life, but under aerobic and neutral-pH conditions, ferric iron (Fe^{3+}) is poorly soluble and therefore biologically scarce. To overcome this limitation, bacteria and fungi secrete small, high-affinity iron chelators called **siderophores** that solubilize ferric iron and transport it into the cell[12].

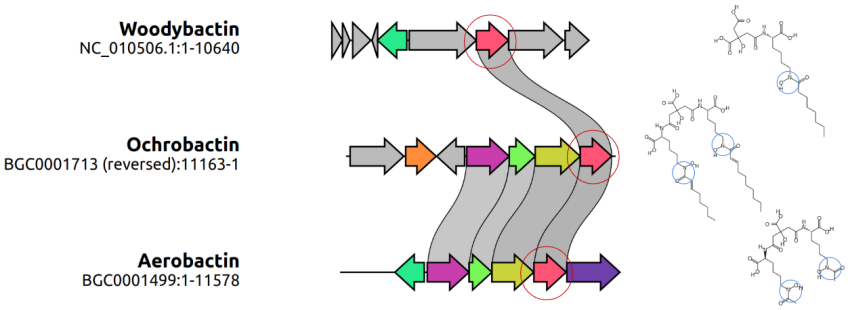


Figure 2: **Representative siderophore biosynthetic gene clusters (BGCs)** from the MiBIG database producing hydroxamate-type siderophores. Gene cluster architectures for woodybactin, ochrobactin, and aerobactin are shown together with their corresponding molecular structures. Despite differences in gene organization and overall scaffold architecture, all three pathways generate siderophores containing hydroxamate functional groups (blue circles), highlighting a shared chemical strategy for iron chelation that can often be linked to conserved enzymatic steps such as lysine monooxygenase (red circles) within the biosynthetic pathways.

This strategy is critical for microbial survival in competitive habitats such as soil, marine systems, and host tissues, where iron availability is tightly restricted by environmental chemistry or host immune sequestration.

Siderophores illustrate how **metabolic innovation** underpins microbial adaptation. Their remarkable chemical diversity arises mainly from two biosynthetic paradigms:

1. **Non-ribosomal peptide synthetase (NRPS) systems**, which assemble peptides modularly from amino acids and derivatives, yielding chelators such as hydroxamate or catecholate (examples include enterobactin and ferrichrome); and

2. **NRPS-independent siderophore (NIS) synthetases**, which use ATP-dependent condensation enzymes (IucA/IucC-type) to form simpler citrate- or ornithine-derived scaffolds like aerobactin or rhizobactin[13].

Together, these pathways span a wide taxonomic range and exemplify the modularity and evolutionary plasticity of microbial metabolism.

Beyond iron, many organisms produce related **metallophores** that chelate other transition metals such as (Zn^{2+}), (Cu^{2+}), and (Mn^{2+})[14]. A well-known example is *Staphylococcus aureus* **staphylopine**, a broad-spectrum metallophore that supports growth under "nutritional immunity" when hosts sequester essential metals[15]. Such systems reveal the ecological reach of metal-binding metabolites and their role in microbial competition

and pathogenesis.

Chemically, siderophores are also examples of multifunctionality. Amphiphilic variants such as marinobactins and amphibactins possess hydrophobic fatty-acid tails that promote self-assembly at interfaces, combining metal-chelation with surfactant properties[16]. This dual behavior links siderophores to other natural-product classes such as biosurfactants, underscoring how evolutionary pressure can repurpose chemical scaffolds for multiple ecological functions.

From a biomedical perspective, siderophores also represent promising **drug-delivery agents**. Their ability to enter bacterial cells via dedicated uptake systems has been exploited in the design of "**Trojan horse antibiotics**," where an antibiotic is covalently linked to a siderophore moiety to hijack iron-transport pathways of resistant pathogens[17]. These applications highlight how understanding siderophore biology can inform both natural-product discovery and antimicrobial innovation.

For these reasons, siderophores and metallophores serve as an ideal model system for connecting genome mining with evolutionary and functional analysis. Their wide distribution, modular enzymology, and ecological significance provide a powerful framework for studying how secondary metabolite biosynthesis evolves, spreads via horizontal transfer, and interfaces with cellular physiology. This dissertation anchors the transition from methodological developments (enhanced detection and phylogenomic frameworks) to biological insight, tracing how siderophore biosynthesis and transport systems diversified across bacterial evolution.

Transport Systems and Functional Coupling

The biological activity of siderophores does not end with metal chelation; the ferric-siderophore complex must be efficiently **transported into the cell**, a process requiring dedicated membrane transport systems. These transporters are as integral to siderophore function as the biosynthetic enzymes themselves and are subject to the same ecological and evolutionary pressures. They also offer unique insight into the link between **metabolic innovation** and **cellular architecture**.

In **Gram-negative bacteria**, uptake typically begins at the outer membrane via

TonB-dependent receptors (TBDRs), β-barrel proteins that recognize specific ferric-siderophore complexes and import them using energy transduced from the inner membrane through the TonB-ExbB-ExbD complex[18, 19].

Once inside the periplasm, siderophores are shuttled by periplasmic binding proteins to **ATP-binding cassette (ABC) importers**, which move the complex across the inner membrane in an ATP-dependent process[20].

In **Gram-positive bacteria**, which lack an outer membrane, siderophore import is mediated directly by ABC-type transporters equipped with extracellular or lipoprotein-anchored substrate-binding domains[21]. **Major Facilitator Superfamily (MFS)** transporters often complement these systems, functioning as efflux or low-affinity uptake channels[12].

The genomic organization of these transporters frequently mirrors that of their cognate biosynthetic genes. Siderophore biosynthetic gene clusters (BGCs) often co-localize with genes encoding import or export proteins, suggesting a history of co-adaptation between production and uptake[22].

However, this association is not universal: many genomes encode siderophore transporters without the corresponding biosynthetic genes, indicating ecological "cheating" or cross-feeding, where organisms exploit siderophores secreted by others[23].

Understanding when these systems evolve together and when they diverge is crucial for reconstructing the **evolutionary logic of microbial cooperation and competition**.

The diversity and modularity of transporter systems also create opportunities for **functional prediction**. Because transporter specificity often parallels siderophore chemistry, the presence of certain transporter domain families (e.g., TBDR + TonB vs ABC FecCD + binding-protein domains) can be used to infer the underlying siderophore type or producer physiology[22].

This connection inspired recent efforts to build **machine-learning classifiers** that predict siderophore classes from transporter profiles, an approach further explored in the comparative analyses of this dissertation.

In summary, siderophore transporters not only complete the biochemical cycle of iron acquisition but also act as evolutionary markers of how bacteria adapt their metabolisms to distinct cell envelopes and ecological niches.

Their distribution, architecture, and co-occurrence with biosynthetic genes provide a rich framework for investigating the **co-evolution of biosynthesis and transport**, the central focus of the final part of this work.

Integrating Phylogenomics: AutoMLST2 as an Enabling Tool

A comprehensive understanding of the distribution and evolution of siderophore biosynthesis and transport systems demands more than accurate gene detection. It requires a *rigorous phylogenomic framework* that can organize thousands of bacterial genomes within a coherent evolutionary hierarchy.

Comparative genomics of biosynthetic traits is only meaningful when each genome can be positioned precisely within a high-quality species tree. Yet, despite the availability of large reference resources such as the **Genome Taxonomy Database (GTDB)**, practical barriers have long limited the accessibility of such analyses to the broader natural-product community.

Existing workflows for bacterial phylogeny reconstruction often rely on small, static sets of marker genes (e.g., the classical *MLST* schemes or single-gene phylogenies), which provide limited resolution for closely related taxa[24].

Moreover, building genome-scale trees for hundreds or thousands of genomes typically requires substantial computational expertise and resources, factors that hinder their integration into genome-mining pipelines. To overcome these limitations, this work developed **AutoMLST2**, a redesigned, scalable platform that provides automated, high-resolution phylogenomic analysis directly from genome assemblies[25]

1. **Dynamic marker selection.** Building on AutoMLST's strategy of selecting single-copy markers present across the dataset, **AutoMLST2** expands the **candidate marker library to >2,800 curated HMMs** and selects an optimized multi-locus set per analysis, improving resolution, especially among closely related genomes, while keeping runs accessible via the web interface.

2. **GTDB integration.** Each analysis is anchored to the GTDB reference tree and

taxonomy, ensuring consistency with the most up-to-date, genome-based bacterial classification[26].

3. **Simplified placement workflow.** A newly implemented rapid-placement algorithm enables users to position novel genomes or MAGs within the GTDB backbone in seconds, maintaining accuracy while drastically reducing computational cost.

4. **Web-accessible and reproducible.** Through its intuitive web interface and Docker-based backend, AutoMLST2 brings robust phylogenomic reconstruction to non-specialist users while maintaining full reproducibility and scalability for large datasets.

These features make AutoMLST2 particularly suited for **phylogeny-aware genome mining**, where biosynthetic and transporter annotations can be interpreted within a robust evolutionary framework.

Within this dissertation, AutoMLST2 provided a **species-tree backbone** for mapping siderophore and metallophore traits across more than 60,000 GTDB representative bacterial genomes, enabling analyses of **trait distribution, evolutionary origin, and co-adaptation** between biosynthetic and transport modules.

Building on this foundation, the subsequent section introduces the **reconciliation framework** used to compare gene and species trees, allowing the inference of **duplication, transfer, and loss (DTL)** events that shape siderophore-system evolution.

Reconciliation Algorithms for Evolutionary Inference

Understanding how siderophore and metallophore systems evolved requires not only a reliable species tree but also a framework for interpreting the histories of the **individual gene families** involved. Biosynthetic enzymes and transporters often show phylogenetic patterns that differ from the species tree, reflecting processes such as **gene duplication**, **horizontal gene transfer (HGT)**, and **gene loss**, all of which are common in modular, ecologically important pathways.

Gene tree-species tree reconciliation provides a principled approach for inferring these events by mapping each gene tree onto the species tree and identifying where and

when evolutionary changes likely occurred. General overviews of reconciliation theory and its application in microbial evolution are provided in recent reviews[27].

Parsimony-based reconciliation A widely used class of algorithms relies on **maximum parsimony**, which seeks the evolutionary scenario with the smallest total number (or cost) of **duplication (D)**, **transfer (T)**, and **loss (L)** events required to reconcile a gene tree with the species tree.

These methods are computationally efficient, scalable, and highly interpretable, qualities essential when analysing hundreds or thousands of gene families.

In this dissertation, we used **eMPRess**, a modern parsimony-based reconciliation tool designed for reproducibility and systematic exploration of near-optimal solutions under alternative cost schemes. eMPRess generates representative reconciliations and quantifies how robust inferred DTL events are to reasonable parameter variation[28].

Probabilistic (likelihood-based) approaches Reconciliation accuracy depends critically on the underlying **gene trees**, which can carry substantial uncertainty, particularly for horizontally transferred, domain-rich, or poorly aligned protein families.

To validate key inferences from parsimony-based reconciliation, a subset of siderophore biosynthetic and transporter families was re-analyzed using **AleRax**, a likelihood-based framework that jointly infers gene trees and their reconciliations under an explicit DTL model.

AleRax samples alternative gene trees and reconciliations to capture uncertainty and evaluate the robustness of major inferred events[29–32].

Practical Implementation in this Dissertation Importantly, reconciliation requires a **fixed species tree**, against which gene-tree histories are interpreted.

In our analyses, we used an **adapted version of the Genome Taxonomy Database (GTDB) species tree**, which provides a consistent, genome-based phylogenomic framework across bacteria[26].

This ensured that all reconciliation results were grounded in a standardized evolutionary backbone that reflects current bacterial taxonomy. Within this framework:

- **GTDB species tree** served as the reference phylogeny for all reconciliations.
- **eMPRess** was used to infer maximum-parsimony DTL scenarios for siderophore

biosynthetic enzymes and transporter domain families across >60,000 genomes.

- **AleRax** was applied to selected families to validate key findings under a probabilistic DTL model and to assess the stability of inferences under gene-tree uncertainty.

Why reconciliation is essential for siderophore-system evolution

Siderophore and metallophore systems are shaped by strong ecological pressures, such as nutrient limitation, competition, and host interaction, and evolve through frequent gain, loss, and exchange of genes. Reconciliation allows us to quantify these processes and to determine:

- where siderophore pathways **originated**,
- how often they were **laterally transferred** across bacterial lineages,
- when transporter families evolved **in concert** with biosynthesis, and
- where biosynthesis and transport **decoupled**, giving rise to pirate or cross-feeding strategies.

Together, these methods form the basis of the evolutionary analysis presented in Manuscript III (Chapter 3), where reconciliation, phylogenomic mapping, and trait correlation are combined to reconstruct the diversification and co-adaptation of siderophore biosynthesis and transport across the bacterial tree of life.

Objectives and Structure of the Dissertation

The overarching objective of this dissertation is to develop and apply a **phylogeny-aware genome-mining framework** to investigate the diversity, distribution, and evolutionary dynamics of **siderophore and metallophore systems** across the bacterial tree of life.

Understanding these systems requires more than mere cataloguing of biosynthetic genes; it requires an analytical approach that can place these traits within their broader evolutionary and ecological contexts. Existing genome-mining tools generally lack the capacity to capture evolutionary relationships, and phylogenetic workflows rarely

incorporate secondary metabolite annotation. To bridge this gap, the present work introduces an integrated, phylogeny-aware genome-mining framework designed to detect, contextualize, and compare siderophore biosynthesis and transport pathways across diverse bacterial lineages.

This required the integration of three complementary components:

1. **Method development**: improving the detection and annotation of metallophore biosynthetic pathways.
2. **Phylogenomic infrastructure**: enabling scalable, high-resolution evolutionary analysis.
3. **Comparative evolutionary synthesis**: reconstructing how siderophore biosynthesis and transport systems diversified and co-adapted.

These components form the scientific arc of the dissertation, which is organized into the following main chapters:

AutoMLST2: a web server for phylogeny and microbial taxonomy

(*Published in Nucleic Acids Research, 2025*) This chapter presents **AutoMLST2**, a redesigned and GTDB-integrated web server for automated species-tree reconstruction and taxonomic placement.

The tool introduces a scalable backend, expanded HMM libraries, and a simplified placement workflow, providing the **phylogenomic foundation** used throughout this dissertation for trait mapping and evolutionary analyses.

Automated genome mining predicts structural diversity and taxonomic distribution of peptide metallophores across bacteria

(Published in eLife, 2026)

This chapter describes the development of chelator-specific HMMs and detection rules to systematically annotate NRP metallophore biosynthetic gene clusters in antiSMASH. Applied to over 60,000 bacterial genomes, this framework reveals previously unrecognised

diversity and distribution patterns of metallophores, providing the functional annotations used in subsequent evolutionary analyses.

Evolution of Siderophore Biosynthesis and Transport Systems in Bacteria

(Advanced Manuscript; ready for submission)

Building on the functional and phylogenomic foundations of Chapters 1 and 2, this chapter investigates the evolutionary history of siderophore biosynthesis and transport systems. Using GTDB species trees, gene tree-species tree reconciliations (eMPRess, AleRax), and large-scale trait-mapping, it reconstructs duplication, transfer, and loss dynamics and proposes a stepwise model for the origin, diversification, and co-adaptation of siderophore systems.

General Discussion

Together, these chapters demonstrate how the combination of **genome mining**, **phylogenomics**, and **reconciliation-based evolutionary inference** provides a coherent framework for studying complex secondary metabolite systems at scale.

The methods and analyses presented here enhance our understanding of microbial metal acquisition, support new strategies for natural-product discovery, and open opportunities for translational applications such as **siderophore-antibiotic (Trojan horse) drug delivery**.

List of publications included in the thesis

AutoMLST2: a web server for phylogeny and microbial taxonomy

Bita Pourmohsenin, Arthur Wiese, Nadine Ziemert, AutoMLST2: a web server for phylogeny and microbial taxonomy, Nucleic Acids Research, Volume 53, Issue W1, 7 July 2025, Pages W45-W50, https://doi.org/10.1093/nar/gkaf397

Automated genome mining predicts structural diversity and taxonomic distribution of peptide metallophores across bacteria

Zachary L Reitz, **Bita Pourmohsenin,** Melanie Susman, Emil Thomsen, Daniel Roth, Alison Butler, Nadine Ziemert, Marnix H Medema,(2026) Automated genome mining predicts structural diversity and taxonomic distribution of peptide metallophores across bacteria. eLife, 14:RP109154. https://doi.org/10.7554/eLife.109154.3

Evolution of Siderophore Biosynthesis and Transport Systems in Bacteria

Bita Pourmohsenin, Nadine Ziemert

List of publications not included in the thesis

MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters

Barbara R Terlouw, Kai Blin, Jorge C Navarro-Muñoz, Nicole E Avalon, Marc G Chevrette, Susan Egbert, Sanghoon Lee, David Meijer, Michael J J Recchia, Zachary L Reitz, Jeffrey A van Santen, Nelly Selem-Mojica, Thomas Tørring, Liana Zaroubi, Mohammad Alanjary, Gajender Aleti, César Aguilar, Suhad A A Al-Salihi, Hannah E Augustijn, J Abraham Avelar-Rivas, Luis A Avitia-Domínguez, Francisco Barona-Gómez, Jordan Bernaldo-Agüero, Vincent A Bielinski, Friederike Biermann, Thomas J Booth, Victor J Carrion Bravo, Raquel Castelo-Branco, Fernanda O Chagas, Pablo Cruz-Morales, Chao Du, Katherine R Duncan, Athina Gavriilidou, Damien Gayrard, Karina Gutiérrez-García, Kristina Haslinger, Eric J N Helfrich, Justin J J van der Hooft, Afif P Jati, Edward Kalkreuter, Nikolaos Kalyvas, Kyo Bin Kang, Satria Kautsar, Wonyong Kim, Aditya M Kunjapur, Yong-Xin Li, Geng-Min Lin, Catarina Loureiro, Joris J R Louwen, Nico L L Louwen, George Lund, Jonathan Parra, Benjamin Philmus, **Bita Pourmohsenin**, Lotte J U Pronk, Adriana Rego, Devasahayam Arokia Balaya Rex, Serina Robinson, L Rodrigo Rosas-Becerra, Eve T Roxborough, Michelle A Schorn, Darren J Scobie, Kumar Saurabh Singh, Nika Sokolova, Xiaoyu Tang, Daniel Udwary, Aruna Vigneshwari, Kristiina Vind, Sophie P J M Vromans, Valentin Waschulin, Sam E Williams, Jaclyn M Winter, Thomas E Witte, Huali Xie, Dong Yang, Jingwei Yu, Mitja Zdouc, Zheng Zhong, Jérôme Collemare, Roger G Linington, Tilmann Weber, Marnix H Medema,

MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Research*, Volume 51, Issue D1, 6 January 2023, Pages D603–D610, https://doi.org/10.1093/nar/gkac1049.

MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration

Mitja M Zdouc, Kai Blin, Nico L L Louwen, Jorge Navarro, Catarina Loureiro, Chantal D Bader, Constance B Bailey, Lena Barra, Thomas J Booth, Kenan A J Bozhüyük, José D D Cediel-Becerra, Zachary Charlop-Powers, Marc G Chevrette, Yit Heng Chooi, Paul M D'Agostino, Tristan de Rond, Elena Del Pup, Katherine R Duncan, Wenjia Gu, Novriyandi Hanif, Eric J N Helfrich, Matthew Jenner, Yohei Katsuyama, Aleksandra Korenskaia, Daniel Krug, Vincent Libis, George A Lund, Shrikant Mantri, Kalindi D Morgan, Charlotte Owen, Chin-Soon Phan, Benjamin Philmus, Zachary L Reitz, Serina L Robinson, Kumar Saurabh Singh, Robin Teufel, Yaojun Tong, Fidele Tugizimana, Dana Ulanova, Jaclyn M Winter, César Aguilar, Daniel Y Akiyama, Suhad A A Al-Salihi, Mohammad Alanjary, Fabrizio Alberti, Gajender Aleti, Shumukh A Alharthi, Mariela Y Arias Rojo, Amr A Arishi, Hannah E Augustijn, Nicole E Avalon, J Abraham Avelar-Rivas, Kyle K Axt, Hellen B Barbieri, Julio Cesar J Barbosa, Lucas Gabriel Barboza Segato, Susanna E Barrett, Martin Baunach, Christine Beemelmanns, Dardan Beqaj, Tim Berger, Jordan Bernaldo-Agüero, Sandra M Bettenbühl, Vincent A Bielinski, Friederike Biermann, Ricardo M Borges, Rainer Borriss, Milena Breitenbach, Kevin M Bretscher, Michael W Brigham, Larissa Buedenbender, Brodie W Bulcock, Carolina Cano-Prieto, João Capela, Victor J Carrion, Riley S Carter, Raquel Castelo-Branco, Gabriel Castro-Falcón, Fernanda O Chagas, Esteban Charria-Girón, Ayesha Ahmed Chaudhri, Vasvi Chaudhry, Hyukjae Choi, Yukyung Choi, Roya Choupannejad, Jakub Chromy, Melinda S Chue Donahey, Jérôme Collemare, Jack A Connolly, Kaitlin E Creamer, Max Crüsemann, Andres Arredondo Cruz, Andres Cumsille, Jean-Felix Dallery, Luis Caleb

Damas-Ramos, Tito Damiani, Martinus de Kruijff, Belén Delgado Martín, Gerardo Della Sala, Jelle Dillen, Drew T Doering, Shravan R Dommaraju, Suhan Durusu, Susan Egbert, Mark Ellerhorst, Baptiste Faussurier, Artem Fetter, Marc Feuermann, David P Fewer, Jonathan Foldi, Andri Frediansyah, Erin A Garza, Athina Gavriilidou, Andrea Gentile, Jennifer Gerke, Hans Gerstmans, Juan Pablo Gomez-Escribano, Luz A González-Salazar, Natalie E Grayson, Claudio Greco, Juan E Gris Gomez, Sebastian Guerra, Shaday Guerrero Flores, Alexey Gurevich, Karina Gutiérrez-García, Lauren Hart, Kristina Haslinger, Beibei He, Teo Hebra, Jethro L Hemmann, Hindra Hindra, Lars Höing, Darren C Holland, Jonathan E Holme, Therese Horch, Pavlo Hrab, Jie Hu, Thanh-Hau Huynh, Ji-Yeon Hwang, Riccardo Iacovelli, Dumitrita Iftime, Marianna Iorio, Sidharth Jayachandran, Eunah Jeong, Jiayi Jing, Jung J Jung, Yuya Kakumu, Edward Kalkreuter, Kyo Bin Kang, Sangwook Kang, Wonyong Kim, Geum Jin Kim, Hyunwoo Kim, Hyun Uk Kim, Martin Klapper, Robert A Koetsier, Cassandra Kollten, Ákos T Kovács, Yelyzaveta Kriukova, Noel Kubach, Aditya M Kunjapur, Aleksandra K Kushnareva, Andreja Kust, Jessica Lamber, Martin Larralde, Niels J Larsen, Adrien P Launay, Ngoc-Thao-Hien Le, Sarah Lebeer, Byung Tae Lee, Kyungha Lee, Katherine L Lev, Shu-Ming Li, Yong-Xin Li, Cuauhtémoc Licona-Cassani, Annette Lien, Jing Liu, Julius Adam V Lopez, Nataliia V Machushynets, Marla I Macias, Taifo Mahmud, Matiss Maleckis, Añadir Maharai Martinez-Martinez, Yvonne Mast, Marina F Maximo, Christina M McBride, Rose M McLellan, Khyati Mehta Bhatt, Chrats Melkonian, Aske Merrild, Mikko Metsä-Ketelä, Douglas A Mitchell, Alison V Müller, Giang-Son Nguyen, Hera T Nguyen, Timo H J Niedermeyer, Julia H O'Hare, Adam Ossowicki, Bohdan O Ostash, Hiroshi Otani, Leo Padva, Sunaina Paliyal, Xinya Pan, Mohit Panghal, Dana S Parade, Jiyoon Park, Jonathan Parra, Marcos Pedraza Rubio, Huong T Pham, Sacha J Pidot, Jörn Piel, **Bita Pourmohsenin**, Malik Rakhmanov, Sangeetha Ramesh, Michelle H Rasmussen, Adriana Rego, Raphael Reher, Andrew J Rice, Augustin Rigolet, Adriana Romero-Otero, Luis Rodrigo Rosas-Becerra, Pablo Y Rosiles, Adriano Rutz, Byeol Ryu, Libby-Ann Sahadeo, Murrel Saldanha, Luca Salvi, Eduardo Sánchez-Carvajal, Christian Santos-Medellin, Nicolau Sbaraini, Sydney M Schoellhorn, Clemens Schumm, Ludek Sehnal, Nelly Selem, Anjali D Shah, Tania K Shishido, Simon Sieber, Velina Silviani, Garima Singh, Hemant Singh, Nika Sokolova, Eva C Sonnenschein, Margherita Sosio, Sven T Sowa, Karin Steffen, Evi Stegmann, Alena B Streiff, Alena Strüder, Frank Surup, Tiziana Svenningsen, Douglas Sweeney, Judit Szenei, Azat Tagirdzhanov, Bin Tan, Matthew J Tarnowski, Barbara R Terlouw, Thomas Rey, Nicola U Thome, Laura Rosina Torres Ortega, Thomas Tørring, Marla Trindade, Andrew W Truman, Marie Tvilum, Daniel W Udwary, Christoph Ulbricht, Lisa Vader, Gilles P van Wezel, Max Walmsley, Randika Warnasinghe, Heiner G Weddeling, Angus N M Weir, Katherine Williams, Sam E Williams, Thomas E Witte, Steffaney M Wood Rocca, Keith Yamada, Dong Yang, Dongsoo Yang, Jingwei Yu, Zhenyi Zhou, Nadine Ziemert, Lukas Zimmer, Alina Zimmermann, Christian Zimmermann, Justin J J van der Hooft, Roger G Linington, Tilmann Weber, Marnix H Medema,

MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Research*, Volume 53, Issue D1, 6 January 2025, Pages D678–D690, https://doi.org/10.1093/nar/gkae1115.

Author Contributions

Manuscript I

AutoMLST2: a web server for phylogeny and microbial taxonomy *Nucleic Acids Research* (Web Server Issue, 2025)

I am the first author of this publication, and contributed the majority of the work. I led the conceptualization of AutoMLST2 together with my supervisor, Prof. Dr. Nadine Ziemert, building on the legacy AutoMLST framework and defining the scope and design of the updated tool. I implemented most of the software development, including large parts of the backend architecture as well as the web-based frontend, and integrated the phylogenomic workflows and GTDB-based taxonomic placement.

I performed the majority of the data analysis, testing, and benchmarking, and was responsible for writing the manuscript, including figures, documentation, and revisions. Arthur Wiese contributed significantly to the development of backend components, data analysis, and supported software implementation. Nadine Ziemert provided conceptual guidance, supervision, and editorial feedback throughout the project.

Manuscript II

Automated genome mining predicts structural diversity and taxonomic distribution of peptide metallophores across bacteria *eLife* (2026)

I am co-first author of this publication, together with Zachary Reitz. Zachary Reitz and Marnix Medema led the initial conceptualization of the study. Zachary Reitz developed the core HMM profiles used for the detection of metallophore-associated biosynthetic domains.

My primary contributions focused on the large-scale computational and evolutionary

analyses. I applied the developed HMMs to genome-wide mining of the GTDB reference genomes, performed the comparative analyses of metallophore diversity and taxonomic distribution, and carried out gene tree-species tree reconciliation analyses to investigate the evolutionary origins and diversification of metallophore-associated gene families. I contributed substantially to data interpretation, figure preparation, and manuscript writing.

Experimental validation and biochemical characterization were conducted by members of the Alison Butler lab and collaborating groups. All authors contributed to discussion of results and revision of the manuscript.

Manuscript III

Evolution of Siderophore Biosynthesis and Transport Systems in Bacteria
(Manuscript in preparation for submission)

I am the primary author of this manuscript and contributed substantially to all stages of the work. The study was conceived jointly with Prof. Dr. Nadine Ziemert, and I led the development of the analytical framework. I carried out the genome mining analyses, transporter profiling, large-scale trait mapping, and gene tree-species tree reconciliation analyses, and was responsible for the development of the evolutionary interpretations presented in the manuscript.

I wrote the manuscript and prepared the figures and supplementary analyses. Prof. Nadine Ziemert is the corresponding author and provided scientific guidance, conceptual input, and critical feedback throughout the study.

Chapter 1

Manuscript I:
AutoMLST2: a web server for phylogeny and microbial taxonomy

Nucleic Acids Research, 2025, 1–6
https://doi.org/10.1093/nar/gkaf397
Web Server issue

AutoMLST2: a web server for phylogeny and microbial taxonomy

Bita Pourmohsenin [1], Arthur Wiese[1], Nadine Ziemert [1,2,*]

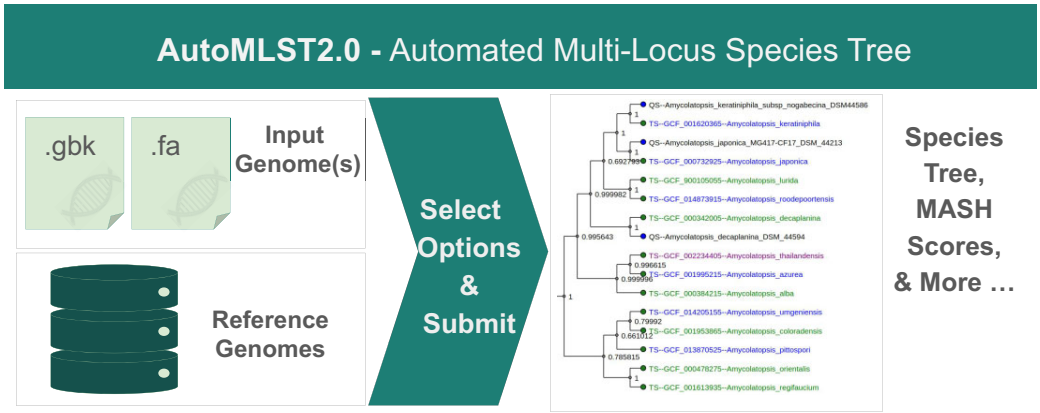
[1]Translational Genome Mining for Natural Products, Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Tübingen, Baden-Württemberg 72076, Germany
[2]German Center for Infection Research (DZIF), Partner Site Tübingen, 72076, Tübingen, Germany

[*]To whom correspondence should be addressed. Email: nadine.ziemert@uni-tuebingen.de

Abstract

Accurate and accessible phylogenetic analysis is essential for understanding microbial taxonomy and evolution, which are integral to microbiology, ecology, and drug discovery, yet it remains a challenging task. AutoMLST2 (https://automlst2.ziemertlab.com) is a web server designed to facilitate automated phylogenetic reconstruction and microbial taxonomy analysis for bacterial and archaeal genomes. It builds on the foundation of AutoMLST, which remains widely used due to its user-friendly interface compared to similar tools. Given its continued popularity and utility, we have enhanced AutoMLST to leverage newer reference databases and computational tools. AutoMLST2 integrates the Genome Taxonomy Database, extends support to archaeal genomes, and improves analytical flexibility. Key improvements include more customizable processing modes, containerization to prevent queue accumulations, and parallel computing for large-scale studies. By incorporating up-to-date databases and workflows, AutoMLST2 continues to provide an accessible and efficient platform for researchers in microbiology, evolutionary ecology, and natural product discovery.

Graphical abstract



Introduction

Understanding microbial phylogeny and taxonomy is fundamental to various biological disciplines, including microbiology, ecology, and drug discovery [1, 2]. Accurate species identification guides comparative genomic analyses and investigations of gene function and metabolic pathways [3, 4]. Traditional methods, such as 16S rRNA gene-based classification, often struggle to resolve closely related species due to their limited phylogenetic resolution [2]. Advances in whole-genome sequencing and computational methods have enabled more robust approaches, such as Average Nucleotide Identity (ANI) analysis and Multi-Locus Sequence Analysis (MLSA) [5–7]. However, implementing these workflows typically requires technical expertise and significant computational resources, making them less accessible to researchers without bioinformatics training [6, 7].

To bridge this gap, AutoMLST was developed as a user-friendly web server for automated phylogenetic analysis based on MLSA [8]. Since its release, AutoMLST has enabled rapid, high-resolution phylogenetic tree generation for bacterial species and has remained widely used due to its ease of use. However, as the field has evolved, so have the expectations for phylogenetic tools [1, 7].

Since the release of AutoMLST, GTDB-Tk has emerged as a highly accurate tool for microbial classification, leveraging the standardized Genome Taxonomy Database (GTDB) [1, 7]. However, GTDB-Tk requires substantial computational infrastructure and technical expertise, limiting its

Received: March 8, 2025. **Revised:** April 14, 2025. **Editorial Decision:** April 24, 2025. **Accepted:** May 1, 2025

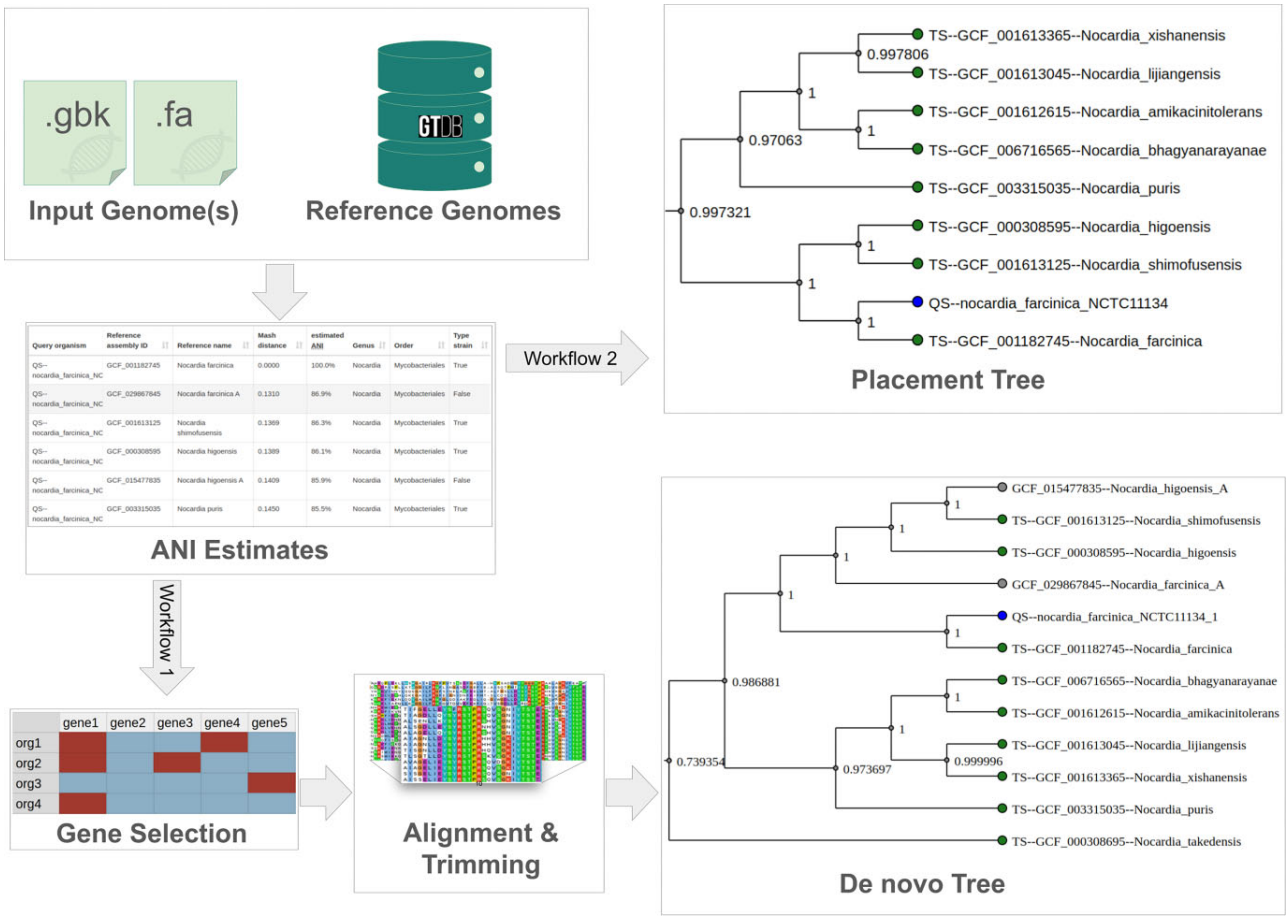


Figure 1. The overall workflow of AutoMLST2, automating microbial phylogenetic analysis through De novo and Placement mode workflows. Users upload genomes, which are compared to GTDB reference genomes using ANI estimation [1, 5]. Relevant genes are dynamically selected, aligned, and trimmed before phylogenetic tree construction with IQ-TREE [16] for concatenated analysis or ASTRAL-Pro3 for coalescent analysis [17]. The final De novo or Placement trees provide high-resolution evolutionary insights.

accessibility for non-expert users [6]. Despite its older database, AutoMLST remains widely used due to its user-friendly interface [8]. To address these challenges, we introduce AutoMLST2, a major update that integrates GTDB reference genomes, combining the accuracy of GTDB-Tk with the ease of use of AutoMLST. Key improvements include support for archaeal genomes, enhanced analytical flexibility, and a more scalable computational framework. By incorporating newer databases and advanced analytical capabilities, AutoMLST2 provides an accessible and efficient platform for microbial genome studies while maintaining the usability that made AutoMLST popular. Here, we outline its enhancements and demonstrate its utility in phylogenetic analysis.

Materials and methods

Overall workflow

AutoMLST2 provides an automated workflow for microbial phylogenetic analysis, taking the genome(s) uploaded by the user as input (Fig. 1). Once the user selects the desired options and submits the job, AutoMLST2 compares the uploaded genomes with a reference database to identify the most similar reference genomes [1, 7]. These reference genomes are then used to construct a phylogenetic species tree, providing an evolutionary context for the query genomes [5, 9]. The platform offers two distinct analysis modes: De novo and Placement modes. In De novo mode [10, 11], phylogenetic trees are constructed entirely from scratch, while Placement mode integrates query genomes into a precomputed reference tree, which is pruned to retain only the most closely related genomes [1].

Input and reference selection

Users can upload up to 50 query genomes in GenBank or FASTA format. To identify appropriate reference genomes, AutoMLST2 utilizes the GTDB representative genome set (release R220, 2024), which contains ∼107 000 bacterial genomes [7]. ANI estimates are rapidly calculated using precomputed MASH sketches [12], enabling the selection of the 50 most similar reference genomes for downstream analysis.

Gene selection and alignment

Gene homologs are detected using >2800 filtered HMMs (Hidden Markov Models) from the PGAP (NCBI Prokaryotic Genome Annotation Pipeline) [13] database, focusing on housekeeping genes for higher resolution [14]. HMM searches are performed on protein sequences, while all alignments for phylogenetic inference are done on the corresponding

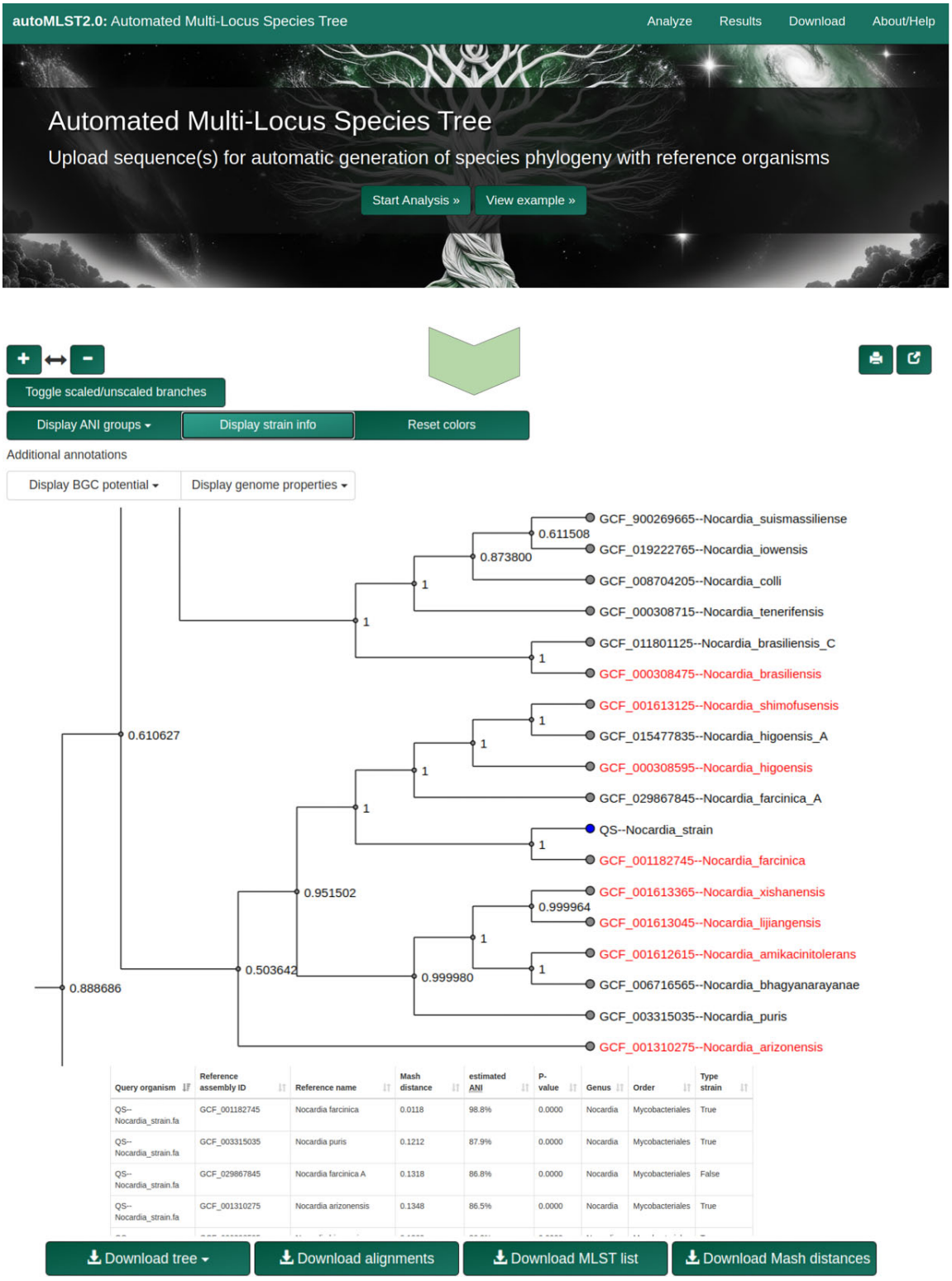


Figure 2. AutoMLST2 web interface, displaying a phylogenetic tree with annotated ANI groups, strain information, and genome properties. Users can visualize relationships, interact with tree elements, and download results.

nucleotide sequences. Sequences are aligned with MAFFT to ensure consistency [15].

Phylogenetic tree construction

AutoMLST2 supports two approaches for phylogenetic tree construction. In the concatenated approach, gene alignments are merged into a supermatrix, which is used to construct a phylogenetic tree with IQ-TREE [16]. Alternatively, in the coalescent approach, individual gene trees are inferred separately, and ASTRAL-PRO3 is employed to estimate the species tree from these independent gene trees [17].

Visualization and output

The final results, including phylogenetic trees and sequence alignments, are accessible through interactive visualization tools on the web interface and can be downloaded for further analysis [18].

Results

AutoMLST2 is a web server designed for automated microbial phylogenetic analysis. Through its analysis page, users can upload genome files, select analysis options, and initiate processing with a single click (Fig. 2). The platform is freely available at https://automlst2.ziemertlab.com and remains open-source, ensuring accessibility for researchers of all expertise levels.

To provide a more accurate and scalable solution, AutoMLST2 incorporates several key enhancements over its predecessor. The integration of the latest GTDB release ensures that taxonomic classifications remain aligned with current genomic standards [1]. The platform also extends support for archaeal genomes, with updated HMM models improving the detection of phylogenetic signals [16]. To enhance computational efficiency, AutoMLST2 employs containerization, enabling simultaneous job processing, increasing throughput, and reducing wait times.

AutoMLST2 further improves phylogenetic accuracy, particularly for closely related genomes. Unlike GTDB-Tk, which relies on a fixed set of 120 conserved marker proteins, AutoMLST2 dynamically selects multi-locus genes from a curated set of >2800 housekeeping genes, providing higher resolution and more precise taxonomic classification. Additionally, the standalone version offers customizable workflows, allowing users to optimize analyses by selecting between fast and accurate modes.

Benchmarking and validation

To evaluate AutoMLST2's performance, we benchmarked it against the GTDB tree using a diverse set of microbial genomes. We tested its accuracy and scalability across different taxonomic levels, including closely related and distantly related genomes. Additionally, we assessed the placement mode by analyzing how well the tool integrates query genomes into existing phylogenies under various scenarios (Table 1). In terms of accuracy, AutoMLST2 performs comparably to GTDB (Fig. 3), offering more precise taxonomic classification than AutoMLST due to its incorporation of an updated reference genome database. In De novo mode, AutoMLST2 is slightly slower than AutoMLST, when analyzing a single genome, as it searches against a much larger set of HMMs. However, it scales significantly better for larger datasets, benefiting from parallel computing. Placement mode is much faster but solely depends on the ANI estimations, therefore its accuracy can decrease with the quality of genomes [especially for Metagenome-Assembled Genomes (MAGs)]. An extra set of four example analyses demonstrating AutoMLST2's performance across different scenarios, including isolates and MAGs, is available at https://automlst2.ziemertlab.com/results/example1 up to https://automlst2.ziemertlab.com/results/example5 as well as in the supplementary data.

Table 1. Taxonomic precision of genome placement across different datasets and tree-building methods

Dataset/scenario	Taxonomic precision
Multiple sets of reference genomes	100% of genome taxonomic assignments match GTDB
10 random isolate genomes from Actinobacteriota	100% of genome genus assignments match GTDB
10 random Gammaproteobacteria genomes, a mix of MAGs and isolates	2 low-quality MAGs are not present in the final tree. Other genomes match GTDB.
Nocardia farcinica NCTC 11134 Placement tree	Correct Placement into the GTDB tree
Nocardia farcinica NCTC 11134 Coalescent tree	Query assignment matches GTDB, final tree slightly different from GTDB
Nocardia farcinica NCTC 11134 Concatenated tree	Query assignment matches GTDB, final tree slightly different from GTDB
Cyanobacterium sp002813895 (randomly chosen)	Taxonomic assignment matches GTDB

Genome assignments were compared against the GTDB, with high accuracy observed for reference genomes and isolate genomes. Low-quality MAGs were sometimes excluded from the final tree. *Nocardia farcinica* NCTC 11134 placement was consistent across different tree-building approaches, with slight variations in final topology.

Discussion

AutoMLST2 advances automated microbial phylogenetic analysis by integrating modern genomic resources and addressing limitations of its predecessor. By incorporating the GTDB and supporting archaeal genomes, AutoMLST2 ensures taxonomic classifications remain up to date with current genomic standards [1]. Its dual-mode workflow—offering both *de novo* phylogenetic reconstruction and genome placement—enhances adaptability across diverse microbial studies [10,11].

A key strength of AutoMLST2 is its user-friendly web interface, which simplifies complex analyses for researchers of all bioinformatics backgrounds [8]. Unlike command-line tools, it provides interactive visualizations, facilitating intuitive exploration of phylogenetic relationships. Automated reference genome selection and dynamic gene filtering improve resolution, particularly for closely related microbial strains, where fixed-marker approaches may be less effective [5,7].

In benchmarking, AutoMLST2 demonstrated high taxonomic precision across diverse datasets. Notably, in the case of *N. farcinica* NCTC 11134, the coalescent tree differed slightly from the GTDB reference tree. This variation likely reflects AutoMLST2's use of a dynamic set of conserved genes, which can provide a higher phylogenetic signal for closely related genomes.

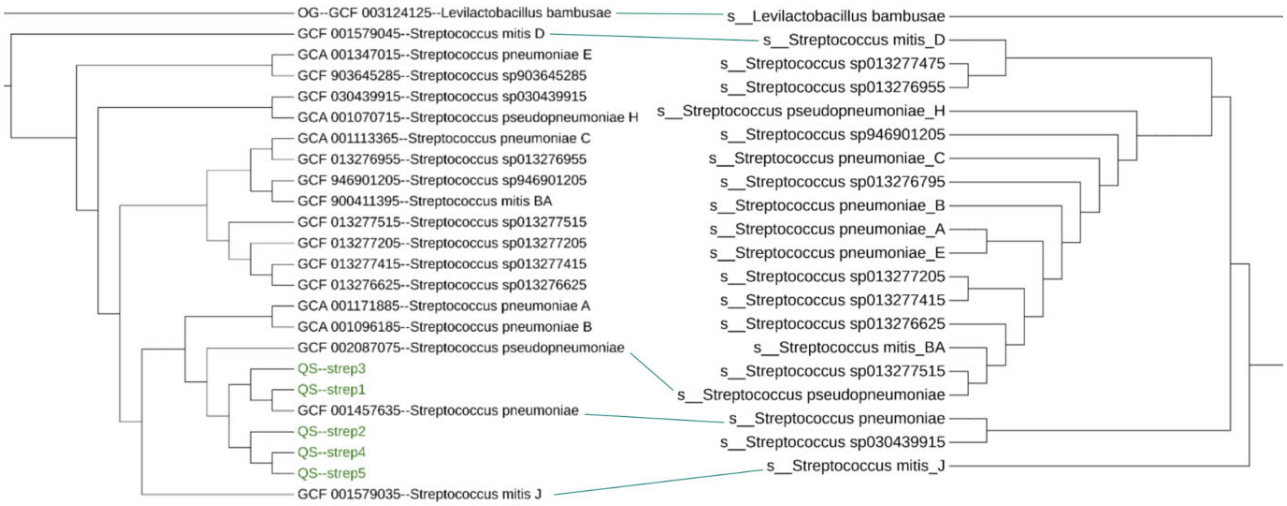


Figure 3. Partial tanglegram of the AutoMLST2 De novo tree for five *Streptococcus pneumoniae* strain genomes, compared to part of the GTDB reference tree. All the genomes were correctly clustered with the *S. pneumoniae* representative genome.

AutoMLST2 efficiently processes large datasets, making it valuable for microbiology, evolutionary biology, and natural product discovery. While it performs well with high-quality MAGs, its accuracy decreases with lower-quality assemblies due to fragmented genomes limiting phylogenetic inference. Future improvements could focus on optimizing gene selection for fragmented datasets and integrating network-based analyses of biosynthetic diversity.

AutoMLST2 is publicly available at https://automlst2.ziemertlab.com, with comprehensive documentation for users at all expertise levels. Its combination of accuracy, flexibility, ease of use, and advanced visualization tools makes it a powerful resource for microbial phylogenetics.

Acknowledgements

The authors would like to thank Mohammad Alanjary, Caner Bagci, Martina Adamek, and Libera Lo Presti for their valuable discussions.

Author contributions: Bita Pourmohsenin (Conceptualization [lead], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Resources [equal], Software [equal], Validation [equal], Visualization [lead], Writing—original draft [lead], Writing—review & editing [equal]), Arthur Wiese (Data curation [equal], Formal analysis [equal], Methodology [equal], Software [equal], Validation [equal], Writing—review & editing [equal]), Nadine Ziemert (Conceptualization [equal], Funding acquisition [lead], Methodology [equal], Project administration [lead], Resources [equal], Supervision [lead], Writing—review & editing [equal])

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

B.P. and N.Z. were supported by H2020-FNR-11-2020: SECRETED—grant agreement: 101000794. N.Z. was supported by the German Center for Infection Research TTU09.717. Structural support was provided by the Cluster of Excellence: EXC 2124: Controlling Microbes to Fight Infection (CMFI, project ID 390838134). Funding to pay the Open Access publication charges for this article was provided by H2020-FNR-11-2020: SECRETED—grant agreement: 101000794.

Data availability

All data generated and analyzed in this study are publicly available. Phylogenetic trees, benchmarking results, and additional supporting datasets, including example analyses and workflow results, are provided in the supplementary data. Genome data can be accessed through NCBI GenBank, with accession numbers listed in the trees. The AutoMLST2 code and reference datasets are available at https://automlst2.ziemertlab.com/download.

References

1. Parks DH, Chuvochina M, Rinke C *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2022;**50**:D785–94. https://doi.org/10.1093/nar/gkab776
2. Yarza P, Yilmaz P, Pruesse E *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* 2014;**12**:635–45. https://doi.org/10.1038/nrmicro3330
3. Hug LA, Baker BJ, Anantharaman K *et al.* A new view of the tree of life. *Nat Microbiol* 2016;**1**:16048. https://doi.org/10.1038/nmicrobiol.2016.48
4. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. *Nature* 2000;**405**:299–304. https://doi.org/10.1038/35012500
5. Jain C, Rodriguez-R LM, Phillippy AM *et al.* High throughput ANI analysis of 90K prokaryotic genomes reveals clear species

boundaries. *Nat Commun* 2018;**9**:5114. https://doi.org/10.1038/s41467-018-07641-9
6. Parks DH, Chuvochina M, Chaumeil P-A *et al.* A complete domain-to-species taxonomy for bacteria and archaea. *Nat Biotechnol* 2020;**38**:1079–86. https://doi.org/10.1038/s41587-020-0501-8
7. Chaumeil P-A, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 2022;**38**:5315–6. https://doi.org/10.1093/bioinformatics/btac672
8. 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**:W276–82. https://doi.org/10.1093/nar/gkz282
9. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* 2005;**102**:2567–72. https://doi.org/10.1073/pnas.0409727102
10. Matsen FA, Kodner RB, Armbrust EV. Pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 2010;**11**:538. https://doi.org/10.1186/1471-2105-11-538
11. Barbera P, Kozlov AM, Czech L *et al.* EPA-Ng: massively parallel evolutionary placement of genetic sequences. *Syst Biol* 2019;**68**:365–9. https://doi.org/10.1093/sysbio/syy054
12. Ondov BD, Treangen TJ, Melsted P *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;**17**:132. https://doi.org/10.1186/s13059-016-0997-x
13. Li W, O'Neill KR, Haft DH *et al.* . RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res* 2021;**49**:D1020–8. https://doi.org/10.1093/nar/gkaa1105
14. Tatusova T, DiCuccio M, Badretdin A *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* 2016;**44**:6614–24. https://doi.org/10.1093/nar/gkw569
15. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform* 2019;**20**:1160–6. https://doi.org/10.1093/bib/bbx108
16. Minh BQ, Schmidt HA, Chernomor O *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020;**37**:1530–4. https://doi.org/10.1093/molbev/msaa015
17. Zhang C, Scornavacca C, Molloy EK *et al.* ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Mol Biol Evol* 2020;**37**:3292–307. https://doi.org/10.1093/molbev/msaa139
18. Letunic I, Bork P. Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 2021;**49**:W293–6. https://doi.org/10.1093/nar/gkab301

Received: March 8, 2025. **Revised:** April 14, 2025. **Editorial Decision:** April 24, 2025. **Accepted:** May 1, 2025

Chapter 2

Manuscript II:
Automated genome mining predicts structural diversity and taxonomic distribution of peptide metallophores across bacteria

RESEARCH ARTICLE

Automated genome mining predicts structural diversity and taxonomic distribution of peptide metallophores across bacteria

Zachary L Reitz[1,2†], Bita Pourmohsenin[3†], Melanie Susman[4], Emil Thomsen[4], Daniel Roth[4], Alison Butler[4], Nadine Ziemert[3*], Marnix Medema[1*]

[1]Bioinformatics Group, Wageningen University, Wageningen, Netherlands; [2]Department of Evolution, Ecology, and Marine Biology, University of California, Santa Barbara, Santa Barbara, United States; [3]Interfaculty Institute of Microbiology and Infection Medicine, Institute for Bioinformatics and Medical Informatics, Tübingen, Germany; [4]Department of Chemistry and Biochemistry, University of California, Santa Barbara, Santa Barbara, United States

***For correspondence:** nadine.ziemert@uni-tuebingen.de (NZ); marnix.medema@wur.nl (MM)

[†]These authors contributed equally to this work

Competing interest: See page 19

Funding: See page 19

Preprint posted 01 August 2025
Sent for Review 16 September 2025
Reviewed preprint posted 30 October 2025
Reviewed preprint revised 03 February 2026
Version of Record published 09 March 2026

Reviewing Editor: Marisa Nicolás, Laboratório Nacional de Computação Científica, Brazil

eLife Assessment

This **important** and **compelling** study establishes a robust computational and experimental framework for the large-scale identification of metallophore biosynthetic clusters. The work advances beyond current standards, providing theoretical and practical value across microbiology, bioinformatics, and evolutionary biology.

Abstract Microbial competition for trace metals shapes their communities and interactions with humans and plants. Many bacteria scavenge trace metals with metallophores, small molecules that chelate environmental metal ions. Metallophore production may be predicted by genome mining, where genomes are scanned for homologs of known biosynthetic gene clusters (BGCs). However, accurately detecting non-ribosomal peptide (NRP) metallophore biosynthesis requires expert manual inspection, stymieing large-scale investigations. Here, we introduce automated identification of NRP metallophore BGCs through a comprehensive algorithm, implemented in antiSMASH, that detects chelator biosynthesis genes with 97% precision and 78% recall against manual curation. We showcase the utility of the detection algorithm by experimentally characterizing metallophores from several taxa. High-throughput NRP metallophore BGC detection enabled metallophore detection across 69,929 genomes spanning the bacterial kingdom. We predict that 25% of all bacterial non-ribosomal peptide synthetases encode metallophore production and that significant chemical diversity remains undiscovered. A reconstructed evolutionary history of NRP metallophores supports that some chelating groups may predate the Great Oxygenation Event. The inclusion of NRP metallophore detection in antiSMASH will aid non-expert researchers and continue to facilitate large-scale investigations into metallophore biology.

Introduction

Across environments, microbes compete for a scarce pool of trace metals. Many microbes scavenge metal ions with small-molecule chelators called *metallophores*, which diffuse through the environment

and chelate metal ions with high affinity (***Hider and Kong, 2010***; ***Kraemer et al., 2015***). A microbe possessing the right membrane transporters will be able to recognize and import a metallophore–metal complex, while other strains are unable to access the chelated metal ions. Thus, the metallophore secreted by one microbe can either support or inhibit growth of a neighboring strain, driving complex community dynamics in marine, freshwater, soil, and host environments (***Kramer et al., 2020***). The most well studied metallophores are the Fe(III)-binding *siderophores*, which have found applications in biocontrol, bioremediation, and medicine (***Soares, 2022***). Two recent studies demonstrated that the disease suppression ability of a rhizosphere microbiome is strongly determined by whether or not the pathogen can use siderophores produced by the community; a microbiome can even encourage pathogen growth when a compatible siderophore is produced (***Gu et al., 2020a***; ***Gu et al., 2020b***). Compared to siderophores, other metallophore classes are relatively understudied, but they likely play equally important biological roles, as exemplified by recent reports of both commensal and pathogenic bacteria relying on zincophores to effectively colonize human hosts (***Behnsen et al., 2021***; ***Mehdiratta et al., 2022***).

Hundreds of unique metallophore structures have been characterized, each with specific chemical properties (e.g., effective pH range, hydrophobicity, and metal selectivity) and biological effects on other microbes (based on membrane transporter compatibility). Experimentally characterizing metallophores can be time-consuming and costly, and thus researchers often use genome mining to predict metallophore production in silico (***Reitz and Medema, 2022***). Taxonomy alone is not sufficient to predict what metallophores will be produced by a microbe, as production can vary significantly even within a single species (***Cézard et al., 2015***). Instead, metallophores must be predicted from each genome based on the presence of biosynthetic gene clusters (BGCs) that encode their biosynthesis. The majority of known metallophores are non-ribosomal peptides (NRPs), a broad class of natural products that also includes many antibiotics, antitumor compounds, and toxins. Specialized chelating moieties bind directly to the metal ion (in the case of siderophores, Fe^{3+}), while other amino acids in the peptide chain give the metallophore the required flexibility for chelation. Nearly all NRP metallophores contain one or more of the substructures shown in ***Figure 1A***: 2,3-dihydroxybenzoate (catechol, 2,3-DHB), hydroxamates, salicylate, β-hydroxyaspartate (β-OHAsp), β-hydroxyhistidine (β-OHHis), gramine, Dmaq (1,1-dimethyl-3-amino-1,2,3,4- tetrahydro-7,8-dihydroxy-quinoline), and the pyoverdine chromophore. Biosynthetic pathways are known for each of the chelating groups (***Figure 1B***), and the presence of a chelator pathway may be used as a marker for metallophore production.

Mining genomes for metallophore BGCs has facilitated the discovery of chemically and biologically diverse metallophore systems; however, automated detection tools are still severely lacking (***Reitz and Medema, 2022***) The peptidic backbones of NRP metallophores are produced by non-ribosomal peptide synthetases (NRPSs), large multi-domain enzymes that activate and condense amino acids and other substrates in an assembly-line manner (***Süssmuth and Mainz, 2017***). In the past two decades, a variety of bioinformatic tools have been developed to identify NRPS BGCs in a genome. One of the most popular is the secondary metabolite prediction platform antiSMASH, which uses a library of profile hidden Markov models (pHMMs) to identify (combinations of) enzyme-coding genes that are indicative of certain classes of specialized metabolite biosynthetic pathways (***Blin et al., 2021b***; ***Blin et al., 2023***). For example, antiSMASH identifies an NRPS BGC region by the minimum requirement of a gene containing at least one condensation and one adenylation domain. NRP metallophore BGCs are technically detected by this rule as well; however, NRPSs also produce many other families of compounds, and additional manual annotation has still been required to identify NRP metallophore BGCs specifically. Accordingly, accurate prediction of BGCs encoding siderophores and other metallophores was limited to experts in natural product biosynthesis, and even experts cannot manually curate the thousands of BGCs produced by high-throughput metagenomic or comparative genomic analyses. To date, no global analysis of NRP metallophores has been performed, and thus the prevalence, combinatorics, and taxonomic distribution of different chelating groups are unknown.

Here, we describe the development and application of a high-accuracy antiSMASH-integrated method for the automated detection of NRP metallophore BGCs, using the presence of chelator biosynthesis genes within NRPS BGCs as key markers for predicting metallophore production. The new detection rule was applied to 15,562 representative bacterial genomes, allowing us to take the first census of NRP metallophore production across bacteria. At least 25% of all NRPS clusters in these representative genomes code for the production of metallophores and significant biosynthetic

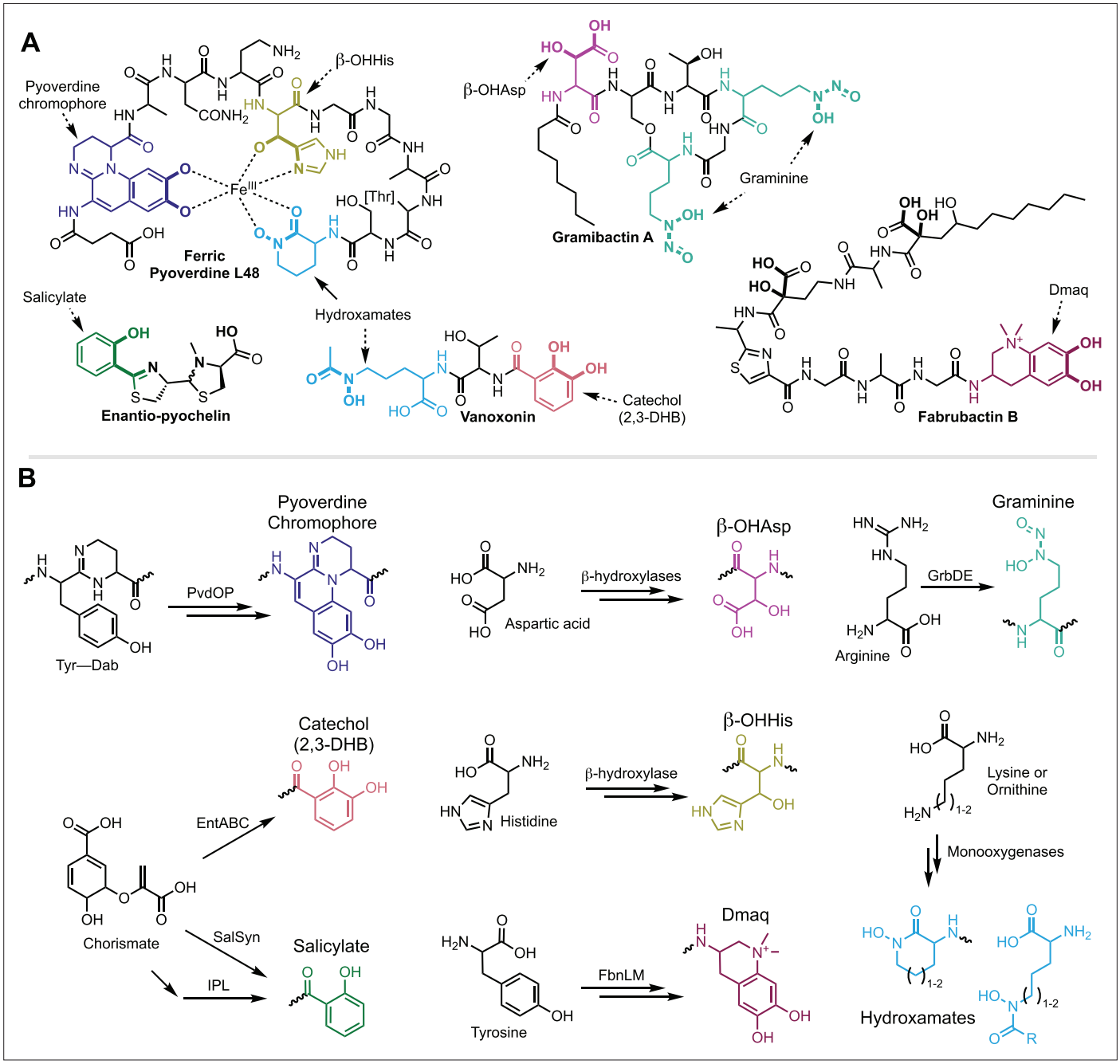


Figure 1. Chelating substructures found in bacterial NRP metallophores and their biosynthetic pathways. (**A**) Representative NRP metallophore structures. Nearly all known NRP metallophores contain one or more of the eight labeled chelating groups. Most chelating groups provide bidentate metal chelation, as shown for ferric pyoverdine L48. (**B**) Chelator biosynthesis pathways that form the basis for the new antiSMASH detection algorithm, as described in the text. The same chelator colors are used in each figure.

The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. Workflow for developing an NRP-metallophore-specific profile hidden Markov model (pHMM) and significance score cutoff for an enzyme (sub-)family.

Figure supplement 2. A maximum-likelihood phylogeny of putative β-hydroxylases found in NRPS BGC regions.

Figure supplement 3. Analysis of a novel putative NRP metallophore BGC from *Sporomusa termitida* DSM 4440.

diversity remains undiscovered. We then leveraged our computational analyses to guide characterization of siderophores from multiple bacterial taxa, finding structures that matched our genome-based predictions. By mapping NRP metallophore BGCs from 59,851 genomes to the Genome Taxonomic Database (GTDB) phylogeny, we identified myxobacterial and cyanobacterial metallophores as understudied and reconstructed a possible evolutionary history of the chelating groups.

Results

A chelator-based strategy for detection of NRP metallophore biosynthetic gene clusters

The specialized chelating moieties found in NRP metallophores are rarely found in other natural products, and thus we sought to automate metallophore BGC prediction by searching for genes encoding their biosynthesis. An extensive review of published NRP metallophore structures revealed that nearly all contain one or more of just eight chelator substructures (***Figure 1A***). Protein domains responsible for their biosyntheses have been reported (***Figure 1B***), and thus pHMMs could be constructed to allow detection of putative chelator biosynthesis genes. Generally, draft pHMMs were built from alignments of known and predicted NRP metallophore biosynthesis genes collected from literature, and cutoffs were manually determined (***Figure 1—figure supplement 1***). The final multiple sequence alignments, pHMMs, and cutoffs are provided in the Supplemental Dataset.

A full description of each biosynthetic pathway detection strategy, including caveats and known limitations, is provided in the Methods and briefly summarized here. The profile HMMs implemented within antiSMASH are given in bold font. The biosynthetic cassette for 2,3-DHB is detected by an isochorismate synthase (**EntC**) and 2,3-dihydro-2,3- dihydroxybenzoate dehydrogenase (**EntA**) (***Raymond et al., 2003***). Two salicylate biosynthesis pathways are detected by the presence of either an isochorismate pyruvate-lyase (**IPL**) (***Serino et al., 1995***) or a bifunctional salicylate synthase (**SalSyn**) (***Pelludat et al., 2003***). We also included detection of two condensation domain subtypes specific to catecholic and phenolic metallophores: VibH-like enzymes (**VibH**) (***Keating et al., 2002***; ***Reitz and Butler, 2020***) and tandem heterocyclization domains (**Cy_tandem**) (***Bloudoff et al., 2017***). Peptidic hydroxamate pathways are detected by an ornithine (Orn) or Lys N-monooxygenase (**Orn_monoox** or **Lys_monoox**, respectively) (***Olucha and Lamb, 2011***). We could not accurately detect the vicibactin hydroxylase VbsO using a pHMM (***Heemstra et al., 2009***), and so the characteristic acyl-hydroxyornithine epimerase **VbsL** is used to detect vicibactin biosynthesis (***Heemstra et al., 2009***). We previously identified three families of siderophore-specific Fe(II)/α-ketoglutarate-dependent enzymes responsible for β-OHAsp (**TBH_Asp** and **IBH_Asp**) or β-OHHis (**IBH_His**) (***Reitz et al., 2019***). Based on the recent discovery of β-OHAsp-containing cyanochelins from cyanobacteria (***Galica et al., 2021***), we have now identified two new clades that are putatively metallophore-specific and tentatively named **CyanoBH_Asp1** and **CyanoBH_Asp2**. The diazeniumdiolate-containing graminine may be detected by the presence of the cryptic necessary enzymes **GrbD** and **GrbE** (***Hermenau et al., 2019***; ***Makris et al., 2022***). The quinoline chelator Dmaq is detected by **FbnL** and **FbnM**, which initiate Dmaq biosynthesis (***Vinnik et al., 2021***). The chromophore of pyoverdines is detected by the presence of a tyrosinase **PvdP** and/or an oxidoreductase **PvdO** (***Nadal-Jimenez et al., 2014***; ***Ringel et al., 2018***).

Several known chelating group pathways are not currently detected. Our detection strategy is limited to clades or combinations of biosynthetic enzymes that are distinct to NRP metallophore pathways. Several chelators are synthesized by the core NRPS and/or polyketide synthase (PKS) machinery and could not be detected without also retrieving many false positives, including NRPS-derived thiazol(id)ine and oxazol(id)ine heterocycles (see pyochelin, ***Figure 1A***) and PKS-derived 5-alkylsalicylate (e.g., in micocacidin ***Kage et al., 2013***). We also did not include detection of a pathway currently only reported in fabrubactins that produces two α-hydroxycarboxylate chelating moieties (***Figure 1A*****,** bolded atoms) (***Vinnik et al., 2021***). Finally, we have not yet designed detection rules for the recently discovered chelating groups 5-aminosalicylate of pseudonochelin (***Zhang et al., 2022***) or 2-napthoate of ecteinamines (***Wu et al., 2023***); however, we expect that their biosyntheses will be amenable to detection by the method used herein (***Figure 1—figure supplement 1***). The NRP metallophore detection algorithm is publicly available in the antiSMASH web server and command line tool (https://antismash.secondarymetabolites.org/, version 7 and upwards).

Table 1. Summary of NRP metallophore BGC detection, comparing the chelator-based rule newly implemented in antiSMASH, the transporter-based method of Crits-Christoph et al., (***Matthijs et al., 2016***) and a combined either/or ensemble.

	Performance metrics*			Number of NRP metallophore BGC regions detected in representative bacterial genomes†		
	Precision	Recall	F1 ‡	Complete NRPS regions n=11,704	Partial NRPS regions n=8,403	Total NRPS regions n=20,107
AntiSMASH rule	0.97	0.78	0.86	2485 (21%) §	725 (8.6%)	3210 (16%)
Transporter genes	0.93	0.56	0.69	1723 (15%)	376 (4.5%)	2099 (10%)
Either/or ensemble	0.92	0.88	0.90	2948 (25%)	855 (10%)	3803 (19%)

*Detection methods were each tested on a set of 758 manually annotated NRPS BGC regions (180 true positives). Full results are given in ***Supplementary file 1b***.

†Detection methods were applied to 15,562 NCBI RefSeq representative bacterial genomes. The full results are given in ***Supplementary file 1c***. A region is 'complete' if it is not on a contig edge, as determined by antiSMASH. An additional 54 BGC regions were detected as NRP metallophores without meeting the requirements for the antiSMASH NRPS rule.

‡F1 score is equal to 2×(Precision ×Recall)/(Precision +Recall).

§Percentages indicate the fraction of NRPS regions that were predicted to encode NRP metallophores.

Validating antiSMASH NRP metallophore detection against manually curated BGCs

In order to assess the performance of our NRP metallophore BGC detection strategy, we manually predicted metallophore production among a large set of antiSMASH-predicted BGCs. A total of 758 NRPS BGC regions from 330 genera were annotated with default antiSMASH v6.1 and inspected for known markers of metallophore production, including genes encoding transporters, iron reductases, chelator biosynthesis, and known metallophore NRPS domain motifs. We thus manually classified 176 BGC regions (23%) as metallophore BGCs (***Supplementary file 1b***). The new antiSMASH detection rule was applied to the same BGC regions, resulting in 145 putative metallophore BGC regions (F1=0.86; ***Table 1*** and ***Supplementary file 1b***). Nine metallophore BGC regions were only detected by antiSMASH. Upon reinvestigation, four were determined to likely represent genuine metallophore BGC regions missed during manual analysis, leaving only five putative false positives in which seemingly unrelated genes matched the pHMMs (97% precision). Conversely, a total of 40 metallophore BGC regions could only be detected manually (78% recall). In the majority of false negatives, NRP metallophore BGCs were missed because chelator biosynthesis genes, on which the detection strategy is based, were not present in the cluster. In 21 cases, genes encoding catechol, salicylate, or hydroxamate biosynthesis were located elsewhere in the genome. In ten cases, chelator biosynthesis pathways were not found anywhere in the genome; these clusters may be non-functional fragments, rely on exogenous precursors (as seen in equibactin biosynthesis; ***Heather et al., 2008***), or have evolved to use novel chelator biosyntheses. Two of the false negatives encoded the 5-alkylsalicylate PKS that is currently undetectable, as described above. Finally, seven manually assigned NRP metallophore BGC regions had no genes corresponding to known chelator pathways (***Supplementary file 1b***); if correctly annotated, they may represent novel structural classes. In one particularly promising case, a salicylate pathway appears to have been replaced with a partial menaquinone pathway to produce a putative 1,4-dihydroxy-2-naphthoate chelating group (***Figure 1—figure supplement 3***). A BGC from *Sporomusa termitida* DSM 4440 was annotated as a putative metallophore due to the presence of a TonB-dependent outer membrane receptor, as well as NRPS domains, methyltransferases, and oxidoreductases homologous to those involved in the biosynthesis of pyochelin and other thiazol(id)ine metallophores. A salicylate synthase gene, present in similar clusters, appears to have been replaced by a partial menaquinone pathway (*menFDEB*). We thus predicted that the salicylate moiety would be replaced by 1,4-dihydroxy-2-naphthoic acid. The Natural Product Atlas contained one family of bacterial compounds with that substructure, karamomycins, which were isolated from an unsequenced strain of *Nonomuraea endophytica* (***Figure 1—figure supplement 3***; ***Winkelmann et al., 1994***). Karamomycins also contain thiazol(id)ines, as predicted for the *S. termitida* DSM 4440

BGC product. Hence, karamomycins are likely produced by a homologous BGC, and both compounds may be involved in trace metal binding and transport.

AntiSMASH outperforms transporter-based detection, although both techniques are complementary

Crits-Christoph et al. found that the presence of transporters could be used to predict siderophore BGCs among other NRPS clusters (***Crits-Christoph et al., 2021***). Specifically, the Pfam families for TonB-dependent receptors, FecCD domains, and periplasmic binding proteins (PF00593, PF01032, and PF01497, respectively) were determined to be highly siderophore-specific, and the authors used the presence of two of the three domains to predict a 'siderophore-like' BGC region (metallophores that transport other metals were also coded as siderophores in their dataset). We used a modified version of antiSMASH to detect the three transporter families among the 758 manually annotated NRPS BGC regions (***Table 1*** and ***Supplementary file 1b***). In total, the transporter-based method detected 108 metallophore clusters (F1=0.69), including 8 putative false positives (93% precision), and had 80 false negatives (56% recall). One false positive was noted in the manual annotation as a likely 'cheater': while several *Bordetella* genomes encode the synthesis of a putative graminine-containing metallophore, *B. petrii* DSM 12804 has retained only the transporter genes alongside a small fragment of the NRPS. In the seven other false positives, BGC regions appeared to coincidentally contain transporter genes in their periphery, as they were not conserved in homologous NRPS clusters. In one case, the triggering genes were part of a putative vitamin B12 import and biosynthesis locus. Combining the two methods in an either/or ensemble approach slightly improved overall performance *versus* the antiSMASH rule alone, achieving 92% precision, 88% recall, and an F1 score of 0.90 (***Table 1***).

Charting NRP metallophore biosynthesis across bacteria

The implementation of NRP metallophore BGC detection into antiSMASH allowed us to take the first bacterial census of NRP metallophore biosynthesis. The finalized detection rule was applied to 15,562 representative bacterial genomes from NCBI RefSeq (June 25, 2022). In total, 3264 NRP metallophore BGC regions were detected (***Table 1*** and ***Supplementary file 1c***), including 54 Type II (non- or semi-modular ***Jaremko et al., 2020***) NRPS regions that would otherwise not be detected by antiSMASH, such as BGCs for acinetobactin and brucebactin (***Mihara et al., 2004***; ***González Carreró et al., 2002***). NRP metallophores comprised 16% of all NRPS BGC regions in the genomes. Among complete regions (not located on a contig edge), 21% of NRPS BGC regions were classified as NRP metallophores, compared to just 8.6% of incomplete NRPS regions. This is consistent with previous reports that low-quality, fragmented genomes result in low-quality BGC annotations in antiSMASH (***Blin et al., 2017***). The transporter-based approach predicted siderophore activity for 15% of complete NRPS regions, including 463 BGC regions without detectable chelator genes; when the two methods are combined, over 25% of NRPS BGCs are predicted to produce NRP metallophores (***Table 1***). Only complete NRP metallophore BGC regions detected by antiSMASH were used for downstream analyses.

Frequency and hybridization of NRP metallophore chelating groups

Complete NRP metallophore regions from the representative genomes were categorized by the type(s) of chelator biosynthesis genes detected within (***Figure 2***). Hydroxamates and catechols were the most common pathways, present in 44% and 36% of BGC regions, respectively. In contrast, β-OHHis, graminine, and Dmaq biosyntheses were rare in representative genomes, each present in less than 2% of detected regions. About 20% of regions contained genes for at least two pathways and putatively produce a hybrid metallophore. Only 42 BGC regions (1.7%) contained three different chelating groups: each encoded genes for the pyoverdine chromophore, a hydroxamate, and either β-OHAsp or β-OHHis. The proportion of hybrid metallophores is likely higher than estimated here. As described above, some chelating moieties could not be captured by the pHMM-based rule. Furthermore, metallophore biosynthesis may require genes from multiple BGCs. Pyoverdine genes may be located in up to five different loci (***Gross and Loper, 2009***), and all 56 regions with only the pyoverdine chromophore pathway are expected to produce hybrid siderophores. Representative characterized siderophores that contain the chelator combinations in our dataset are shown in ***Figure 3***.

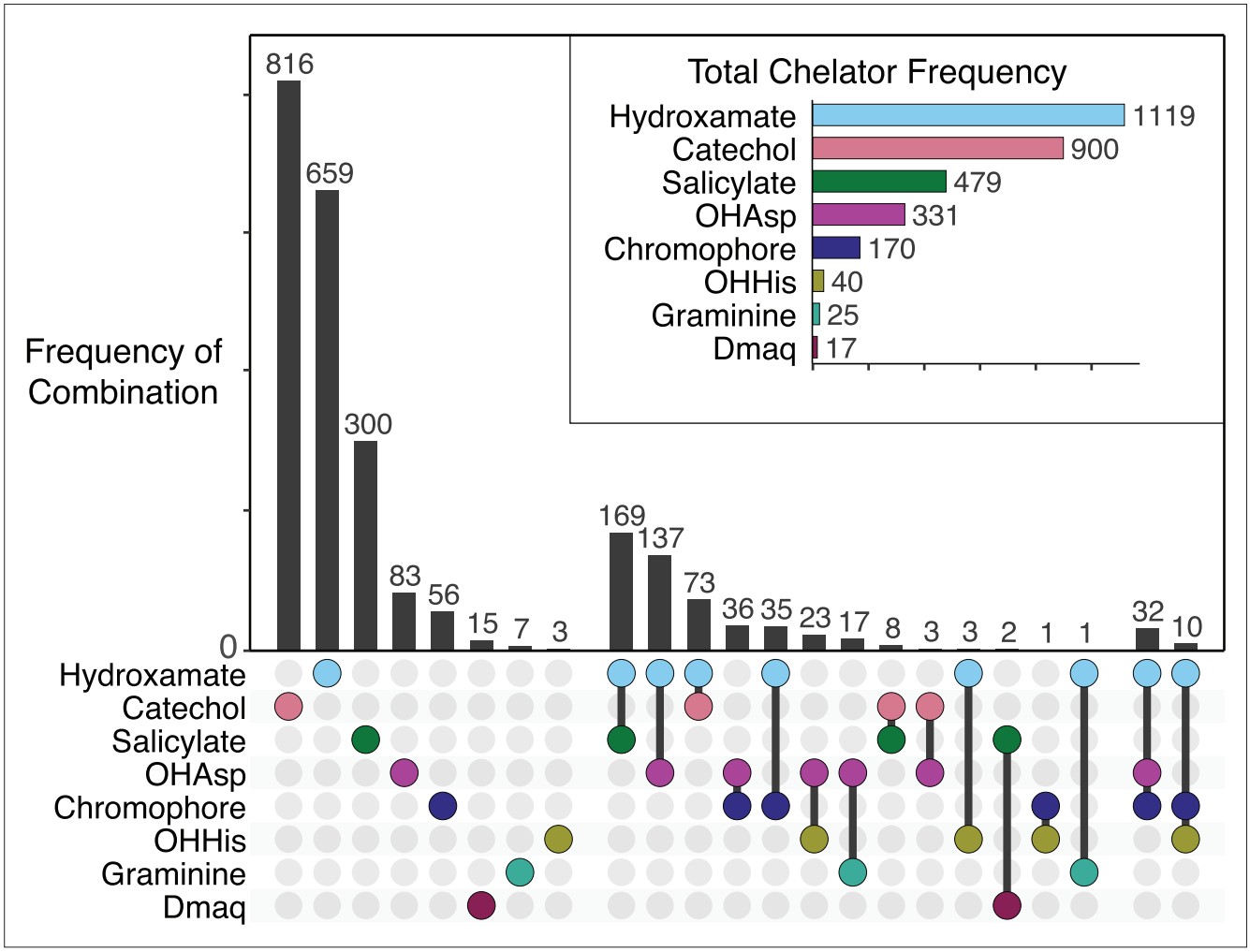


Figure 2. An upset plot of chelator frequency among 2,489 complete NRP metallophore BGC regions from RefSeq representative genomes. An additional 38 BGC regions were detected by metallophore-specific NRPS domains (VibH-like or tandem Cy) rather than chelator biosynthesis, and may produce catechol and/or salicylate metallophores using biosyntheses encoded elsewhere in the genome.

The most widespread NRP metallophore families have likely been found, yet significant diversity remains unexplored

Different species of bacteria can contain highly similar metallophore BGCs. To gauge the biosynthetic diversity of the putative NRP metallophores (and thereby the structural diversity), the complete BGC regions were organized into a sequence similarity network using BiG-SCAPE, which clusters BGCs based on their shared gene content and identity. An additional 75 reported NRP metallophore BGCs were included as reference nodes (***Supplementary file 1a***), and a distance cutoff of 0.5 was chosen to allow highly similar reference BGCs to form connected components (gene cluster families; GCFs) in the network. The final network, colored and organized by chelator type, is presented in ***Figure 3***. The majority of BGC regions (57%) clustered with the reference BGCs in just 45 GCFs, suggesting that many of the most widespread NRP metallophore families with known chelating groups already have characterized representatives (***Figure 3—figure supplement 1***). However, 1093 BGC regions did not cluster with any reference BGC, forming 619 separate GCFs in the network (93% of all GCFs). Some of these may encode orphan metallophores previously isolated from unsequenced strains, or be similar to known BGCs that were not included in our non-exhaustive literature search. Nevertheless, significant NRP metallophore structural diversity remains undiscovered, particularly among the 484 BGC regions distinct enough to form isolated nodes in the network.

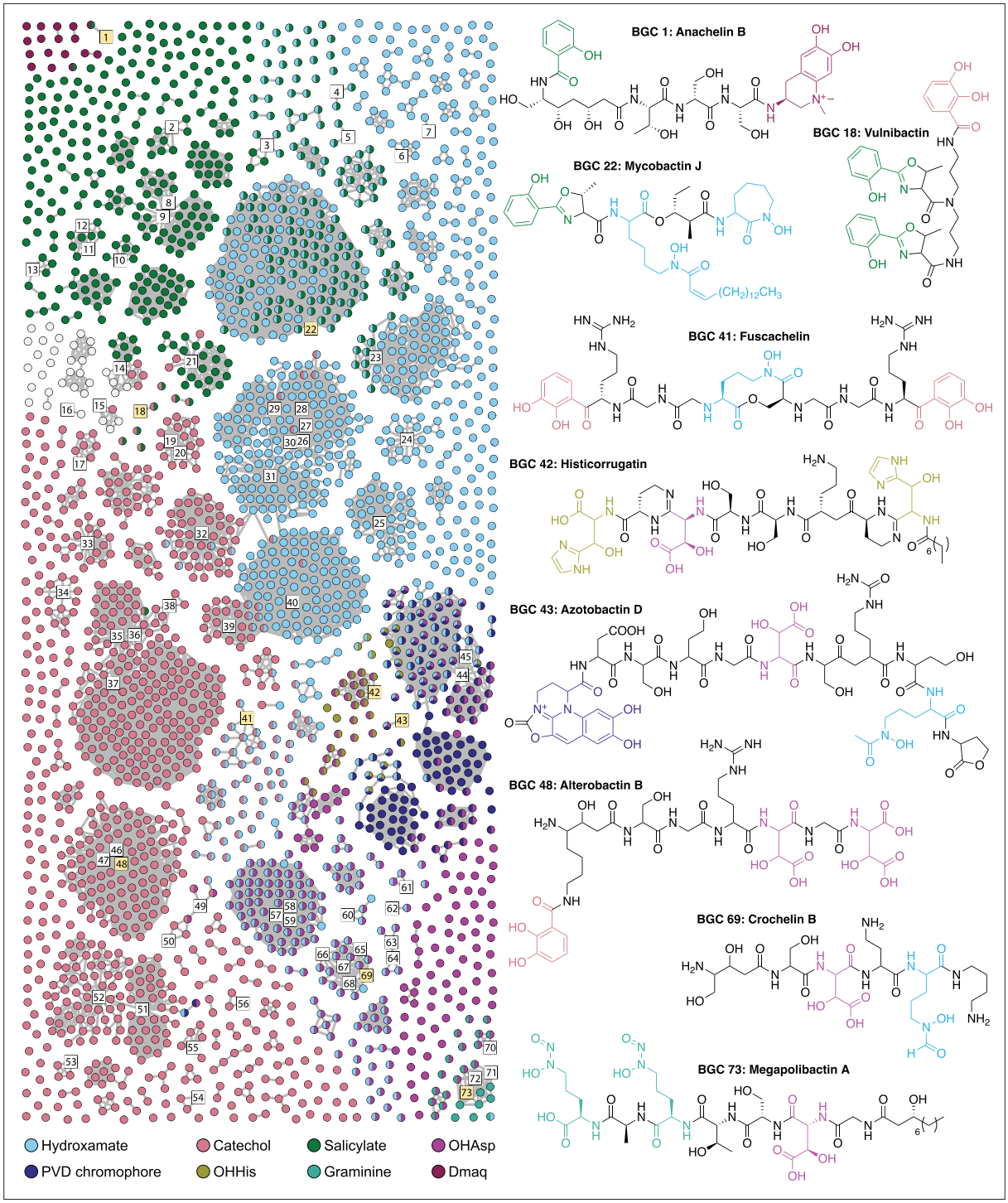


Figure 3. BiG-SCAPE similarity network of complete NRP metallophore BGC regions from RefSeq representative genomes. Numbered square nodes indicate published BGCs, as given in ***Supplementary file 1a***. Select hybrid metallophore BGC nodes are highlighted yellow, and their corresponding structures are shown. Nodes are colored by the type(s) of chelator biosynthesis detected therein. BGC regions colored light gray contain only metallophore-specific NRPS domains (VibH-like or tandem Cy) and may produce catechol and/or salicylate metallophores using biosyntheses encoded elsewhere in the genome. The network was constructed in BiG-SCAPE v1.1.2 using 2596 BGC regions as input, including 78 reference BGCs, and a distance cutoff of 0.5.

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Similarity network of complete NRP metallophore BGC regions from RefSeq representative genomes.

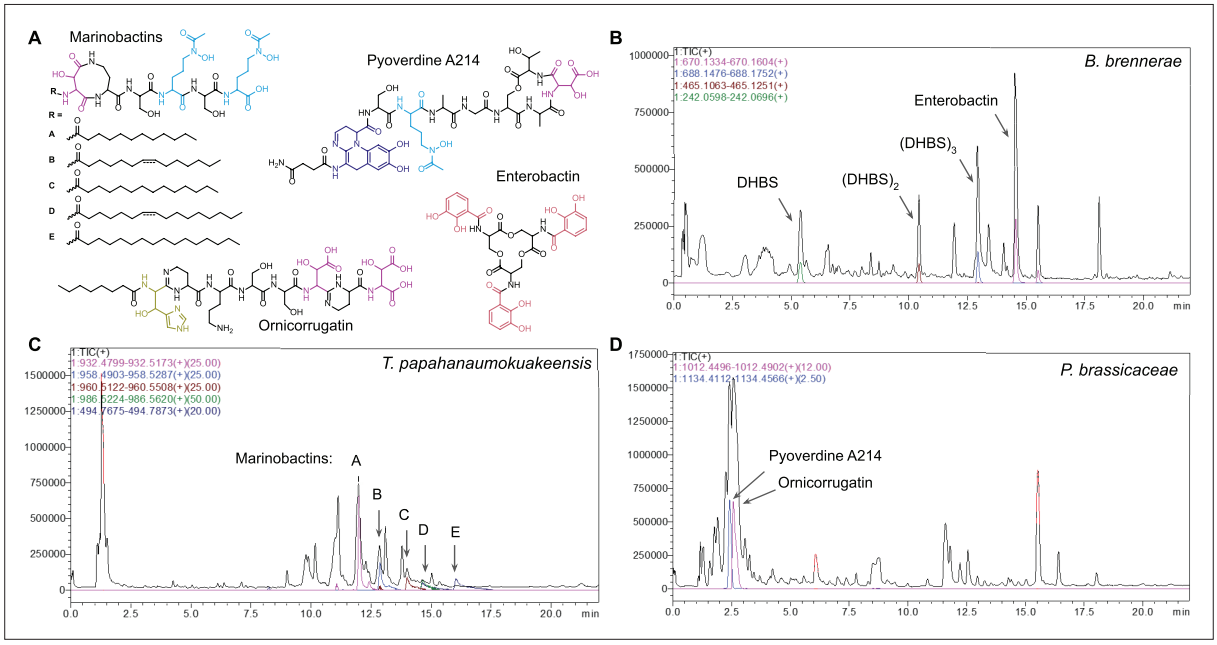


Figure 4. Identification of siderophores predicted from genome mining. (**A**) Chemical structures of marinobactins **A**–**E**, (***Winkelmann et al., 1994***) produced by *Terasakiispira papahanaumokuakeensis* DSM 29361; enterobactin, (***Martinez and Butler, 2007***) produced by *Buttiauxella brennerae* DSM 9396; and pyoverdine A214 (***Uría Fernández et al., 2003***) and ornicorrugatin, (***Matthijs et al., 2008***) both produced by *Pseudomonas brassicacearum* DSM 13227. The position and orientation of the fatty acid desaturation in marinobactins B and D was not determined in this work. (**B–D**) High pressure liquid chromatography / high-resolution mass spectrometry (HPLC-HRMS) total ion chromatograms of culture supernatant extracts, overlaid with extracted ion chromatograms for siderophore features.

The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. Semi-preparative RP HPLC chromatogram of *B. brennerae* DSM 9396 crude extract.

Figure supplement 2. UPLC-ESI-MS TIC of the purified catechol compounds from *B. brennerae* DSM 9396.

Figure supplement 3. Mass spectra of the purified catechol compounds from *B. brennerae* DSM 9396.

Figure supplement 4. UPLC-ESI-MS of *T. papahanaumokuakeensis* DSM 29361.

Figure supplement 5. ESI-MS/MS fragmentation of marinobactin A $[M+H]^+$, 932 *m/z*, from *T. papahanaumokuakeensis* DSM 29361.

Figure supplement 6. ESI-MS/MS fragmentation of marinobactin B $[M+H]^+$, 958 *m/z*, from *T. papahanaumokuakeensis* DSM 29361.

Figure supplement 7. ESI-MS/MS fragmentation of marinobactin C $[M+H]^+$, 960 *m/z*, from *T. papahanaumokuakeensis* DSM 29361.

Figure supplement 8. ESI-MS/MS of marinobactin D $[M+H]^+$, 986 *m/z*, from *T. papahanaumokuakeensis* DSM 29361.

Figure supplement 9. UPLC-MS spectrum of ornicorrugatin siderophore from *Pseudomonas brassicacearum* DSM 13227.

Figure supplement 10. LC-MS tandem MS spectrum of ornicorrugatin siderophore from *Pseudomonas brassicacearum* DSM 13227.

Figure supplement 11. UPLC-MS spectrum of pyoverdine siderophore from *Pseudomonas brassicacearum* DSM 13227.

Figure supplement 12. LC-MS tandem MS spectrum of pyoverdine siderophore, pyoverdine A214, from *Pseudomonas brassicacearum* DSM 13227.

Chemical identification of genome-predicted siderophores

To showcase how our large-scale automated genome mining methodology can be used to effectively predict functional metallophore biosynthetic pathways, we experimentally characterized the metallophores of three bacterial strains. Two strains belong to genera that have no reported metallophores: *Buttiauxella brennerae* DSM 9396 was predicted to produce enterobactin (***Figure 4A***), and *Terasakiispira papahanaumokuakeensis* DSM 29361 was predicted to produce both marinobactin(s) (***Figure 4A***) and enantio-pyochelin (***Figure 1A***). The third strain, *Pseudomonas brassicacearum* DSM 13227, was selected because its genome contains a BGC that clustered with the histicorrugatin reference BGC in the BiG-SCAPE network (***Figure 3***, node 42). We predicted that the BGC may encode the biosynthesis of ornicorrugatin (***Figure 4A***, ***Matthijs et al., 2008***) a previously reported siderophore with no known BGC. A fragmented pyoverdine BGC was also present in the strain's genome, which

was predicted to produce the known siderophore pyoverdine A214 (***Figure 4A***; ***Uría Fernández et al., 2003***; ***Matthijs et al., 2016***).

The crude supernatant extract from the predicted enterobactin producer *B. brennerae* was subjected to semi-preparative reverse-phase high-pressure liquid chromatography (HPLC) coupled to a UV/visible spectrophotometer. Four catecholic compounds were identified by their characteristic absorbance of UV light at 310 nm (***Figure 4—figure supplement 1***). Their relative retention times were consistent with compounds found in supernatant extracts of the known enterobactin producer *Escherichia coli*, which produces the enterobactin fragments 2,3-DHB–Ser (DHBS), $(DHBS)_2$ and linear $(DHBS)_3$, in addition to enterobactin (cyclic $DHBS_3$; ***Figure 4B***; ***Santichaivekin et al., 2021***). The four compounds were purified and subjected to ultra-performance liquid chromatography (UPLC) coupled to electrospray ionization mass spectrometry (ESI-MS) (***Figure 4—figure supplements 2 and 3***). The molecular ions of the purified compounds were each consistent with the enterobactin family (*m/z* 242.0647, 465.1157, 688.1614, and 670.1469, respectively).

Marinobactins A-E, differing in the identity of their fatty acid tails (***Figure 4A***), were previously isolated from *Marinobacter* spp (***Morel et al., 2024***). We searched for the compounds in the crude extract of *T. papahanaumokuakeensis* by UPLC-ESI-MS. Molecular ions consistent with all five of the marinobactins were detected (*m/z* 932.4986, 958.5159, 960.5315, 986.5422, and 988.5421, respectively, ***Figure 4C***, ***Figure 4—figure supplement 4***). The crude extract was then analysed by tandem ESI-MS/MS, yielding fragmentation for the peaks putatively corresponding to marinobactins A–D (the peak at *m/z* 988.5421, putatively marinobactin E, was low abundance and did not give a clear spectrum). Each had similar fragmentation patterns, with b_4, b_5, y_1, and y_2 fragments further supporting the assignments of the marinobactins (***Figure 4—figure supplements 5–8***). No peaks consistent with enantio-pyochelin (*m/z* 324.4; ***Figure 1A***) could be observed.

Siderophores were likewise identified from the crude extract of *P. brassicacearum* by UPLC-ESI-MS and -MS/MS. As hypothesised, ornicorrugatin (***Figure 4A***) was indeed identified by searching the chromatogram for the expected molecular weight (*m/z* 1012.5). A candidate peak was identified with a singly, doubly, and triply charged molecular ion consistent with ornicorrugatin (*m/z* 1012.46, 506.74, and 338.17, respectively, ***Figure 4D***, ***Figure 4—figure supplement 9***). The tandem MS/MS spectrum closely matched that previously reported for ornicorrugatin (***Figure 4—figure supplement 10***

Table 2. Taxonomic distribution of 4953 NRP-metallophore BGC regions detected in 59,851 GTDB representative bacterial genomes.
Phylum nomenclature is preserved from GTDB r207. An additional 413 BGC regions with 'unknown' taxonomy are not included here. Phyla not listed had zero detected regions.

Phylum	Number of detected NRP metallophore BGC regions	Percentage of total detected NRP-met regions	Proportion of genomes with ≥1 NRP-met regions
Proteobacteria	2439	49	2042/16536 (12%)
Actinomycetota	1986	40	1561/6931 (23%)
Cyanobacteria	200	4.0	176/1318 (13%)
Firmicutes_I	192	3.9	191/4013 (4.8%)
Myxococcota	55	1.1	52/418 (12%)
Firmicutes	28	0.6	28/9026 (0.3%)
Chloroflexota	18	0.4	14/1317 (1.1%)
Nitrospirota	16	0.3	15/307 (4.9%)
Acidobacteriota	9	0.2	9/836 (1.1%)
Desulfobacterota	5	0.1	5/847 (0.6%)
Verrucomicrobiota	2	<0.1	2/1304 (0.2%)
Planctomycetota	1	<0.1	1/1034 (0.1%)
Bdellovibrionota	1	<0.1	1/248 (0.4%)
Gemmatimonadota	1	<0.1	1/345 (0.3%)

and ***Supplementary file 1d***; ***Parks et al., 2022***). A chromatographic peak corresponding to pyoverdine was identified by the characteristic absorbance of the chromophore at 400 nm. A candidate compound was identified with singly and doubly charged molecular ions at *m/z* 1134.41 and 567.72, respectively (***Figure 4D***, ***Figure 4—figure supplement 11***). As predicted from the BGC analysis, this mass was identical to pyoverdine A214, the previously characterized siderophore of *P. brassicacearum* subsp. *brassicacearum* NFM 421 and *Pseudomonas* sp. A214 (***Figure 4A***; ***Parks et al., 2018***; ***Gavriilidou et al., 2022***). MS/MS fragmentation yielded B-fragments identical to those previously reported (***Figure 4—figure supplement 12***; ***Parks et al., 2018***).

Thus, our method was able to successfully identify the putative BGC for the orphan siderophore ornicorrugatin and also correctly predict the potential to produce known siderophores by taxa that were not yet studied for their metallophore biosynthetic capacities.

Taxonomic distribution of NRP-metallophores

We investigated the taxonomic distribution and evolution of NRP siderophore biosynthesis within the bacterial kingdom by applying our antiSMASH detection rule to 59,851 representative bacterial genomes from the Genome Taxonomy Database (GTDB) (***Parks et al., 2022***). Among these, 4,098 genomes (6.8%) were predicted to contain at least one NRP metallophore BGC. A total of 5366 BGC regions were detected, representing 14% of all detected NRPS regions. Taxonomic distribution analysis using the GTDB phylogeny highlighted the uneven prevalence of NRP-metallophores across bacterial phyla (***Table 2***). Proteobacteria and Actinomycetota were overrepresented in the GTDB representatives, together accounting for 89% of all detected NRP metallophore regions. After correcting for the number of representative genomes in each phylum, NRP metallophore BGCs were most abundant in Actinomycetota, with 23% of genomes containing at least one detectable region. Proteobacteria, Cyanobacteria, and Myxococcota each had similar proportions of genomes with NRP metallophore BGCs; however, due to biased coverage in the GTDB database, 49% of the detected BGC regions were from Proteobacteria, compared to only 4% and 1.1% found in Cyanobacteria and Myxococcota. Thus, we expect that further sequencing efforts directed at these two phyla will yield many new NRP metallophore BGCs.

To map the distribution of NRP-metallophore producers across the bacterial kingdom, we employed Relative Evolutionary Divergence (RED) values, a framework proposed by ***Parks et al., 2018*** and utilized within the GTDB. Building on this, ***Gavriilidou et al., 2022*** defined REDgroups—phylogenetically consistent clusters based on RED values—that provide a standardized framework analogous to genera. Unlike traditional genera, which can vary significantly in their evolutionary distances, REDgroups offer greater consistency in evolutionary relationships among their members. This framework allowed us to summarize the data as the average number of NRP-metallophore BGC regions per genome within each group, enabling effective visualization and more equitable comparative analyses of biosynthetic potential across bacterial lineages. By collapsing the GTDB tree to the REDgroup level, we annotated each group with the average number of putative NRP-metallophore clusters (***Figure 5***, Ring C). The analysis revealed that 16% of REDgroups encoded detected NRP-metallophores, and within each REDgroup, the number of NRP-metallophores was relatively consistent (standard deviation: 0.3425). This observation aligns with the findings of ***Gavriilidou et al., 2022***, who demonstrated that BGC diversity is consistent at the genus level. While most REDgroups with NRP-metallophores averaged one per genome, several REDgroups, particularly within Actinomycetota, Proteobacteria, and Cyanobacteria exhibited higher averages, with some exceeding two per genome (***Figure 5***, Ring C). These results reveal lineage-specific patterns in siderophore biosynthesis and highlight the utility of REDgroups as an alternative to traditional taxonomic units.

Evolution of gene families and phylogenetic reconciliation to uncover the evolutionary history of NRP-metallophores

To investigate the evolution and origins of NRP-metallophores, we conducted a detailed phylogenetic analysis of each chelator group. Elucidating the evolutionary history of bacterial gene families is complicated by gene duplications, horizontal gene transfers (HGTs), and deletions that cause discordance between the bacterial species phylogeny and each chelator gene phylogeny. To reconcile the trees, we used the software package eMPRess, which infers the most likely series of duplication, HGT, and deletion events (maximum parsimony reconciliation) to reconstruct the evolutionary history

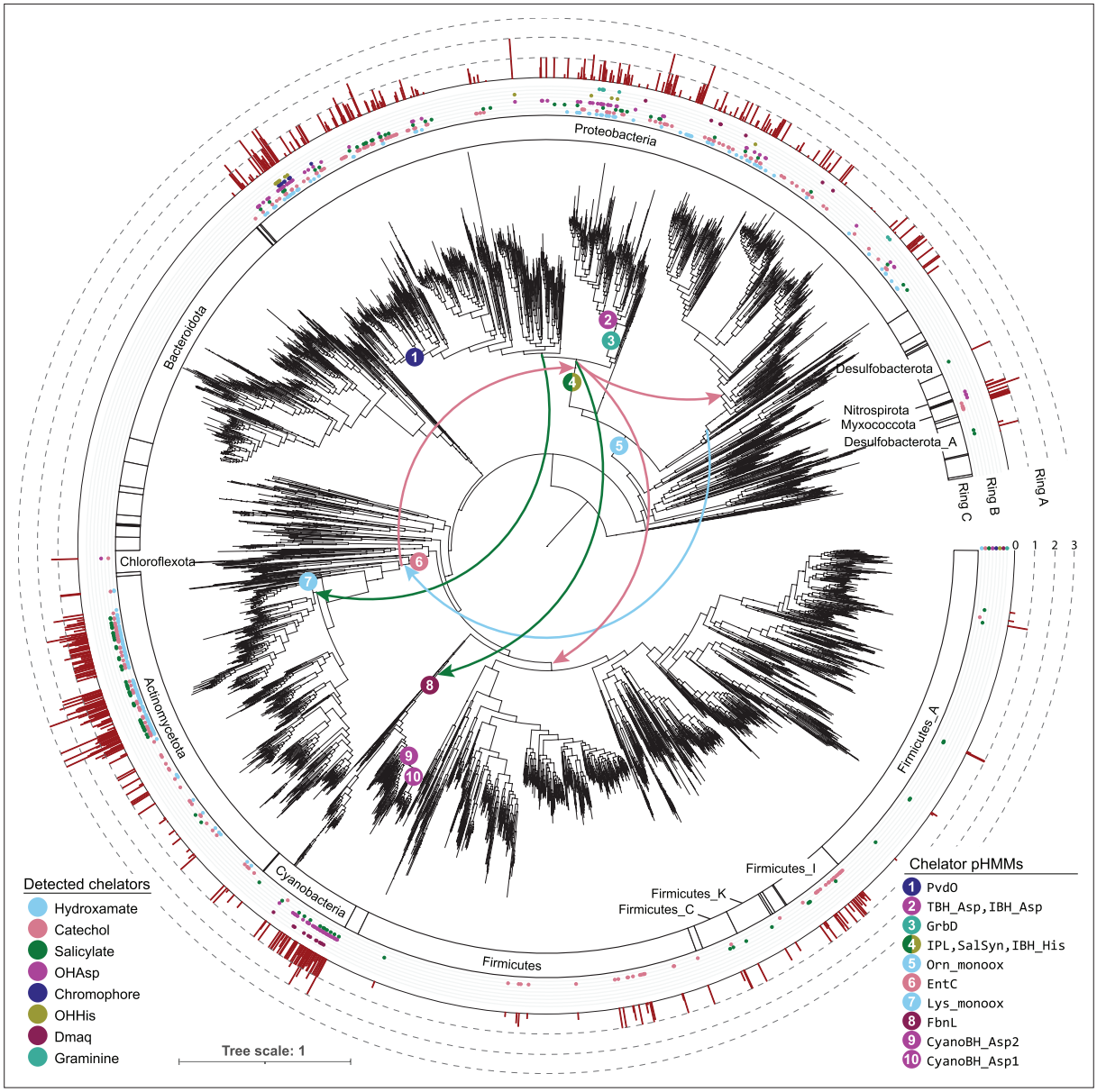


Figure 5. NRP metallophore biosynthesis across the bacterial kingdom. **Center:** The Genome Taxonomy Database (GTDB) phylogenetic tree (version r207), with strains collapsed to the REDgroup level (***Gavriilidou et al., 2022***). Numbered circles indicate the most parsimonious origins of chelator pathways, as determined by reconciliation with eMPRess (***Santichaivekin et al., 2021***). The bottom-right legend lists the specific hidden Markov models (pHMMs) associated with each estimated origin. Arrows indicate ancient horizontal gene transfers predicted by eMPRess. **Ring A:** Phylal divisions. Phyla with detected chelating groups are labeled using nomenclature from GTDB r207. **Ring B:** Chelator biosynthetic pathways detected in at least one member of each REDgroup. **Ring C:** Average number of detected NRP metallophore BGC regions per genome for each REDgroup. Annotations were mapped to the phylogenetic tree using iTOL v6 (***Letunic and Bork, 2024***).

of the gene family (***Santichaivekin et al., 2021***). We first extracted non-fragmented BGC regions from the GTDB representative genomes, then clustered them with BiG-SCAPE to yield 1108 representative BGCs. From these BGCs, we extracted chelator (***Morel et al., 2024***) biosynthesis genes and reconstructed gene trees, which were then compared to the GTDB species tree with eMPRess (***Santichaivekin et al., 2021***). We validated the eMPRess results by analyzing two gene families with the likelihood-based tool AleRax (***Morel et al., 2024***), which produced duplication, transfer, and loss (DTL) patterns consistent with those obtained from eMPRess.

Estimates for the origins and early HGTs of the chelating groups are presented in the center of ***Figure 5***. Reconciliation indicates that the most wide-spread chelating groups—catechols, hydroxamates, and salicylates—are among the most ancient. Genes for producing 2,3-DHB may have originated in a common ancestor of Actinobacteriota (ca. 2.7 Ga, according to rough estimates from TimeTree; ***Marin et al., 2017***) and were then transferred stepwise to Proteobacteria and to Firmicutes. Salicylate biosynthesis genes have an estimated origin in a common ancestor of Gammaproteobacteria (ca. 1.9 Ga; ***Marin et al., 2017***), with early HGT to Cyanobacteria and Actinobacteriota. Hydroxamate NRP metallophores appear to have originated in the common ancestor of Alpha- and Gammaproteobacteria (ca. 2.3 Ga; ***Marin et al., 2017***) and were transferred into Actinobacteria, while Lys-based hydroxamates evolved in Actinobacteriota. The other chelator groups display a more phylum-specific distribution, with HGT predominantly occurring within the same phylum (see Supplemental Dataset, empress_reconciliations). Dmaq is predicted to be among the oldest chelating groups and may have been produced by the common ancestor of Cyanobacteria (ca. 2.7 Ga; ***Marin et al., 2017***), while the pyoverdine chromophore, exclusively observed within the order Pseudomonadales, likely represents one of the most recent siderophore biosynthetic pathways (ca. 1.2 Ga; ***Marin et al., 2017***).

Discussion

Trace metal starvation shapes interactions within microbial communities and between bacteria and the host; therefore, natural and synthetic microbiomes cannot be understood without knowing the metallophore biosynthetic potential of the community. High-throughput biotechnological applications will benefit from in silico metallophore prediction due to the prohibitively high cost of isolation and characterization. To date, distinguishing peptidic metallophore BGCs from other NRPS BGCs has been largely limited to manual expert analyses, leading to blind spots in our understanding of microbes and their communities. We have now automated bacterial NRP metallophore prediction by extending the secondary metabolite prediction platform antiSMASH to detect genes involved in the biosynthesis of metal chelating moieties, enabling the first global analysis of bacterial metallophore biosynthetic diversity.

The presence of genes encoding catechols, hydroxamates, and other chelating groups is one of the most frequently used markers of a metallophore BGC (***Reitz and Medema, 2022***). We have formalized and automated the identification of eight chelator pathways, allowing us to detect 78% of NRP metallophore BGCs with a 3% false-positive rate against a manually annotated set of NRPS clusters. Biosynthetic genes are detected with custom pHMMs and significance score cutoffs calibrated for accurate metallophore discovery, diminishing the ambiguity of interpreting gene annotations, protein families, and BLAST results. We acknowledge that human biases may have influenced which clusters were coded as putative metallophores during both algorithm development and testing; however, expert manual curation remains the most reliable method for NRP metallophore BGC detection. Unfortunately, 22% of manually identifiable metallophore BGCs could not be automatically distinguished from other NRPS clusters, as the algorithm developed (for the purpose of being easily integrated into antiSMASH) relies on the presence of one or more known chelator biosynthesis genes colocalized with the NRPS genes. Detection rates were also lower for fragmented genomes, a limitation (inherent to antiSMASH itself; ***Blin et al., 2017***) that hinders the identification of metallophore biosynthesis in metagenomes. As long-read sequencing of metagenomes becomes more common, we expect that detection will improve.

Recently, Crits-Christoph et al. demonstrated the use of transporter families to predict that a BGC encodes siderophore (or metallophore) biosynthesis (***Crits-Christoph et al., 2021***). Among our test dataset, the biosynthesis-based antiSMASH rule outperformed the transporter-based approach (F1=0.86 versus F1=0.69). However, some putative metallophore BGCs were only found using the transporter-based approach, and a combined either/or ensemble approach slightly outperformed the antiSMASH rule alone (F1=0.90). Biosynthetic- and transporter-based techniques are thus complementary, and future work could incorporate transporter genes into antiSMASH metallophore prediction. We note that the reported transporter-based approach uses just three pHMMs from Pfam, while our biosynthetic detection requires many custom pHMMs. An extended set of metallophore-specific transporter pHMMs designed according to the same principles as those followed for the biosynthesis-related pHMMs could significantly improve detection by reducing false positives and capturing other

families of transporters. The NRP metallophore BGCs discovered in this study could serve as a dataset for developing a more comprehensive model for metallophore transporter detection.

The diverse enzyme families responsible for the biosynthesis of NRP metallophore chelating groups (***Figure 1B***) evince that metal chelation has evolved multiple times, and we expect that more NRPS chelator substructures remain undiscovered. In fact, during manuscript preparation, the novel chelator 5-aminosalicylate was reported in the structure of the *Pseudonocardia* NRP siderophore pseudonochelin (***Zhang et al., 2022***), and we found several unexplored clades of Fe(II)/α-ketoglutarate-dependent amino acid β-hydroxylases that are likely involved in metallophore biosyntheses (***Figure 1—figure supplement 2***). Additionally, we have likely identified a new biosynthetic pathway in the genome of *Sporomusa termitida* DSM 4440, which encodes a partial menaquinone pathway in place of a salicylate synthase to putatively produce a novel karamomycin-like metallophore (***Figure 1—figure supplement 3***; ***Shaaban et al., 2019***). The modular nature of the pHMM-based detection rule will allow for new chelating groups to be added as their biosyntheses are experimentally characterized.

Metallophore BGC regions from representative genomes were compared to reference BGCs and de-replicated into gene cluster families (GCFs) with BiG-SCAPE (***Figure 3***). We found 1093 metallophore BGC regions that were dissimilar from any reference BGC, and almost 500 distinct BGC regions were found in only a single strain. Although significant biosynthetic diversity remains undiscovered, cluster de-replication will become increasingly important to avoid re-isolating known compounds. We also assessed the taxonomic distribution of NRP-metallophore BGC regions by mapping their presence onto a GTDB REDgroup phylogeny. We found that Cyanobacteria and Myxococcota were underrepresented in our analyses due to a relatively low number of published genomes. Considering that only a handful of NRP metallophores have been isolated from these two phyla, we suggest that Cyanobacteria and Myxococcota deserve coordinated efforts of genomic sequencing and experimental work to further characterize their metallophore diversity.

Finally, we used our dataset of detected BGCs and paired taxonomic data from GTDB to investigate the complex evolutionary history of chelating group biosynthesis by reconstructing the most likely origin and major HGT events for each pathway with eMPRess (***Figure 5***; ***Santichaivekin et al., 2021***). Catechols, hydroxamates, and salicylates were among the most widespread and ancient chelators with evidence of HGT between phyla. This widespread distribution suggests significant ecological relevance for these chelators in diverse bacterial lineages. Intriguingly, our timeline estimates place the origin of 2,3-DHB and Dmaq prior to the Great Oxygenation Event (~2.4–2.1 Ga), during an era of abundant, soluble ferrous iron. This result leads credence to the hypothesis that chelating molecules first evolved as metal detoxification mechanisms and were repurposed when oxidized iron became scarce (***Kramer et al., 2020***). Tracing ancient evolutionary events, particularly those involving multiple gene gains and losses, remains challenging due to the exponential increase in complexity as the number of possible events grows. More detailed examinations dedicated to each individual chelating group may yield deeper insights into the complex evolutionary history of these pathways. For example, the origin of hydroxamates must consider the homologous enzymes in NRPS-independent siderophore pathways, and we cannot yet state if metallophore-specific β-OHAsp biosynthesis is polyphyletic due to repeated incorporation into metallophores or a single incorporation followed by repeated transfer into non-chelating roles. Nevertheless, this study represents, to our knowledge, the first attempt at a large-scale phylogenetic analysis into the origin of chelating groups in bacteria.

By integrating chelator detection into antiSMASH, we have taken a major step towards accurate, automated NRP metallophore BGC detection. The new strategy affords a clear practical improvement over manual curation, and has already allowed for the high-throughput identification of thousands of likely NRP metallophore BGC regions, both in this study and in several other recently published analyses that have been enabled by early availability of our methodology (***Mohite et al., 2024***; ***Jørgensen et al., 2024***). A future antiSMASH module might predict metallophore activity more accurately with a machine learning algorithm that considers multiple forms of genomic evidence, including the presence of transporter genes, NRPS domain architecture and sequence, metal-responsive regulatory elements, and other markers of metallophore biosynthesis that are still limited to manual inspection (***Reitz and Medema, 2022***). In particular, regulatory elements will likely be required to accurately distinguish siderophores, zincophores, and other classes of metallophores. We also envision a later version that predicts specific structural features of the metallophore by identifying accessory biosynthetic genes. Implementation of NRP metallophore BGC detection into antiSMASH will enable

scientists of diverse expertises to identify and quantify NRP metallophore biosynthetic pathways in their bacterial genomes of interest and promote large-scale investigations into the chemistry and biology of metallophores.

Methods

For all software, default parameters were used unless otherwise specified. All python, R, and bash scripts used in this paper, as well as underlying data, is available in the Supplemental Dataset, published to Zenodo: 10.5281/zenodo.17806971. All strains were obtained from the Leibniz Institute DSMZ. Unless otherwise stated, all water is doubly deionized (dd H_2O), 18 MΩ.

Profile hidden Markov model construction

Profile hidden Markov models (pHMMs) were built from biosynthetic genes in known metallophore pathways, supplemented with putative BGC genes where required (***Figure 1—figure supplement 1*** and Supplemental Dataset, 1_development/). Amino acid sequences were aligned with MUSCLE (v3.8) (***Edgar, 2004***) and pHMMs were constructed using hmmbuild (HMMER3) (***Eddy, 2011***). pHMMs were tested against the MIBiG database (v2.0) (***Kautsar et al., 2020***) and an additional 37 NRP siderophore BGCs from literature (***Supplementary file 1a***) using hmmsearch (HMMER3). Rough bitscore significance cutoffs were determined for each pHMM. More precise cutoffs were assigned by testing against 28,688 NRPS BGC regions from the antiSMASH database (v3) (***Blin et al., 2021a***). BGC regions containing genes near the rough cutoff were manually inspected to determine if these were likely metallophore BGCs based on the presence of genes encoding membrane transport, metal acquisition (ex: ferric reductases), and multiple chelating groups. If no clear bitscore cutoff could be discerned, representative low-scoring putative true hits were added to the pHMM seed alignment. This process was repeated until a precise bitscore cutoff could be determined.

Detection of catechols and phenols

Like the aromatic amino acids, 2,3-dihydroxybenzoic acid (2,3-DHB) and salicylic acid are derived from chorismate (***Figure 1B***; ***Raymond et al., 2003***; ***Serino et al., 1995***; ***Pelludat et al., 2003***). The three-step synthesis of 2,3-DHB is catalyzed by an isochorismate synthase (**EntC**), an isochorismatase (EntB), and a 2,3-dihydro-2,3- dihydroxybenzoate dehydrogenase (**EntA**) (***Raymond et al., 2003***). The sole presence of **EntA** matches in a cluster was sufficient to detect 2,3-DHB-containing metallophore BGCs; however, only **EntC** could be used to eliminate the nearly-identical 3-hydroxyanthranilic acid pathway (***Pavlikova et al., 2018***), and thus both **EntA** and **EntC** were required to accurately detect 2,3-DHB production. Salicylate biosynthesis was detected by the presence of either an isochorismate pyruvate-lyase (**IPL**) (***Serino et al., 1995***) or a bifunctional salicylate synthase (**SalSyn**) (***Pelludat et al., 2003***). Two condensation domain subtypes specific to catecholic and phenolic metallophores were also included as independent detection rules: VibH-like enzymes (**VibH**), which condense 2,3-DHB to diamines and polyamines, (***Keating et al., 2002***; ***Reitz and Butler, 2020***) and tandem heterocyclization domains (**Cy_tandem**) are sometimes responsible for Ser, Thr, or Cys cyclization (***Bloudoff et al., 2017***).

Detection of hydroxamic acids

Hydroxamates are all produced by the hydroxylation and acylation of a primary amine. In peptidic metallophores, the amine is usually the sidechain of ornithine (Orn) or Lys, which is first hydroxylated by a flavin-dependent monooxygenase (**Orn_monoox** or **Lys_monoox**, respectively). Vicibactin is an unusual hydroxamate siderophore (***Heemstra et al., 2009***) that could not be accurately captured by either **Orn_monoox** or **Lys_monoox**. The Orn hydroxylase VbsO is more similar to those found in NIS pathways, and a VbsO pHMM was non-specific. Instead, the acyl-hydroxyornithine epimerase (***Heemstra et al., 2009***) **VbsL** is used to detect vicibactin. In two non-metallophore pathways (***Du et al., 2017***; ***Matsuda et al., 2022***), hydroxyornithine or hydroxylysine are the substrates for N-N bond-forming enzymes. No clear bitscore separation could be achieved to eliminate these false positives, so two negative constraints were added: the presence of **KtzT** associated with biosynthesis of piperazates (***Du et al., 2017***), and **MetRS-like**, associated with several hydrazines (***Matsuda et al., 2022***). Unfortunately, false positives may still arise if the ornithine monooxygenase and the negative

constraint genes are distantly located within the same BGC region (>30 kbp), such as in the himastatin locus (MIBiG BGC0001117).

Detection of graminine, Dmaq, and the pyoverdine chromophore

Although the biosynthesis of the chelating amino acid graminine has not been fully elucidated, gene knockouts and stable isotope studies have revealed two enzymes, **GrbD** and **GrbE**, responsible for diazeniumdiolate formation from arginine (***Hermenau et al., 2019***; ***Vinnik et al., 2021***) The quinoline chelator Dmaq was first identified in anachelins (***Beiderbeck et al., 2000***) and the biosynthesis was recently established for fabrubactins (***Vinnik et al., 2021***). Synthesis is initialized by **FbnL** and **FbnM**, which form a two-protein heme peroxidase that oxidizes L-tyrosine to L-DOPA (***Vinnik et al., 2021***). GrbDE and FbnLM were previously used as handles for genome mining (***Hermenau et al., 2019***; ***Vinnik et al., 2021***), and our pHMMs gave similar results. We did not include detection of a pathway currently only reported in fabrubactins that produces two α-hydroxycarboxylate chelating moieties (***Figure 1A*****,** bolded atoms) (***Vinnik et al., 2021***). Both substructures are proposed to be synthesized by the flavin-dependent monooxygenase FbnE. A profile HMM was built for FbnE (Supplemental Dataset) but was not included in antiSMASH, as all BGCs with **FbnE** hits were either captured independently by **FbnL** and **FbnM** (Dmaq biosynthesis), or appeared to be false positives based on a lack of other metallophore-related genes. The diverse pyoverdine family of siderophores is defined by a fluorescent chelating chromophore (***Figure 1***). Chromophore maturation is driven by the tyrosinase **PvdP** and the oxidoreductase **PvdO** (***Nadal-Jimenez et al., 2014***; ***Ringel et al., 2018***). The two genes are sometimes in separate loci, and including both as independent pHMMs increased the number of pyoverdine BGCs detected. These pHMMs also capture the biosynthesis pathway for azotobactin, which contains a slightly different chromophore (***Figure 3***; ***Demange et al., 1988***).

Detection of β-hydroxyamino acids and phylogenetic analysis of Asp and His β-hydroxylases

Our previous analysis of siderophores containing β-hydroxy-aspartate (β-OHAsp) and β-hydroxy-histidine (β-OHHis) delineated three families of siderophore-specific Fe(II)/α-ketoglutarate-dependent enzymes responsible for β-hydroxylation of Asp (**TBH_Asp** and **IBH_Asp**) or His (**IBH_His**) (***Reitz et al., 2019***). More recently, β-OHAsp-containing siderophores named cyanochelins were isolated from a diverse group of cyanobacteria (***Galica et al., 2021***). The BGCs contain two additional classes of β-hydroxylases, which we have tentatively named **CyanoBH_Asp1** and **CyanoBH_Asp2**. Because of the polyphyletic nature of these hydroxylases, well-performing pHMMs could not be made using the standard workflow used for other pathways (***Figure 1—figure supplement 1***); even after many iterations, non-metallophore NRPs were captured as false positives, such as the β-OHAsp-containing lipopeptide turnercyclamycin (***Miller et al., 2021***).

An expanded phylogenetic analysis was performed to serve as a guide for pHMM construction (Supplemental Dataset, 1_development/hydroxylase_tree/). NRPS BGC regions from the antiSMASH database (v3) were scanned for matches to previously reported β-hydroxylase pHMMs (***Reitz et al., 2019***) and Pfam pHMMs for siderophore-related transporters (PF00593, PF01032, and PF01497; ***Crits-Christoph et al., 2021***; ***El-Gebali et al., 2019***) using a modified version of antiSMASH v6.0. β-Hydroxylase genes meeting a relaxed bitscore cutoff of 300 (1070 genes total) were dereplicated with CD-HIT web server (***Huang et al., 2010***) and a sequence identity cutoff of 70%, giving 425 representative amino acid sequences. A multiple sequence alignment was created using hmmalign (HMMER3) and the TauD Pfam (PF02668) (***El-Gebali et al., 2019***) and a maximum-likelihood phylogenetic tree was reconstructed with IQ-TREE (multicore v2.2.0-beta) (***Nguyen et al., 2015***) using the WAG +F + I+G4 evolutionary model. The presence of nearby transporters was mapped onto the phylogenetic tree to identify clades or paraphyletic groups putatively involved in siderophore biosynthesis. Sequences in groups corresponding to previously reported TBH_Asp, IBH_Asp, and IBH_His subtypes and the novel putative CyanoBH_Asp1 and CyanoBH_Asp2 subtypes were extracted, and pHMMs were constructed and tested as described above. Several unreported clades of amino acid β-hydroxylases were identified that may also be involved in metallophore biosynthesis based on their co-occurrence with transporter genes (***Figure 1—figure supplement 2***). However, these putative β-hydroxylases were not included in the current NRP metallophore rules due to a lack of experimental evidence for metallophore production. A negative constraint (in the form of a competing pHMM for

SBH_Asp) was used to exclude the syringomycin family of BGCs, which contain a clade of β-hydroxylases that sit within the IBH_Asp clade (***Reitz et al., 2019***).

Incorporation into antiSMASH

The pHMMs and cutoffs were added to antiSMASH as a single detection rule called 'NRP-metallophore' with the following logic:

```
VibH_like or Cy_tandem or
(cds(Condensation and AMP-binding) and (
      (IBH_Asp and not SBH_Asp) or IBH_His or TBH_Asp or
      CyanoBH_Asp1 or CyanoBH_Asp2 or
      IPL or SalSyn or (EntA and EntC) or
      (GrbD and GrbE) or (FbnL and FbnM) or PvdO or PvdP or
      (Orn_monoox and not (KtzT or MetRS-like))
      Lys_monoox or VbsL))
```

Manual validation

A subset of RefSeq representative bacterial genomes was generated by randomly selecting one genome for each of the 330 genera determined by GTDB (Supplemental Dataset, 3_manual_testing/). All NRPS BGC regions in the genomes were annotated with antiSMASH v6.1, yielding 758 BGC regions in the final testing dataset (***Supplementary file 1b***). The antiSMASH output for each BGC was manually inspected for evidence of NRP metallophore production, including genes encoding transporters, iron reductases, chelator biosynthesis, and known metallophore NRPS domain motifs. The same 758 BGC regions were classified as NRP metallophores using the chelator-based strategy described above. We compared the performance against the transporter-based strategy of ***Crits-Christoph et al., 2021***. Genes matching Pfam pHMMs for siderophore-related transporters (PF00593, PF01032, and PF01497) (***Crits-Christoph et al., 2021***; ***El-Gebali et al., 2019***) were identified using a custom version of antiSMASH v6.1, in which the three pHMMs were added to the list of domains that are identified and labeled as 'biosynthetic-additional'. BGC regions were classified as metallophores if two of the three transporter families were present (***Crits-Christoph et al., 2021***). Each putative false positive was re-investigated before performance statistics were calculated, resulting in the reannotation of four BGCs.

BIG-SCAPE clustering

The 3264 NRP metallophore BGC regions from RefSeq representative genomes (Supplemental Dataset, 2_refseq_reps_results/metallophores_Jun25.tar.gz) were filtered to remove BGC regions on contig boundaries. The resulting 2523 BGC regions, as well as 78 previously reported BGCs (***Supplementary file 1b***) were clustered using BiG-SCAPE v1.1.2 with the following settings: "`--no_classify --mix --cutoffs 0.3 0.4 0.5 --clans-off`". The network (Supplemental Dataset, 6_bigscape/mix_c0.50.network) was imported to Cytoscape for figure preparation.

Bacterial culturing and siderophore isolations

Terasakiispira papahanaumokuakeensis DSM 29361 and *Buttiauxella brennerae* DSM 9396 were cultivated in a 4 L Erlenmeyer flask acid washed with 4 M HCl. *T. papahanaumokuakeensis* was grown in 2 L of the following medium: glycerol phosphate disodium 4 g/L, NaCl 25 g/L, KCl 0.67 g/L, $CaCl_2$ * 2 H_2O 1.36 g/L, NH_4Cl 4 g/L, L sodium succinate hexahydrate 5 g/L and 25 mM MOPS pH 7.5. For *B. brennerae*, the medium was 2 L: K_2HPO_4 7 g/L, KH_2PO_4 2 g/L, NaCl 0.6 g/L, $MgSO_4$ * 7 H_2O 1 g/L, NH_4SO_4 1 g/L sodium succinate hexahydrate 5 g/L. After autoclaving 20 mL of filter sterilized D-glucose 50% (w/v) was added, completing the medium. The cultures were incubated at room temperature on an orbital shaker, 180 RPM. After 41 h and 48 h, the cultures of *T. papahanaumokuakeensis* and *B. brennerae*, respectively, were centrifuged at 6,000xg for 10 min at 4°C. The supernatant was decanted and approximately 100 g of XAD-4 resin added. The solution was agitated gently at room temperature for approximately 4 h. Organic compounds were eluted from the resin with 100% methanol, which after rotary evaporation yielded the crude extract. Four catechol compounds produced by

B. brennerae were isolated using reverse phase (RP) HPLC. The HPLC system was comprised of a YMC AQ12S05-2520WT column connected to two Waters 515 HPLC pumps and a waters 2487 dual wavelength absorbance detector. The mobile phase consisted of dd H_2O and methanol, both amended with 0.05% trifluoroacetic acid (w/v). To separate the compounds a method of 10–100% methanol, 3% per min was utilized. The crude extract of *T. papahanaumokuakeensis* and the catechol compounds of *B. brennerae* were analyzed by ESI-MS and MS/MS on a Waters XevoG2-XS QTof instrument in positive mode electrospray ionization coupled to an ACQUITY UPLC H-Class system with a Waters BEH C18 column.

Pseudomonas brassicacearum DSM 13227 was cultured in a modified M9 minimal medium composed of 3.0 g/L disodium hydrogen phosphate heptahydrate, 1.5 g/L potassium dihydrogen phosphate, 2.5 g/L sodium chloride, 0.50 g/L ammonium chloride, 2.5 g/L disodium succinate, and 5.0 g sodium pyruvate in ultrapure (e.g., 18 mW) water, amended after autoclave sterilization with 20 mL/L 50% w/v Steri-filtered glucose solution, 1 mL/L 1 M magnesium chloride, and 0.05 mL/L 1 M calcium chloride. Culturing glassware was rinsed with 4 M HCl before use to remove adsorbed iron(III). Seed cultures were inoculated in Luria-Bertani (LB) broth with single colonies of *P. brassicacearum* DSM 13227 grown on LB agar and grown for at least 24 h at 30°C. Cultures were monitored by OD_{600}, and the culture supernatant was harvested at late log/early stationary phase (OD_{600}~1.2 au) with a positive Fe(III)-CAS response. Culture supernatants were obtained by centrifugation at 6500 rpm for 20 min at 4°C. To extract the siderophores, the culture supernatant was decanted and shaken with 100 g of Amberlite XAD-4 resin. The XAD-4 resin was prepared by washing with methanol and then equilibrating with ultrapure water. The resin and supernatant were orbitally shaken to equilibrate for 2 h at 150 rpm. The resin was filtered from the supernatant and washed with 0.5 L of ultrapure water. The adsorbed organics were eluted with 80% aqueous methanol. The eluent was concentrated under vacuum and stored at 4°C. Ornicorrugatin and the pyoverdine siderophore were identified by UPLC-MS and LC-MS/MS analysis of the concentrated eluent.

Phylogenetic mapping

Genome mining was performed on 62,291 GTDB representative genomes (59,851 after filtering; version r207) (***Parks et al., 2022***) using AntiSMASH v7.0beta, (***Blin et al., 2023***; ***Parks et al., 2022***) with the inclusion of the NRP metallophore detection module. The outputs were analysed to identify predicted NRP-metallophore producers and categorized into distinct chelator groups based on predefined detection criteria. A total of 5366 NRP-metallophores were identified, representing approximately 14% of all detected NRPS regions. To map the distribution patterns of these producers, the results were integrated with the GTDB tree. Due to the size of the tree, visualization tools such as iTOL (***Letunic and Bork, 2024***) were impractical, prompting dereplication to a higher taxonomic rank. The GTDB tree was collapsed to the REDgroup level—a phylogenetically defined rank analogous to genera—allowing normalization to reflect the average number of NRP-metallophore biosynthetic gene clusters (BGCs) per genome within each REDgroup (***Gavriilidou et al., 2022***).

To uncover the evolutionary history of siderophore biosynthesis, phylogenetic analyses and reconciliation were performed. Gene sequences for each chelator group were extracted from 4060 complete BGCs, filtered to exclude BGC regions on contig boundaries, and clustered into Gene Cluster Families (GCFs) using BiG-SCAPE (***Navarro-Muñoz et al., 2020***) with a 0.5 cutoff. From each GCF, one representative BGC was selected, resulting in a dataset of 1108 clusters. Multiple sequence alignments (MSAs) were conducted using MAFFT v7, (***Katoh and Standley, 2013***) and phylogenetic trees were constructed using FastTree 2 with the WAG model (***Price et al., 2010***). Evolutionary events, including gene duplication, loss, and horizontal gene transfer, were identified using phylogenetic reconciliation in eMPRess (***Santichaivekin et al., 2021***) by comparing gene trees to species trees. To validate the robustness of the inferred evolutionary scenarios, SalSyn and CyanoBH_Asp1 were also reconciled using the likelihood-based tool AleRax (***Morel et al., 2024***). Reconciliation results were annotated using iTOL v6 (***Letunic and Bork, 2024***) for visualization, manually mapping key evolutionary events onto the GTDB tree. Individual gene tree reconciliations are available in the Supplementary Dataset.

Acknowledgements

This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (Starting Grant 948770-DECIPHER; ZR and MM), as

well as from the US National Science Foundation (CHE-2108596; AB). BT and NZ were supported by H2020-FNR-11-2020: SECRETED—Grant agreement: 101000794. NZ was supported by the German Center for Infection Research TTU09.716. We thank Allegra Aron for providing useful feedback on the manuscript.

Additional information

Competing interests

Marnix Medema: M.H.M. is a member of the Scientific Advisory Boards of Hexagon Bio and Hothouse Therapeutics. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
European Research Council	948770-DECIPHER	Marnix Medema
National Science Foundation	CHE-2108596	Alison Butler
HORIZON EUROPE Framework Programme	H2020-FNR-11-2020	Nadine Ziemert
Deutsches Zentrum für Infektionsforschung	TTU09.716	Nadine Ziemert

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Zachary L Reitz, Conceptualization, Data curation, Software, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing – original draft; Bita Pourmohsenin, Conceptualization, Data curation, Formal analysis, Investigation, Visualization, Methodology, Writing – original draft; Melanie Susman, Formal analysis, Investigation, Methodology, Writing – review and editing; Emil Thomsen, Daniel Roth, Formal analysis, Investigation, Writing – review and editing; Alison Butler, Supervision, Funding acquisition, Writing – review and editing; Nadine Ziemert, Marnix Medema, Conceptualization, Supervision, Funding acquisition, Project administration, Writing – review and editing

Author ORCIDs

Zachary L Reitz https://orcid.org/0000-0003-1964-8221
Alison Butler https://orcid.org/0000-0002-3525-7864
Nadine Ziemert https://orcid.org/0000-0002-7264-1857
Marnix Medema https://orcid.org/0000-0002-2191-2821

Peer review material

Reviewer #1 (Public review): https://doi.org/10.7554/eLife.109154.3.sa1
Reviewer #2 (Public review): https://doi.org/10.7554/eLife.109154.3.sa2
Author response https://doi.org/10.7554/eLife.109154.3.sa3

Additional files

Supplementary files

Supplementary file 1. Tabular data for reference BGCs, rule validation, RefSeq genome results, and ornicorrugatin MS/MS. An Excel (xlsx) file containing four sheets: (a) Reference BGCs used in this work. Numbering matches ***Figure 3***. BGCs marked with an asterisk are non-metallophore BGCs that emerged as false positives during rule development. (b) Data for the rule validation. The sheet contains BGC metadata, notes used for manual classification, and individual results from detection rule strategies, as well as summary confusion matrices and performance metrics. (c) Results for applying detection rules to BGCs from NCBI RefSeq representative bacterial genomes, including

bitscore values for significant matches to profile HMMs. (d) Selected MS/MS fragment ions observed for ornicorrugatin produced by Pseudomonas brassicacearum DSM 13227. All given fragment ions agree with previously reported masses for ornicorrugatin (***Matthijs et al., 2008***).

MDAR checklist

Data availability

All python, R, and bash scripts used in this paper, as well as underlying data, is available in the Supplemental Dataset, published to Zenodo: 10.5281/zenodo.18866949. The enterobactin, marinobactin, and ornicorrugatin BGCs have been submitted to the MIBiG repository (***Blin et al., 2021a***) with accession numbers BGC0003172, BGC0003173, BGC0003174, respectively, and are also available in our Supplemental Dataset at Zenodo. The antiSMASH v8 outputs for the BGCs are also available in the Supplemental Dataset.

The following datasets were generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Reitz ZL	2026	Supplemental Dataset for Automated genome mining predicts structural diversity and taxonomic distribution of peptide metallophores across bacteria	https://doi.org/10.5281/zenodo.18866949	Zenodo, 10.5281/zenodo.18866949
Reitz ZL	2026	Enterobactin biosynthetic gene cluster from Buttiauxella brennerae DSM 9396	https://mibig.secondarymetabolites.org/go/BGC0003172	MIGBiG, BGC0003172
Reitz ZL	2026	Marinobactin biosynthetic gene cluster from Terasakiispira papahanaumokuakeensis DSM 29361	https://mibig.secondarymetabolites.org/go/BGC0003173	MIGBiG, BGC0003173
Reitz ZL	2026	Ornicorrugatin biosynthetic gene cluster from Pseudomonas brassicacearum DSM 13227	https://mibig.secondarymetabolites.org/go/BGC0003174	MIGBiG, BGC0003174

References

Behnsen J, Zhi H, Aron AT, Subramanian V, Santus W, Lee MH, Gerner RR, Petras D, Liu JZ, Green KD, Price SL, Camacho J, Hillman H, Tjokrosurjo J, Montaldo NP, Hoover EM, Treacy-Abarca S, Gilston BA, Skaar EP, Chazin WJ, et al. 2021. Siderophore-mediated zinc acquisition enhances enterobacterial colonization of the inflamed gut. *Nature Communications* **12**:7016. DOI: https://doi.org/10.1038/s41467-021-27297-2, PMID: 34853318

Beiderbeck H, Taraz K, Budzikiewicz H, Walsby AE. 2000. Anachelin, the siderophore of the cyanobacterium Anabaena cylindrica CCAP 1403/2A. *Zeitschrift Fur Naturforschung. C, Journal of Biosciences* **55**:681–687. DOI: https://doi.org/10.1515/znc-2000-9-1002, PMID: 11098815

Blin K, Medema MH, Kottmann R, Lee SY, Weber T. 2017. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Research* **45**:D555–D559. DOI: https://doi.org/10.1093/nar/gkw960, PMID: 27924032

Blin K, Shaw S, Kautsar SA, Medema MH, Weber T. 2021a. The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Research* **49**:D639–D643. DOI: https://doi.org/10.1093/nar/gkaa978, PMID: 33152079

Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, Weber T. 2021b. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **49**:W29–W35. DOI: https://doi.org/10.1093/nar/gkab335, PMID: 33978755

Blin K, Shaw S, Augustijn HE, Reitz ZL, Biermann F, Alanjary M, Fetter A, Terlouw BR, Metcalf WW, Helfrich EJN, van Wezel GP, Medema MH, Weber T. 2023. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Research* **51**:W46–W50. DOI: https://doi.org/10.1093/nar/gkad344

Bloudoff K, Fage CD, Marahiel MA, Schmeing TM. 2017. Structural and mutational analysis of the nonribosomal peptide synthetase heterocyclization domain provides insight into catalysis. *PNAS* **114**:95–100. DOI: https://doi.org/10.1073/pnas.1614191114

Cézard C, Farvacques N, Sonnet P. 2015. Chemistry and biology of pyoverdines, Pseudomonas primary siderophores. *Current Medicinal Chemistry* **22**:165–186. DOI: https://doi.org/10.2174/0929867321666141011194624, PMID: 25312210

Crits-Christoph A, Bhattacharya N, Olm MR, Song YS, Banfield JF. 2021. Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity. *Genome Research* **31**:239–250. DOI: https://doi.org/10.1101/gr.268169.120, PMID: 33361114

Demange P, Bateman A, Dell A, Abdallah MA. 1988. Structure of azotobactin D, a siderophore of Azotobacter vinelandii strain D (CCM 289). *Biochemistry* **27**:2745–2752. DOI: https://doi.org/10.1021/bi00408a014

Du YL, He HY, Higgins MA, Ryan KS. 2017. A heme-dependent enzyme forms the nitrogen-nitrogen bond in piperazate. *Nature Chemical Biology* **13**:836–838. DOI: https://doi.org/10.1038/nchembio.2411, PMID: 28628093

Eddy SR. 2011. Accelerated Profile HMM Searches. *PLOS Computational Biology* **7**:e1002195. DOI: https://doi.org/10.1371/journal.pcbi.1002195, PMID: 22039361

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**:1792–1797. DOI: https://doi.org/10.1093/nar/gkh340, PMID: 15034147

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. *Nucleic Acids Research* **47**:D427–D432. DOI: https://doi.org/10.1093/nar/gky995, PMID: 30357350

Galica T, Borbone N, Mareš J, Kust A, Caso A, Esposito G, Saurav K, Hájek J, Řeháková K, Urajová P, Costantino V, Hrouzek P. 2021. Cyanochelins, an overlooked class of widely distributed Cyanobacterial Siderophores, discovered by silent gene cluster awakening. *Applied and Environmental Microbiology* **87**:e0312820. DOI: https://doi.org/10.1128/AEM.03128-20, PMID: 34132591

Gavriilidou A, Kautsar SA, Zaburannyi N, Krug D, Müller R, Medema MH, Ziemert N. 2022. Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nature Microbiology* **7**:726–735. DOI: https://doi.org/10.1038/s41564-022-01110-2, PMID: 35505244

González Carreró MI, Sangari FJ, Agüero J, Garcı A Lobo JM. 2002. Brucella abortus strain 2308 produces brucebactin, a highly efficient catecholic siderophore. *Microbiology* **148**:353–360. DOI: https://doi.org/10.1099/00221287-148-2-353, PMID: 11832499

Gross H, Loper JE. 2009. Genomics of secondary metabolite production by Pseudomonas spp. *Natural Product Reports* **26**:1408–1446. DOI: https://doi.org/10.1039/b817075b, PMID: 19844639

Gu S, Wei Z, Shao Z, Friman VP, Cao K, Yang T, Kramer J, Wang X, Li M, Mei X, Xu Y, Shen Q, Kümmerli R, Jousset A. 2020a. Competition for iron drives phytopathogen control by natural rhizosphere microbiomes. *Nature Microbiology* **5**:1002–1010. DOI: https://doi.org/10.1038/s41564-020-0719-8, PMID: 32393858

Gu S, Yang T, Shao Z, Wang T, Cao K, Jousset A, Friman VP, Mallon C, Mei X, Wei Z, Xu Y, Shen Q, Pommier T. 2020b. Siderophore-mediated interactions determine the disease suppressiveness of microbial consortia. *mSystems* **5**:e00811–e00819. DOI: https://doi.org/10.1128/mSystems.00811-19, PMID: 32606030

Heather Z, Holden MTG, Steward KF, Parkhill J, Song L, Challis GL, Robinson C, Davis-Poynter N, Waller AS. 2008. A novel streptococcal integrative conjugative element involved in iron acquisition. *Molecular Microbiology* **70**:1274–1292. DOI: https://doi.org/10.1111/j.1365-2958.2008.06481.x, PMID: 18990191

Heemstra JR, Walsh CT, Sattely ES. 2009. Enzymatic tailoring of ornithine in the biosynthesis of the Rhizobium cyclic trihydroxamate siderophore vicibactin. *Journal of the American Chemical Society* **131**:15317–15329. DOI: https://doi.org/10.1021/ja9056008, PMID: 19778043

Hermenau R, Mehl JL, Ishida K, Dose B, Pidot SJ, Stinear TP, Hertweck C. 2019. Genomics-driven discovery of no-donating diazeniumdiolate siderophores in diverse plant-associated bacteria. *Angewandte Chemie* **58**:13024–13029. DOI: https://doi.org/10.1002/anie.201906326, PMID: 31276269

Hider RC, Kong X. 2010. Chemistry and biology of siderophores. *Natural Product Reports* **27**:637–657. DOI: https://doi.org/10.1039/b906679a, PMID: 20376388

Huang Y, Niu B, Gao Y, Fu L, Li W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**:680–682. DOI: https://doi.org/10.1093/bioinformatics/btq003, PMID: 20053844

Jaremko MJ, Davis TD, Corpuz JC, Burkart MD. 2020. Type II non-ribosomal peptide synthetase proteins: structure, mechanism, and protein-protein interactions. *Natural Product Reports* **37**:355–379. DOI: https://doi.org/10.1039/c9np00047j, PMID: 31593192

Jørgensen TS, Mohite OS, Sterndorff EB, Alvarez-Arevalo M, Blin K, Booth TJ, Charusanti P, Faurdal D, Hansen TØ, Nuhamunada M, Mourched AS, Palsson BØ, Weber T. 2024. A treasure trove of 1034 actinomycete genomes. *Nucleic Acids Research* **52**:7487–7503. DOI: https://doi.org/10.1093/nar/gkae523, PMID: 38908028

Kage H, Kreutzer MF, Wackler B, Hoffmeister D, Nett M. 2013. An iterative type I polyketide synthase initiates the biosynthesis of the antimycoplasma agent micacocidin. *Chemistry & Biology* **20**:764–771. DOI: https://doi.org/10.1016/j.chembiol.2013.04.010, PMID: 23790487

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**:772–780. DOI: https://doi.org/10.1093/molbev/mst010, PMID: 23329690

Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, van Santen JA, Tracanna V, Suarez Duran HG, Pascal Andreu V, Selem-Mojica N, Alanjary M, Robinson SL, Lund G, Epstein SC, Sisto AC, Charkoudian LK, Collemare J, Linington RG, Weber T, et al. 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research* **48**:D454–D458. DOI: https://doi.org/10.1093/nar/gkz882, PMID: 31612915

Keating TA, Marshall CG, Walsh CT, Keating AE. 2002. The structure of VibH represents nonribosomal peptide synthetase condensation, cyclization and epimerization domains. *Nature Structural Biology* **9**:522–526. DOI: https://doi.org/10.1038/nsb810, PMID: 12055621

Kraemer SM, Duckworth OW, Harrington JM, Schenkeveld WDC. 2015. Metallophores and trace metal biogeochemistry. *Aquatic Geochemistry* **21**:159–195. DOI: https://doi.org/10.1007/s10498-014-9246-7

Kramer J, Özkaya Ö, Kümmerli R. 2020. Bacterial siderophores in community and host interactions. *Nature Reviews. Microbiology* **18**:152–163. DOI: https://doi.org/10.1038/s41579-019-0284-4, PMID: 31748738

Letunic I, Bork P. 2024. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research* **52**:W78–W82. DOI: https://doi.org/10.1093/nar/gkae268, PMID: 38613393

Makris C, Carmichael JR, Zhou H, Butler A. 2022. *C*-Diazeniumdiolate graminine in the siderophore gramibactin is photoreactive and originates from arginine. *ACS Chemical Biology* **17**:3140–3147. DOI: https://doi.org/10.1021/acschembio.2c00593, PMID: 36354305

Marin J, Battistuzzi FU, Brown AC, Hedges SB. 2017. The timetree of prokaryotes: new insights into their evolution and speciation. *Molecular Biology and Evolution* **34**:437–446. DOI: https://doi.org/10.1093/molbev/msw245, PMID: 27965376

Martinez JS, Butler A. 2007. Marine amphiphilic siderophores: marinobactin structure, uptake, and microbial partitioning. *Journal of Inorganic Biochemistry* **101**:1692–1698. DOI: https://doi.org/10.1016/j.jinorgbio.2007.07.007, PMID: 17868890

Matsuda K, Arima K, Akiyama S, Yamada Y, Abe Y, Suenaga H, Hashimoto J, Shin-Ya K, Nishiyama M, Wakimoto T. 2022. A natural dihydropyridazinone scaffold generated from a unique substrate for a hydrazine-forming enzyme. *Journal of the American Chemical Society* **144**:12954–12960. DOI: https://doi.org/10.1021/jacs.2c05269, PMID: 35771530

Matthijs S, Budzikiewicz H, Schäfer M, Wathelet B, Cornelis P. 2008. Ornicorrugatin, a new siderophore from Pseudomonas fluorescens AF76. *Zeitschrift Für Naturforschung C* **63**:8–12. DOI: https://doi.org/10.1515/znc-2008-1-202

Matthijs S, Brandt N, Ongena M, Achouak W, Meyer JM, Budzikiewicz H. 2016. Pyoverdine and histicorrugatin-mediated iron acquisition in Pseudomonas thivervalensis. *Biometals* **29**:467–485. DOI: https://doi.org/10.1007/s10534-016-9929-1, PMID: 27007713

Mehdiratta K, Singh S, Sharma S, Bhosale RS, Choudhury R, Masal DP, Manocha A, Dhamale BD, Khan N, Asokachandran V, Sharma P, Ikeh M, Brown AC, Parish T, Ojha AK, Michael JS, Faruq M, Medigeshi GR, Mohanty D, Reddy DS, et al. 2022. Kupyaphores are zinc homeostatic metallophores required for colonization of *Mycobacterium tuberculosis PNAS* **119**:e2110293119. DOI: https://doi.org/10.1073/pnas.2110293119, PMID: 35193957

Mihara K, Tanabe T, Yamakawa Y, Funahashi T, Nakao H, Narimatsu S, Yamamoto S. 2004. Identification and transcriptional organization of a gene cluster involved in biosynthesis and transport of acinetobactin, a siderophore produced by Acinetobacter baumannii ATCC 19606T. *Microbiology* **150**:2587–2597. DOI: https://doi.org/10.1099/mic.0.27141-0, PMID: 15289555

Miller BW, Lim AL, Lin Z, Bailey J, Aoyagi KL, Fisher MA, Barrows LR, Manoil C, Schmidt EW, Haygood MG. 2021. Shipworm symbiosis ecology-guided discovery of an antibiotic that kills colistin-resistant Acinetobacter. *Cell Chemical Biology* **28**:1628–1637. DOI: https://doi.org/10.1016/j.chembiol.2021.05.003, PMID: 34146491

Mohite OS, Jørgensen TS, Booth TJ, Charusanti P, Phaneuf PV, Weber T, Palsson BO. 2024. Pangenome mining of the Streptomyces genus redefines species' biosynthetic potential. *bioRxiv*. DOI: https://doi.org/10.1101/2024.02.20.581055

Morel B, Williams TA, Stamatakis A, Szöllősi GJ. 2024. AleRax: a tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss. *Bioinformatics* **40**:btae162. DOI: https://doi.org/10.1093/bioinformatics/btae162, PMID: 38514421

Nadal-Jimenez P, Koch G, Reis CR, Muntendam R, Raj H, Jeronimus-Stratingh CM, Cool RH, Quax WJ. 2014. PvdP is a tyrosinase that drives maturation of the pyoverdine chromophore in *Pseudomonas aeruginosa*. *Journal of Bacteriology* **196**:2681–2690. DOI: https://doi.org/10.1128/JB.01376-13, PMID: 24816606

Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH. 2020. A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology* **16**:60–68. DOI: https://doi.org/10.1038/s41589-019-0400-9, PMID: 31768033

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**:268–274. DOI: https://doi.org/10.1093/molbev/msu300, PMID: 25371430

Olucha J, Lamb AL. 2011. Mechanistic and structural studies of the N-hydroxylating flavoprotein monooxygenases. *Bioorganic Chemistry* **39**:171–177. DOI: https://doi.org/10.1016/j.bioorg.2011.07.006, PMID: 21871647

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* **36**:996–1004. DOI: https://doi.org/10.1038/nbt.4229, PMID: 30148503

Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-

based taxonomy. *Nucleic Acids Research* **50**:D785–D794. DOI: https://doi.org/10.1093/nar/gkab776, PMID: 34520557

Pavlikova M, Kamenik Z, Janata J, Kadlcik S, Kuzma M, Najmanova L. 2018. Novel pathway of 3-hydroxyanthranilic acid formation in limazepine biosynthesis reveals evolutionary relation between phenazines and pyrrolobenzodiazepines. *Scientific Reports* **8**:7810. DOI: https://doi.org/10.1038/s41598-018-26179-w, PMID: 29773836

Pelludat C, Brem D, Heesemann J. 2003. Irp9, encoded by the high-pathogenicity island of Yersinia enterocolitica, is able to convert chorismate into salicylate, the precursor of the siderophore yersiniabactin. *Journal of Bacteriology* **185**:5648–5653. DOI: https://doi.org/10.1128/JB.185.18.5648-5653.2003, PMID: 12949119

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**:e9490. DOI: https://doi.org/10.1371/journal.pone.0009490, PMID: 20224823

Raymond KN, Dertz EA, Kim SS. 2003. Enterobactin: An archetype for microbial iron transport. *PNAS* **100**:3584–3588. DOI: https://doi.org/10.1073/pnas.0630018100

Reitz ZL, Hardy CD, Suk J, Bouvet J, Butler A. 2019. Genomic analysis of siderophore β-hydroxylases reveals divergent stereocontrol and expands the condensation domain family. *PNAS* **116**:19805–19814. DOI: https://doi.org/10.1073/pnas.1903161116, PMID: 31527229

Reitz ZL, Butler A. 2020. Precursor-directed biosynthesis of catechol compounds in *Acinetobacter bouvetii* DSM 14964 . *Chemical Communications* **56**:12222–12225. DOI: https://doi.org/10.1039/D0CC04171H

Reitz ZL, Medema MH. 2022. Genome mining strategies for metallophore discovery. *Current Opinion in Biotechnology* **77**:102757. DOI: https://doi.org/10.1016/j.copbio.2022.102757, PMID: 35914390

Ringel MT, Dräger G, Brüser T. 2018. PvdO is required for the oxidation of dihydropyoverdine as the last step of fluorophore formation in *Pseudomonas fluorescens The Journal of Biological Chemistry* **293**:2330–2341. DOI: https://doi.org/10.1074/jbc.RA117.000121, PMID: 29208656

Santichaivekin S, Yang Q, Liu J, Mawhorter R, Jiang J, Wesley T, Wu YC, Libeskind-Hadas R. 2021. eMPRess: a systematic cophylogeny reconciliation tool. *Bioinformatics* **37**:2481–2482. DOI: https://doi.org/10.1093/bioinformatics/btaa978, PMID: 33216126

Serino L, Reimmann C, Baur H, Beyeler M, Visca P, Haas D. 1995. Structural genes for salicylate biosynthesis from chorismate in *Pseudomonas aeruginosa*. *Molecular & General Genetics* **249**:217–228. DOI: https://doi.org/10.1007/BF00290369, PMID: 7500944

Shaaban KA, Shaaban M, Rahman H, Grün-Wollny I, Kämpfer P, Kelter G, Fiebig HH, Laatsch HKAC. 2019. Karamomycins A-C: 2-Naphthalen-2-yl-thiazoles from Nonomuraea endophytica. *Journal of Natural Products* **82**:870–877. DOI: https://doi.org/10.1021/acs.jnatprod.8b00928, PMID: 30907593

Soares EV. 2022. Perspective on the biotechnological production of bacterial siderophores and their use. *Applied Microbiology and Biotechnology* **106**:3985–4004. DOI: https://doi.org/10.1007/s00253-022-11995-y, PMID: 35672469

Süssmuth RD, Mainz A. 2017. Nonribosomal peptide synthesis-principles and prospects. *Angewandte Chemie* **56**:3770–3821. DOI: https://doi.org/10.1002/anie.201609079, PMID: 28323366

Uría Fernández D, Geoffroy V, Schäfer M, Meyer JM, Budzikiewicz H. 2003. Bacterial constituents CXIII structure revision of several pyoverdins produced by plant-growth promoting and plant-deleterious pseudomonas species. Monatshefte fuer Chemie/Chemical Monthl **134**:1421–1431. DOI: https://doi.org/10.1007/s00706-003-0065-6

Vinnik V, Zhang F, Park H, Cook TB, Throckmorton K, Pfleger BF, Bugni TS, Thomas MG. 2021. Structural and biosynthetic analysis of the fabrubactins, unusual siderophores from *Agrobacterium fabrum* Strain C58. *ACS Chemical Biology* **16**:125–135. DOI: https://doi.org/10.1021/acschembio.0c00809, PMID: 33373180

Winkelmann G, Cansier A, Beck W, Jung G. 1994. HPLC separation of enterobactin and linear 2,3-dihydroxybenzoylserine derivatives: a study on mutants of *Escherichia coli* defective in regulation (fur), esterase (fes) and transport (fepA). *Biometals* **7**:149–154. DOI: https://doi.org/10.1007/BF00140485, PMID: 8148617

Wu Q, Bell BA, Yan JX, Chevrette MG, Brittin NJ, Zhu Y, Chanana S, Maity M, Braun DR, Wheaton AM, Guzei IA, Ge Y, Rajski SR, Thomas MG, Bugni TS. 2023. Metabolomics and genomics enable the discovery of a new class of nonribosomal peptidic metallophores from a marine *Micromonospora Journal of the American Chemical Society* **145**:58–69. DOI: https://doi.org/10.1021/jacs.2c06410, PMID: 36535031

Zhang F, Ramos Alvarenga RF, Throckmorton K, Chanana S, Braun DR, Fossen J, Zhao M, McCrone S, Harper MK, Rajski SR, Rose WE, Andes DR, Thomas MG, Bugni TS. 2022. Genome mining and metabolomics unveil pseudonochelin: A siderophore containing 5-aminosalicylate from a marine-derived *Pseudonocardia* sp. Bacterium. *Organic Letters* **24**:3998–4002. DOI: https://doi.org/10.1021/acs.orglett.2c01408, PMID: 35649263

Chapter 3

Manuscript III:
Evolution of Siderophore Biosynthesis and Transport Systems in Bacteria

Evolution of Siderophore Biosynthesis and Transport Systems in Bacteria

Bita Pourmohsenin[1,2] and Nadine Ziemert[1,2*]

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

[2]German Center for Infection Research (DZIF), Partner Site Tübingen, Tübingen, Germany.

Abstract

Iron is essential for bacterial growth but often scarce in natural and host-associated environments. To cope with this limitation, bacteria deploy siderophores, specialized metabolites that bind iron and are produced through two distinct biosynthetic strategies: nonribosomal peptide synthetase pathways and NRPS-independent enzymatic routes. While many individual siderophore systems have been studied in detail, how biosynthesis and uptake systems have co-evolved across bacteria remains poorly understood.

Here, we present a genome-scale analysis of siderophore biosynthesis and transport across more than 62,000 representative bacterial genomes. By integrating domain-informed biosynthetic classification, transporter profiling, phylogenetic mapping, and duplication-transfer-loss reconciliation, we reveal distinct evolutionary trajectories for NRPS-derived metallophores and NRPS-independent siderophores (NISs). NRPS-based pathways are evolutionarily ancient and phylogenetically widespread, shaped by extensive gene duplication and horizontal transfer, and frequently coupled to outer-membrane uptake systems in Gram-negative bacteria. In contrast, NIS pathways show narrower, lineage-biased distributions and are more often associated with inner-membrane transporters, reflecting stronger constraints imposed by cell-envelope architecture.

Transport systems emerge as a highly dynamic and modular component of iron acquisition. TonB-dependent receptors are largely restricted to diderm lineages, whereas ABC-type transporters span both monoderm and diderm bacteria and exhibit particularly fluid evolutionary histories. As a result, siderophore uptake capacity is substantially more widespread than biosynthesis itself, leading to the frequent emergence of "pirate" genomes that retain transport systems while lacking detectable siderophore production.

Together, these findings support a modular model of siderophore system evolution in which biosynthesis and transport follow partially decoupled trajectories shaped by phylogeny, cell-envelope structure, and ecological opportunity. By linking siderophore chemistry, transporter architecture, and evolutionary history at genome scale, this study provides a unified framework for understanding microbial iron acquisition, public-good dynamics, and the repeated evolution of dependency strategies across bacterial lineages.

Keywords: siderophores; iron acquisition; biosynthetic gene clusters; transporters; horizontal gene transfer; bacterial evolution

1 Introduction

Iron is an essential micronutrient for nearly all living organisms, playing a crucial role in cellular processes such as respiration, DNA synthesis, and metabolism[1, 2]. Despite its abundance in the Earth's crust, the bioavailability of ferric iron (Fe^{3+}) is severely constrained under aerobic, neutral pH conditions due to its low solubility, posing a major challenge in diverse environments, including soil, seawater, and host tissues. To overcome this limitation, bacteria have evolved multiple iron acquisition strategies, most notably the secretion of siderophores: small, high-affinity iron-chelating specialized metabolites that solubilize and transport iron into the cell (Figure 1a)[3–5]. Siderophore systems are thus not only crucial for microbial survival, but also have profound implications for microbial ecology, plant and animal health, and biogeochemical cycling.

1.1 Bacterial Siderophore Biosynthesis Strategies

Bacterial siderophores are synthesized via two principal biosynthetic strategies:

(i) Nonribosomal peptide synthetase (NRPS) siderophores are assembled by large, modular NRPS enzymes that link precursors such as amino acids into structurally diverse macrocyclic or linear compounds[6]. Some NRPS-derived metabolites can also chelate non-iron metals (e.g., the NRP-S/PKS siderophore yersiniabactin binding Cu^{2+}), underscoring broader "metallophore" functionality, giving rise to the broader class of NRP-metallophores[6–9]. Recent advances in automated genome-mining approaches have substantially improved the detection and structural annotation of NRP-derived siderophores, highlighting their diversity and extensive taxonomic distribution[10, 11].

(ii) NRPS-independent siderophores (NISs) are generally structurally simpler and are synthesized by IucA/IucC-family adenylate-forming ligases (NIS synthetases), which activate carboxylate substrates via adenylation and catalyze amide bond formation to assemble hydroxamate or catecholate siderophores from metabolic intermediates such as citrate or ornithine (Figure 1b)[7, 12–16].

1.2 Transport Systems for Siderophores

Efficient uptake of ferric-siderophore complexes is critical for siderophore function, and therefore, bacteria have evolved specialized transporter systems tailored to their cell envelope architecture[17–19].

In Gram-negative bacteria, uptake begins with TonB-dependent receptors located in the outer membrane (e.g., FepA, FhuA). These β-barrel proteins possess N-terminal plug domains that confer substrate specificity and require energy transduced from the inner membrane via the TonB-ExbB-ExbD complex. After transport into the periplasm, ferric-siderophores are captured by periplasmic-binding proteins and subsequently imported across the inner membrane by ATP-binding cassette (ABC) transporters (Figure 1c)[20–22].

In Gram-positive bacteria, which lack an outer membrane, the uptake of ferric-siderophore occurs directly via lipoprotein-associated ABC transporters (e.g., Pia/Piu systems), which constitute the predominant, well-characterized route of siderophore import [23–25]. Reports of major facilitator superfamily (MFS) transporters largely concern siderophore export/secretion rather than import[26, 27].

1.3 Linking Transporters to Biosynthetic Gene Clusters (BGCs)

Siderophore biosynthetic gene clusters (BGCs) frequently co-localize with genes encoding cognate uptake and utilization functions, an organization that aids in predicting siderophore pathways from genomes. In Gram-negative bacteria, siderophore BGCs often include or occur near TonB-dependent receptor genes, whereas in Gram-positives, ABC importer genes are common neighbors; however, receptor/importer genes can also reside elsewhere in the genome, and co-localization is not universal[28–30]. Genomic modularity, horizontal gene transfer, and environmental selection produce considerable architectural plasticity. For example, marine *Vibrio spp.* produce amphiphilic siderophores (amphibactins), and comparative studies indicate transporter diversity within the genus, even for chemically related siderophores, complicating function inference from transporter annotations alone[17, 31–33].

Machine learning and feature-based genome mining approaches (e.g., antiSMASH, DeepBGC) combined with curated BGC databases (MIBiG) have markedly improved the sensitivity and precision of siderophore/metallophore BGC detection, but accurately linking encoded transporters to specific siderophore substrates remains a key challenge[28, 29, 34, 35].

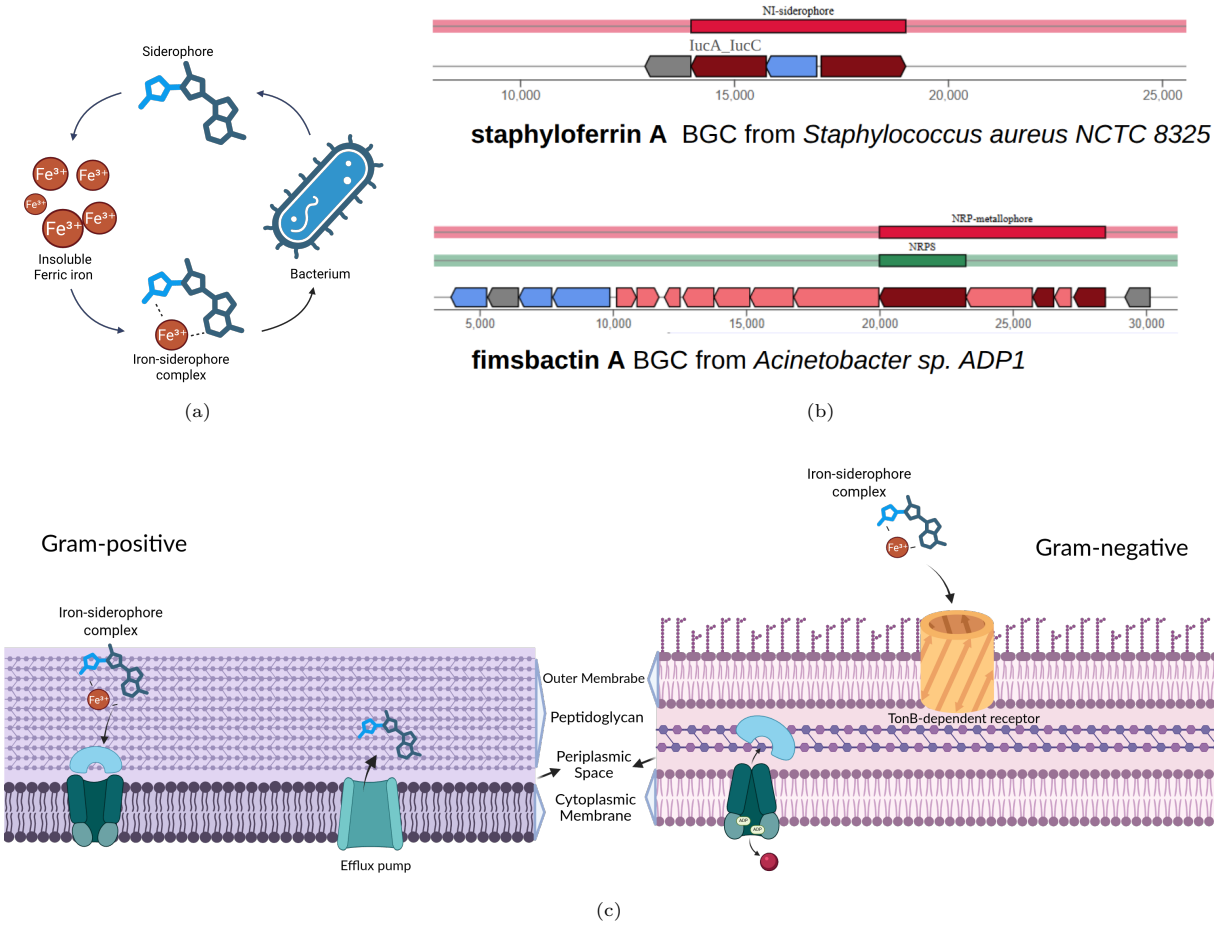


Fig. 1: **Overview of bacterial siderophore systems. (a)** Iron is abundant in many environments but poorly soluble as Fe^{3+}, necessitating the secretion of siderophores that chelate iron and return it to the cell as ferric–siderophore complexes. **(b)** Two distinct biosynthetic strategies for siderophore production, illustrated using representative biosynthetic gene clusters from the MiBIG database. The upper panel shows an NRPS-independent siderophore (NIS) cluster encoding staphyloferrin A from *Staphylococcus aureus* NCTC 8325 (BGC0000944), while the lower panel shows an NRPS-based siderophore cluster encoding fimsbactin A from *Acinetobacter* sp. ADP1 (BGC0000352). Gene cluster architectures are displayed using antiSMASH visualizations, highlighting the compact IucA/IucC-based NIS pathway versus the extended modular NRPS assembly line. **(c)** Schematic comparison of ferric-siderophore uptake architectures in Gram-positive and Gram-negative bacteria. In Gram-positive bacteria (left), ferric–siderophore complexes are captured at the cell surface and transported directly across the cytoplasmic membrane via ATP-binding cassette (ABC) transporters, typically involving surface-anchored substrate-binding proteins (SBPs), in the absence of an outer membrane. In Gram-negative bacteria (right), uptake proceeds in two steps: ferric-siderophore complexes are first translocated across the outer membrane through TonB-dependent receptors, followed by import across the inner membrane via ABC transporters.

1.4 Motivation and Knowledge Gaps

Although individual siderophore systems have been studied and characterized, their broader evolutionary relationships remain insufficiently resolved. In particular, the co-evolutionary dynamics between siderophore biosynthesis pathways and their associated transport systems have not been systematically examined at genome scale. Several key issues remain unresolved, including the timing and diversification of siderophore biosynthetic pathways and transporter systems across bacterial lineages, the extent to which biosynthetic and uptake modules have co-evolved versus arisen independently, and the presence of phylogenetic or ecological associations linking specific siderophore classes to distinct transport architectures.

Addressing these gaps requires a comparative evolutionary framework that integrates biosynthetic diversity, transporter repertoires, and bacterial taxonomy within a robust phylogenetic context. In this

study, we combine large-scale phylogenetic profiling, gene tree/species tree reconciliation, and trait correlation analyses to investigate the origin, distribution, and evolutionary coordination of siderophore production and uptake systems across the bacterial tree of life.

2 Methods

2.1 Genome Selection and Biosynthetic Gene Cluster (BGC) Annotation

To investigate the distribution of siderophore biosynthesis and uptake across Bacteria, we analyzed the GTDB release r207 species-representative genomes ($n \approx 62,291$), spanning all major bacterial phyla (GTDB r207 statistics[36]). All genomes were screened for biosynthetic gene clusters (BGCs) using antiSMASH v7.0 (command-line, default parameters). We parsed GenBank and JSON outputs to extract BGC coordinates, cluster types, and domain annotations. GTDB taxonomy and available metadata were retained for downstream comparative analyses[28].

2.2 Classification of Siderophore Biosynthetic Strategies

Siderophore BGCs were classified into two major biosynthetic strategies based on domain composition and established biochemical markers:

NRPS-dependent siderophores (hereafter NRP-metallophores) were defined as BGCs containing at least one core NRPS module (i.e., condensation [C], adenylation [A], and thiolation [T] domains), along with metal-chelation-related tailoring enzymes (e.g., N-hydroxylating monooxygenases, cyclization domains, salicylate synthases)[10].

NRPS-independent siderophores (NISs) were identified based on the presence of IucA/IucC-like enzymes, typically associated with biosynthesis of hydroxamate or catecholate siderophores assembled from intermediates like citrate or ornithine. We used Pfam domain PF04183 to detect these systems within antiSMASH BGC regions.

Ambiguous cases were excluded from comparative analyses to maintain clear class definitions.

2.3 Identification of Siderophore Transport Systems

We used two complementary approaches to characterize siderophore transport.

2.3.1 Transporter domains within BGC regions

To identify siderophore transport systems co-localized with BGCs, we scanned all open reading frames (ORFs) within BGC boundaries using HMMER (v3.3.2) against a curated set of Pfam and TIGRFAM models for transporter-related families. These included ABC transport components (ATPase/permease/substrate-binding proteins), TonB-dependent receptor families and associated energy components, and additional transporter superfamilies used in the downstream correlation analyses (Full accessions in Supplementary Table S1)

HMMER was run using hmmscan –domtblout, applying profile-specific thresholds derived from trusted cutoffs or profile documentation, with domain coverage $\geq$50%. Overlapping hits were resolved using custom parsing scripts. The resulting domain presence/absence matrices were used for statistical analyses.

2.3.2 Genome-wide screening for siderophore uptake systems (Crits-Christoph model)

To assess siderophore uptake independently of BGC detection, we performed a genome-wide scan using the five Pfam families defined in the Crits-Christoph siderophore uptake model: PF01497, PF00496, PF01032, PF00593, and PF07660. For each genome, we recorded presence/absence of each family and computed a binary uptake call (CC_sid = 1/0) using the model rules.

This genome-wide transporter screen covers all GTDB representative genomes and is used for finding the “pirate” lineages and analysing their evolution and distribution.

2.4 Phylogenetic Inference of Core Siderophore Biosynthetic Genes

To compare evolutionary histories across chelator classes, we targeted marker gene families representing distinct siderophore pathways:

- Catecholates: EntC (isochorismate synthase) + EntA (2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase. Only genomes encoding both were kept to permit concatenation.
- Hydroxamates (NRP-dependent): lysine N6-monooxygenase (Lys monooxygenase) and ornithine N5-monooxygenase (Orn monooxygenase).
- Salicylates: isochorismatase-like (IPL) + salicylate synthase (SalSyn).
- NIS marker: IucA/IucC.

Protein sequences were retrieved via HMM searches restricted to antiSMASH BGC regions (±10 kb), using curated models per family. Per-profile cutoffs and coverage filters were applied. To keep trees tractable while retaining diversity, we performed taxonomy-aware subsampling (one sequence per order when present). Alignments were produced with MAFFT v7, trimming used trimAl (-automated1), and gene trees were inferred with IQ-TREE 2 using ModelFinder (-m MFP) and ultrafast bootstrap (-bb 1000). Trees were midpoint-rooted and visualized in iTOL[37–40].

2.5 Phylogenomic Mapping and Gene Tree Reconciliation

To infer evolutionary events shaping siderophore genes, we performed duplication-transfer-loss (DTL) reconciliations with eMPRess. Species trees were derived from the GTDB r207 backbone phylogeny and pruned to include only taxa present in each gene tree. Trees were analyzed at the order level to reduce noise from species-level discordance and to emphasize broader evolutionary patterns.

For each biosynthetic marker and transporter gene family, we followed a consistent workflow:

1. Protein sequences were aligned with MAFFT and trimmed with trimAl (-automated1);
2. Maximum-likelihood gene trees were inferred with IQ-TREE 2, using model selection (-m MFP) and ultrafast bootstrap support (-bb 1000);
3. GTDB species trees were pruned to match gene tree taxa, and one-to-one mappings between gene and species tree tips were generated;
4. Reconciliations were performed in eMPRess using a parsimonious DTL cost scheme (duplication = 2, transfer = 3, loss = 1).

To ensure compatibility with eMPRess, gene trees containing rare multifurcations were converted to strictly bifurcating trees by resolving polytomies prior to reconciliation. Tip labels were standardized to genome accessions, and internal node labels were simplified. Mapping files were generated directly from the intersection of gene and species tree tips to ensure consistency.

Reconciliations were carried out for core biosynthetic marker genes (NRP-metallophore and NIS pathways) and for five transporter-associated Pfam families identified using the Crits-Christoph transporter model: PF01497, PF00496, PF01032, PF00593, and PF07660. For each family, reconciliation outputs were summarized as counts of duplications, transfers, losses, and cospeciations. Events were annotated with GTDB taxonomy to assign them to specific bacterial lineages.

2.6 Visualization of Trait Distribution on Collapsed Phylogenies

To visualize siderophore-related traits across the bacterial tree of life, we collapsed the GTDB r207 reference phylogeny to the order level using ETE3. For each order, we computed the fraction of genomes encoding NRP-metallophore BGCs and NIS BGCs (Figure 2).

Trait frequencies were visualized as quantitative bar tracks mapped onto the collapsed phylogeny using iTOL[40]. To improve the visibility of low-frequency traits across orders with highly uneven prevalence, values were transformed using a power transformation ($x^{0.25}$) prior to visualization. This transformation enhances the dynamic range of small values while preserving the relative ordering of trait frequencies.

To account for uneven taxonomic sampling across the tree, an additional bar track indicating the number of genomes per order was included. Together, these visualizations summarize the large-scale phylogenetic distribution of siderophore biosynthetic potential across bacterial lineages.

A complementary visualization was generated to illustrate the presence or absence of siderophore-related traits across the same phylogenetic framework (Figure 3). Instead of quantitative frequencies, this representation displays binary trait annotations for each order, including (i) the presence of any BGC detected by antiSMASH, (ii) NRP-metallophore biosynthetic clusters, (iii) NIS biosynthetic clusters, (iv) conserved siderophore transporter systems detected across the genome (CC_sid), and (v) putative siderophore pirate lineages, defined as genomes encoding siderophore uptake systems but lacking detectable NRP-metallophore or NIS BGCs.

In addition, inferred evolutionary origins of major siderophore biosynthetic systems were annotated on the tree based on gene tree/species tree reconciliation analyses performed with eMPRess. These annotations indicate the most likely ancestral nodes where specific biosynthetic systems emerged, as well as alternative placements supported across near-optimal reconciliation solutions, displayed as branch symbols.

2.7 Trait correlation and Enrichment Analysis

To quantify associations between siderophore biosynthetic strategies and transporter architectures, we analyzed genome-level presence–absence matrices for five conserved transporter Pfam families (PF01497, PF00496, PF01032, PF00593, PF07660). Contingency tables comparing transporter presence with NRP-metallophore or NRPS-independent siderophore (NIS) biosynthetic status were constructed for each transporter family.

Associations were evaluated using two-sided Fisher's exact tests. Odds ratios were computed using Haldane–Anscombe correction to avoid zero-count artifacts, and statistical significance was assessed after multiple testing correction using the Benjamini–Hochberg procedure. Associations with adjusted p-values < 0.01 were considered significant. Effect sizes were visualized as $\log_2$ odds ratios to highlight enrichment or depletion of transporter families in genomes encoding each biosynthetic strategy (Figure 4a).

To investigate whether transporter genes form recurring uptake architectures, we further analyzed co-occurrence patterns among transporter families using pairwise Fisher's exact tests on the genome-level presence/absence matrix. Higher-order combinations of transporter domains were visualized using UpSet plots to identify common transporter modules across genomes (Figure 4c).

2.8 Metadata Integration and Cell Envelope Classification

To examine how cell envelope architecture shapes transporter repertoires, genomes were annotated with Gram category based on GTDB taxonomy-derived rules distinguishing diderm (Gram-negative) from monoderm (Gram-positive) bacteria. Taxa with ambiguous envelope structure were retained as "Unknown" and excluded from Gram-specific comparisons.

For each transporter family, we calculated the fraction of genomes encoding the corresponding domain within Gram-positive and Gram-negative groups. Differences in transporter prevalence between envelope types were visualized to assess how uptake architectures are constrained by cell envelope structure (Figure 4b).

In addition, genomes were classified as putative siderophore pirates when conserved siderophore transporter domains were detected but no NRP-metallophore or NIS biosynthetic genes were present in the genome. This operational definition allowed the identification of lineages that retain siderophore uptake capacity despite lacking detectable biosynthetic pathways.

3 Results

3.1 Phylogenetic Distribution of Siderophore Biosynthesis

We analyzed ∼62,000 species-representative bacterial genomes from GTDB release r207 to examine the global distribution of siderophore biosynthesis and uptake systems. Mapping these traits onto an order-level bacterial phylogeny revealed pronounced lineage-specific patterns in siderophore biosynthetic potential (Figure 2).

NRP-metallophore biosynthetic gene clusters (BGCs), defined by NRPS modules and metal-chelating tailoring enzymes, were widespread across the bacterial tree. They were particularly prevalent in major radiations such as Actinomycetota, Alphaproteobacteria, and Gammaproteobacteria, where in many orders more than half of genomes encoded at least one NRP-metallophore. In contrast, NRPS-independent siderophores (NISs), identified by the presence of IucA/IucC homologs, exhibited a more restricted and lineage-specific distribution. NIS biosynthesis was especially prominent in Bacillota and Bacteroidota, whereas several phyla, including Acidobacteriota, Bacillota_B, Bacillota_C, and Planctomycetota, were dominated by NRP-metallophore systems with little or no detectable NIS biosynthesis.

These large-scale patterns are summarized quantitatively in Figure 2b, which shows the proportion of genomes encoding NRP-metallophore or NIS biosynthesis across major bacterial phyla. While some

phyla predominantly encode a single biosynthetic strategy, others, most notably Actinomycetota, Pseudomonadota, and parts of Bacillota, harbor both systems, indicating long-term coexistence of distinct siderophore biosynthetic solutions within the same evolutionary lineages.

Together, these results demonstrate that siderophore biosynthetic potential is widely distributed but strongly structured by bacterial phylogeny, with different bacterial lineages favoring distinct biochemical strategies for iron acquisition.

3.2 Evolutionary Landscape of Siderophore Biosynthesis

3.2.1 NRP-metallophores

Gene tree/species tree reconciliations revealed complex evolutionary histories for NRP-metallophore pathways, shaped by both deep lineage-specific retention and frequent horizontal transfer. Across the marker enzymes examined, numerous duplication and transfer events were inferred, particularly among Actinomycetota and Proteobacterial lineages (Figure 3)

Lysine monooxygenase, a key enzyme in hydroxamate-type siderophore biosynthesis, consistently traced back to an origin within Actinomycetota. This supports an actinobacterial origin for hydroxamate-producing NRP systems, followed by subsequent diversification and transfer into other bacterial groups. In contrast, ornithine monooxygenase exhibited a more polyphyletic pattern, with multiple inferred transfers into Proteobacterial orders such as Burkholderiales and Pseudomonadales.

Enzymes associated with pyoverdine-type chromophore biosynthesis (including PvdO and PvdP) showed a more restricted evolutionary pattern. Reconciliation analyses consistently placed their origins within Gammaproteobacteria, where these systems remain strongly conserved. This pattern suggests a lineage-specific innovation followed by diversification within this clade.

Salicylate-associated biosynthetic enzymes exhibited the most reticulate evolutionary histories. Multiple independent origins and repeated horizontal transfers were inferred across distant bacterial lineages, highlighting the evolutionary flexibility of salicylate-based siderophore systems and their repeated recruitment into different biosynthetic contexts.

Together, these results indicate that NRP-metallophore pathways behave as modular and mobile traits: some biosynthetic components remain deeply rooted within specific bacterial lineages, whereas others repeatedly cross phylogenetic boundaries through horizontal gene transfer.

3.2.2 NRP-Independent Siderophores

Compared with NRP systems, NRPS-independent siderophore (NIS) biosynthetic genes showed more constrained evolutionary patterns. Reconciliation analyses of the core IucA/IucC enzyme family consistently placed the most likely ancestral origin within Proteobacteria (Figure 3), with strong representation in lineages such as Enterobacterales and Neisseriales.

Relative to NRP-metallophore pathways, NIS systems exhibited fewer duplication and transfer events, suggesting a larger contribution of vertical inheritance within established lineages. Nevertheless, several transfer events into other phyla, including Bacillota and Actinomycetota, were detected. These transfers were often associated with specialized or host-associated taxa, indicating that NIS pathways can still move across phylogenetic boundaries under appropriate ecological pressures.

Overall, the evolutionary histories of NRP and NIS systems differ markedly. Whereas NRP-metallophore pathways display highly reticulate histories with frequent gene transfers and lineage-specific innovations, NIS systems appear more phylogenetically constrained, consistent with a more limited but recurrent spread across bacterial lineages.

3.3 Evolution and diversification of Siderophore Transport Systems

We next examined the distribution of five transporter-associated Pfam families implicated in siderophore uptake (PF01497, PF00496, PF01032, PF00593, PF07660) across the genome dataset. In contrast to siderophore biosynthesis, which shows strong phylogenetic structuring, transporter genes were broadly distributed across the bacterial tree (Figure 3b).

Reconciliation analyses were also performed for the five transporter-associated Pfam families. In contrast to biosynthetic marker genes, these families showed highly reticulate evolutionary histories with numerous duplication and horizontal transfer events across the bacterial tree. Across near-optimal reconciliation solutions, no strongly supported single ancestral origin could be inferred for any transporter

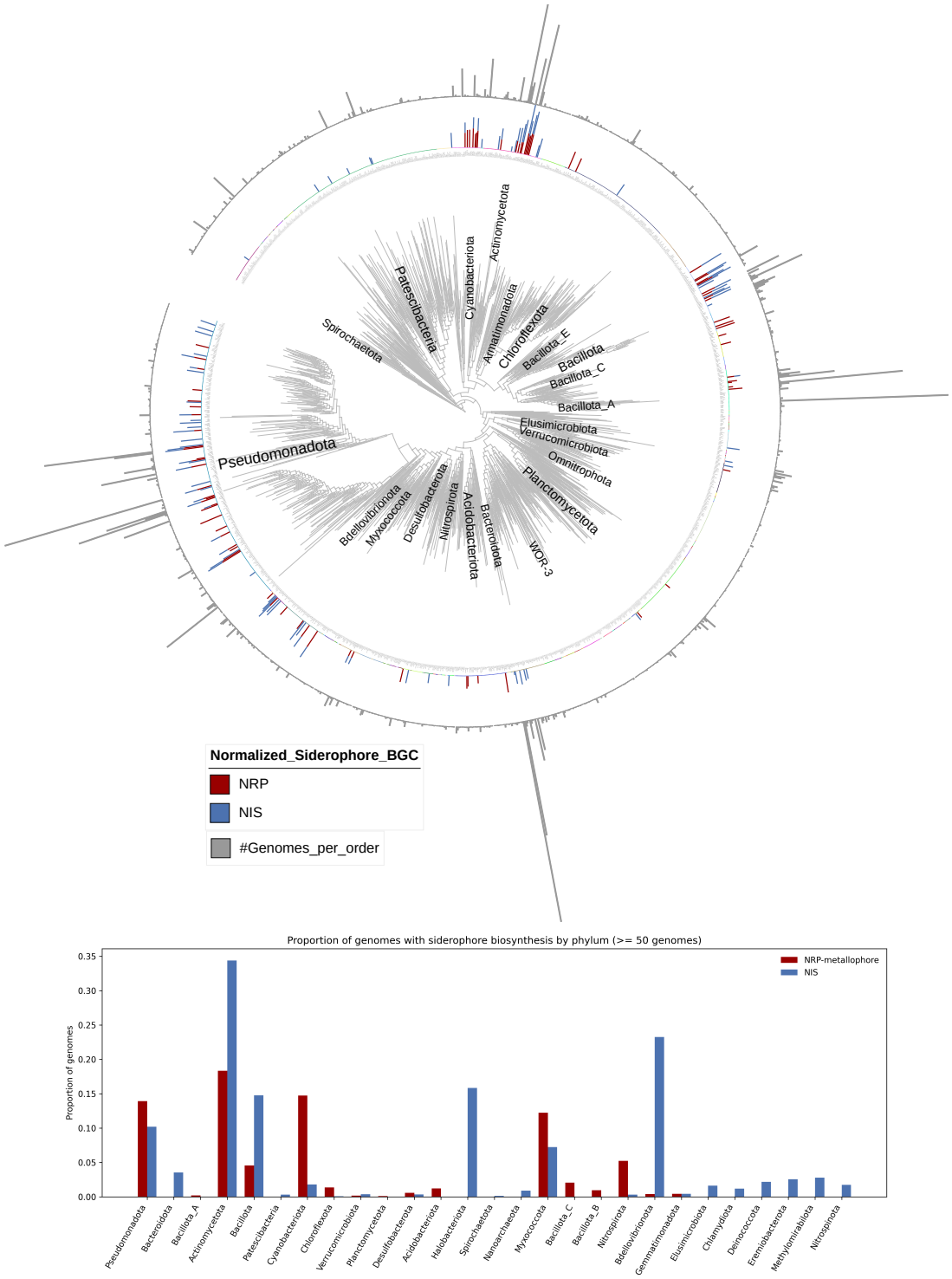


Fig. 2: **Phylogenetic distribution of siderophore biosynthesis and biosynthetic potential across bacterial phyla. (Up)** Order-level GTDB r207 phylogeny with major phyla labelled on internal branches. Inner bars show the power–transformed ($x^{0.25}$), genome-normalized frequency of NRP-metallophore (red) and NRPS-independent siderophore (NIS; blue) BGCs per order; the outer bar track shows genome counts per order, highlighting sampling bias. **(Down)** Proportion of genomes with NRP-metallophore (red) and NIS (blue) biosynthetic potential by phylum. Only phyla with $\geq$50 genomes and at least one siderophore BGC are shown.

family, consistent with the broad evolutionary mobility and functional reuse of membrane transport proteins.

TonB-dependent receptors (PF00593), together with the associated plug domain (PF07660), were almost exclusively confined to Gram-negative lineages. These proteins mediate outer-membrane import of ferric-siderophore complexes and therefore require the presence of a diderm envelope architecture. Consistent with this structural constraint, these domains were largely absent from monoderm bacteria (Figure 4b). In contrast, ABC transporter components involved in siderophore uptake (PF01497, PF00496, PF01032) were detected across both Gram-positive and Gram-negative lineages, reflecting their role in ATP-dependent transport across the cytoplasmic membrane.

Statistical association analysis revealed clear links between transporter families and siderophore biosynthetic strategies (Figure 4a). Several transporter domains showed strong enrichment in genomes encoding NRP-metallophore biosynthetic pathways, particularly PF01497, PF00496, and PF01032. In contrast, NIS-containing genomes showed consistent depletion of TonB-dependent receptors, reflecting the predominance of NIS systems in monoderm bacterial lineages.

Despite these associations, transporter genes were far more widespread than biosynthetic gene clusters. Many genomes encoded siderophore uptake systems without detectable NRP-metallophore or NIS biosynthetic markers (Figure 3b). These genomes represent siderophore "pirates," capable of importing siderophores produced by other organisms. Pirate genomes occurred across most major bacterial phyla and were particularly common in Proteobacterial lineages.

Co-occurrence analysis further revealed that siderophore transporters form recurrent architectural modules (Figure 4c). The strongest association was observed between PF00593 and PF07660, consistent with their structural coupling in TonB-dependent outer-membrane receptors. A second tightly linked module was formed by the ABC transporter components PF01497 and PF01032. Additional components showed weaker associations, indicating more flexible assembly of transporter architectures across different bacterial lineages.

These patterns indicate that siderophore uptake systems are both modular and phylogenetically constrained by cell envelope structure. While biosynthetic pathways show lineage-specific distributions and evolutionary origins, uptake systems are broadly retained across bacteria and can persist even in the absence of local siderophore production, enabling the widespread emergence of siderophore piracy strategies.

3.4 Pirate Genomes: Distribution and Ecological Implication

The widespread distribution of siderophore transporters raised the question of how often uptake systems occur in genomes that lack the capacity to produce siderophores themselves. To address this, we identified "pirate" genomes: genomes encoding siderophore uptake systems but lacking detectable NRP-metallophore or NIS biosynthetic gene clusters.

Across the full GTDB r207 dataset, pirate genomes were common and taxonomically widespread (Figure 3b). They were particularly abundant in Pseudomonadota, but were also detected across diverse bacterial phyla including Bacteroidota, Fusobacteriota, and Nitrospirota. In contrast, pirate genomes were comparatively rare in Actinomycetota, where siderophore biosynthesis, especially NRP-metallophore pathways, was more frequently retained.

Where environmental metadata were available, pirate genomes were more frequently associated with host-associated and aquatic environments than with soil habitats. Although environmental annotations were incomplete for a substantial fraction of genomes, this trend is consistent with ecological settings where siderophores are abundant, diffusible, and shared among community members. In such environments, the ability to import siderophores produced by neighboring organisms may provide a selective advantage, allowing bacteria to exploit existing public goods without incurring the metabolic cost of synthesis. These observations indicate that siderophore uptake capacity is frequently retained even after loss of biosynthetic pathways, suggesting that siderophore piracy represents a recurrent ecological strategy across bacterial communities.

3.5 Associations between Biosynthesis, Transport, and Taxonomy

To quantify how siderophore biosynthesis, transport architectures, and taxonomy are linked, we performed a series of statistical association and co-occurrence analyses across the full genome dataset (Figure 4).

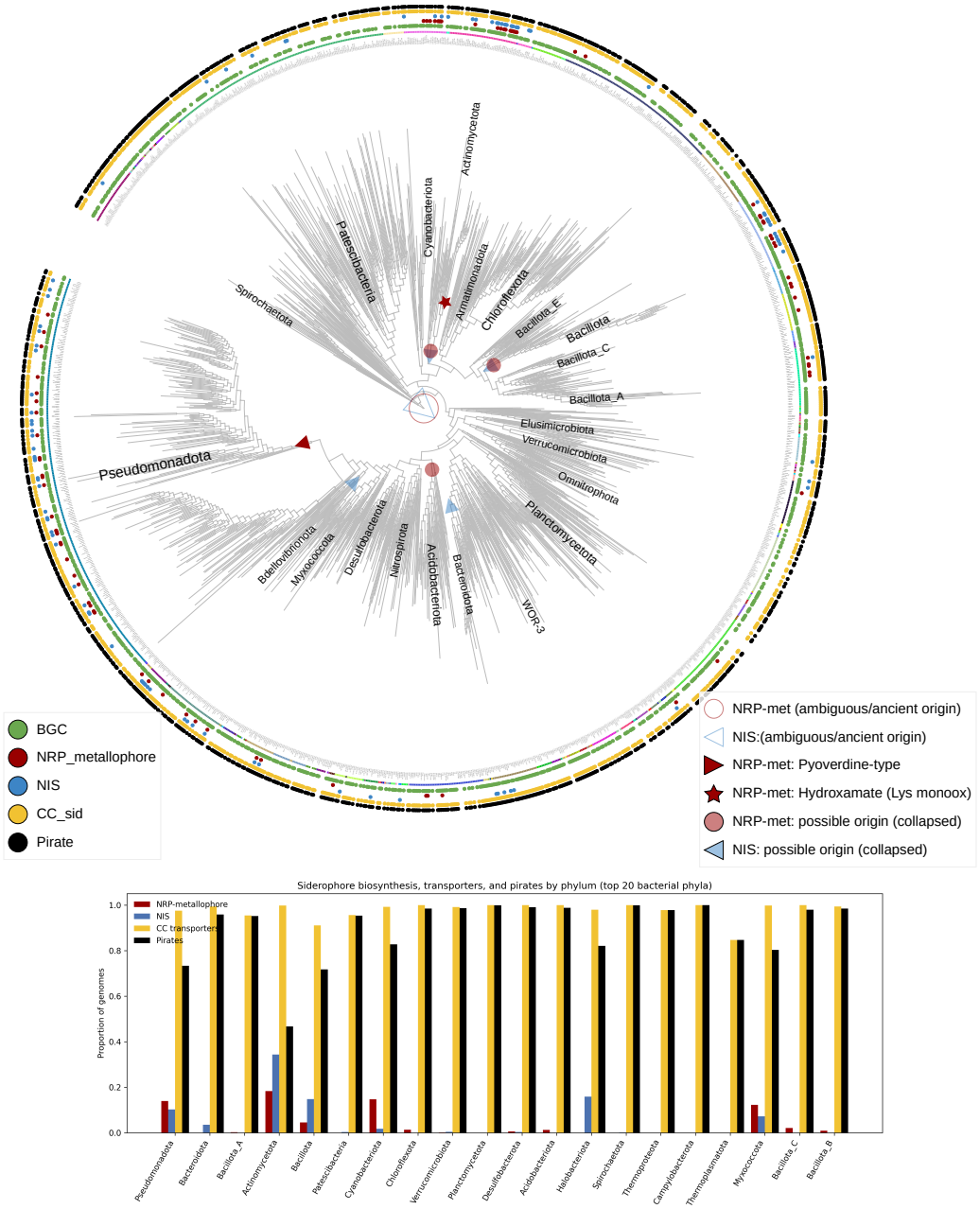


Fig. 3: **Evolutionary origins and modular decoupling of siderophore biosynthesis and transport. (Up)** Order-level GTDB r207 phylogeny annotated with inferred origins of siderophore biosynthetic systems based on gene tree-species tree reconciliation. Solid symbols indicate confidently supported origins (pyoverdine-type NRP systems in Gammaproteobacteria; hydroxamate-type NRP systems synthesized by Lys monooxygenase in Actinomycetota), whereas semi-transparent symbols denote alternative ancestral placements consistent across near-optimal solutions. Outer annotation tracks show presence of any siderophore BGC, NRP-metallophore BGCs, NIS BGCs, conserved siderophore transporters (CC_sid), and pirate genomes (transport-positive, biosynthesis-negative). Transport capacity is substantially more widespread than biosynthesis, and pirate genomes occur across most major phyla. **(Down)** Phylum-level summary showing the proportion of genomes encoding NRP-metallophore BGCs, NIS BGCs, siderophore transporters, and pirate genomes across the most represented bacterial phyla.

3.5.1 Taxonomic Predictors of Biosynthesis Strategy

Fisher's exact tests across taxonomic ranks revealed strong associations between lineage and siderophore biosynthetic strategy (Supplementary Table S3). At the phylum-level, Actinomycetota showed a pronounced enrichment for NRP-metallophore biosynthesis (odds ratio$\approx$ 6.8, $p < 10^{-300}$), whereas Firmicutes (particularly Bacillota) were significantly enriched for NISs (odds ratio$\approx$ 9.4, $p < 10^{-300}$).

At finer taxonomic resolution, many orders and families displayed consistent biosynthetic strategies, reflecting vertical inheritance of siderophore pathways. In contrast, lineages such as Pseudomonadaceae exhibited substantial within-family heterogeneity, reflecting frequent gains and losses of biosynthetic capacity. Notably, several of these heterogeneous lineages also harbored high fractions of pirate genomes, suggesting that loss of biosynthesis can occur without loss of siderophore uptake capability.

3.5.2 Transporter associations with biosynthesis and piracy

To quantify the relationship between siderophore biosynthesis and uptake systems, we tested associations between five transporter-associated Pfam families and biosynthetic traits across the genome dataset using Fisher's exact tests (Figure 4a). The analysis was performed on a dataset collapsed to one row per genome.

Several transporter families showed strong enrichment in genomes encoding NRP-metallophore biosynthetic pathways. In particular, PF01497, PF00496, and PF01032 were significantly associated with NRP-metallophore producers (odds ratios $\approx$ 3.08, 3.61, and 3.16, respectively; all $p \ll 0.001$). TonB-dependent receptors (PF00593) and the associated plug domain (PF07660) were also enriched in these genomes, although with smaller effect sizes (odds ratios $\approx$ 1.43 and 1.57).

In contrast, genomes encoding NRPS-independent siderophores showed a consistent depletion of TonB-linked outer membrane transport components. Both PF00593 and PF07660 were significantly underrepresented in NIS-containing genomes (odds ratios $\approx$ 0.35 and 0.54, respectively; $p \ll 0.001$). PF01497 and PF01032 were also depleted in NIS genomes, whereas PF00496 showed no detectable association (odds ratio $\approx$ 1.01, $p \approx 0.92$).

We next examined transporter repertoires in genomes classified as siderophore pirates. Despite sharing the same functional definition, pirate genomes showed clear differences in transporter composition across bacterial lineages. In Gram-negative taxa, pirates predominantly retained TonB-dependent receptors together with the associated plug domain, consistent with uptake through the outer membrane. In contrast, pirate genomes in Gram-positive lineages more frequently relied on ABC-type transporter components, reflecting the absence of an outer membrane and the reliance on ATP-driven transport across the cytoplasmic membrane.

These results indicate that siderophore piracy does not involve a single conserved transporter architecture. Instead, pirate genomes retain uptake systems that are compatible with the envelope structure and transporter repertoires characteristic of their respective lineages.

3.5.3 Transporter co-occurrence patterns

To determine whether siderophore uptake genes form recurring architectural modules, we tested pairwise co-occurrence among the five transporter-associated Pfam families using Fisher's exact tests with Benjamini–Hochberg correction (Figure 4c).

Two strongly associated transporter modules were evident. First, PF00593 (TonB-dependent receptor) co-occurred tightly with PF07660 (plug domain) (odds ratio $\approx$ 17.3, adjusted $p \ll 0.001$), consistent with their structural coupling in TonB-dependent outer membrane receptors. Second, the ABC-associated domains PF01497 and PF01032 also showed very strong co-occurrence (odds ratio $\approx$ 20, adjusted $p \ll 0.001$), indicating that these components frequently function together within the same ATP-driven import system.

Associations among ABC transporter components were weaker but still significant, suggesting partial modular assembly of uptake systems. Negative associations were comparatively modest; for example, PF00496 showed depletion with TonB-linked components, indicating that some genomes preferentially encode one transporter architecture rather than combining multiple systems. Together, these results support a modular organization of siderophore uptake systems, in which tightly linked domain pairs form core functional units while other components vary more flexibly across bacterial lineages.

3.5.4 Cell envelope constraints on siderophore transport

The distribution of transporter architectures closely mirrored bacterial cell envelope structure (Figure 4b). TonB-dependent receptors were effectively restricted to Gram-negative taxa, where they mediate siderophore uptake across the outer membrane, whereas ABC transporter components were broadly distributed across both Gram-positive and Gram-negative bacteria.

These structural constraints were also reflected in biosynthetic strategies. NRPS-independent siderophores were significantly enriched in Gram-positive lineages, whereas NRP-metallophore systems were more frequently associated with Gram-negative bacteria (Fisher's exact test, $p \ll 0.001$).

With the widespread occurrence of transporter-only genomes described above, these patterns highlight how siderophore uptake systems can persist independently of biosynthesis while remaining constrained by the envelope architecture of each lineage.

4 Discussion and Broader Implications

Iron acquisition is a universal challenge across bacterial lineages, and siderophore systems represent a major evolutionary innovation to overcome this constraint. By combining large-scale genome mining, phylogenetic mapping, and gene tree reconciliation, this study provides a comprehensive view of how siderophore biosynthesis and transport systems have diversified, decoupled, and been repeatedly reshaped across the bacterial tree of life. Together, our results highlight siderophore systems as modular traits shaped by deep evolutionary history, cell envelope constraints, and ecological interactions centered on shared public goods.

4.1 Divergent Evolutionary Paths of NRP and NIS Siderophores

Our analyses reveal markedly different evolutionary trajectories for NRP-metallophores and NRPS-independent siderophores (NISs). NRP-metallophores are phylogenetically widespread and evolutionarily dynamic. Their broad distribution across major bacterial radiations, together with repeated horizontal transfer inferred for multiple marker enzymes, supports a model in which NRP-based metallophore pathways represent ancient and flexible solutions to iron limitation. At the same time, some pathway components remain strongly lineage-associated, such as actinobacterial hydroxamate systems and gammaproteobacterial pyoverdine-associated functions. This combination of conserved lineage-specific cores and mobile auxiliary components underscores the modular nature of NRP-metallophore evolution.

In contrast, NIS systems appear more phylogenetically constrained. Their distribution is narrower, their inferred histories involve fewer duplication and transfer events, and their occurrence is concentrated in particular bacterial lineages. Rather than representing a broadly mobile strategy, NIS systems appear to have spread more selectively, with a stronger contribution from vertical inheritance. Together, these patterns suggest that siderophore evolution does not follow a single route. Instead, bacteria have repeatedly arrived at distinct biosynthetic solutions that differ in age, mobility, and phylogenetic breadth.[6].

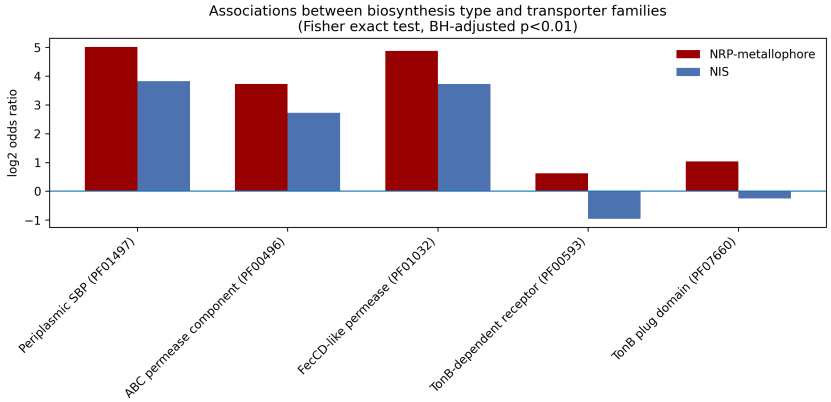
4.2 Transport System Evolution is Constrained by Cell Envelope Architecture

Compared with biosynthetic pathways, siderophore uptake systems are much more broadly distributed. Transporter-associated Pfam families were detected across much of the bacterial tree, including many lineages that lack detectable siderophore biosynthetic gene clusters. At the same time, this apparent flexibility is not unconstrained. The strongest organizing factor was cell envelope architecture.

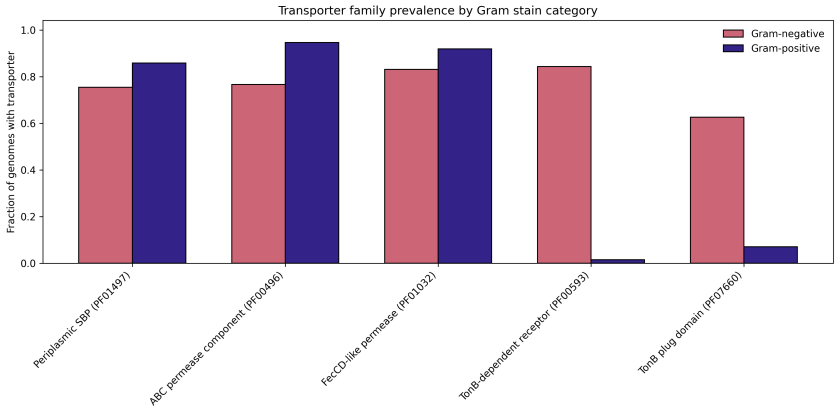
TonB-dependent receptors and their associated plug domains were effectively restricted to Gram-negative lineages, consistent with their requirement for an outer membrane. By contrast, ABC-associated transporter components were compatible with both monoderm and diderm bacteria and therefore showed a much broader phylogenetic distribution. These structural constraints also parallel the biosynthetic patterns observed here: NIS systems were enriched in Gram-positive lineages, whereas NRP-metallophore pathways were more often associated with Gram-negative taxa. In this sense, the envelope does not just constrain transport; it also shapes which siderophore strategies are more likely to persist in different parts of the bacterial tree[16, 22].

Our co-occurrence analyses reinforce this picture by showing that uptake systems are organized into recurring functional modules. The strong pairing of PF00593 with PF07660, and of PF01497 with PF01032, indicates that siderophore transport is built around a limited number of recurrent architectural

(a)



(b)



(c)

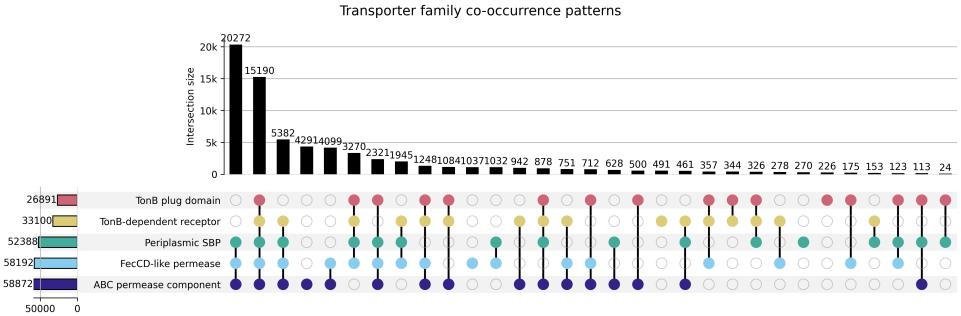


Fig. 4: **Statistical associations linking siderophore biosynthesis, transporter architecture, and cell envelope structure. (a)** Associations between siderophore biosynthetic strategy and transporter families. Bars show $\log_2$ odds ratios from Fisher exact tests comparing the presence of transporter Pfam domains in genomes encoding NRP metallophore biosynthetic gene clusters or NRPS-independent siderophores (NIS). Positive values indicate enrichment of the transporter family in genomes encoding the corresponding biosynthetic strategy. **(b)** Prevalence of transporter families across Gram-stain categories. Bars show the fraction of genomes encoding each transporter family in Gram-negative and Gram-positive bacteria. TonB-dependent receptors and associated plug domains are strongly enriched in Gram-negative lineages, whereas ABC transporter components are present in both envelope types. **(c)** Co-occurrence patterns among the five transporter-associated Pfam families. The UpSet plot shows the most common combinations of transporter domains across genomes. Two dominant modules are evident: the TonB-dependent outer membrane system (TonB receptor with plug domain) and the ABC-type import system (periplasmic binding protein with FecCD-like permease and ABC permease component), indicating modular organization of siderophore uptake architectures.

solutions rather than a diffuse collection of interchangeable parts. This modularity likely helps explain how uptake systems can be retained and reused across diverse genomic and ecological backgrounds.

4.3 Decoupling of Biosynthesis and Uptake Enables Widespread Siderophore Piracy

One of the clearest outcomes of this study is that siderophore uptake is far more widespread than siderophore biosynthesis. Many genomes encode transporter repertoires consistent with siderophore uptake despite lacking detectable NRP-metallophore or NIS biosynthetic genes. These transporter-only genomes were widespread across bacterial phyla and particularly common in several Proteobacterial lineages.

This pattern suggests that the decoupling of production and uptake is not an anomaly, but a common evolutionary state. In ecological terms, these genomes are best understood as potential siderophore pirates: organisms that can exploit extracellular siderophores without paying the metabolic cost of producing them. The environmental trends we observed, although limited by metadata completeness, are consistent with this interpretation. Pirate genomes were more frequently associated with host-associated and aquatic environments than with soil, pointing to ecological settings where diffusible public goods are likely to be abundant and accessible.

Importantly, piracy does not appear to arise through the invention of novel uptake strategies. Instead, pirate genomes retain the same broad transporter architectures found in producers from the same lineages. Gram-negative pirates predominantly use TonB-linked systems, whereas Gram-positive pirates rely more heavily on ABC-type transporters. This suggests that biosynthesis can be lost relatively easily, but uptake remains constrained by the structural and evolutionary framework already present in each lineage.

4.4 Siderophore Piracy and Black Queen Hypothesis

The widespread occurrence of transporter-positive, biosynthesis-negative genomes fits naturally within a public-goods framework. Because siderophores are secreted molecules whose benefits can extend beyond the producing cell, they create opportunities for exploitation by neighboring organisms. In this context, our data are broadly consistent with social-evolutionary models in which a subset of the community maintains a costly cooperative trait while others benefit from it at lower cost. This logic also aligns with the Black Queen Hypothesis, which predicts that costly and leaky functions can be lost when they remain available through the activity of other community members. Siderophore production is a particularly plausible example: it is metabolically expensive, extracellular, and potentially exploitable. Our results suggest that the repeated emergence of pirate genomes across distantly related bacterial groups is not random, but reflects a recurring and potentially stable evolutionary outcome of this asymmetry between production and uptake[41].

At the same time, this dependency has obvious limits. Pirate strategies are only viable in communities where producers remain present. Understanding how this balance is maintained, and under which ecological conditions piracy becomes advantageous, will be important for linking siderophore evolution to community structure and microbial interaction networks.

4.5 A Modular View of Siderophore System Evolution

Taken together, our findings support a modular model of siderophore system evolution. Biosynthesis and uptake are clearly linked, but they are not evolutionarily inseparable. Biosynthetic pathways retain stronger lineage-specific signatures and more clearly traceable origins, whereas uptake systems are more widely retained, reused, and recombined within the constraints imposed by cell envelope architecture.

This modularity offers a useful way to interpret the diversity of siderophore systems across bacteria. Rather than evolving as fixed, self-contained units, siderophore systems appear to be assembled from partially independent components with different evolutionary dynamics. In this framework, the loss of biosynthesis while retaining uptake is not exceptional, but an expected outcome under ecological conditions where siderophores can be accessed as shared resources.

4.6 Limitations and Future Directions

Several limitations should be considered when interpreting these findings. First, taxonomic representation in GTDB remains uneven. Some well-studied lineages are heavily sampled, whereas many deep-branching or candidate phyla remain sparsely represented. Although we accounted for this by normalizing trait frequencies and explicitly visualizing sampling depth, uneven representation remains an unavoidable source of bias.

Second, ecological metadata were available for only a subset of genomes. This limited our ability to rigorously connect siderophore strategies to habitat, host association, or other environmental variables. Better genome-linked metadata will be essential for testing whether particular siderophore strategies are consistently associated with specific ecological settings.

Third, transporter identification was based on domain-level HMM detection and does not by itself establish substrate specificity. Some detected domains may belong to transport systems that are not dedicated to siderophore uptake, particularly when they occur outside recognizable biosynthetic contexts. Incorporating gene neighborhood, operon structure, regulation, or experimental validation would improve confidence in assigning transporter function.

Finally, reconciliation analyses are most informative for biosynthetic markers with relatively clear phylogenetic signal. For transporter families, the combination of broad distribution, repeated reuse, and possible functional promiscuity makes deeper evolutionary interpretation more difficult. Future analyses that incorporate synteny, substrate specificity, or network-based representations of gene flow may help resolve these histories more fully.

Despite these limitations, the patterns described here point to a broader principle: siderophore systems evolve not only under biochemical constraints, but also under ecological and social ones. This has implications beyond evolutionary biology, including the use of siderophore uptake pathways in antimicrobial delivery, Trojan-horse antibiotic design, and the engineering of microbial communities.

5 Conclusion

This study provides a genome-scale view of siderophore biosynthesis and uptake across the bacterial tree of life by integrating biosynthetic classification, transporter profiling, and phylogenetic analysis across ∼62,000 representative genomes. Our results reveal contrasting evolutionary trajectories for major siderophore classes: NRP-metallophore pathways are broadly distributed and evolutionarily dynamic, whereas NRPS-independent siderophores show more lineage-restricted patterns consistent with stronger vertical inheritance.

Across bacteria, siderophore transport systems are far more widespread than biosynthetic pathways and are strongly constrained by cell envelope architecture. TonB-dependent uptake is largely restricted to Gram-negative bacteria, while ABC-type transporters occur across both envelope types and form recurrent modular uptake architectures. The frequent occurrence of genomes encoding transporters but lacking biosynthetic pathways highlights the widespread emergence of siderophore "pirate" strategies, in which organisms exploit siderophores produced by neighboring microbes.

Together, these findings support a modular view of siderophore system evolution in which biosynthesis and uptake follow partially independent trajectories shaped by cellular structure, metabolic costs, and ecological interactions. Beyond evolutionary insight, this framework provides a basis for predicting iron-acquisition strategies from genomic data and may inform the discovery of novel siderophores and the targeting of siderophore uptake pathways in antimicrobial development.

References

[1] Rout, G.R., Sahoo, S.: ROLE OF IRON IN PLANT GROWTH AND METABOLISM. Reviews in Agricultural Science **3**(0), 1–24 (2015) https://doi.org/10.7831/ras.3.1

[2] Hider, R.C., Kong, X.: Chemistry and biology of siderophores. Natural Product Reports **27**(5), 637 (2010) https://doi.org/10.1039/b906679a

[3] Ratledge, C., Dover, L.G.: Iron Metabolism in Pathogenic Bacteria. Annual Review of Microbiology **54**(1), 881–941 (2000) https://doi.org/10.1146/annurev.micro.54.1.881

[4] Neilands, J.B.: Siderophores: Structure and Function of Microbial Iron Transport Compounds. Journal of Biological Chemistry **270**(45), 26723–26726 (1995) https://doi.org/10.1074/jbc.270.45.26723

[5] Butler, A., Theisen, R.M.: Iron(III)–siderophore coordination chemistry: Reactivity of marine siderophores. Coordination Chemistry Reviews **254**(3-4), 288–296 (2010) https://doi.org/10.1016/j.ccr.2009.09.010

[6] Süssmuth, R.D., Mainz, A.: Nonribosomal Peptide Synthesis—Principles and Prospects. Angewandte Chemie International Edition **56**(14), 3770–3821 (2017) https://doi.org/10.1002/anie.201609079

[7] Oliveira, F., Lima, T., Correia, A., Silva, A.M., Soares, C., Morais, S., Weißelberg, S., Vilanova, M., Rohde, H., Cerca, N.: Siderophore-Mediated Iron Acquisition Plays a Critical Role in Biofilm Formation and Survival of Staphylococcus epidermidis Within the Host. Frontiers in Medicine **8** (2021) https://doi.org/10.3389/fmed.2021.799227

[8] Chaturvedi, K.S., Hung, C.S., Giblin, D.E., Urushidani, S., Austin, A.M., Dinauer, M.C., Henderson, J.P.: Cupric Yersiniabactin Is a Virulence-Associated Superoxide Dismutase Mimic. ACS Chemical Biology **9**(2), 551–561 (2014) https://doi.org/10.1021/cb400658k

[9] Koh, E.-I., Henderson, J.P.: Microbial Copper-binding Siderophores at the Host-Pathogen Interface. Journal of Biological Chemistry **290**(31), 18967–18974 (2015) https://doi.org/10.1074/jbc.R115.644328

[10] Reitz, Z.L., Pourmohsenin, B., Susman, M., Thomsen, E., Roth, D., Butler, A., Ziemert, N., Medema, M.H.: Automated genome mining predicts structural diversity and taxonomic distribution of peptide metallophores across bacteria. Biochemistry (2025). https://doi.org/10.1101/2025.07.29.667501

[11] Alanjary, M., Kronmiller, B., Adamek, M., Blin, K., Weber, T., Huson, D., Philmus, B., Ziemert, N.: The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. Nucleic Acids Research **45**(W1), 42–48 (2017) https://doi.org/10.1093/nar/gkx360

[12] Carroll, C.S., Moore, M.M.: Ironing out siderophore biosynthesis: a review of non-ribosomal peptide synthetase (NRPS)-independent siderophore synthetases. Critical Reviews in Biochemistry and Molecular Biology **53**(4), 356–381 (2018) https://doi.org/10.1080/10409238.2018.1476449

[13] Challis, G.L.: A Widely Distributed Bacterial Pathway for Siderophore Biosynthesis Independent of Nonribosomal Peptide Synthetases. ChemBioChem **6**(4), 601–611 (2005) https://doi.org/10.1002/cbic.200400283

[14] Crosa, J.H., Walsh, C.T.: Genetics and Assembly Line Enzymology of Siderophore Biosynthesis in Bacteria. Microbiology and Molecular Biology Reviews **66**(2), 223–249 (2002) https://doi.org/10.1128/MMBR.66.2.223-249.2002

[15] Lamont, I.L., Martin, L.W.: Identification and characterization of novel pyoverdine synthesis genes in Pseudomonas aeruginosa. Microbiology **149**(4), 833–842 (2003) https://doi.org/10.1099/mic.0.26085-0

[16] Chu, B.C., Garcia-Herrero, A., Johanson, T.H., Krewulak, K.D., Lau, C.K., Peacock, R.S., Slavinskaya, Z., Vogel, H.J.: Siderophore uptake in bacteria and the battle for iron with the host; a bird's eye view. BioMetals **23**(4), 601–611 (2010) https://doi.org/10.1007/s10534-010-9361-x

[17] Martinez, J.S., Butler, A.: Marine amphiphilic siderophores: Marinobactin structure, uptake, and microbial partitioning. Journal of Inorganic Biochemistry **101**(11-12), 1692–1698 (2007) https://doi.org/10.1016/j.jinorgbio.2007.07.007

[18] Sandy, M., Butler, A.: Microbial Iron Acquisition: Marine and Terrestrial Siderophores. Chemical Reviews **109**(10), 4580–4595 (2009) https://doi.org/10.1021/cr9002787

[19] Miethke, M., Marahiel, M.A.: Siderophore-Based Iron Acquisition and Pathogen Control. Microbiology and Molecular Biology Reviews **71**(3), 413–451 (2007) https://doi.org/10.1128/mmbr.00012-07

[20] Noinaj, N., Guillier, M., Barnard, T.J., Buchanan, S.K.: TonB-Dependent Transporters: Regulation, Structure, and Function. Annual Review of Microbiology **64**(1), 43–60 (2010) https://doi.org/10.1146/annurev.micro.112408.134247

[21] Silale, A., Van Den Berg, B.: TonB-Dependent Transport Across the Bacterial Outer Membrane. Annual Review of Microbiology **77**(1), 67–88 (2023) https://doi.org/10.1146/annurev-micro-032421-111116

[22] Delepelaire, P.: Bacterial ABC transporters of iron containing compounds. Research in Microbiology **170**(8), 345–357 (2019) https://doi.org/10.1016/j.resmic.2019.10.008

[23] Fukushima, T., Allred, B.E., Raymond, K.N.: Direct Evidence of Iron Uptake by the Gram-Positive Siderophore-Shuttle Mechanism without Iron Reduction. ACS Chemical Biology **9**(9), 2092–2100 (2014) https://doi.org/10.1021/cb500319n

[24] Hanks, T.S., Liu, M., McClure, M.J., Lei, B.: ABC transporter FtsABCD of Streptococcus pyogenes mediates uptake of ferric ferrichrome. BMC Microbiology **5**(1), 62 (2005) https://doi.org/10.1186/1471-2180-5-62

[25] Cheng, W., Li, Q., Jiang, Y.-L., Zhou, C.-Z., Chen, Y.: Structures of Streptococcus pneumoniae PiaA and Its Complex with Ferrichrome Reveal Insights into the Substrate Binding and Release of High Affinity Iron Transporters. PLoS ONE **8**(8), 71451 (2013) https://doi.org/10.1371/journal.pone.0071451

[26] Miethke, M., Schmidt, S., Marahiel, M.A.: The Major Facilitator Superfamily-Type Transporter YmfE and the Multidrug-Efflux Activator Mta Mediate Bacillibactin Secretion in *Bacillus subtilis*. Journal of Bacteriology **190**(15), 5143–5152 (2008) https://doi.org/10.1128/JB.00464-08

[27] Liu, L., Wang, W., Wu, S., Gao, H.: Recent Advances in the Siderophore Biology of Shewanella. Frontiers in Microbiology **13**, 823758 (2022) https://doi.org/10.3389/fmicb.2022.823758

[28] Blin, K., Shaw, S., Augustijn, H.E., Reitz, Z.L., Biermann, F., Alanjary, M., Fetter, A., Terlouw, B.R., Metcalf, W.W., Helfrich, E.J.N., van Wezel, G.P., Medema, M.H., Weber, T.: antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Research **51**(W1), 46–50 (2023) https://doi.org/10.1093/nar/gkad344

[29] Hannigan, G.D., Prihoda, D., Palicka, A., Soukup, J., Klempir, O., Rampula, L., Durcak, J., Wurst, M., Kotowski, J., Chang, D., Wang, R., Piizzi, G., Temesi, G., Hazuda, D.J., Woelk, C.H., Bitton, D.A.: A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Research **47**(18), 110–110 (2019) https://doi.org/10.1093/nar/gkz654

[30] Crits-Christoph, A., Bhattacharya, N., Olm, M.R., Song, Y.S., Banfield, J.F.: Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity. Genome Research **31**(2), 239–250 (2021) https://doi.org/10.1101/gr.268169.120

[31] Martinez, J.S., Carter-Franklin, J.N., Mann, E.L., Martin, J.D., Haygood, M.G., Butler, A.: Structure and membrane affinity of a suite of amphiphilic siderophores produced by a marine bacterium. Proceedings of the National Academy of Sciences **100**(7), 3754–3759 (2003) https://doi.org/10.1073/pnas.0637444100

[32] Thode, S.K., Rojek, E., Kozlowski, M., Ahmad, R., Haugen, P.: Distribution of siderophore gene systems on a Vibrionaceae phylogeny: Database searches, phylogenetic analyses and evolutionary perspectives. PLOS ONE **13**(2), 0191860 (2018) https://doi.org/10.1371/journal.pone.0191860

[33] Payne, S.M., Mey, A.R., Wyckoff, E.E.: Vibrio Iron Transport: Evolutionary Adaptation to Life in Multiple Environments. Microbiology and Molecular Biology Reviews **80**(1), 69–90 (2016) https://doi.org/10.1128/MMBR.00046-15

[34] Terlouw, B.R., Blin, K., Navarro-Muñoz, J.C., Avalon, N.E., Chevrette, M.G., Egbert, S., Lee, S., Meijer, D., Recchia, M.J., Reitz, Z., van Santen, J., Selem-Mojica, N., Tørring, T., Zaroubi, L., Alanjary, M., Aleti, G., Aguilar, C., Al-Salihi, S.A., Augustijn, H., Avelar-Rivas, J., Avitia-Domínguez, L., Barona-Gómez, F., Bernaldo-Agüero, J., Bielinski, V.A., Biermann, F., Booth, T., Carrion Bravo, V., Castelo-Branco, R., Chagas, F., Cruz-Morales, P., Du, C., Duncan, K., Gavriilidou, A., Gayrard,

D., Gutiérrez-García, K., Haslinger, K., Helfrich, E.N., van der Hooft, J.J., Jati, A., Kalkreuter, E., Kalyvas, N., Kang, K., Kautsar, S., Kim, W., Kunjapur, A., Li, Y.-X., Lin, G.-M., Loureiro, C., Louwen, J.R., Louwen, N.L., Lund, G., Parra, J., Philmus, B., Pourmohsenin, B., Pronk, L.U., Rego, A., Rex, D., Robinson, S., Rosas-Becerra, L., Roxborough, E., Schorn, M., Scobie, D., Singh, K., Sokolova, N., Tang, X., Udwary, D., Vigneshwari, A., Vind, K., Vromans, S.J.M., Waschulin, V., Williams, S., Winter, J., Witte, T., Xie, H., Yang, D., Yu, J., Zdouc, M., Zhong, Z., Collemare, J., Linington, R., Weber, T., Medema, M.: MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Research **51**(D1), 603–610 (2023) https://doi.org/10.1093/nar/gkac1049

[35] Zdouc, M., Blin, K., Louwen, N.L., Navarro, J., Loureiro, C., Bader, C., Bailey, C., Barra, L., Booth, T., Bozhüyük, K.J., Cediel-Becerra, J.D., Charlop-Powers, Z., Chevrette, M., Chooi, Y.H., D'Agostino, P., de Rond, T., Del Pup, E., Duncan, K., Gu, W., Hanif, N., Helfrich, E.N., Jenner, M., Katsuyama, Y., Korenskaia, A., Krug, D., Libis, V., Lund, G., Mantri, S., Morgan, K., Owen, C., Phan, C.-S., Philmus, B., Reitz, Z., Robinson, S., Singh, K.S., Teufel, R., Tong, Y., Tugizimana, F., Ulanova, D., Winter, J., Aguilar, C., Akiyama, D., Al-Salihi, S.A., Alanjary, M., Alberti, F., Aleti, G., Alharthi, S., Rojo, M.A., Arishi, A., Augustijn, H., Avalon, N., Avelar-Rivas, J., Axt, K., Barbieri, H., Barbosa, J., Barboza Segato, L.G., Barrett, S., Baunach, M., Beemelmanns, C., Beqaj, D., Berger, T., Bernaldo-Agüero, J., Bettenbühl, S., Bielinski, V., Biermann, F., Borges, R., Borriss, R., Breitenbach, M., Bretscher, K., Brigham, M., Buedenbender, L., Bulcock, B., Cano-Prieto, C., Capela, J., Carrion, V., Carter, R., Castelo-Branco, R., Castro-Falcón, G., Chagas, F., Charria-Girón, E., Chaudhri, A.A., Chaudhry, V., Choi, H., Choi, Y., Choupannejad, R., Chromy, J., Donahey, M.C., Collemare, J., Connolly, J., Creamer, K., Crüsemann, M., Cruz, A., Cumsille, A., Dallery, J.-F., Damas-Ramos, L., Damiani, T., de Kruijff, M., Martín, B.D., Sala, G.D., Dillen, J., Doering, D., Dommaraju, S., Durusu, S., Egbert, S., Ellerhorst, M., Faussurier, B., Fetter, A., Feuermann, M., Fewer, D., Foldi, J., Frediansyah, A., Garza, E., Gavriilidou, A., Gentile, A., Gerke, J., Gerstmans, H., Gomez-Escribano, J.P., González-Salazar, L., Grayson, N., Greco, C., Gomez, J.G., Guerra, S., Flores, S.G., Gurevich, A., Gutiérrez-García, K., Hart, L., Haslinger, K., He, B., Hebra, T., Hemmann, J., Hindra, H., Höing, L., Holland, D., Holme, J., Horch, T., Hrab, P., Hu, J., Huynh, T.-H., Hwang, J.-Y., Iacovelli, R., Iftime, D., Iorio, M., Jayachandran, S., Jeong, E., Jing, J., Jung, J., Kakumu, Y., Kalkreuter, E., Kang, K.B., Kang, S., Kim, W., Kim, G.J., Kim, H., Kim, H., Klapper, M., Koetsier, R., Kollten, C., Kovács, , Kriukova, Y., Kubach, N., Kunjapur, A., Kushnareva, A., Kust, A., Lamber, J., Larralde, M., Larsen, N., Launay, A., Le, N.-T.-H., Lebeer, S., Lee, B.T., Lee, K., Lev, K., Li, S.-M., Li, Y.-X., Licona-Cassani, C., Lien, A., Liu, J., Lopez, J., Machushynets, N., Macias, M., Mahmud, T., Maleckis, M., Martinez-Martinez, A.M., Mast, Y., Maximo, M., McBride, C., McLellan, R., Bhatt, K.M., Melkonian, C., Merrild, A., Metsä-Ketelä, M., Mitchell, D., Müller, A., Nguyen, G.-S., Nguyen, H., Niedermeyer, T., O'Hare, J., Ossowicki, A., Ostash, B., Otani, H., Padva, L., Paliyal, S., Pan, X., Panghal, M., Parade, D.S., Park, J., Parra, J., Rubio, M.P., Pham, H., Pidot, S., Piel, J., Pourmohsenin, B., Rakhmanov, M., Ramesh, S., Rasmussen, M., Rego, A., Reher, R., Rice, A., Rigolet, A., Romero-Otero, A., Rosas-Becerra, L., Rosiles, P., Rutz, A., Ryu, B., Sahadeo, L.-A., Saldanha, M., Salvi, L., Sánchez-Carvajal, E., Santos-Medellin, C., Sbaraini, N., Schoellhorn, S., Schumm, C., Sehnal, L., Selem, N., Shah, A., Shishido, T., Sieber, S., Silviani, V., Singh, G., Singh, H., Sokolova, N., Sonnenschein, E., Sosio, M., Sowa, S., Steffen, K., Stegmann, E., Streiff, A., Strüder, A., Surup, F., Svenningsen, T., Sweeney, D., Szenei, J., Tagirdzhanov, A., Tan, B., Tarnowski, M., Terlouw, B., Rey, T., Thome, N., Torres Ortega, L.R., Tørring, T., Trindade, M., Truman, A., Tvilum, M., Udwary, D., Ulbricht, C., Vader, L., van Wezel, G., Walmsley, M., Warnasinghe, R., Weddeling, H., Weir, A.M., Williams, K., Williams, S., Witte, T., Rocca, S.W., Yamada, K., Yang, D., Yang, D., Yu, J., Zhou, Z., Ziemert, N., Zimmer, L., Zimmermann, A., Zimmermann, C., van der Hooft, J.J., Linington, R., Weber, T., Medema, M.: MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. Nucleic Acids Research **53**(D1), 678–690 (2025) https://doi.org/10.1093/nar/gkae1115

[36] Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., Hugenholtz, P.: GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Research **50**(D1), 785–794 (2022) https://doi.org/10.1093/nar/gkab776

[37] Eddy, S.R.: Accelerated profile HMM searches. PLoS Computational Biology **7**(10), 1002195 (2011) https://doi.org/10.1371/journal.pcbi.1002195

[38] Katoh, K., Rozewicki, J., Yamada, K.D.: MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. Briefings in Bioinformatics **20**(4), 1160–1166 (2019) https://doi.org/10.1093/bib/bbx108

[39] Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Haeseler, A., Lanfear, R.: IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution **37**(5), 1530–1534 (2020) https://doi.org/10.1093/molbev/msaa015

[40] Letunic, I., Bork, P.: Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research **49**(W1), 293–296 (2021) https://doi.org/10.1093/nar/gkab301

[41] Morris, J.J., Lenski, R.E., Zinser, E.R.: The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. mBio **3**(2), 00036–12 (2012) https://doi.org/10.1128/mBio.00036-12

Discussion

Overview and Integration of Findings

This dissertation brings together three interconnected lines of research that address complementary aspects of siderophore and metallophore biology in bacteria, while also contributing substantial methodological innovations to the field of phylogeny-aware genome mining. Each component focuses on a different level of analysis, ranging from phylogenomic infrastructure to functional genome mining to evolutionary interpretation. Taken together, these contributions enable a broad view of how bacterial iron acquisition systems are detected, distributed, and diversified across the tree of life.

The first major contribution is the development of AutoMLST2, a redesigned and expanded platform based on the original AutoMLST framework[24, 25], for automated phylogenomic reconstruction. By modernizing the backend into a modular, Docker-based architecture and expanding the marker HMM library, AutoMLST2 enables scalable, high-resolution species-tree inference directly from genome assemblies and integrates results with the Genome Taxonomy Database (GTDB)[26]. Rather than serving as the primary species tree for the siderophore evolution analyses, AutoMLST2 generalizes the phylogenomic principles used there into an accessible, web-based tool that allows the wider community to perform GTDB-consistent phylogeny reconstruction and trait mapping at scale.

The second component is the study, titled "**Automated genome mining predicts structural diversity and taxonomic distribution of peptide metallophores across bacteria**", which introduced new chelator-centric detection rules into antiSMASH, the most widely used genome mining tool for biosynthetic gene clusters[5]. By distinguishing NRP-derived metallophore pathways from generic NRPS BGCs, these rules add a critical

functional layer to existing annotations and reveal that more than a quarter of NRPS clusters in the analyzed genome set encode metallophore-like chelators[33]. This work builds on and contributes to community resources such as MIBiG, where I helped curate siderophore and metallophore entries; these experimentally validated BGCs provided important reference points for designing and benchmarking the new detection rules[7, 34]. Together, these efforts substantially extend the annotated space of metal-binding metabolites beyond the small set of classic model pathways.

The third component expands from peptide metallophores to the full landscape of siderophore biosynthesis and transport systems, integrating biosynthetic annotation, transporter profiling, large-scale trait mapping, and reconciliation-based evolutionary inference. Using the GTDB reference species tree[26] and its pruned derivatives, including REDgroup-level representations[2], this study examines how NRPS- and NIS-type siderophore pathways, TonB-dependent receptors, ABC transporters, periplasmic binding proteins, and MFS families are distributed across >60,000 bacterial genomes. Gene tree-species tree reconciliations then quantify the duplication, transfer, and loss dynamics of these systems, revealing where biosynthesis and transport have diversified together, where they have become uncoupled, and how social interactions and cell-envelope constraints have shaped their evolution[27, 35, 36].

Collectively, these three strands, phylogenomics, improved functional annotation, and comparative evolutionary analysis, form a unified narrative. The phylogenetic and curation work provides the infrastructure and reference datasets needed to interpret siderophore traits in an evolutionary context; the metallophore-focused genome mining demonstrates the scale and modularity of metal-binding chemistry; and the siderophore evolution study brings these insights into a broader view of how complex metabolic traits evolve under ecological constraint, how biosynthesis and transport co-adapt or decouple, and how dependency strategies repeatedly emerge in microbial systems. The following sections build on this integrated framework to discuss the broader evolutionary, ecological, and methodological implications of the findings.

Importance of Accurate Phylogeny and the Role of AutoMLST2

A central premise of this dissertation is that large-scale genome mining can only yield meaningful evolutionary insight when embedded in a robust and consistent phylogenomic framework. While thousands of bacterial genomes are now available, comparative analyses of biosynthetic and transporter traits remain fundamentally limited if genomes are not placed into a shared species-tree context. Throughout this work, the GTDB reference phylogeny and its pruned derivatives provided such a framework, enabling siderophore traits to be mapped, compared, and interpreted across the bacterial tree of life[26]. The GTDB taxonomy, derived from a standardized set of universally conserved marker genes, offers an important advantage for evolutionary analysis: it reflects genome-based relationships rather than historical naming conventions or lineage-specific traditions. Using this reference tree ensured that trait distributions, enrichment patterns, and reconciliation results were evaluated against a consistent and up-to-date representation of bacterial diversity[37]. This consistency was particularly important given the scale of the analyses performed here and the uneven taxonomic representation inherent in large genome collections.

Working within this framework also made clear how strongly evolutionary inference depends on phylogenetic accuracy. Small differences in topology, especially at deeper taxonomic levels such as classes or orders, can influence interpretations of lineage-specific innovation, inferred pathway origins, or the apparent enrichment of traits in particular clades. Pruning the GTDB tree to different levels of granularity proved essential for balancing resolution with interpretability, allowing broad evolutionary patterns to be identified without overinterpreting noise introduced by uneven sampling or recent diversification[2].

The importance of accurate phylogeny becomes particularly evident when applying reconciliation algorithms. Gene tree-species tree reconciliation does not infer evolutionary events in isolation; instead, it explains gene tree discordance relative to a fixed species tree. As a result, the biological plausibility of inferred duplication, transfer, and loss events depends directly on the quality and coherence of the species tree[27]. In this dissertation, anchoring reconciliation analyses to the GTDB framework provided a stable evolutionary

reference that allowed broad trends in siderophore biosynthetic and transporter evolution to be compared across gene families and lineages.

The practical and conceptual challenges encountered during these analyses directly motivated the redesign of the original AutoMLST pipeline into AutoMLST2, presented in Manuscript I. Although the siderophore evolution analyses in Manuscript III (Chapter 3) rely on the GTDB reference tree rather than species trees reconstructed de novo with AutoMLST2, the experience of scaling phylogenetic analyses to tens of thousands of genomes, maintaining taxonomic consistency, and ensuring reproducibility across workflows informed many of the design decisions underlying the updated tool. AutoMLST2 was developed to address these challenges by providing a GTDB-consistent, user-friendly platform for species-tree reconstruction and taxonomic placement that can be applied outside specialized high-performance computing environments[25].

In this sense, AutoMLST2 represents a methodological contribution that complements the biological findings of this dissertation. By lowering the barrier to accurate phylogenomic reconstruction and standardizing species-tree generation for genome mining studies, it supports the growing need for evolutionary context in large comparative analyses. More broadly, this work highlights that phylogeny is not merely a visualization tool, but a prerequisite for interpreting the evolutionary history of biosynthetic pathways, transport systems, and other complex genomic traits.

Expanding the Landscape of Siderophore and Metallophore Diversity

Manuscript II (Chapter 2) expanded the detectable diversity of bacterial metallophore biosynthetic systems through the development of new genome-mining rules for peptide metallophore gene clusters. When interpreted within the phylogenomic framework established in Manuscript I (Chapter 1), these expanded annotations reveal that metal-chelating biosynthetic pathways are broadly distributed across bacterial lineages but vary strongly in their taxonomic enrichment and evolutionary trajectory. These patterns provide an important comparative background for the evolutionary analyses of siderophore biosynthesis and transport systems presented in Manuscript III (Chapter 3).

By introducing chelator-centric detection rules for NRPS-derived metallophore pathways and integrating these with phylogeny-aware analyses, this work expands the observable landscape of metal-binding metabolites beyond the limited set of pathways traditionally emphasized in the literature[5, 14, 38]. Classical views of siderophore diversity are strongly shaped by a small number of intensively studied model organisms and experimentally tractable systems. While these systems provided critical biochemical insight, they also created a detection bias: many NRPS clusters with metal-binding potential were annotated only as generic peptide synthetases, obscuring their functional role in iron or metal acquisition. The metallophore-focused genome mining presented in this dissertation demonstrates that metal-binding chemistry is not a rare specialization, but a widespread feature of bacterial secondary metabolism[14].

This expansion is primarily functional. The chelator-centric framework does not redefine NRPS architecture, but instead identifies characteristic domain combinations and tailoring enzymes that confer metal-binding capacity. This distinction is central for evolutionary analysis. Treating metallophores as a functional subset of NRPS pathways allows their distribution, diversity, and evolutionary history to be analyzed in a unified way, rather than fragmenting them into pathway-specific case studies[5, 33].

Once these pathways are systematically detected, broader patterns become visible. Metallophore biosynthesis spans a wide range of bacterial phyla and ecological contexts, and frequently exhibits modular organization, including the combination of multiple chelating moieties within a single biosynthetic locus. Such modularity suggests that chelation properties may be tuned through recombination and domain shuffling, providing bacteria with flexible solutions to metal limitation under different environmental conditions. These observations reinforce the view of siderophore systems as evolvable assemblies rather than fixed biochemical entities[14, 33].

Expanding the detectable biosynthetic landscape also changes how evolutionary questions can be posed. Without improved annotation, the apparent rarity or lineage restriction of certain siderophore strategies could be misinterpreted as true biological absence. By contrast, the framework developed here reveals that many lineages possess latent or previously unrecognized metal-binding capacity, which in turn influences how transporter repertoires, uptake strategies, and dependency patterns are interpreted in later sections of this dissertation.

In this regard, improved functional detection is not an endpoint but a prerequisite. It enables subsequent analyses of evolutionary trajectories, co-adaptation with transport systems, and the repeated emergence of biosynthesis-uptake decoupling to be grounded in a more complete and less biased representation of siderophore diversity. The following sections build on this expanded landscape to examine how different biosynthetic strategies have evolved over time, and how their evolutionary dynamics differ depending on chemical complexity, cellular architecture, and ecological context.

Divergent Evolutionary Trajectories of NRP and NIS Siderophores

One of the clearest patterns emerging from this dissertation is that NRPS-derived siderophores and NRPS-independent siderophores (NISs) follow markedly different evolutionary trajectories. Although both strategies solve the same physiological problem, which is iron acquisition under limiting conditions, their histories, modes of diversification, and degrees of evolutionary flexibility differ substantially[12, 14].

NRP-metallophores are broadly distributed across the bacterial tree and show strong signatures of evolutionary plasticity. Their biosynthetic genes frequently exhibit histories shaped by gene duplication and horizontal transfer, consistent with repeated recruitment into new genomic and ecological contexts. At the same time, some components of NRP pathways retain deep lineage-specific signatures, suggesting that ancient cores are repeatedly modified rather than replaced wholesale. This combination of stability and flexibility is consistent with the modular organization of NRPS systems and helps explain their persistence across diverse bacterial lineages[39–41].

In contrast, NIS pathways display more restricted phylogenetic distributions and stronger signals of vertical inheritance. Their concentration within specific proteobacterial and firmicute lineages, together with fewer inferred transfer events, suggests that NISs represent a more specialized solution that emerged later in bacterial evolution. Rather than being broadly mobile, these systems appear to be optimized within particular ecological or physiological niches, where simpler biosynthetic routes may provide sufficient chelation capacity without the metabolic overhead of large NRPS assemblies[12, 14].

These contrasting trajectories highlight that siderophore evolution is not governed by a single dominant strategy. Instead, different biosynthetic paradigms coexist because they occupy distinct evolutionary and ecological regimes. NRP systems favor flexibility, recombination, and repeated redeployment, whereas NIS systems favor economy, lineage stability, and constrained dissemination[39]. Recognizing this divergence is essential for interpreting later analyses of transporter evolution and biosynthesis-uptake decoupling, as it frames why some siderophore systems readily spread and recombine, while others remain tightly associated with specific clades.

Transport Systems, Cell Envelope Constraints, and Evolutionary Channeling

The analyses presented in Manuscript III (Chapter 3) reveal that siderophore biosynthesis and transport systems are tightly linked evolutionarily and must be interpreted within their broader cellular and ecological context. Identifying these associations requires both comprehensive genome annotation and scalable comparative frameworks. These methodological foundations were established in Manuscript II (Chapter 2), which expanded the detectable diversity of metal-chelating biosynthetic pathways through improved genome-mining strategies, and in Manuscript I (Chapter 1), which introduced the AutoMLST2 phylogenomic framework for consistent species-tree reconstruction and large-scale comparative analyses.

Within this integrated framework, an important insight emerging from this dissertation is that siderophore transport systems impose stronger evolutionary constraints than the biosynthetic pathways themselves. While biosynthetic genes can be gained, modified, or lost with relatively high evolutionary flexibility, uptake systems are tightly coupled to bacterial cell-envelope architecture. As a consequence, transport mechanisms act as long-term structural constraints that define which siderophore strategies are feasible within a given lineage[12].

This constraint is most evident for TonB-dependent uptake systems, which require a diderm cell envelope and associated energy-transduction machinery. Their confinement to Gram-negative bacteria effectively restricts TonB-linked siderophore strategies to these

lineages, independent of the potential biochemical utility of the siderophore. In contrast, ABC-type transporters operate at the cytoplasmic membrane and are compatible with both monoderm and diderm bacteria, allowing them to persist and diversify across a broader phylogenetic range[18, 20, 42, 43].

These differences lead to a form of evolutionary channeling. Although siderophore biosynthesis explores diverse chemical solutions, uptake systems are limited to a smaller set of envelope-compatible architectures. Once established, these architectures tend to be retained over long evolutionary timescales, even as biosynthetic pathways change. This helps explain why similar transport strategies recur within lineages and why shifts between fundamentally different uptake mechanisms are rare[44].

These constraints also frame later patterns of biosynthesis-uptake decoupling. When siderophore production is lost, the retained uptake systems reflect pre-existing, envelope-compatible solutions rather than novel innovations. Transport systems therefore do not simply respond to biosynthetic change, but actively channel the evolutionary trajectories available to siderophore systems as a whole[35].

All in all, this perspective highlights transporters as more than functional complements to biosynthesis. They represent structural and evolutionary filters that shape the long-term organization, persistence, and diversification of siderophore strategies across bacteria.

Decoupling of Biosynthesis and Uptake: Siderophore Piracy as an Evolutionary Strategy

A recurring pattern across the analyses in this dissertation is the frequent decoupling of siderophore biosynthesis and uptake. Many bacterial genomes retain the ability to import siderophore–iron complexes while lacking detectable biosynthetic pathways, indicating that siderophore production and uptake do not necessarily evolve as a single, inseparable unit. Rather than representing rare exceptions, such genomes appear across diverse bacterial lineages, suggesting that this decoupling reflects a common evolutionary outcome[12, 14].

From an evolutionary perspective, this pattern reflects the unequal costs and benefits of siderophore production and uptake. Siderophore biosynthesis is metabolically expensive and often involves large, multi-gene pathways, whereas uptake allows direct access to

an essential resource. In environments where siderophores are abundant and diffusible, selection may therefore favor the retention of uptake capacity even after biosynthetic functions are lost. This interpretation aligns with broader models of microbial public goods, in which costly, leaky functions are maintained by a subset of community members while others persist by exploiting shared resources[35, 36].

Siderophore piracy does not appear to involve the evolution of novel uptake mechanisms. Instead, genomes that lack biosynthesis typically retain transport systems already compatible with their phylogenetic background and cell-envelope architecture. As a result, piracy strategies differ between lineages in predictable ways: Gram-negative bacteria rely on outer-membrane-associated uptake systems, whereas Gram-positive bacteria depend on ABC-type transporters. This indicates that decoupling occurs within existing structural constraints rather than through the emergence of new transport solutions[18, 20, 45].

Taken together, these observations suggest that siderophore piracy is best viewed not as a transient anomaly, but as a stable and repeatedly accessible evolutionary strategy. The widespread occurrence of such genomes underscores that loss of biosynthesis can represent an adaptive response under appropriate ecological conditions, while continued reliance on conserved uptake architectures maintains access to communal iron pools[23].

Methodological Limitations and Interpretational Boundaries

While this dissertation integrates large-scale genome mining, phylogenomics, and evolutionary inference to investigate bacterial metal acquisition systems, several methodological limitations should be considered when interpreting the results. These limitations affect the phylogenomic analyses presented in Manuscript I (Chapter 1), the genome-mining approaches developed in Manuscript II (Chapter 2), and the evolutionary interpretations of siderophore systems discussed in Manuscript III (Chapter 3). Acknowledging these boundaries is essential for placing the findings in an appropriate methodological and biological context.

First, genome availability and taxonomic representation remain uneven across bacteria. Certain lineages are heavily overrepresented, whereas many deep-branching or candidate

phyla are sparsely sampled. Although analyses were normalized across taxonomic ranks and interpreted with respect to sampling depth, residual bias is unavoidable and may influence apparent trait prevalence or lineage-specific patterns[46, 47].

Second, the identification of siderophore transport systems is based primarily on domain-level HMM detection. While this approach enables consistent screening across tens of thousands of genomes, it does not guarantee functional specificity. Transporter domains associated with siderophore uptake can also occur in systems with broader or alternative substrate ranges, and some detected components may represent incomplete, repurposed, or inactive transport systems. As a result, a subset of transporter calls, particularly in genomes lacking detectable biosynthetic clusters, may constitute false positives with respect to siderophore uptake. Incorporating additional contextual information, such as gene neighborhood organization, operon structure, regulatory signals, or experimental validation, would help refine functional assignments in future studies[48–51].

Third, evolutionary inferences based on gene tree-species tree reconciliation necessarily simplify complex evolutionary histories. Parsimony-based reconciliation models summarize gene histories in terms of duplication, transfer, and loss events, but cannot capture all aspects of reticulate evolution, recombination, or uncertainty in gene-tree reconstruction. In this work, reconciliation results are therefore interpreted as indicators of broad evolutionary trends rather than precise reconstructions of individual events[52–55].

Finally, environmental and ecological metadata remain incomplete for many genomes, limiting the ability to directly link siderophore strategies to habitat, community composition, or iron availability. Although broad associations are discussed, more fine-grained ecological interpretation will require improved genome-linked metadata and integration with metagenomic or experimental studies[56–58].

Overall, these limitations do not undermine the central conclusions of this dissertation, but they define the scope within which the results should be interpreted. The analyses presented here are best viewed as a framework for identifying large-scale evolutionary patterns and generating testable hypotheses, rather than as definitive functional annotations for individual genomes or pathways.

Future Directions and Broader Implications

Beyond the specific insights into siderophore biology, this dissertation establishes a general framework for studying complex metabolic traits at scale[2, 8, 59]. Siderophores were chosen as a model system because they combine chemical diversity, ecological relevance, and well-defined transport mechanisms. However, the conceptual and methodological approach developed here is not limited to iron acquisition systems[48, 60–62].

The integration of improved functional genome mining with phylogeny-aware analysis and reconciliation-based evolutionary inference can be applied to a wide range of secondary metabolites and other complex genomic traits[61]. Many biosynthetic systems share key features with siderophores, including modular organization, frequent horizontal transfer, and partial decoupling between biosynthesis and associated functions such as transport or regulation[63]. Extending this framework to additional metabolite classes offers an opportunity to move beyond cataloging biosynthetic diversity toward understanding the evolutionary principles that shape secondary metabolism more broadly.

At the same time, future work will benefit from richer genomic context and improved functional resolution[64]. Incorporating gene neighborhood information, regulatory signals, metabolomic data, and experimental validation will allow more precise links between predicted pathways, ecological function, and evolutionary history. Improved genome-linked metadata will further enable direct connections between metabolic strategies and environmental conditions[65].

In conclusion, this dissertation demonstrates how phylogeny-aware genome mining can transform large genomic datasets into evolutionary insights. By treating phylogeny as an analytical foundation rather than a descriptive backdrop, it provides a path toward a more integrative understanding of how complex metabolic systems evolve, diversify, and persist across the bacterial tree of life.

Bibliography

[1] Atanas G. Atanasov et al. "Natural products in drug discovery: advances and opportunities". en. In: *Nature Reviews Drug Discovery* 20.3 (Mar. 2021), pp. 200–216. ISSN: 1474-1776, 1474-1784. DOI: 10.1038/s41573-020-00114-z.

[2] Athina Gavriilidou et al. "Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes". en. In: *Nature Microbiology* 7.5 (May 2022), pp. 726–735. ISSN: 2058-5276. DOI: 10.1038/s41564-022-01110-2.

[3] Yannik K. Schneider. "Bacterial Natural Product Drug Discovery for New Antibiotics: Strategies for Tackling the Problem of Antibiotic Resistance by Efficient Bioprospecting". en. In: *Antibiotics* 10.7 (July 2021), p. 842. ISSN: 2079-6382. DOI: 10.3390/antibiotics10070842.

[4] Marcus Miethke et al. "Towards the sustainable discovery and development of new antibiotics". en. In: *Nature Reviews Chemistry* 5.10 (Aug. 2021), pp. 726–749. ISSN: 2397-3358. DOI: 10.1038/s41570-021-00313-1.

[5] Kai Blin et al. "antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation". en. In: *Nucleic Acids Research* 51.W1 (July 2023), W46–W50. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkad344.

[6] Heonjun Jeong et al. "A guide to genome mining and genetic manipulation of biosynthetic gene clusters in Streptomyces". en. In: *Journal of Microbiology* 63.4 (Apr. 2025), e2409026. ISSN: 1225-8873, 1976-3794. DOI: 10.71150/jm.2409026.

[7] Barbara R Terlouw et al. "MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters". en. In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D603–D610. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkac1049.

[8] Kai Blin et al. “antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline”. en. In: *Nucleic Acids Research* 47.W1 (July 2019), W81–W87. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz310.

[9] Mehmet Direnç Mungan et al. “ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining”. en. In: *Nucleic Acids Research* 48.W1 (July 2020), W546–W552. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkaa374.

[10] Jessie Malit, Hiu Leung, and Pei-Yuan Qian. “Targeted Large-Scale Genome Mining and Candidate Prioritization for Natural Product Discovery”. en. In: *Marine Drugs* 20.6 (June 2022), p. 398. ISSN: 1660-3397. DOI: 10.3390/md20060398.

[11] Qilong Lai et al. “Deciphering the biosynthetic potential of microbial genomes using a BGC language processing neural network model”. en. In: *Nucleic Acids Research* 53.7 (Apr. 2025), gkaf305. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkaf305.

[12] Marcus Miethke and Mohamed A. Marahiel. “Siderophore-Based Iron Acquisition and Pathogen Control”. en. In: *Microbiology and Molecular Biology Reviews* 71.3 (Sept. 2007). Publisher: American Society for Microbiology, pp. 413–451. ISSN: 1092-2172, 1098-5557. DOI: 10.1128/mmbr.00012-07.

[13] Cassandra S. Carroll and Margo M. Moore. “Ironing out siderophore biosynthesis: a review of non-ribosomal peptide synthetase (NRPS)-independent siderophore synthetases”. en. In: *Critical Reviews in Biochemistry and Molecular Biology* 53.4 (July 2018), pp. 356–381. ISSN: 1040-9238, 1549-7798. DOI: 10.1080/10409238.2018.1476449.

[14] Robert C. Hider and Xiaole Kong. “Chemistry and biology of siderophores”. en. In: *Natural Product Reports* 27.5 (2010), p. 637. ISSN: 0265-0568, 1460-4752. DOI: 10.1039/b906679a.

[15] Ghassan Ghssein et al. “Biosynthesis of a broad-spectrum nicotianamine-like metallophore in *Staphylococcus aureus*”. en. In: *Science* 352.6289 (May 2016), pp. 1105–1109. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaf1018.

[16] Jennifer S. Martinez et al. “Structure and membrane affinity of a suite of amphiphilic siderophores produced by a marine bacterium”. en. In: *Proceedings of the National*

Academy of Sciences 100.7 (Apr. 2003), pp. 3754–3759. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0637444100.

[17] Carsten Peukert et al. "Trojan Horse Siderophore Conjugates Induce Pseudomonas aeruginosa Suicide and Qualify the TonB Protein as a Novel Antibiotic Target". In: *Journal of Medicinal Chemistry* 66.1 (Jan. 2023). Publisher: American Chemical Society, pp. 553–576. ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.2c01489.

[18] Augustinas Silale and Bert Van Den Berg. "TonB-Dependent Transport Across the Bacterial Outer Membrane". en. In: *Annual Review of Microbiology* 77.1 (Sept. 2023), pp. 67–88. ISSN: 0066-4227, 1545-3251. DOI: 10.1146/annurev-micro-032421-111116.

[19] Volkmar Braun. "Substrate Uptake by TonB -Dependent Outer Membrane Transporters". en. In: *Molecular Microbiology* 122.6 (Dec. 2024), pp. 929–947. ISSN: 0950-382X, 1365-2958. DOI: 10.1111/mmi.15332.

[20] Philippe Delepelaire. "Bacterial ABC transporters of iron containing compounds". en. In: *Research in Microbiology* 170.8 (Nov. 2019), pp. 345–357. ISSN: 09232508. DOI: 10.1016/j.resmic.2019.10.008.

[21] Tatsuya Fukushima, Benjamin E. Allred, and Kenneth N. Raymond. "Direct Evidence of Iron Uptake by the Gram-Positive Siderophore-Shuttle Mechanism without Iron Reduction". en. In: *ACS Chemical Biology* 9.9 (Sept. 2014), pp. 2092–2100. ISSN: 1554-8929, 1554-8937. DOI: 10.1021/cb500319n.

[22] Alexander Crits-Christoph et al. "Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity". en. In: *Genome Research* 31.2 (Feb. 2021). Publisher: Cold Spring Harbor Laboratory, pp. 239–250. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.268169.120.

[23] Jos Kramer, Özhan Özkaya, and Rolf Kümmerli. "Bacterial siderophores in community and host interactions". en. In: *Nature Reviews Microbiology* 18.3 (Mar. 2020). Publisher: Springer Science and Business Media LLC, pp. 152–163. ISSN: 1740-1526, 1740-1534. DOI: 10.1038/s41579-019-0284-4.

[24] Mohammad Alanjary, Kat Steinke, and Nadine Ziemert. "AutoMLST: an automated web server for generating multi-locus species trees highlighting natural product potential". en. In: *Nucleic Acids Research* 47.W1 (July 2019), W276–W282. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz282.

[25] Bita Pourmohsenin, Arthur Wiese, and Nadine Ziemert. "AutoMLST2: a web server for phylogeny and microbial taxonomy". en. In: *Nucleic Acids Research* 53.W1 (July 2025), W45–W50. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkaf397.

[26] Donovan H Parks et al. "GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy". en. In: *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D785–D794. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkab776.

[27] Gergely J. Szöllősi et al. "The Inference of Gene Trees with Species Trees". en. In: *Systematic Biology* 64.1 (Jan. 2015), e42–e62. ISSN: 1076-836X, 1063-5157. DOI: 10.1093/sysbio/syu048.

[28] Santi Santichaivekin et al. "eMPRess: a systematic cophylogeny reconciliation tool". en. In: *Bioinformatics* 37.16 (Aug. 2021). Ed. by Schwartz Russell, pp. 2481–2482. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btaa978.

[29] Benoit Morel et al. "AleRax: a tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss". en. In: *Bioinformatics* 40.4 (Mar. 2024). Ed. by Russell Schwartz, btae162. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btae162.

[30] Zhi-Zhong Chen, Fei Deng, and Lusheng Wang. "Simultaneous Identification of Duplications, Losses, and Lateral Gene Transfers". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9.5 (Sept. 2012), pp. 1515–1528. ISSN: 1545-5963. DOI: 10.1109/TCBB.2012.79.

[31] Misagh Kordi and Mukul S. Bansal. "Exact Algorithms for Duplication-Transfer-Loss Reconciliation with Non-Binary Gene Trees". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16.4 (July 2019), pp. 1077–1090. ISSN: 1545-5963, 1557-9964, 2374-0043. DOI: 10.1109/TCBB.2017.2710342.

[32] Örjan Åkerborg et al. “Simultaneous Bayesian gene tree reconstruction and reconciliation analysis”. en. In: *Proceedings of the National Academy of Sciences* 106.14 (Apr. 2009), pp. 5714–5719. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0806251106.

[33] Zachary L. Reitz et al. *Automated genome mining predicts structural diversity and taxonomic distribution of peptide metallophores across bacteria*. en. Aug. 2025. DOI: 10.1101/2025.07.29.667501.

[34] Mitja M Zdouc et al. “MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration”. en. In: *Nucleic Acids Research* 53.D1 (Jan. 2025), pp. D678–D690. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkae1115.

[35] J. Jeffrey Morris, Richard E. Lenski, and Erik R. Zinser. “The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss”. en. In: *mBio* 3.2 (May 2012), e00036–12. ISSN: 2161-2129, 2150-7511. DOI: 10.1128/mBio.00036-12.

[36] Otto X. Cordero et al. “Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations”. en. In: *Proceedings of the National Academy of Sciences* 109.49 (Dec. 2012), pp. 20059–20064. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1213344109.

[37] Pierre-Alain Chaumeil et al. “GTDB-Tk v2: memory friendly classification with the genome taxonomy database”. en. In: *Bioinformatics* 38.23 (Nov. 2022). Ed. by Karsten Borgwardt, pp. 5315–5316. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btac672.

[38] Zachary L Reitz et al. “Automated genome mining predicts structural diversity and taxonomic distribution of peptide metallophores across bacteria”. en. In: *eLife* 14 (Mar. 2026), RP109154. ISSN: 2050-084X. DOI: 10.7554/eLife.109154.3.

[39] Nadine Ziemert, Mohammad Alanjary, and Tilmann Weber. “The evolution of genome mining in microbes – a review”. en. In: *Natural Product Reports* 33.8 (2016), pp. 988–1005. ISSN: 0265-0568, 1460-4752. DOI: 10.1039/C6NP00025H.

[40] Roderich D. Süssmuth and Andi Mainz. “Nonribosomal Peptide Synthesis—Principles and Prospects”. en. In: *Angewandte Chemie International Edition* 56.14 (Mar. 2017). Publisher: Wiley, pp. 3770–3821. ISSN: 1433-7851, 1521-3773. DOI: 10.1002/anie.201609079.

[41] Robert Finking and Mohamed A. Marahiel. "Biosynthesis of Nonribosomal Peptides". en. In: *Annual Review of Microbiology* 58.1 (Oct. 2004), pp. 453–488. ISSN: 0066-4227, 1545-3251. DOI: 10.1146/annurev.micro.58.030603.123615.

[42] Andrew D. Ferguson and Johann Deisenhofer. "TonB-dependent receptors—structural perspectives". en. In: *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1565.2 (Oct. 2002), pp. 318–332. ISSN: 00052736. DOI: 10.1016/S0005-2736(02)00578-3.

[43] Jessica R. Sheldon and David E. Heinrichs. "Recent developments in understanding the iron acquisition strategies of gram positive pathogens". en. In: *FEMS Microbiology Reviews* 39.4 (July 2015). Ed. by Chris Whitfield, pp. 592–630. ISSN: 1574-6976. DOI: 10.1093/femsre/fuv009.

[44] Iain C. Sutcliffe. "A phylum level perspective on bacterial cell envelope architecture". en. In: *Trends in Microbiology* 18.10 (Oct. 2010), pp. 464–470. ISSN: 0966842X. DOI: 10.1016/j.tim.2010.06.005.

[45] Nicholas Noinaj et al. "TonB-Dependent Transporters: Regulation, Structure, and Function". en. In: *Annual Review of Microbiology* 64.1 (Oct. 2010), pp. 43–60. ISSN: 0066-4227, 1545-3251. DOI: 10.1146/annurev.micro.112408.134247.

[46] Sage Albright and Stilianos Louca. "Trait biases in microbial reference genomes". en. In: *Scientific Data* 10.1 (Feb. 2023), p. 84. ISSN: 2052-4463. DOI: 10.1038/s41597-023-01994-7.

[47] Samuel D. Chorlton. "Ten common issues with reference sequence databases and how to mitigate them". In: *Frontiers in Bioinformatics* 4 (Mar. 2024), p. 1278228. ISSN: 2673-7647. DOI: 10.3389/fbinf.2024.1278228.

[48] Shaohua Gu et al. "Feature sequence-based genome mining uncovers the hidden diversity of bacterial siderophore pathways". en. In: *eLife* 13 (Oct. 2024), RP96719. ISSN: 2050-084X. DOI: 10.7554/eLife.96719.3.

[49] Arkadiy I. Garber et al. "FeGenie: A Comprehensive Tool for the Identification of Iron Genes and Iron Gene Neighborhoods in Genome and Metagenome Assemblies". In: *Frontiers in Microbiology* 11 (Jan. 2020), p. 37. ISSN: 1664-302X. DOI: 10.3389/fmicb.2020.00037.

[50] Hugh Gifford et al. “Xenosiderophore transporter gene expression and clade-specific filamentation in Candida auris killifish (Aphanius dispar) infection”. en. In: *Communications Biology* 8.1 (Dec. 2025), p. 1790. ISSN: 2399-3642. DOI: 10.1038/s42003-025-09321-z.

[51] Miguel A Andrade et al. “NEAT: a domain duplicated in genes near the components of a putative Fe3+siderophore transporter from Gram-positive pathogenic bacteria”. en. In: *Genome Biology* 3.9 (Aug. 2002), research0047.1. ISSN: 1474-760X. DOI: 10.1186/gb-2002-3-9-research0047.

[52] Mukul S. Bansal, Eric J. Alm, and Manolis Kellis. “Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss”. en. In: *Bioinformatics* 28.12 (June 2012), pp. i283–i291. ISSN: 1367-4811, 1367-4803. DOI: 10.1093/bioinformatics/bts225.

[53] Krister M Swenson, Andrea Doroftei, and Nadia El-Mabrouk. “Gene tree correction for reconciliation and species tree inference”. en. In: *Algorithms for Molecular Biology* 7.1 (Dec. 2012), p. 31. ISSN: 1748-7188. DOI: 10.1186/1748-7188-7-31.

[54] Tom A. Williams et al. *The power and limitations of species tree-aware phylogenetics*. en. Mar. 2023. DOI: 10.1101/2023.03.17.533068.

[55] Yi-Chieh Wu et al. “Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees”. en. In: *Genome Research* 24.3 (Mar. 2014), pp. 475–486. ISSN: 1088-9051. DOI: 10.1101/gr.161968.113.

[56] Oleksandr M Maistrenko et al. “Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity”. en. In: *The ISME Journal* 14.5 (May 2020), pp. 1247–1259. ISSN: 1751-7362, 1751-7370. DOI: 10.1038/s41396-020-0600-z.

[57] Eric D. Crandall et al. “Importance of timely metadata curation to the global surveillance of genetic diversity”. en. In: *Conservation Biology* 37.4 (Aug. 2023), e14061. ISSN: 0888-8892, 1523-1739. DOI: 10.1111/cobi.14061.

[58] Aylin Caliskan, Seema Dangwal, and Thomas Dandekar. “Metadata integrity in bioinformatics: Bridging the gap between data and knowledge”. en. In: *Computational*

and Structural Biotechnology Journal 21 (2023), pp. 4895–4913. ISSN: 20010370. DOI: 10.1016/j.csbj.2023.10.006.

[59] Marnix H. Medema, Tristan De Rond, and Bradley S. Moore. "Mining genomes to illuminate the specialized chemistry of life". en. In: *Nature Reviews Genetics* 22.9 (Sept. 2021), pp. 553–571. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-021-00363-7.

[60] Kirstin Scherlach and Christian Hertweck. "Mining and unearthing hidden biosynthetic potential". en. In: *Nature Communications* 12.1 (June 2021), p. 3864. ISSN: 2041-1723. DOI: 10.1038/s41467-021-24133-5.

[61] Hilke Bruns et al. "Function-related replacement of bacterial siderophore pathways". en. In: *The ISME Journal* 12.2 (Feb. 2018), pp. 320–329. ISSN: 1751-7362, 1751-7370. DOI: 10.1038/ismej.2017.137.

[62] Zhaochao Wang et al. "The Deep Mining Era: Genomic, Metabolomic, and Integrative Approaches to Microbial Natural Products from 2018 to 2024". en. In: *Marine Drugs* 23.7 (June 2025), p. 261. ISSN: 1660-3397. DOI: 10.3390/md23070261.

[63] Zachary L Reitz and Marnix H Medema. "Genome mining strategies for metallophore discovery". en. In: *Current Opinion in Biotechnology* 77 (Oct. 2022). Publisher: Elsevier BV, p. 102757. ISSN: 0958-1669. DOI: 10.1016/j.copbio.2022.102757.

[64] Tilmann Weber and Hyun Uk Kim. "The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production". en. In: *Synthetic and Systems Biotechnology* 1.2 (June 2016), pp. 69–79. ISSN: 2405805X. DOI: 10.1016/j.synbio.2015.12.002.

[65] Justin J. J. Van Der Hooft et al. "Linking genomics and metabolomics to chart specialized metabolic diversity". en. In: *Chemical Society Reviews* 49.11 (2020), pp. 3297–3314. ISSN: 0306-0012, 1460-4744. DOI: 10.1039/D0CS00162G.