From sequence to molecules: Feature sequence-based genome mining uncovers the hidden diversity of bacterial siderophore pathways

Shaohua Gu[1,2#], Yuanzhe Shao[2#], Karoline Rehm[3], Laurent Bigler[3], Di Zhang[1], Ruolin He[1], Jiqi Shao[1], Alexandre Jousset[4], Ville-Petri Friman[5], Zhong Wei[4*], Rolf Kümmerli[6*], Zhiyuan Li[1,2*]

[1] Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China

[2] Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China

[3] University of Zurich, Department of Chemistry, Winterthurerstr. 190, 8057 Zurich, Switzerland

[4] Jiangsu Provincial Key Lab for Organic Solid Waste Utilization, Key lab of organic-based fertilizers of China, Nanjing Agricultural University, Nanjing, P R China

[5] University of Helsinki, Department of Microbiology, 00014, Helsinki, Finland

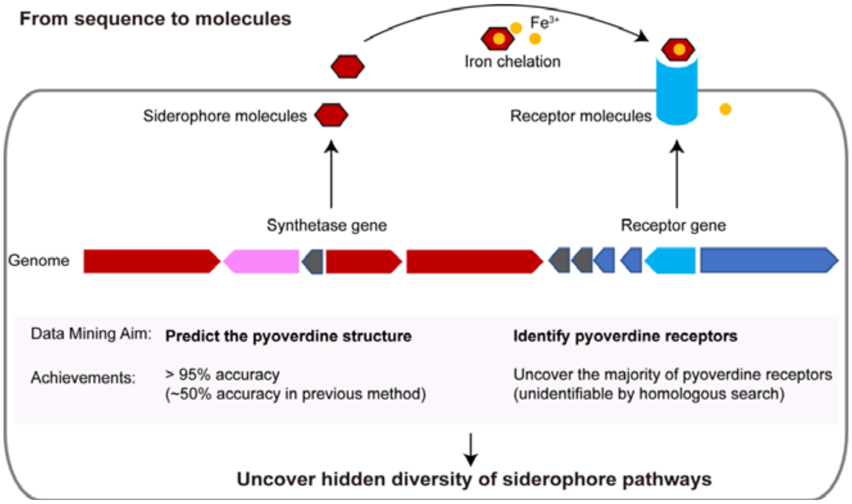
[6] University of Zurich, Department of Quantitative Biomedicine, Winterthurerstr. 190, 8057 Zurich, Switzerland

[#] These authors contributed equally to this article.

[*]Corresponding authors (email: weizhong@njau.edu.cn; rolf.kuemmerli@uzh.ch; zhiyuanli@pku.edu.cn)

Abstract

Microbial secondary metabolites have long been recognized as a rich source for pharmaceutical compound discovery and to have crucial ecological functions. However, the sequence-to-function mapping in microbial secondary metabolism pathways remains challenging because neither protein function nor substrate specificity can accurately be predicted from genome data. Here we focus on the iron-scavenging pyoverdines, siderophores of *Pseudomonas* bacteria, as model system to develop a knowledge-guided bioinformatic pipeline that extracts functional information of both the pyoverdine synthesis machinery and uptake receptors from 1928 draft genomes. For pyoverdine synthesis, our approach predicts the chemical structure of 188 different pyoverdines with nearly 100% accuracy. For pyoverdine uptake, our pipeline uncovers 94 different pyoverdine receptor groups. Our results demonstrate that combining feature sequence and phylogenetic approaches is a powerful way to reconstruct bacterial secondary metabolism pathways based on sequence data, unveiling an enormous yet overlooked diversity of siderophores and their receptors.


From sequence to molecules
Fe3+
Iron chelation
Siderophore molecules
Receptor molecules
Synthetase gene
Receptor gene
Genome
Data Mining Aim:
Predict the pyoverdine structure
Identify pyoverdine receptors
Achievements:
> 95% accuracy
(~50% accuracy in previous method)
Uncover the majority of pyoverdine receptors
(unidentifiable by homologous search)
Uncover hidden diversity of siderophore pathways

Introduction

Rapid advancements in sequencing technologies have revolutionized our view on microbial communities. While amplicon sequencing provides information on community composition and diversity, shotgun and whole genome sequencing allow us to reliably anticipate evolutionary and ecological relationships between microbes and to obtain functional information on communities. Computational models assessing the metabolic capacity of individual members, or an entire consortium, have become very popular and powerful[1-3]. The major focus of such modelling approaches is typically on the primary metabolism of bacteria, as genes involved in core metabolic pathways are highly conserved and can be identified with relative ease[2,4]. Conversely, analysis of the secondary metabolites has attracted less attention, even though they include compounds such as antibiotics, toxins, siderophores, biosurfactants, all known to have important implications for community assembly[5,6] and to be important sources for pharmaceutical discoveries[7,8 9,10].

There are multiple challenges that currently prevent a detailed unravelling of secondary metabolism of bacteria based on genome data[5,11]. First, most secondary metabolites are produced by pathways comprised of modular enzymes such as non-ribosomal peptide synthetases (NRPSs) or polyketide synthases (PKS)[12,13]. Locating complete synthesis clusters and identifying all enzyme-encoding genes is challenging from highly fragmented metagenomic sequences or draft genomes with a high number of contigs. Second, functional predictions for coding regions within a cluster rely on homologous comparisons with experimentally characterized genes. Such information is often restricted to a limited number of model organisms, meaning that only a small portion of the existing secondary metabolism pathways is covered by current data bases. Finally, given the complex multi-modular synthesis machineries, it is challenging to precisely predict the secondary metabolites produced even with accurately annotated NRPS or PKS clusters. The main challenge is that a large pool of non-proteinogenic amino acids is used as substrates and the specificity of an enzyme's A domain, connecting these unusual amino acids, is often poorly understood[14]. As a result, new computational methods are needed to accurately reconstruct bacterial secondary metabolism from sequence data.

Here, we present a new bioinformatic pipeline that overcomes these challenges. We specifically focus on a particular class of secondary metabolites (iron-scavenging siderophores) as a case study to develop a bioinformatic workflow that predicts the chemical structure of the produced metabolites with near 100% accuracy. Our pipeline is based on improved gene annotation combined with a phylogeny- and feature sequence-based substrate prediction techniques (Figure 1). In comparison with the currently available databases and bioinformatic tools[14-16], the main advancement of our method is the more accurate prediction of synthesized products based on NRPS clusters identified in genome data.

Among siderophores, we focus on NPRS machineries that are responsible for the synthesis of pyoverdines, a class of chemically diverse siderophores with high iron affinity, produced by *Pseudomonas* bacteria[17,18]. While each *Pseudomonas* strain produces a single type of pyoverdine, an enormous structural diversity has been described across strains and species[19-22]. Pyoverdine types differ in their peptide backbone, meaning that the diversity should be mirrored in NPRS enzyme diversity and their selectivity for the different amino acid substrates[18]. Based on this knowledge, our pipeline entails the following steps (Figure 1): (i) identification of the complete sequences of pyoverdine synthetase genes from fragmented draft genomes, (ii) building the pyoverdine synthesis machinery *in silico* by extracting the feature sequences for substrate specificity from motif-standardized NRPSs, and (iii) predicting the precise chemical structure of pyoverdines followed by empirical verification.

An additional element of iron metabolism is that when siderophores are secreted and bound to iron, bacteria rely on a specific receptor for their uptake into the cell. Pyoverdine receptors are annotated as FpvA and it is known that receptor diversity matches pyoverdine diversity[22,23]. Moreover, FpvA belongs to the family of TonB-dependent receptors and a single *Pseudomonas* species often has many gene copies encoding these receptors. This poses an additional bioinformatic challenge: how to find the gene encoding the specific pyoverdine receptor among several potential receptor genes? To overcome this, we develop an algorithm that focuses on sequence regions involved in pyoverdine recognition and translocation across the outer membrane with supervised learning methods that locate the *fpvA* genes in the fragmented genomes based on these regions (Figure 1). Altogether, our bioinformatic pipeline uses knowledge-guided insights empowered by supervised learning to construct a first systematic sequence-to-function mapping of a family of secondary metabolites (pyoverdine) and their corresponding receptors. Our analysis unveils a yet unrecognized extraordinary diversity of iron-scavenging machineries in pseudomonads.

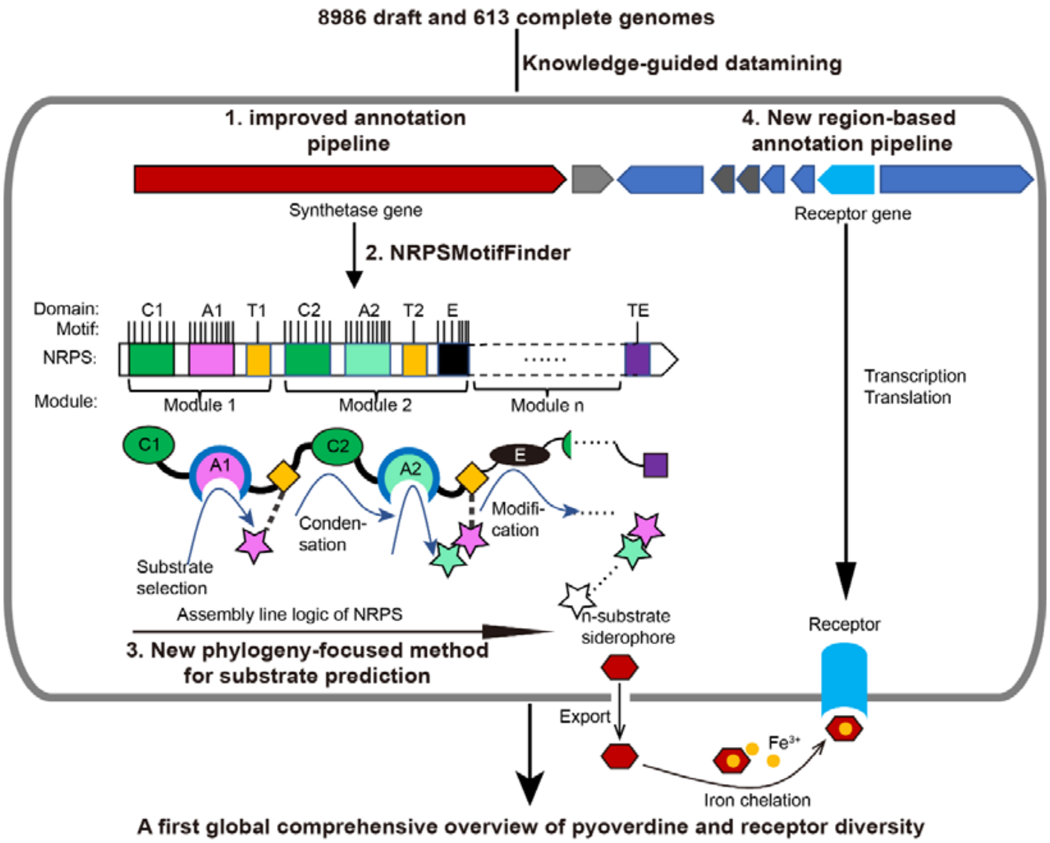

8986 draft and 613 complete genomes
Knowledge-guided datamining
1. improved annotation pipeline
4. New region-based annotation pipeline
Synthetase gene
Receptor gene
2. NRPSMotifFinder
Domain:
Motif:
NRPS:
Module:
C1
A1
T1
C2
A2
T2
E
TE
Module 1
Module 2
Module n
Transcription
Translation
C1
A1
C2
A2
E
Condensation
Modification
Substrate selection
Assembly line logic of NRPS
n-substrate siderophore
3. New phylogeny-focused method for substrate prediction
Receptor
Export
Fe3+
Iron chelation
A first global comprehensive overview of pyoverdine and receptor diversity

Figure 1 Scheme depicting our new genome mining pipeline to precisely predict the synthesis, the molecular structure and the uptake machinery of pyoverdines, a family of iron-scavenging siderophores produced by members of the *Pseudomonas* genus. The grey rounded outer rectangle represents a bacterial cell. The red and blue arrow-shaped boxes stand for the synthetase and receptor genes for pyoverdines, respectively. Synthetase genes are transcribed and translated to form the *n*-modular NRPS enzymes. These enzymes synthesize the peptide backbone of pyoverdine through an assembly line using their repeating module units, with the A domain being responsible for substrate selection and the E domain for chirality. The *n*-substrate siderophores are then exported to the extracellular space for iron chelation. Membrane-embedded TonB-dependent receptors recognize the ferri-siderophore complex and import it into the cell. Bold black text and black arrows describe our multi-step computational methods developed to reconstruct the entire

process from genome sequence data. First, the annotation pipeline was improved (from antiSMASH) to extract the complete sequence of pyoverdine synthetase genes from draft genomes. Second, NRPSMotifFinder was used to define A- and E-domains and to determine the exact motif-intermotif structure of the pyoverdine assembly line. Third, intermotif regions most indicative of substrate specificity were used to develop a phylogeny-focused method for precise product prediction. Fourth, a sequence-region-based annotation method was combined with genome architecture features to identify the FpvA, receptors responsible for ferri-pyoverdine import.

Results

Section 1: Improved annotation pipeline reveals a vast reservoir of pyoverdine synthetase genes

The first step of our bioinformatic pipeline was to improve the annotation of pyoverdine synthetase genes. The pyoverdine molecules is composed of a conserved fluorescent chromophore (Flu) and a peptide chain (Pep), which are both synthesized by NRPS enzymes[24]. There are already existing tools, such as antiSMASH, that can find and annotate NRPS clusters in microbial genomes[25]. However, antiSMASH (and other popular annotation platforms[15,16]) rely on accurate gene predictions, which are typically problematic for fragmented genomes. Consequently, while antiSMASH can recognize and annotate certain genes of an NRPS cluster, the precise reconstruction of a complete NRPS assembly line often fails. This is particularly problematic because most available genomes are drafts and any analysis suffers from the unavoidable issue of incomplete or misannotation of gene fragments.

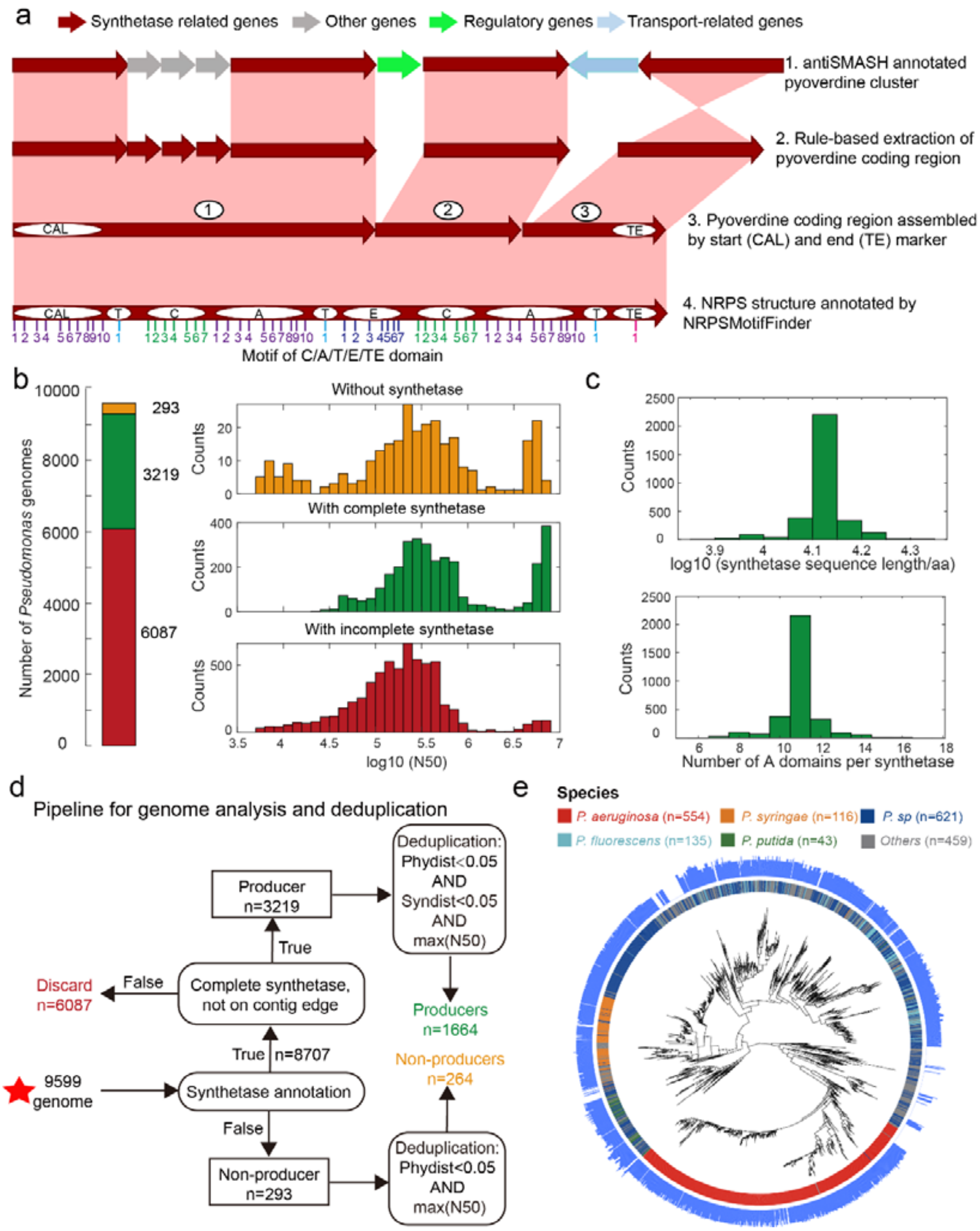


Figure 2 Improved annotation pipeline reveals a vast diversity of pyoverdine synthetase genes. a. Improved annotation pipeline based on the raw annotation from antiSMASH. **b.** The annotation pipeline was applied to 9599 *Pseudomonas* genomes (94% draft genomes). Genomes could be separated into three categories. Yellow: genomes without pyoverdine cluster. Green: genomes

with a complete pyoverdine cluster. Red: genomes with incomplete pyoverdine synthetase cluster. The red category involved genomes with truly incomplete clusters (lacking Flu or Pep synthetic genes) or genomes with likely truncated synthetic genes at the edge of contigs. **c.** Distributions of the sequence length (upper panel) and the number of A domains (lower panel) across all the genomes with a complete synthetase cluster. **d.** Workflow applied to separate the 9599 *Pseudomonas* genomes into the three categories described in b and removing of redundant genomes with high phylogenic similarity and showing high similarity in pyoverdine synthetases. Red star indicates the start of the workflow. **e.** Phylogenetic tree depicting the relationship among the 1928 non-redundant *Pseudomonas* strains (1664 producers and 264 non-producers) based on the concatenated alignment of 400 single-copy conserved genes in their genomes. The inner ring depicts the taxonomical classification including the four most prevalent species. The outer ring shows the number of A domains present in the pyoverdine synthetase assembly line in each strain.

To overcome these issues, we developed an improved four-step annotation pipeline starting with the raw annotation of the pyoverdine cluster obtained from antiSMASH (Figure 2a). First, we implemented a NRPS Hidden Markov Model (HMM) to re-annotate and extract the entire nucleotide sequence of the pyoverdine synthetase cluster[26], including the genes missed by antiSMASH. For this step, the nucleotide sequences were converted into amino acid sequences to avoid erroneous gene predictions typically associated with antiSMASH. Second, we assembled the entire re-annotated pyoverdine coding region into a single sequence with a defined start (CAL) and/or end (TE) markers. Third, we used NRPSMotifFinder to identify the C, A, T, E and TE motifs that are characteristic for the NRPS structure of pyoverdine[14]. Finally, we applied a safety measure to ensure that the recovered

NRPS assembly line is complete (contains both Flu and Pep) and is not truncated, which can occur when a synthetase coding region is at the edge of a contig. Consequently, we dismissed all pyoverdine synthesis clusters located within 100 bp proximity to contigs' edges and either lacked Flu or Pep synthetic genes.

Next, we applied our improved pyoverdine synthesis annotation pipeline to 9599 *Pseudomonas* genomes (including 613 complete and 8986 draft genomes) retrieved from the Pseudomonas Genome Database[27]. We found the pyoverdine synthesis machinery in 97% of the genomes (Figure 2b), indicating that the machinery is ubiquitous in *Pseudomonas*. However, since 94% of the analyzed genomes were in draft form, the pyoverdine synthesis machinery was likely truncated (i.e., on the edge of the contig) in 63.4% (6087) of the genomes. These genomes were excluded from further analysis. Around 3.1% of retained genomes (293) with high assembly completeness were missing pyoverdine synthetic genes, indicating that these *Pseudomonas* strains were not able to produce pyoverdine ('non-producers'). The rest of the genomes (33.5%; 3219 genomes) were classified as 'producers' with complete pyoverdine NRPS assembly lines that meet all our quality controls. For these 3219 genomes, we used NRPSMotifFinder to find boundaries between the various synthesis domains and to determine amino acid length and the number of A domains. The lengths of pyoverdine synthetic genes ranged between 7690 and 21333 amino acids, and the number of A domains per synthetase ranged between 6 and 17, with a total of 35,281 A domains being present across all strains (Figure 2c and Figure S1). Overall, our analysis pipeline unveiled a vast diversity of pyoverdine synthetase that goes far beyond of what has previously been described in the literature.

Finally, we conducted a phylogenetic analysis based on 400 conserved genes with the 293 non-producers and the 3219 producers. We first removed redundant non-producers by retaining the most integrative genome among strains with high phylogenic similarity. Then, we removed redundant producers by retaining the most integrative genome among strains with high phylogenic and pyoverdine synthetase similarity (Figure 2d). This data cleaning yielded a total of 1928 *Pseudomonas* strains (403 complete and 1525 incomplete genomes), segregating into 1664 pyoverdine producers and 264 non-producers. The phylogenetic tree revealed that all major *Pseudomonas* species clades were present in our data set (Figure 2e). Moreover, the number of A domains varied widely among species and even between strains within species. For example, the number of *Pseudomonas aeruginosa* A domains ranges between 7 and 14. In summary, by improving the synthetase annotation method, we successfully obtained 1664 highly reliable pyoverdine synthetases (with a total of 18,292 A domains) and 264 non-producers.

Section 2: Phylogeny-focused substrate prediction for pyoverdine A domains

Our next goal was to precisely predict the molecular structure of the pyoverdines produced by the 1664 strains with complete synthetase gene clusters. The first essential step towards this goal was to reliably predict the substrate selectivity of all A domains in the NRPS assembly line. The A domain of each module selects for a single substrate among 22 proteinogenic and hundreds of non-proteinogenic amino acids[28,29]. Moreover, whenever an E domain exists downstream of an A domain, the chirality of the amino acid incorporated into the peptide chain gets modified from L to D. Thus, the modularity combined with the selectivity of A domains can promote an enormous diversity of pyoverdine molecule structures. To date, 73 pyoverdine structures have been reported (Supplementary_table1) out of which 13 have their synthetase genes sequenced (Supplementary_table2). In order to make reliable predictions, two challenges must be addressed: (i) the extraction of relevant information from A domain sequences for which the substrate is known, and (ii) the effective application of this information to predict specificity of A domain sequences for which the substrate is unknown.

To address the first challenge, we built our analysis on the NRPS assembly lines of the known 13 pyoverdines to extract relevant information from the A domain sequences. From this dataset, we could identify 101 A domains that could be experimentally linked to 13 amino acid substrates (Supplementary_table3). We next performed multisequence alignment of the 101 A domains to determine the “feature sequence distance”, which is the most informative for the substrate selectivity. To this end, we tested three different A domain regions, three different sequence similarity measurements and seven different clustering methods for their predictive power (Figure 3a). We found that the full A domain sequence is not informative for substrate prediction (Figure 3b, left panel). Instead, our analysis indicated that information-rich positions start with motif A4 and end before motif A5, consistent with the known role of the A domain pocket in substrate selectivity[14]. Overall, the sequence region from motifs A4 to A5 (termed “Amotif4-5”), in conjunction with Jukes-Cantor distance and Ward linkage clustering, performed best in accurately distinguishing between different substrates and maintaining homogeneity for identical substrates (Figure 3b).

To address the second challenge, we developed a “phylogeny-focused method” to apply the feature sequence distance derived in the preceding paragraph to the 18,292 discovered A domains. We realized that a direct construction of a phylogenetic tree including all 18,292 query A domains and the 101 reference A domains would be computationally too demanding and impossible to scale up. Furthermore, such an approach would result in phylogeny-interference issues, where domains would cluster not only based on their substrate similarities but also based on overall species relatedness[14]. To minimize the effect of phylogeny and speed up calculation, we took each of the 18,292 query A domains and identified the two most similar A domain clusters within the 101 reference A domain set. We then compared the feature distance between each query A domain and the two most similar reference A domains in different clusters and assigned the query A domain to a substrate specificity using the following rules (Figure 3c). (1) If the feature distance is below the 0.7 threshold (corresponding to 50% identity) for only one of the two reference A domains, then the substrate of the query A domain is matched to the substrate of the more similar (lower distance) reference A domain. (2a) If the feature distance is below the 0.7 threshold for both reference A domains, then we considered the relative difference of the query A domain towards the two reference A domains. If the relative difference is larger than 0.2, the query A domain is matched to the substrate of the more similar reference A domain. (2b) If the relative difference is smaller than 0.2, the substrate of the query A domain cannot unambiguously be determined and is thus matched with both reference substrates. (3) If the feature distance is above the 0.7 threshold (below 50% identity) for both reference A domains, then the substrate of the query A domain is marked as “unknown”. For most query A domains, rule (1) could be applied (17880 cases), whereas rules (2) and (3) had to be used rarely (133 and 279 cases, respectively). We applied our methodology termed “phylogeny-focused method” to all following substrate and pyoverdine structure predictions.

Section 3: Experimental validation of the annotation and prediction pipeline

We tested whether our bioinformatic pipeline can reliably predict the structure of a set of yet uncharacterized pyoverdines. To achieve this objective, we selected 20 *Pseudomonas* strains, all known to produce pyoverdines, from a natural strain collection that was previously isolated from soil and water[30]. We sequenced their genomes and subsequently applied our annotation and prediction pipeline to generate predicted pyoverdine structures for all 20 strains harboring a total of 237 A domains. Then, we elucidated the chemical structure of the 20 pyoverdines using culture-based methods combined with UHPLC-HR-MS/MS[31]. We found a near-perfect match (96.2%) between the predicted and the observed pyoverdine chemical structures and were able to accurately assign amino acids in 228 out of 237 cases (Figure 3d). Our method demonstrated a substantial improvement comparing to the prediction accuracy of AntiSMASH in pyoverdines (46.0%), which could accurately assign correct amino acids only in 109 out of 237 cases (Supplementary_table4). The nine non-matching cases of our method segregated into three groups. In three cases (1.3%), our algorithm could not distinguish between the substrates Lysine and Ornithine, as these two amino acids are highly similar both in terms of their chemical structures and corresponding A domain sequences. This is the only sensitivity issue that is associated with our approach. In four cases (1.7%), our technique assigned an “unknown” substrate to amino acids that turned out to be valine, citrulline and histidine. Indeed, these three amino acids have not been reported in pyoverdines before and are therefore not yet present in the reference dataset. These cases show that our analysis pipeline can be used to identify new substrates. Once experimentally verified, the new A domains and their substrates can expand the reference dataset, allowing targeted improvement of our phylogeny-focused prediction technique. Finally, there were only two cases (0.8%) that represented true mismatches between observed and predicted amino acids. Altogether, our phylogeny-focused method is highly accurate in predicting pyoverdine peptide structures and in identifying unknown substrates in *Pseudomonas*.

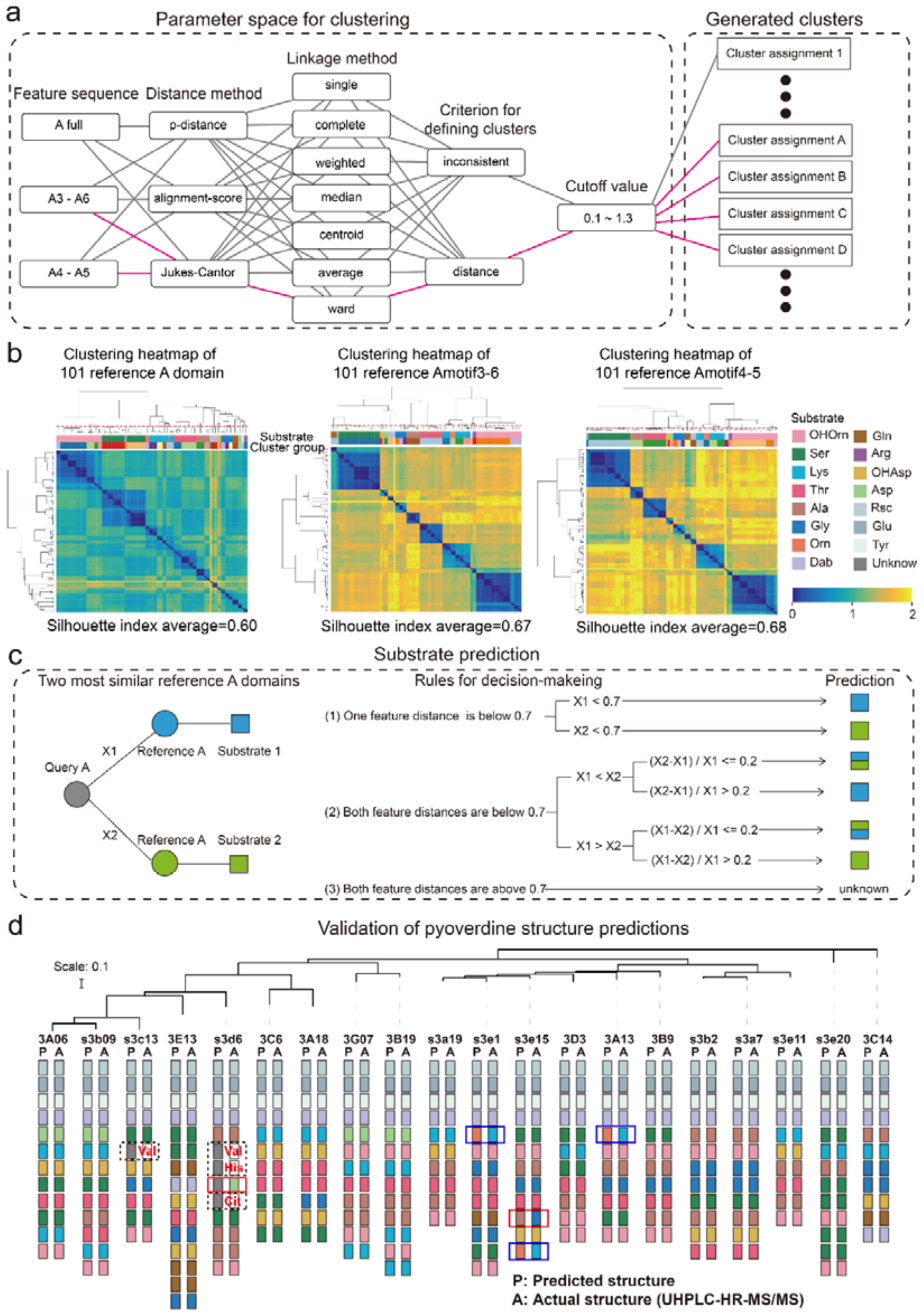


Figure 3 Phylogeny-focused substrate prediction for pyoverdine synthetase assembly lines. **a.** Information from 101 reference A domains with known amino acid substrates were used to develop an algorithm that predicts substrates from A domain sequence data with high accuracy. The

challenge is to group the variable A domains into clusters that predict the same substrate (captured by the silhouette index). To find the most distinctive algorithm, we combined different feature sequences of A domains (Amotif) with different distance and linkage methods in our hierarchical clustering analyses. The best performing path is shown in pink. **b.** Heatmap showing the hierarchically clustered distances of the 101 reference A domains as a function of the feature sequence used. Left panel: complete A domain sequences. Middle panel: Amotif3-6 sequences. Right panel: Amotif4-5 sequences. The experimentally validated substrates are shown on top of the heatmaps. The heatmaps show that hierarchical clustering, reliably associating sequence distances with substrate, worked best with the Amotif4-5. **c.** Phylogeny-focused substrate prediction pipeline for query A domains (grey circle) based on Amotif4-5 feature sequence comparisons. X1 and X2 represent the feature distance between the query A domain and two closest reference A domains (blue and green circles), respectively. Three rules are used, based on the feature distances X1 and X2 and a threshold value of 0.7 (50% similarity), to make substrate predictions for the query A domain. There are three possible outcomes: unambiguous substrate prediction (blue or green squares), ambiguous substrate prediction (dual-colored squares), and no prediction (“unknown). **d.** Phylogenetic tree of 20 *Pseudomonas* strains and visualization of their predicted and actual pyoverdine structures to validate our phylogeny-focused substrate prediction pipeline. 228 out of the 237 substrates (96.2%) were correctly predicted. The nine inconsistencies are boxed in blue (Lysine and Ornithine are indistinguishable), in dashed black (correct detection of “unknown” substrates), and in red (true mismatches). Note that our prediction pipeline (as any other pipeline) cannot distinguish between modified variants of the same amino acid.

Section 4: Application of the annotation and prediction pipelines to a full dataset

After successful validation, we applied our bioinformatic pipeline to the 1664 complete NRPS assembly lines annotated in our genome analysis (Figure 2). Across all assembly lines, we were able to predict the substrates of 17,880 A domains (97.75%) without ambiguity, whereas 133 A domains (0.73%) were associated with two different substrates, and 279 A domains (1.52%) predicted an unknown substrate (similar to the case of valine above). After considering the presence/absence of an E domain in each module, we derived the structure of 1664 pyoverdines according to method at section 2 (Figure 4). Our prediction yielded 188 different pyoverdine molecules, out of which only 37 structures had been previously reported. However, these 37 reported structures were highly abundant across strains (1103 out of 1664). Agreeing with previous studies, we observed that the fluorophore is highly conserved among the 188 predicted structures. Moreover, our analysis confirmed that 13 amino acid substrates form the core of all the 188 pyoverdine structures, with most of the variation being attributable to different substrate combinations, peptide lengths, and substrate chirality (Figure 4). In addition, the 279 unknown substrates will significantly increase the repertoire of pyoverdine amino acids if could be characterized by future experiments, despite that they were much rarer than the 13 main substrates. Notably, pyoverdine structural diversity was not strongly linked to phylogeny because the same pyoverdine structure could be found in completely unrelated species, while closely related species often had different pyoverdine structures (Figure 4). These observations suggest that there may be both frequent recombination and horizontal gene transfer of pyoverdine synthetase clusters between species. Taken together, the bioinformatics methods developed in our study can predict a suit of secondary metabolites (pyoverdines) from sequence data with high accuracy, revealing an unprecedented richness and evolutionary history of siderophores within pseudomonads and the discovery of 151 putative novel pyoverdine variants.

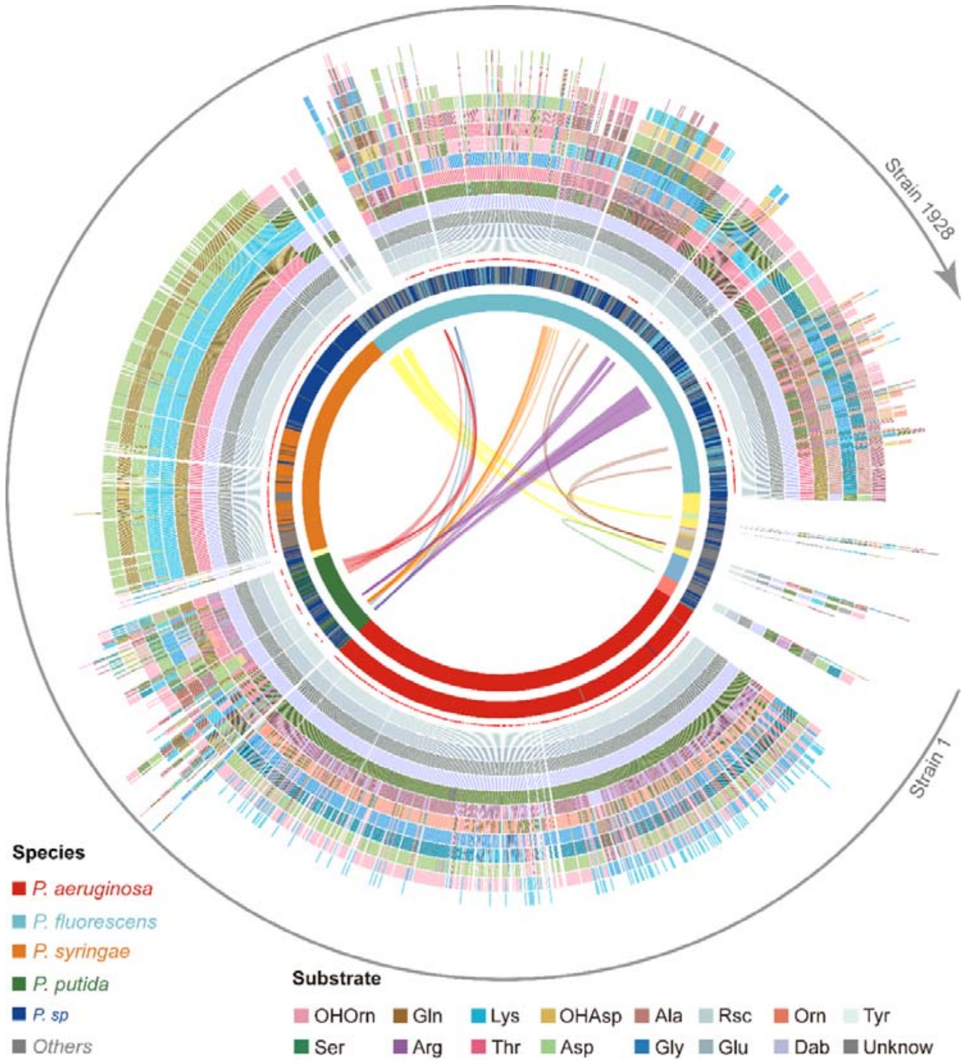


Figure 4 Predicted pyoverdine structural diversity based on our developed algorithm mapped onto the phylogenetic tree comprising all 1928 (non-redundant) *Pseudomonas* strains. The stacked boxes in the outermost circle show the predicted structure of pyoverdines, whereby each color represents a specific amino acid substrate. Strains without boxes represent non-producers (n = 264). Boxes with two colors indicate cases of ambiguous (dual) substrate prediction. The red dots at the basis of the stacked boxes indicate experimentally validated pyoverdine structures. The inner circle shows the taxonomic species classification following Figure 2e. Because the allocation of strains to species names is often imprecise, we divided the 1928 strains by their

phylogenetic distance into 18 clades (color shadings in inner-most circle), out of which 13 contained more than 1 strain. Lines within the inner-most circle link strains from different clades that share the same pyoverdine structures, whereby line colors represent the shared unique pyoverdines. The bending of the lines represents the phylogenetic sequence distances of the connected strain pairs.

Section 5: Development of a region-based identification method for annotation of the FpvA receptors

In pseudomonads, iron-loaded pyoverdines are recognized by FpvA, a TonB-dependent receptor, that transports the ferri-siderophore into the periplasm[19,32,33] . The protein structure of characterized FpvA variants consists of three domains: The Secretin and TonB N-terminus short domain (STN), the Plug domain (Plug), and the TonB dependent receptor domain (TonB)[19]. While these domains are conserved across FpvA variants and other siderophore receptors, there is substantial variation at the sequence level. This makes it challenging to reliably identify FpvA receptors from sequence data by homologous search. As an example, we were unable to find FpvA genes (with a 60% identity threshold) by homologous search in several genomes although they had complete pyoverdine synthesis machineries. Moreover, there are many other TonB-dependent receptors with fairly high sequence identity to FpvA but that transport other siderophores than pyoverdine (e.g. FpvB, 55% identity, transporting pyoverdine, ferrichrome and ferrioxamine B[34]. Therefore, it is imperative to develop a new comprehensive method for identifying FpvA receptors in *Pseudomonas* genomes.

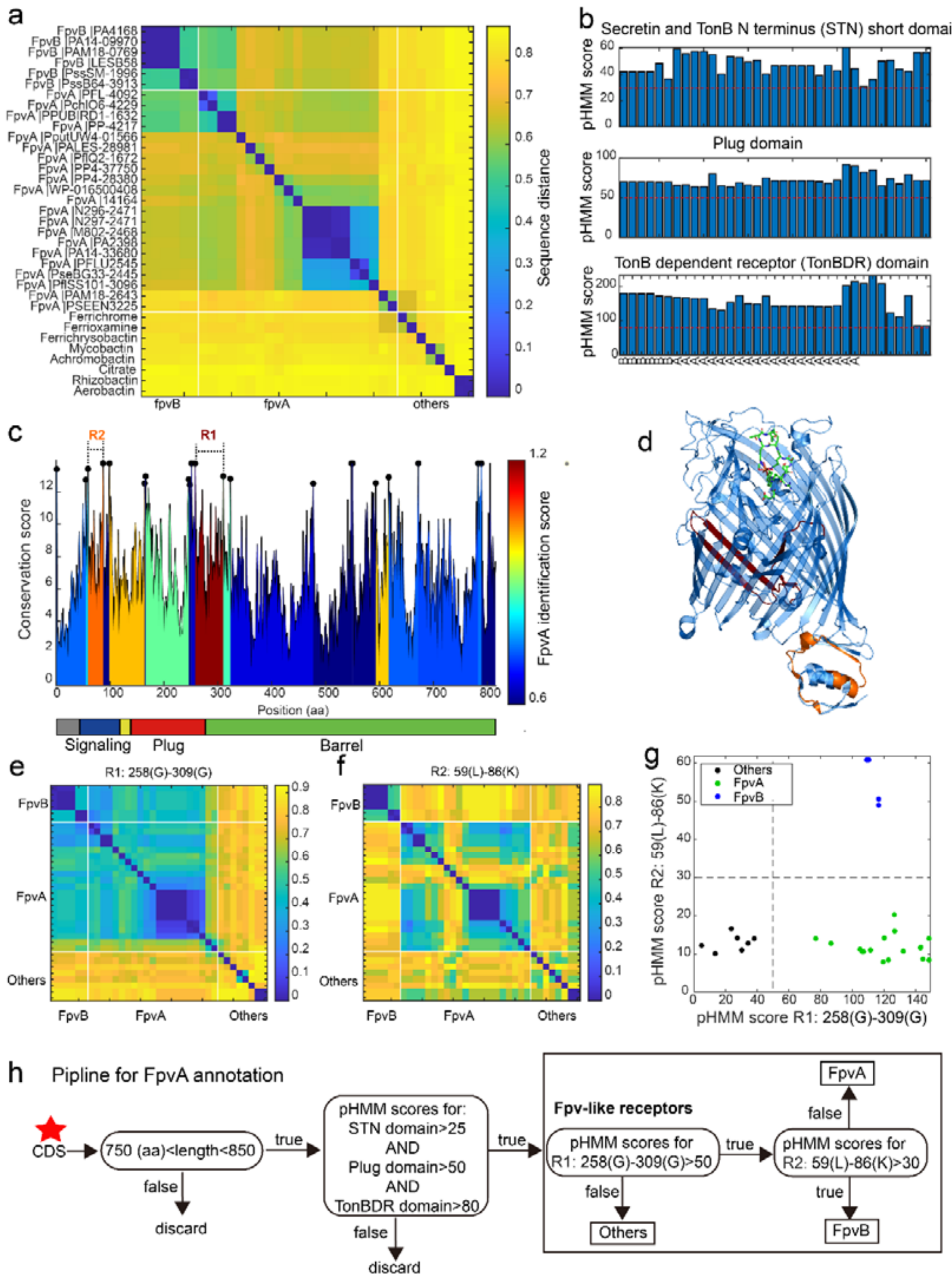


Figure 5 A sequence-region-based identification pipeline for annotating FpvA receptors.

a. Heatmap displaying the hierarchically clustered sequence distances (p-distance calculation method, identity (%) = (1-sequence distance) * 100) of 35 reference siderophore receptors identified in *Pseudomonas* spp., based on full sequences. No clear discrimination between FpvA, FpvB and other receptors is possible. The order of receptors is consistent across panels (**b**), (**e**), and (**f**). **b.** The pHMM scores of the three standard receptor domains (STN, Plug, and TonBDR) vary across the 35 reference sequences (A: FpvA, B: FpvB and NA: others), but do not allow to distinguish between receptor groups. **c.** FpvA region-based conservation scores from a multi-alignment of the 35 reference sequences mapped to the FpvA sequence of strain *P. aeruginosa* PAO1. All residues within the top 10% of the conservation score denoted with black dots. For each region flanked by two black dots, we calculated the FpvA identification score (heatmap), representing the ability to distinguish FpvA from non-FpvA receptors. **d.** Mapping of the two regions with the highest FpvA identification scores R1(dark red) and R2 (orange) to the crystal structure of FpvA from PAO1 conjugated with pyoverdine (PDB 2IAH). **e.** Heatmap showing the hierarchically clustered sequence distances of 35 reference siderophore receptors based on the R1 sequence region. A clear discrimination between FpvA/FpvB and other receptors emerges. **f.** Heatmap showing the hierarchically clustered sequence distances of 35 reference siderophore receptors based on the R2 sequence region. A clear discrimination between FpvA and FpvB receptors emerges. **e.** The pHMM scores of regions R1 and R2 for the 35 siderophore reference receptors are contrasted against each other, yielding a clear separation between FpvA, FpvB and other receptors. Dashed lines indicate the pHMM threshold scores used for later analysis. **f.** Flowchart showing all steps involved in the FpvA annotation from genome sequence data. The red star indicates the start of the workflow.

We started our approach by comparing the sequences of 35 reported siderophore receptors, including 21 FpvA, 6 FpvB, and 8 TonB-dependent siderophore receptor sequences often found in *Pseudomonas* genomes, encoding receptors for the uptake of heterologous siderophores (Supplementary_table5). We found that all receptor sequences share a similar length of around 800 amino acids (FpvA and FpvB sequences: 809 ± 10 amino acids). We then used the complete sequences to calculate the pair-wise distances by global alignment before applying hierarchical clustering (Figure 5a). We found substantial divergence between FpvA variants to an extent that was comparable to the distance between FpvA and other siderophore receptors. Moreover, FpvB variants clustered with FpvA variants, showing that FpvA identification based on full sequence distances is unachievable. We hence focused on the three typical receptor domains (TonB, Plug, and STN, retrieved from the Pfam database) and applied Profile Hidden Markov Models (pHMM) to calculate the pHMM probability scores for each domain and reference sequence. The probability scores (calculated as the log-odd ratios for emission probabilities and log probabilities for state transitions) had reasonably high scores but no distinction was apparent between the three receptor classes (Figure 5b).

We next asked whether there are specific regions within the receptor sequences that are characteristic of FpvA. To address this, we conducted a multiple sequence alignment (MSA) with all 35 reference receptor sequences and mapped them onto the sequence of the well-characterized FpvA of *P. aeruginosa* PAO1 (Figure 5c). MSA allows to identify conserved sites (Figure 5c, black dots representing the top 10% most conserved sites) that are shared by the majority of the reference sequences. We then used these conserved sites to partition the MSA into variable regions which were flanked by two conserved sites. For each variable region, we assessed its predictive power to differentiate FpvA from non-FpvA sequences. For this we defined the "FpvA identification score" analogous to the intercluster-vs-intracluster Calinski-Harabasz variance ratio, as

$$I_{\mathrm{FpvA}} = d_{\mathrm{A:non}}/d_{\mathrm{A:A}}$$

where $d_{\mathrm{A:A}}$ is the sequence distance among all 21 FpvA sequences, and $d_{\mathrm{A:non}}$ is the sequence distance between all 21 FpvA and the 14 non-FpvA sequences.

Our analysis yielded two locations with noticeably high FpvA identification scores (Figure 5c). The region with the highest FpvA identification score (referred to as R1) locates at the intersection of the Plug domain and the barrel structure of the TonB domain (Figure 5d, between 258 Gly and 309 Gly in the PAO1 FpvA). According to the sequence distance matrix, the R1 region allows to distinguish heterologous siderophore receptors from FpvA and FpvB receptors (Figure 5e). The region with the second highest FpvA identification score (referred to as R2) was located in the C-terminal signaling domain (Figure 5d, between 59 Leu and 86 Lys in the PAO1 FpvA). The sequence distance matrix revealed that R2 allows to distinguish FpvB from FpvA receptors (Figure 5f).

We then constructed two pHMM by (i) the alignment of the 21 FpvA sequences in the R1 region, termed R1(FpvA), and (ii) the alignment of the 6 FpvB sequences in the R2 region, termed R2(FpvB). Running R1(FpvA) and R2(FpvB) against all 35 reference sequences revealed a clear separation between the three receptor categories (Figure 5g). Along the R1(FpvA) axis, FpvA and FpvB reference sequences have high R1 scores (minimal score 77.0) that separate them from other siderophore receptors (maximal score 38.1), whereas FpvAs references have substantially lower R2 scores (maximal 20.2) than FpvBs (minimal 49.0) along the R2(FpvB) axis.

Based on these insights, we developed a decision flow chart for annotating FpvAs in *Pseudomonas* genomes (Figure 5h): First, we considered sequences as Fpv-like receptors that share similar properties to the ones identified in our reference database. Particularly, protein coding sequence (CDS) length has to be between 750 and 850 amino acids and the phMM scores for the three typical receptor domains STN, Plug, and TonB have to be greater than 25, 50, and 80, respectively (Figure 5b, red dashed lines). Second, we used the pHMM threshold scores obtained for R1(FpvA) and R2(FpvB) (Figure 5g) to differentiate other siderophore receptors (R1(FpvA) score < 50) from FpvB receptors (R1(FpvA) score > 50 and R2(FpvB) score > 30) and FpvA receptors (R1(FpvA) score > 50 and R2(FpvB) score < 30). Our method effectively identifies FpvA receptors from sequence data and can be readily applied to the entire *Pseudomonas* dataset.

Section 6: Application of the receptor annotation pipeline to the full dataset.

The region-based receptor identification pipeline was applied to all 1928 *Pseudomonas* genomes. The analysis identified 4547 FpvAs, 615 FpvBs, and 9139 other TonB-dependent Fpv-like receptors across the dataset (Figure 6a). The 4547 FpvA sequences clustered hierarchically into 114 groups, defined by an identity threshold of 60%. When comparing to the 21 reference FpvAs (Figure 6b), we found that 2293 FpvA sequences have close homologues in the reference data base, while 2254 FpvA sequences lack such close homologues (sequence identity < 50%). These latter sequences, termed as "dissimilar to reference", may represent novel subtypes of FpvA receptors that could not be found by simple homology search. Our analysis further shows that many strains have more than one FpvA receptor.

We then asked whether the 4547 FpvAs are found in proximity of pyoverdine Pep synthetase genes as it is commonly the case for cognate FpvA receptors [35]. We thus calculated the proximity between pyoverdine Pep genes and the Fpv-like receptor genes by counting the number of base pairs between the two coding regions. All TonB-dependent receptors that have not been classified as FpvAs were more than 20 kb away from the Pep genes (Figure 6c). In contrast, 92% of the nearest FpvA genes were indeed located within 20 kb of their pyoverdine Pep genes (Figure 6d, called proximate receptors). These proximate receptors encompassed both those with close (66%) and more distant (34%) resemblance to the reference receptor types. Overall, this proximity analysis confirmed that our region-based gene identification method can reliably identify FpvA receptors.

We next explored the diversity among FpvA receptors in more detail by focusing on the 1534 strains that had proximate-receptors within 20 kb of the pyoverdine Pep genes (Figure 6d) and using high-confidence FpvAs for sequence feature extraction. When considering the whole gene sequences, these receptors segregated into 44 groups according to single-linkage clustering with an identity threshold of 60% (Figure S2a). To investigate which sequence regions were the most informative for reliable clustering, we used a similar approach as with FpvAs detection by quantifying the "group identification score" for variable regions flanked by highly conserved sites. The higher the score, the stronger a region's capacity to discriminate between FpvA groups. We found that the four regions with the top discrimination capacities all located near the Plug domain surrounding the pyoverdine transmission channel (Figure 6e). The plug domain is known to undergo conformational changes and is involved in pyoverdine selectivity and import[36,37], suggesting that the four high-score regions are responsible for pyoverdine specificity.

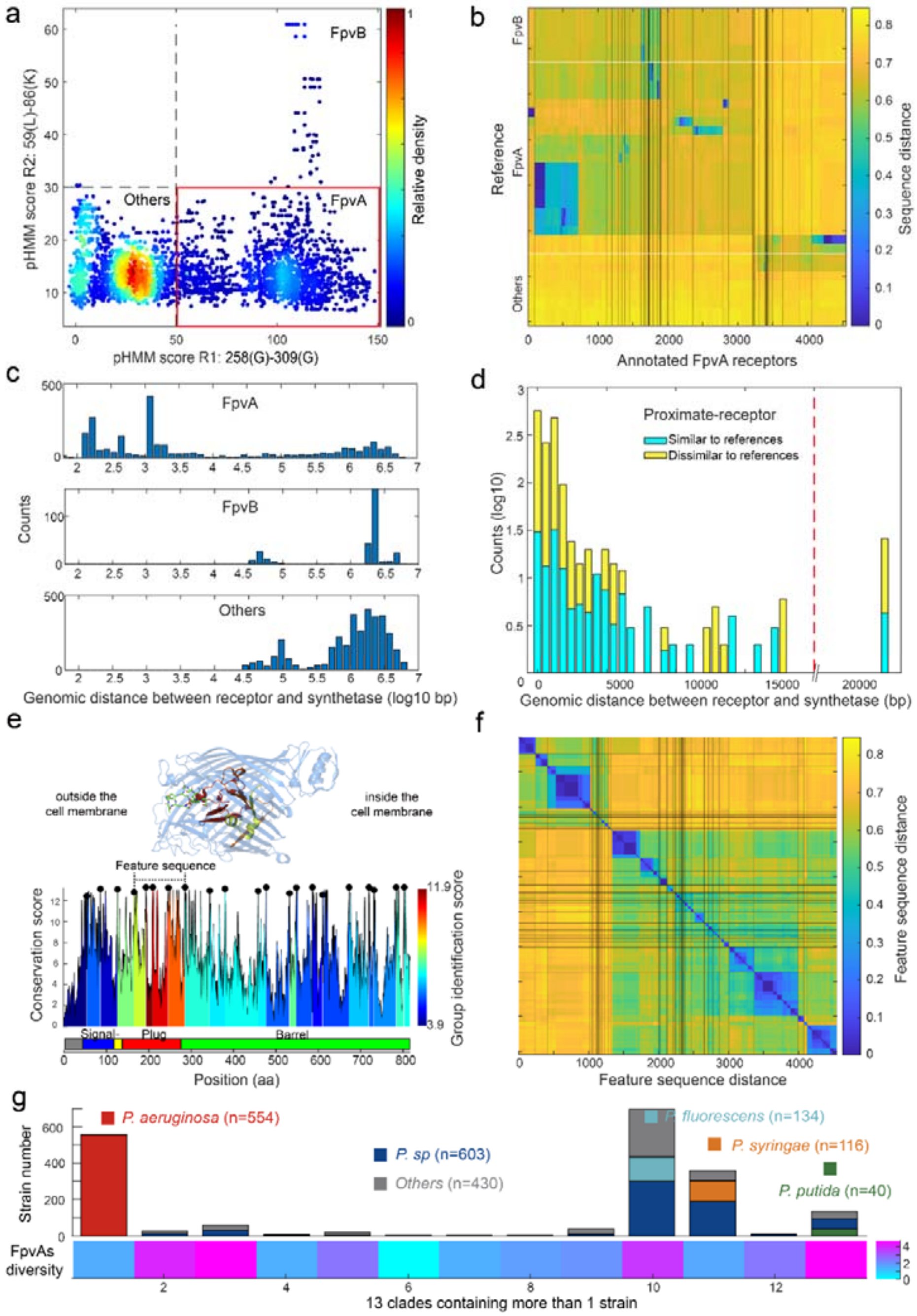


Figure 6 Application of the receptor annotation pipeline to the full database. a. Applying the receptor annotation pipeline to the genomes of the 1928 non-redundant *Pseudomonas* strains yields 14301 Fpv-like receptors, which segregate into 4547 FpvA receptors (red box), 615 FpvB receptors, and 9139 other receptors, based on the pHMM score thresholds for regions R1 and R2. The heatmap

indicates receptor density. **b.** Sequence distance matrix between the 35 reference sequences (y-axis) and the 4547 annotated FpvA sequences in the full database (x-axis). Database sequences were ordered by hierarchically clustering and segregated into 114 groups. 2254 of the annotated FpvA sequences have sequence identity < 60% compared to the reference receptors, pointing at novel subtypes of FpvA receptors. **c.** Genomic distance (in base pairs) between each Fpv-like receptor sequence and its pyoverdine peptide synthetase gene (Pep) for annotated FpvA receptors (upper panel), FpvB receptors (middle panel) and other receptors (lower panel). **d.** Distribution of the genomic distance between each FpvA receptor and its nearest pyoverdine peptide synthetase depending on whether the annotated FpvA receptor has high sequency similarity (blue, ≥ 50%) or low sequence similarity (yellow, < 50%) with at least one of the 21 reference FpvAs. **e.** FpvA region-based conservation scores from a multi-alignment of all the annotated FpvA receptors that are proximate (< 20 kbp) to the pyoverdine synthetase cluster mapped to the FpvA sequence of strain *P. aeruginosa* PAO1. All residues within the top 10% of the conservation score denoted with black dots. For each region flanked by two black dots, we calculated the group identification score (heatmap, lower panel), representing the ability of the region to distinguish between different groups of FpvA receptors. Four regions in the plug domains had a particularly high group identification score (called the feature sequence). They are mapped to the crystal structure of FpvA from PAO1 conjugated with pyoverdine (PDB 2IAH, up panel). All four regions surround the pyoverdine transmission channel and are shown in the respective heatmap color. **f.** Heatmap showing the hierarchically clustered distances between the 4547 annotated FpvA receptors based on the feature sequence (comprising the four groups with the highest identification scores). The analysis identifies 94 receptor groups with a 70% identity threshold. **e.** The diversity of FpvA receptors along the 13 phylogeny clades containing more than 1 strain. Receptor diversity was calculated by the

Shannon entropy, similar to the alpha-diversity in microbial community.

Based on the above insights, we concatenated the four high-score regions (from 168 Pro to 295 Ala in PAO1) into a single "feature sequence". The feature sequence could characterize 98% of the distance matrix compared to the whole sequence (1534 FpvAs, r=0.98, Figure S2a-b) and substantially reduced within-group distance. We applied the concatenated feature sequence approach to all the 4547 annotated FpvAs to calculate the sequence distance matrix. Single-linkage clustering with an identity threshold of 70% revealed a total of 94 groups, out of which 43 groups contained more than 10 members (Figure 6f). The diversity of receptors is hence much larger than currently anticipated as only 3 groups of FpvAs have previously been reported. Finally, we calculated the diversity of receptor FpvAs for each of the 13 phylogenetic clades with more than one strain by the Shannon entropy, which is similar to the alpha-diversity in microbial community (Figure 6g). We noticed large differences in FpvA diversity across the clades and species: clades with *P. aeruginosa* and *P. syringae* species had lower FpvAs diversity (1.55 and 1.60) than clades containing *P. putida* and *P. fluorescens* species (4.82 and 3.77). Taken together, the region-based identification method developed in our study can reliably mine the FpvAs (pyoverdines receptors) from genome data, revealing undiscovered diversity of FpvA pyoverdine receptors that are unequally distributed across the different phylogenetic clades of pseudomonads.

Discussion

The rapid expansion of sequencing data offers exciting opportunities for microbiology[38-40]. One key challenge of current research in the field is to infer biological functions of microbial communities from genome sequence data[41-43]. While this endeavor is increasingly successful for biological functions involving the primary metabolism and the associated complex metabolic flux, reconstructing aspects of the secondary metabolism is much more challenging. The main issue is that neither the function of a secondary metabolite enzyme nor the resulting metabolite can be precisely predicted from gene sequence data. In our study, we tackled this challenge and developed a bioinformatic pipeline to reconstruct the complete secondary metabolism pathway of pyoverdines, a class of iron-scavenging siderophores produced by *Pseudomonas* spp. These secondary metabolites are synthetized by a series of non-ribosomal peptide synthetases and require a specific receptor (FpvA) for uptake. We combined knowledge-guided learning with phylogeny-based methods to predict with high accuracy: (i) the full pyoverdine assembly line, (ii) the substrate specificity for each enzyme within the assembly lines, (iii) the complete chemical structure of pyoverdines, and (iv) the FpvA receptors from genome sequences. After validation, we tested our pipeline with sequence data from 1664 phylogenetically distinct *Pseudomonas* strains and were able to determine 18,292 enzymatic A domains involved in pyoverdine synthesis, reliably predicted 97.8% of their substrates, identified 188 different pyoverdine molecule structures and 4547 FpvA receptor variants belonging to 94 distinct groups. The uncovered diversity is stunning and goes far beyond currently known levels of variation (73 pyoverdines and 3 FpvA groups). The molecular diversity of iron scavenging capacity highlights its importance among pseudomonads.

We show that knowledge-guided learning is an extremely powerful tool to predict enzyme, metabolite, and receptor properties. The establishment of our entire pipeline is based on only

101 previously known enzymatic A domains (from 13 known pyoverdine assembly lines) and 21 FpvA receptor sequences. Even with this limited amount of information, we were able to predict the substrates of almost all the 18,292 enzymatic A domains and to identify 4547 FpvA receptors from the sequence data. A key insight from our knowledge-guided learning is that comparisons based on the full gene sequences (e.g., for pyoverdine synthetase or receptor) are likely non-informative and unsuitable for obtaining functional information. This is because overall diversity does not stand for functional diversity, meaning that A domains recognizing the same substrate can diverge substantially in their full sequences. The same holds true for receptor sequences: whole-sequence alignments can neither accurately identify FpvA receptors nor reliably separate them into functional groups. Instead, it is imperative to extract informative feature sequences that are defined as sequence stretches within a gene whose diversity is tightly linked to variation in its functioning. We successfully extracted and applied feature sequence comparisons for both A domain substrate prediction and FpvA identification. It is important to note that a knowledge-guided pipeline does not have to be perfect right from the start. For example, our pipeline for pyoverdine structure prediction returned unknowns for several amino acid positions within the PEP. Our experimental verifications then revealed indeed new substrates such as valine and citrulline. This information can then be used to refine our prediction algorithm in a feedback loop.

Another main advantage of our bioinformatic pipeline is that it can be applied to draft genomes. This reflects a major improvement compared to existing annotation tools such as antiSMASH[25], which typically has difficulties in recognizing NRPS structures in fragmented genome assemblies. However, draft genomes are the most common data source in microbiology. While our pipeline shows high performance, we need to acknowledge that we still lose many genomes (6087 out of 9599, 63.4%). The reason for the loss is that the pyoverdine synthesis machinery is large, which increases the probability that it is positioned at the end of a contig. We decided to exclude those cases because the annotated synthesis machinery might be truncated and thus incomplete. Thus, the high loss rate of draft genomes is rather due to limitations in sequence quality (too many short contigs) and not due to a limitation of our bioinformatic pipeline. We believe that this limitation will be lesser of a proble in the future as long-read sequencing technologies are quickly becoming cheaper and more reliable.

We further show that knowledge-guided learning combined with a phylogeny-focused approach is a powerful tool for predicting the substrate specificity of A domains of synthetases. It outperforms currently known bioinformatics prediction tools of NRPS substrates such as antiSMASH[25]. Most current algorithms[44-48]perform poorly when applied to pyoverdines, particularly when encountering non-proteogenic amino acids. The high accuracy of our algorithm can largely be attributed to our reference set, composing only 13 pyoverdines from *Pseudomonas spp*., yet capturing most of the substrate diversity. Similarly accurate predictions based on a handful of known substrates among closely related species were observed in several fungal NRPS systems[49]. It is worth noting that when the algorithm output is "unknown," it actually signifies uncharacterized A domains not yet incorporated into the reference data set. This should prompt researchers to pay attention to these A domains, and like in our case, subject them to further experimental investigation. This approach helped us discover new substrates (valine, histidine, citrulline), which had not been previously documented in pyoverdines and were therefore absent from the reference A domains. The novel substrates identified through our structural assessment and mass spectrometry experiment can subsequently be used to enhance the precision of our phylogeny-centered substrate prediction technique in the future, creating a progressive feedback loop of expanding knowledge. Taken together, supervised learning based on a few known compounds produced by species from the same genus probably outperforms generalized prediction algorithms trained on many products from a diverse set of microbes for NRPS substrate predictions.

Our results show that both pyoverdine and receptor diversity has been vastly underestimated. While considerable pyoverdine diversity (n=73) has already been captured in previous studies, here we discovered 151 new variants. On the receptor side, the uncovered novel diversity is more dramatic. One reason for this is that research on receptors has mainly focused on the pathogen *P. aeruginosa*[19-22,50]. For this species, three different pyoverdine types were described[21] together with three structurally different FpvA receptor types that each recognize one of the pyoverdine types[22]. While our study confirmed that *P. aeruginosa* strains (n = 554) indeed have only 3 pyoverdine-receptor systems, we also discovered 91 new FpvA groups among environmental *Pseudomonas* spp. Our findings raise the question why there are so many different pyoverdine and receptor variants. One potential explanation is that the benefit of specific siderophores could be context-dependent and locally adapted to multitude of different environmental conditions pseudomonads are exposed to. For example, experimental work has revealed that pyoverdines can be cooperatively shared among strains with matching receptors[33,51], or conversely, pyoverdines can serve as competitive agents by locking away iron from species that have non-matching receptors[52]. Given that bioavailable iron is limited in most natural and host-associated habitats[53-55], the unraveled functional diversity is likely a direct evolutionary consequence of the struggle and competition of microbes for iron. While experimental work is often restricted to a low number of strains, we propose that our bioinformatic pipeline can be used to predict pyoverdine-mediated interaction networks across thousands of strains and across different habitats. We will address this point in a future study.

We believe our pipeline could be easily expanded to study iron competition in multi-species communities in the future and perhaps in plant-microbe ecosystems, as siderophores exist ubiquitously and are shared among microbes[56]. To move further, a key question is whether our knowledge-guided approach can be applied to other important secondary metabolites, such as antibiotics, toxins, biosurfactants and pigments? This answer is: not directly but the pipeline development strategies are translational between different types of compounds. As soon as sufficient case-by-case knowledge on a specific system is available, the annotation strategies together with the feature sequence extraction and the phylogeny-focused approach developed in our paper can be applied. For most of the secondary metabolites listed above, there are no receptors as the compounds have purely extra-cellular functions, which substantially simplifies the development of bioinformatic pipelines. In the long run, it will certainly be possible to automate the steps implemented in our workflow so that the algorithms can be applied to a large set of secondary metabolites when fed with an appropriate training set.

Data Availability

The source code and parameters used are available in the supplementary material.

Acknowledgements

We thank Richard Allen for genome sequencing of the 20 strains. We also appreciated Vera Vollenweider for sample preparation for the experiment of pyoverdine structure elucidation.

Funding

This work was supported by the National Key Research and Development Program of China (No. 2021YFF1200500, 2021YFA0910700), National Natural Science Foundation of China (No. 42107140, No. 32071255, No.41922053), National Postdoctoral Program for Innovative Talents (No. BX2021012), and Clinical Medicine Plus X - Young Scholars Project, Peking University (7100603954). R.K. is supported jointly by a grant from the Swiss National Science Foundation no. 310030_212266. V-P.F. is supported jointly by a grant from UKRI, Defra, and the Scottish Government, under the Strategic Priorities Fund Plant Bacterial Diseases program (BB/T010606/1), Research Council of Finland, and The Finnish Research Impact Foundation.

Author contributions

Shaohua Gu performed the majority of computational analysis in this research and drafted the manuscript. Yuanzhe Shao built the pipeline for retrieving synthetase sequence information and developed the phylogeny-centered method for predicting substrates. Karoline Rehm and Laurent Bigler performed the experiment of pyoverdine structure elucidation. Di Zhang depicted the circular plot. Ruolin He standardized of all NRPS structures into motif-intermotif format. Jiqi Shao assisted in revising the manuscript. Alexandre Jousset and Ville-Petri Friman offered insightful comments and assisted in revising and writing of the manuscript. Rolf Kümmerli and Zhong Wei oversaw the project, designed experiments and revised the manuscript. Zhiyuan Li conceptualized and oversaw the project, conducted the analysis of the receptor annotated method, and revised the manuscript.

Competing interests

The authors declare no competing interests.

References

1 Zengler, K. & Palsson, B. O. A road map for the development of community systems (CoSy) biology. *Nature Reviews Microbiology* **10**, 366-372, doi:10.1038/nrmicro2763 (2012).

2 Gu, C., Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models. *Genome Biology* **20**, 121, doi:10.1186/s13059-019-1730-3 (2019).

3 García-Jiménez, B., Torres-Bacete, J. & Nogales, J. Metabolic modelling approaches for describing and engineering microbial communities. *Computational and Structural Biotechnology Journal* **19**, 226-246, doi:10.1016/j.csbj.2020.12.003 (2021).

4 Colarusso, A. V., Goodchild-Michelman, I., Rayle, M. & Zomorrodi, A. R. Computational modeling of metabolism in microbial communities on a genome-scale.

Current Opinion in Systems Biology **26**, 46-57, doi:10.1016/j.coisb.2021.04.001 (2021).

5 Scherlach, K. & Hertweck, C. Mediators of mutualistic microbe–microbe interactions. *Natural Product Reports* **35**, 303-308, doi:10.1039/C7NP00035A (2018).

6 Trivedi, P., Leach, J. E., Tringe, S. G., Sa, T. & Singh, B. K. Plant–microbiome interactions: from community assembly to plant health. *Nature Reviews Microbiology* **18**, 607-621, doi:10.1038/s41579-020-0412-1 (2020).

7 Price-Whelan, A., Dietrich, L. E. P. & Newman, D. K. Rethinking 'secondary' metabolism: physiological roles for phenazine antibiotics. *Nature Chemical Biology* **2**, 71-78, doi:10.1038/nchembio764 (2006).

8 Thirumurugan, D., Cholarajan, A., Raja, S. & Vijayakumar, R. An introductory chapter: secondary metabolites. (2018).

9 Quinn, G. A., Banat, A. M., Abdelhameed, A. M. & Banat, I. M. Streptomyces from traditional medicine: sources of new innovations in antibiotic discovery. *J Med Microbiol* **69**, 1040-1048, doi:10.1099/jmm.0.001232 (2020).

10 Durand, G. A., Raoult, D. & Dubourg, G. Antibiotic discovery: history, methods and perspectives. *International Journal of Antimicrobial Agents* **53**, 371-382, doi:10.1016/j.ijantimicag.2018.11.010 (2019).

11 Penn, K. *et al.* Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. *The ISME Journal* **3**, 1193-1203, doi:10.1038/ismej.2009.58 (2009).

12 Andryukov, B., Mikhailov, V. & Besednova, N. The Biotechnological Potential of

Secondary Metabolites from Marine Bacteria. *Journal of Marine Science and Engineering* **7** (2019).

13 Kautsar, S. A., Blin, K., Shaw, S., Weber, T. & Medema, M. H. BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Research* **49**, D490-D497, doi:10.1093/nar/gkaa812 (2021).

14 He, R. *et al.* Knowledge-guided data mining on the standardized architecture of NRPS: Subtypes, novel motifs, and sequence entanglements. *PLOS Computational Biology* **19**, e1011100, doi:10.1371/journal.pcbi.1011100 (2023).

15 Xu, Z., Park, T. J. & Cao, H. Advances in mining and expressing microbial biosynthetic gene clusters. *Crit Rev Microbiol*, 1-20, doi:10.1080/1040841x.2022.2036099 (2022).

16 Keller, N. P. Fungal secondary metabolism: regulation, function and drug discovery. *Nature Reviews Microbiology* **17**, 167-180, doi:10.1038/s41579-018-0121-1 (2019).

17 Ringel, M. T. & Brüser, T. The biosynthesis of pyoverdines. *Microb Cell* **5**, 424-437, doi:10.15698/mic2018.10.649 (2018).

18 Hopkinson, B. M. & Morel, F. M. M. The role of siderophores in iron acquisition by photosynthetic marine microorganisms. *BioMetals* **22**, 659-669, doi:10.1007/s10534-009-9235-2 (2009).

19 Cobessi, D. *et al.* The Crystal Structure of the Pyoverdine Outer Membrane Receptor FpvA from Pseudomonas aeruginosa at 3.6Å Resolution. *Journal of Molecular Biology* **347**, 121-134, doi:10.1016/j.jmb.2005.01.021 (2005).

20 Diggle, S. P. & Whiteley, M. Microbe Profile: Pseudomonas aeruginosa: opportunistic

pathogen and lab rat. *Microbiology (Reading)* **166**, 30-33, doi:10.1099/mic.0.000860 (2020).

21 Meyer, J.-M. *et al.* Use of Siderophores to Type Pseudomonads: The Three Pseudomonas Aeruginosa Pyoverdine Systems. *Microbiology* **143**, 35-43, doi:10.1099/00221287-143-1-35 (1997).

22 Bodilis, J. *et al.* Distribution and evolution of ferripyoverdine receptors in Pseudomonas aeruginosa. *Environmental Microbiology* **11**, 2123-2135, doi:10.1111/j.1462-2920.2009.01932.x (2009).

23 Kümmerli, R. Iron acquisition strategies in pseudomonads: mechanisms, ecology, and evolution. *BioMetals*, doi:10.1007/s10534-022-00480-8 (2022).

24 Meyer, J. M. Pyoverdines: pigments, siderophores and potential taxonomic markers of fluorescent Pseudomonas species. *Arch Microbiol* **174**, 135-142, doi:10.1007/s002030000188 (2000).

25 Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research* **47**, W81-W87, doi:10.1093/nar/gkz310 (2019).

26 Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Research* **32**, D138-D141, doi:10.1093/nar/gkh121 (2004).

27 Winsor, G. L. *et al.* Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database. *Nucleic Acids Research* **44**, D646-D653, doi:10.1093/nar/gkv1227 (2016).

28 Felnagle, E. A. *et al.* Nonribosomal peptide synthetases involved in the production of medically relevant natural products. *Mol Pharm* **5**, 191-211, doi:10.1021/mp700137g

(2008).

29 Süssmuth, R. D. & Mainz, A. Nonribosomal Peptide Synthesis—Principles and Prospects. *Angewandte Chemie International Edition* **56**, 3770-3821, doi:10.1002/anie.201609079 (2017).

30 Butaitė, E., Kramer, J., Wyder, S. & Kümmerli, R. Environmental determinants of pyoverdine production, exploitation and competition in natural Pseudomonas communities. *Environmental Microbiology* **20**, 3629-3642, doi:10.1111/1462-2920.14355 (2018).

31 Rehm, K., Vollenweider, V., Kümmerli, R. & Bigler, L. A comprehensive method to elucidate pyoverdines produced by fluorescent Pseudomonas spp. by UHPLC-HR-MS/MS. *Analytical and Bioanalytical Chemistry* **414**, 2671-2685, doi:10.1007/s00216-022-03907-w (2022).

32 Butaitė, E., Baumgartner, M., Wyder, S. & Kümmerli, R. Siderophore cheating and cheating resistance shape competition for iron in soil and freshwater Pseudomonas communities. *Nature Communications* **8**, 414, doi:10.1038/s41467-017-00509-4 (2017).

33 Kramer, J., Özkaya, Ö. & Kümmerli, R. Bacterial siderophores in community and host interactions. *Nature Reviews Microbiology* **18**, 152-163, doi:10.1038/s41579-019-0284-4 (2020).

34 Chan, D. C. K. & Burrows, L. L. &lt;em&gt;Pseudomonas aeruginosa&lt;/em&gt; FpvB is a high-affinity transporter for xenosiderophores ferrichrome and ferrioxamine B. *bioRxiv*, 2022.2009.2020.508722, doi:10.1101/2022.09.20.508722 (2022).

35 González, J. *et al.* Loss of a pyoverdine secondary receptor in Pseudomonas aeruginosa results in a fitter strain suitable for population invasion. *The ISME Journal* **15**, 1330-1343, doi:10.1038/s41396-020-00853-2 (2021).

36 Schalk, I. J., Mislin, G. L. & Brillet, K. Structure, function and binding selectivity and stereoselectivity of siderophore-iron outer membrane transporters. *Curr Top Membr* **69**, 37-66, doi:10.1016/B978-0-12-394390-3.00002-1 (2012).

37 Greenwald, J. *et al.* FpvA bound to non-cognate pyoverdines: molecular basis of siderophore recognition by an iron transporter. *Molecular Microbiology* **72**, 1246-1259, doi:10.1111/j.1365-2958.2009.06721.x (2009).

38 Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499-504, doi:10.1038/s41586-019-0965-1 (2019).

39 Schloss, P. D. & Handelsman, J. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biology* **6**, 229, doi:10.1186/gb-2005-6-8-229 (2005).

40 Handelsman, J. Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews* **68**, 669-685, doi:10.1128/MMBR.68.4.669-685.2004 (2004).

41 Tsilimigras, M. C. B. & Fodor, A. A. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of Epidemiology* **26**, 330-335, doi:10.1016/j.annepidem.2016.03.002 (2016).

42 Weiss, S. *et al.* Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J* **10**, 1669-1681, doi:10.1038/ismej.2015.235 (2016).

43 Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nature Reviews Microbiology* **10**, 538-550, doi:10.1038/nrmicro2832 (2012).

44 Khayatt, B. I., Overmars, L., Siezen, R. J. & Francke, C. Classification of the Adenylation and Acyl-Transferase Activity of NRPS and PKS Systems Using Ensembles of Substrate Specific Hidden Markov Models. *PLOS ONE* **8**, e62136, doi:10.1371/journal.pone.0062136 (2013).

45 Minowa, Y., Araki, M. & Kanehisa, M. Comprehensive Analysis of Distinctive Polyketide and Nonribosomal Peptide Structural Motifs Encoded in Microbial Genomes. *Journal of Molecular Biology* **368**, 1500-1517, doi:10.1016/j.jmb.2007.02.099 (2007).

46 Prieto, C., García-Estrada, C., Lorenzana, D. & Martín, J. F. NRPSsp: non-ribosomal peptide synthase substrate predictor. *Bioinformatics* **28**, 426-427, doi:10.1093/bioinformatics/btr659 (2012).

47 Röttig, M. *et al.* NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Research* **39**, W362-W367, doi:10.1093/nar/gkr323 (2011).

48 Zierep, P. F., Ceci, A. T., Dobrusin, I., Rockwell-Kollmann, S. C. & Günther, S. SeMPI 2.0-A Web Server for PKS and NRPS Predictions Combined with Metabolite Screening in Natural Product Databases. *Metabolites* **11**, doi:10.3390/metabo11010013 (2020).

49 Fan, J. *et al.* Biosynthetic diversification of peptaibol mediates fungus-mycohost interactions. *bioRxiv* (2022).

50 Smith, E. E., Sims, E. H., Spencer, D. H., Kaul, R. & Olson, M. V. Evidence for diversifying selection at the pyoverdine locus of Pseudomonas aeruginosa. *J Bacteriol* **187**, 2138-2147, doi:10.1128/jb.187.6.2138-2147.2005 (2005).

51 Gu, S. *et al.* Competition for iron drives phytopathogen control by natural rhizosphere microbiomes. *Nature Microbiology* **5**, 1002-1010, doi:10.1038/s41564-020-0719-8 (2020).

52 Figueiredo, A. R. T., Özkaya, Ö., Kümmerli, R. & Kramer, J. Siderophores drive invasion dynamics in bacterial communities through their dual role as public good versus public bad. *Ecology Letters* **25**, 138-150, doi:10.1111/ele.13912 (2022).

53 Andrews, S. C., Robinson, A. K. & Rodríguez-Quiñones, F. Bacterial iron homeostasis. *FEMS Microbiology Reviews* **27**, 215–237, doi:10.1016/s0168-6445(03)00055-x (2003).

54 Boyd, P. W. & Ellwood, M. J. The biogeochemical cycle of iron in the ocean. *Nat Geosci* **3**, 675-682, doi:10.1038/ngeo964 (2010).

55 Emerson, D., Roden, E. & Twining, B. The microbial ferrous wheel: iron cycling in terrestrial, freshwater, and marine environments. *Frontiers in Microbiology* **3**, doi:10.3389/fmicb.2012.00383 (2012).

56 Ruolin, H. *et al.* SIDERITE: Unveiling Hidden Siderophore Diversity in the Chemical Space Through Digital Exploration. *bioRxiv*, 2023.2008.2031.555687, doi:10.1101/2023.08.31.555687 (2023).