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], Ruichen Xu[4], Jiqi Shao[1], Alexandre Jousset[5], Ville-Petri Friman[6], Xiaoying Bian[7], Zhong Wei[5#], Rolf Kümmerli[8#], 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] School of Life Science, Shandong University, Qingdao, Shandong 266237, China

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

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

[7] Helmholtz International Lab for Anti-infectives, State Key Laboratory of Microbial Technology, Shandong University, Qingdao, Shandong 266237, China

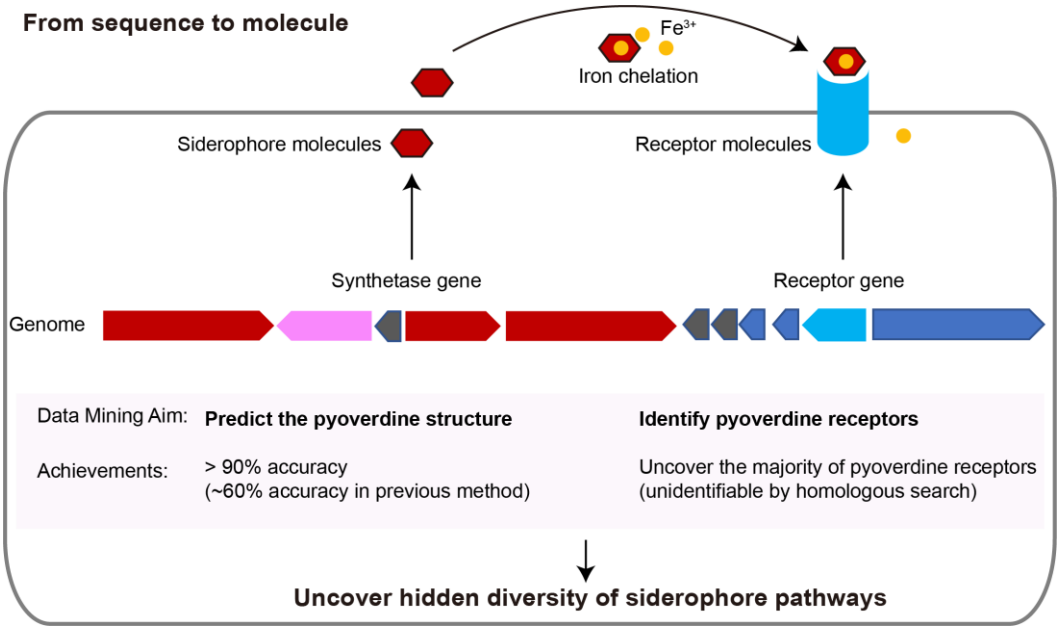
[8] University of Zurich, Department of Quantitative Biomedicine, Winterthurerstr. 190, 8057 Zurich, Switzerland

[*] These authors contributed equally to this article.

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

Abstract (146 words count)

Microbial secondary metabolites are a rich source for pharmaceutical discoveries and play crucial ecological functions. While tools exist to identify secondary metabolite clusters in genomes, precise sequence-to-function mapping remains challenging because neither function nor substrate specificity of biosynthesis enzymes can accurately be predicted. Here we developed a knowledge-guided bioinformatic pipeline to solve these issues. We analyzed 1928 genomes of *Pseudomonas* bacteria and focused on iron-scavenging pyoverdines as model metabolites. Our pipeline predicted 188 chemically different pyoverdines with nearly 100% structural accuracy and the presence of 94 distinct receptor groups required for the uptake of iron-loaded pyoverdines. Our pipeline unveils an enormous yet overlooked diversity of siderophores (151 new structures) and receptors (91 new groups). Our approach, combining feature sequence with phylogenetic approaches, is extendable to other metabolites and microbial genera, and thus emerges as powerful tool to reconstruct bacterial secondary metabolism pathways based on sequence data.



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 microbial community assembly[5,6] and to be important sources for pharmaceutical discoveries[7-9] .

There are multiple challenges that currently prevent a detailed unravelling of secondary metabolism of bacteria based on genome data[10-12]. First, most secondary metabolites are produced by pathways comprised of modular enzymes such as non-ribosomal peptide synthetases (NRPSs) or polyketide synthases (PKS)[13,14]. Locating complete biosynthesis 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

61 pathways is covered by current data bases. Finally, given the complex multi-modular
62 biosynthesis machineries, it is challenging to precisely predict the secondary metabolites
63 produced even with accurately annotated NRPS or PKS clusters. The main challenge is that a
64 large pool of non-proteinogenic amino acids is used as substrates and the specificity of an
65 enzyme's A domain, connecting these unusual amino acids, is often poorly understood[15]. As a
66 result, new computational methods are needed to accurately reconstruct bacterial secondary
67 metabolism from sequence data.

68 Here, we present a new bioinformatic pipeline that overcomes these challenges. We
69 specifically focus on a particular class of secondary metabolites (iron-scavenging siderophores)
70 as a case study to develop a bioinformatic workflow that predicts the chemical structure of the
71 produced metabolites with nearly 100% accuracy. Our pipeline is based on improved gene
72 annotation combined with a phylogeny- and feature sequence-based substrate prediction
73 techniques (Figure 1). In comparison with the currently available databases and bioinformatic
74 tools[16], the main advancement of our method is the more accurate prediction of synthesized
75 products based on NRPS clusters identified in genome data. In addition, we show that our
76 workflow is capable of correcting previous mis-annotations and is extendable to other
77 secondary metabolites (e.g., toxins, biosurfactants, virulence factors) and microbial genera.

78 Among siderophores, we focus on NRPS machineries that are responsible for the
79 biosynthesis of pyoverdines, a class of chemically diverse siderophores with high iron affinity,
80 produced by *Pseudomonas* bacteria[17,18]. While each *Pseudomonas* strain produces a single
81 type of pyoverdine, an enormous structural diversity has been described across strains and
82 species[19-22]. Pyoverdine types differ in their peptide backbone, meaning that the diversity

83 should be mirrored in NRPS enzyme diversity and their selectivity for the different amino acid

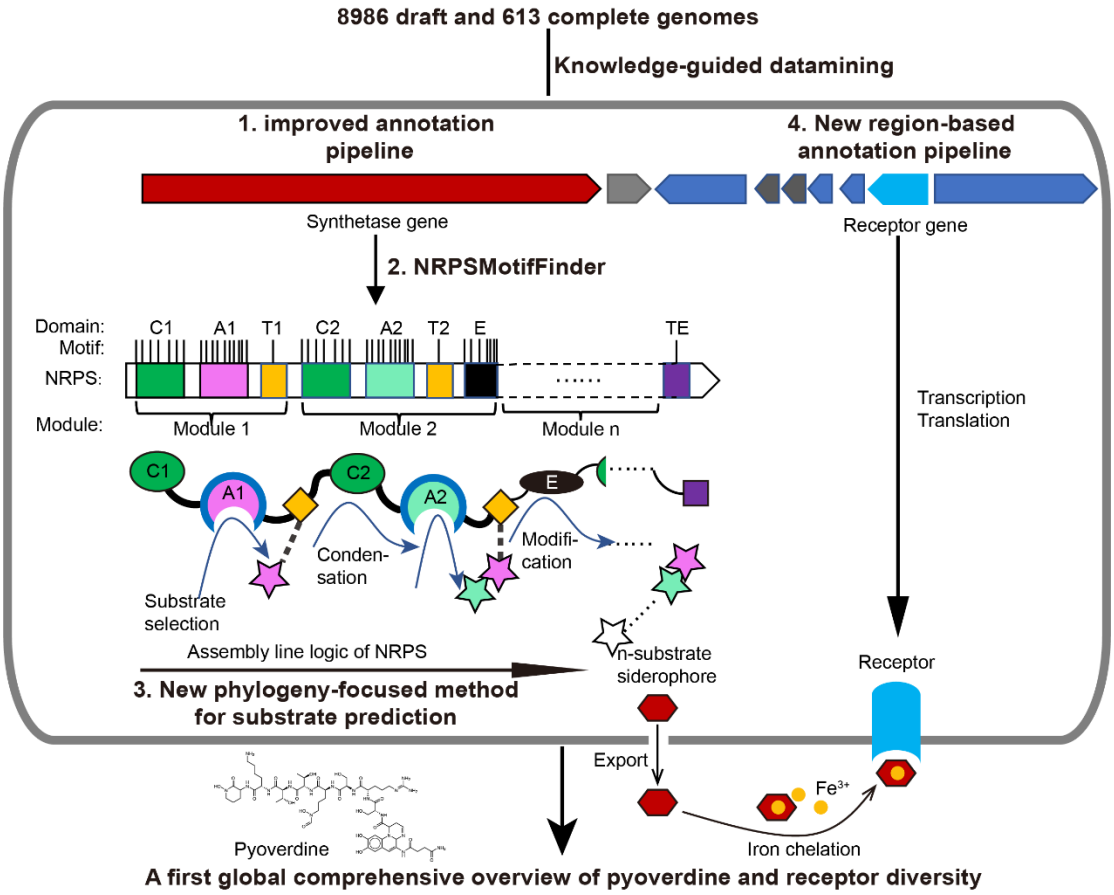
84 substrates[18]. Based on this knowledge, our pipeline entails the following steps (Figure 1): (i)

85 identification of the complete sequences of pyoverdine synthetase genes from fragmented draft

86 genomes, (ii) building the pyoverdine biosynthesis machinery *in silico* by extracting the feature

87 sequences for substrate specificity from motif-standardized NRPSs, and (iii) predicting the

88 precise chemical structure of pyoverdines followed by empirical verification.



89

90 **Figure 1 Scheme depicting our new genome mining pipeline to precisely predict the**
91 **biosynthesis, the molecular structure and the uptake machinery of pyoverdines, a family**
92 **of iron-scavenging siderophores produced by members of the *Pseudomonas* genus.** The
93 grey rounded outer rectangle represents a bacterial cell. The red and blue arrow-shaped boxes stand for
94 the synthetase and receptor genes for pyoverdines, respectively. Synthetase genes are transcribed and
95 translated to form the *n*-modular NRPS enzymes. These enzymes synthesize the peptide backbone of
96 pyoverdine through an assembly line using their repeating module units, with the A domain being
97 responsible for substrate selection and the E domain for chirality. The *n*-substrate siderophores are then
98 exported to the extracellular space for iron chelation. Membrane-embedded TonB-dependent receptors
99 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[23]) 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.

An additional element of iron acquisition involves specific receptors that recognize and trigger the uptake of iron-loaded siderophores. Pyoverdine receptors are annotated as FpvA and it is known that receptor diversity matches pyoverdine diversity[22,24]. Moreover, FpvA belongs to the family of TonB-dependent receptors and a single *Pseudomonas* species often has multiple gene copies encoding different 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 receptor sequence regions involved in pyoverdine recognition and translocation across the outer membrane with supervised learning methods that reliably locate *fpvA* genes in fragmented genomes based on these regions (Figure 1). Not only did our method unveil a substantial number of previously unrecognized pyoverdine receptors, but it also revealed that certain receptor groups are widely distributed among *Pseudomonas* strains., opening the possibility of pyoverdine sharing and exploitation across the strain and species boundaries. 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.

Results

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

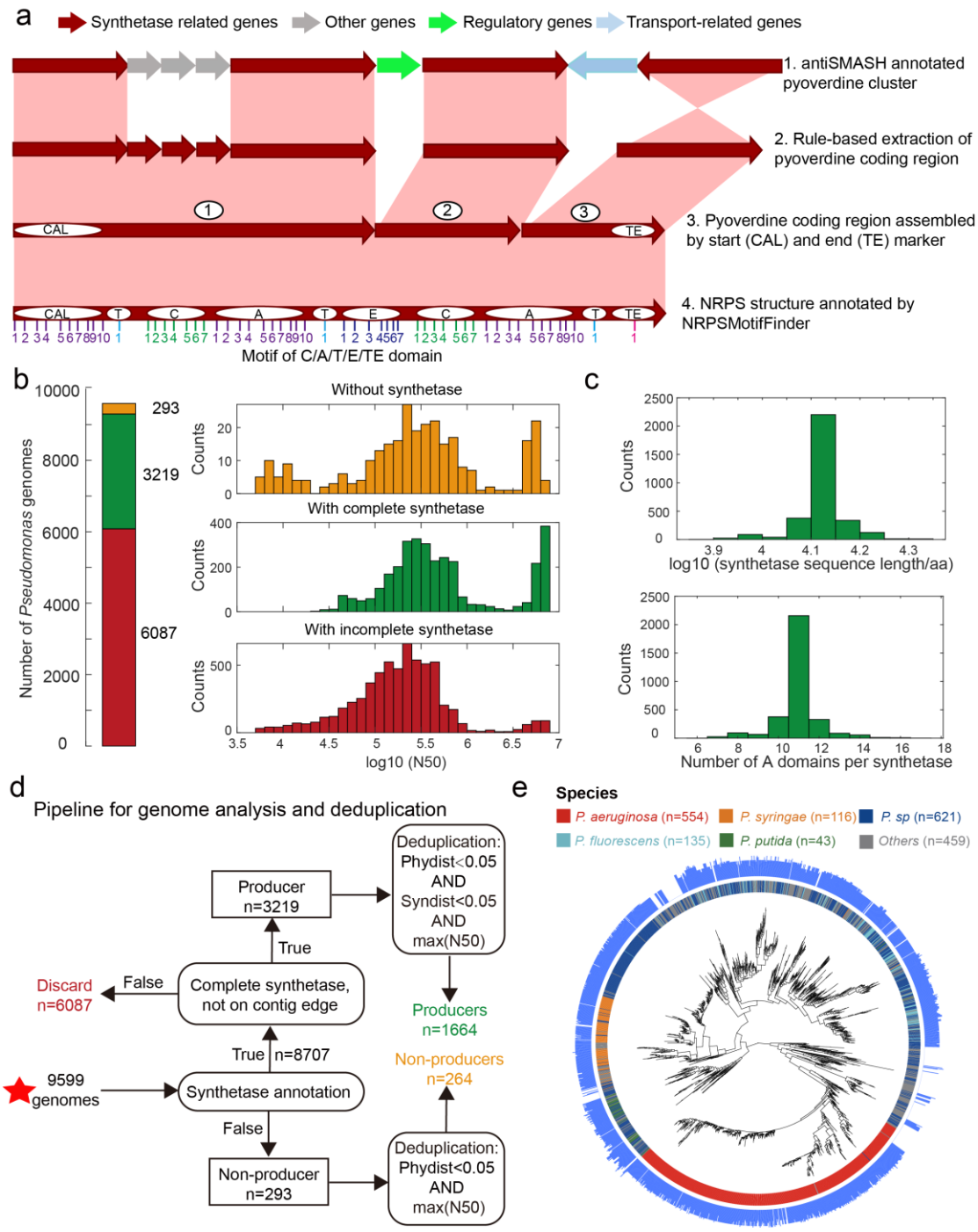


Figure 2 Improved annotation pipeline reveals a vast diversity of pyoverdine synthetase genes. a. Improved annotation pipeline based on the raw annotation from antiSMASH[23]. **b.** The annotation pipeline was applied to 9599 *Pseudomonas* genomes (94% draft genomes). Genomes could be separated into 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.

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[25]. There exist already tools, such as antiSMASH that can find and annotate NRPS clusters in microbial genomes[23]. However, antiSMASH (and other popular annotation platforms[26,27]) 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.

To overcome these issues, we developed an improved four-step annotation pipeline starting with the raw annotation of the pyoverdine cluster obtained from antiSMASH[23] (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[28], 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,

164 we assembled the entire re-annotated pyoverdine coding region into a single sequence with
165 defined start (co-enzyme A ligase, CAL) and end (thioesterase, TE) markers. In most cases
166 (~90%), there are two separate biosynthetic gene clusters for Flu and Pep. Their ordering is
167 easy as Flu always precedes Pep. In the remaining cases (~10%), there is either one or three
168 biosynthetic gene clusters. No ordering is required for the former. For the latter, we put Flu first
169 and the cluster with TE domain at the end. Third, we used NRPSMotifFinder to identify the C,
170 A, T, E and TE motifs that determine the NRPS structure of pyoverdine[15]. Finally, we applied a
171 safety measure to ensure that the recovered NRPS assembly line is complete (contains both
172 Flu and Pep) and is not truncated, which can occur when a synthetase coding region is at the
173 edge of a contig. Consequently, we dismissed all pyoverdine biosynthesis clusters located
174 within 100 bp proximity to contigs' edges and either lacked Flu or Pep synthetic genes.

175 Next, we applied our improved pyoverdine biosynthesis annotation pipeline to 9599
176 *Pseudomonas* genomes (including 613 complete and 8986 draft genomes) retrieved from the
177 *Pseudomonas* Genome Database[29]. We found the pyoverdine biosynthesis machinery in 97%
178 of the genomes (Figure 2b), indicating that the machinery is ubiquitous in *Pseudomonas*.
179 However, since 94% of the analyzed genomes were in draft form, the pyoverdine biosynthesis
180 machinery was likely truncated (i.e., on the edge of the contig) in 63.4% (6087) of the genomes.
181 These genomes were excluded from further analysis. Around 3.1% of retained genomes (293)
182 with high assembly completeness were missing pyoverdine synthetic genes, indicating that
183 these *Pseudomonas* strains were not able to produce pyoverdine ('non-producers'). The rest
184 of the genomes (33.5%; 3219 genomes) were classified as 'producers' with complete
185 pyoverdine NRPS assembly lines that meet all our quality controls. For these 3219 genomes,

186 we used NRPSMotifFinder to find boundaries between the various biosynthesis domains and
187 to determine the amino acid length of the cluster and the number of A domains. The lengths of
188 pyoverdine synthetic clusters ranged between 7690 and 21333 amino acids, and the number
189 of A domains per synthetase ranged between 6 and 17, with a total of 35,281 A domains being
190 present across all strains (Figure 2c and Figure S1). Overall, our analysis pipeline unveiled a
191 vast diversity of pyoverdine synthetases that goes far beyond of what has previously been
192 described in the literature.

193 Finally, we utilized the PhyloPhlAn3 pipeline[30] to conduct a phylogenetic analysis based
194 on 400 conserved genes with the 293 non-producers and the 3219 producers. We first removed
195 redundant non-producers by retaining the most integrative genome among strains with high
196 phylogenic similarity (>95% identity). Then, we removed redundant producers by retaining the
197 most integrative genome among strains with high phylogenic and pyoverdine synthetase
198 similarity (Figure 2d, >95% identity). This data cleaning yielded a total of 1928 *Pseudomonas*
199 strains (403 complete and 1525 incomplete genomes), segregating into 1664 pyoverdine
200 producers and 264 non-producers. The phylogenetic tree revealed that all major *Pseudomonas*
201 species clades were present in our data set (Figure 2e). Moreover, the number of A domains
202 varied widely among species and even between strains within species. For example, the
203 number of *Pseudomonas aeruginosa* A domains ranges between 7 and 14. In summary, by
204 improving the synthetase annotation method, we successfully obtained 1664 highly reliable
205 pyoverdine synthetases (with a total of 18,292 A domains) and 264 non-producers.
206

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

208 Our next goal was to precisely predict the molecular structure of the pyoverdines produced by
209 the 1664 strains with complete synthetase gene clusters. The first essential step towards this
210 goal was to reliably predict the substrate selectivity of all A domains in the NRPS assembly line.
211 The A domain of each module selects for a single substrate among 22 proteinogenic and
212 hundreds of non-proteinogenic amino acids[31,32]. Moreover, whenever an E domain exists
213 downstream of an A domain, the chirality of the amino acid incorporated into the peptide chain
214 gets modified from L to D. Thus, the modularity combined with the selectivity of A domains can
215 promote an enormous diversity of pyoverdine molecule structures. To date, at least 73
216 pyoverdine structures have been reported (Supplementary_table1). Among those, we could
217 identify 13 cases for which both structures (obtained from structural analysis experiments, we
218 wanted to avoid any form of biases that bioinformatic pre-predictions could introduce to
219 downstream analyses) and completely sequenced synthetase genes were available
220 (Supplementary_table2). In order to make reliable molecule predictions, two challenges must
221 be addressed: (i) the extraction of relevant information from A domain sequences for which the
222 substrate is known, and (ii) the effective application of this information to predict specificity of A
223 domain sequences for which the substrate is unknown.

224 To address the first challenge, we built our analysis on the NRPS assembly lines of the 13
225 known pyoverdines identified above to extract relevant information from the A domain
226 sequences. From this dataset, we could identify 101 A domains that are linked to 13
227 experimentally confirmed amino acid substrates (Supplementary_table3). We next performed
228 a multisequence alignment of the 101 A domains to determine (i) the “feature sequence”,

229 defined as the most informative sequence for substrate selectivity and (ii) the “feature sequence
230 distance”, which allows to reliably differentiate between two different substrates. To this end,
231 we tested three different A domain regions, three different sequence similarity measurements
232 and seven different clustering methods for their predictive power (Figure 3a). We found that the
233 full A domain sequence is not informative for substrate prediction (Figure 3b, left panel). Instead,
234 our analysis indicated that information-rich positions start with motif A4 and end before motif
235 A5, consistent with the known role of the A domain pocket in substrate selectivity[15]. Overall, the
236 sequence region from motifs A4 to A5 (termed “Amotif4-5”), in conjunction with Jukes-Cantor
237 distance and Ward linkage clustering, performed best in accurately distinguishing between
238 different substrates and maintaining homogeneity for identical substrates (Figure 3b). Utilizing
239 these parameters, we conducted a comprehensive analysis of all A domains and their
240 respective substrate information within the *Pseudomonas* NRPS biosynthetic gene cluster
241 available in the MIBiG database[33] (Supplementary_table4). To our delight, our method can find
242 and correct previously misclassified substrate information of A domains in the MIBiG database
243 (Figure S2 and Supplementary_table5). This demonstrates the effectiveness of performing
244 substrate predictions for the A domain based on feature sequences and the selected parameter
245 combination.

246 To address the second challenge, we developed a “phylogeny-focused method” to apply
247 the above-derived feature sequence distance approach to the 18,292 discovered A domains.
248 We realized that a direct construction of a phylogenetic tree including all 18,292 query A
249 domains and the 101 reference A domains would be computationally too demanding and
250 impossible to scale up. Furthermore, such an approach would result in phylogeny-interference

251 issues, where domains would cluster not only based on their substrate similarities but also
252 based on overall species relatedness[15]. To minimize the effect of phylogeny and speed up
253 calculations, we took each of the 18,292 query A domains and identified the two most similar A
254 domain clusters among our 101 reference A domains. We then compared the feature distance
255 between each query A domain and the two most similar reference A domains and assigned a
256 substrate to the query A domain using the following rules (Figure 3c). (1) If the feature sequence
257 distance is below the 0.7 threshold (corresponding to 50% identity) for only one of the two
258 reference A domains, then the substrate of the query A domain is matched to the substrate of
259 the more similar (lower distance) reference A domain. (2a) If the feature sequence distance is
260 below the 0.7 threshold for both reference A domains, then we considered the relative difference
261 of the query A domain towards the two reference A domains. If the relative difference is larger
262 than 0.2, the query A domain is matched to the substrate of the more similar reference A domain.
263 (2b) If the relative difference is smaller than 0.2, the substrate of the query A domain cannot
264 unambiguously be determined and is thus matched with both reference substrates. (3) If the
265 feature distance is above the 0.7 threshold (below 50% identity) for both reference A domains,
266 then the substrate of the query A domain is marked as “unknown”. For most query A domains,
267 rule (1) could be applied (17880 cases), whereas rules (2) and (3) had to be used rarely (133
268 and 279 cases, respectively). We applied our methodology termed “phylogeny-focused method”
269 to all following substrate and pyoverdine structure predictions.
270
271 **Section 3: Experimental validation of the annotation and prediction pipeline**
272 To test whether our bioinformatic pipeline can reliably predict pyoverdine structures, we

273 sequenced the genomes of 20 pyoverdine-producing *Pseudomonas* strains previously isolated
274 from soil and water[34]. We applied our annotation and prediction pipeline to generate predicted
275 pyoverdine structures for all 20 strains harboring a total of 237 A domains. Notably, none of the
276 predicted structures matched any of the 13 reference pyoverdine structures, and 9 out of the
277 20 structures were predicted to be novel (not yet characterized) pyoverdines. We then used the
278 UHPLC-HR-MS/MS method developed in 2022[35] to elucidate the chemical structures of these
279 pyoverdines. Six out of the 20 structures had already been elucidated in our previous work[35],
280 while we here used the same method to elucidate the chemical structure of the remaining 14
281 pyoverdines (see Supplementary Materials for MS/MS profiles).

282 We found a near-perfect match (94.4%) between the predicted and the observed
283 pyoverdine chemical structures and were able to accurately assign amino acids in 151 out of
284 160 cases (Figure 3d). Our method demonstrated a substantial improvement compared to
285 commonly used approaches or platforms integrating several methods (Supplementary_table6),
286 such as antiSMASH 7.0[16], which could only accurately assign 94 out of 160 amino acids
287 (accuracy: 58.8%). The nine non-matching cases in our analysis segregated into three groups.
288 In three cases (1.9%), our algorithm could not distinguish between the substrates Lysine and
289 Ornithine, as these two amino acids are highly similar both in terms of their chemical structures
290 and corresponding A domain sequences. This is the only sensitivity issue that is associated
291 with our approach. In four cases (2.5%), our technique assigned an “unknown” substrate to
292 amino acids that turned out to be valine, citrulline and histidine. Indeed, these three amino acids
293 were not yet present in our reference dataset and therefore no correct prediction was possible.
294 While citrulline has never been reported in pyoverdines before, valine has been found in a

295 pyoverdine structure but no sequence data from the corresponding strain was available. These
296 cases show that our analysis pipeline can be used to highlight new substrates. Once
297 experimentally verified, the new A domains and their substrates can expand the reference
298 dataset, allowing targeted improvement of our phylogeny-focused prediction technique. Finally,
299 there were only two cases (0.8%) that represented true mismatches between observed and
300 predicted amino acids. Altogether, our phylogeny-focused method is highly accurate in
301 predicting pyoverdine peptide structures and in identifying unknown substrates in
302 *Pseudomonas*.

303 To assess the extendibility of our pipeline, we chose *Burkholderiales* as a test case,
304 leveraging a manually curated dataset that we built up from the literature (see Methods for
305 detail). This dataset contains 203 A-domain sequences and corresponding substrate
306 information for 34 NRPS metabolites (e.g., toxins, biosurfactants, virulence factors) across 7
307 genera of *Burkholderiales* (Supplementary_table7). We further integrated experimentally
308 validated biosynthetic gene cluster data from *Burkholderiales* in MIBiG 3.0[33]
309 (Supplementary_table8) into our manual dataset leading to a total of 302 A-domains. We then
310 split the available data into a training set (124 A domains, Supplementary_table12) and a test
311 set (178 A domains, Supplementary_table9). As for the pyoverdines, we first trained our
312 algorithm on the training data set and then applied it to the test data set. We observed that our
313 pipeline consistently yields high prediction accuracy within *Burkholderiales* (83%), which is
314 considerably higher than the accuracy obtained with antiSMASH 7.0[16], (67%, Figure S3 and
315 Supplementary_table9). This outcome underscores the robustness of our developed prediction
316 algorithm, highlighting its potential extension to other NRPS secondary metabolites and

microbial genera.

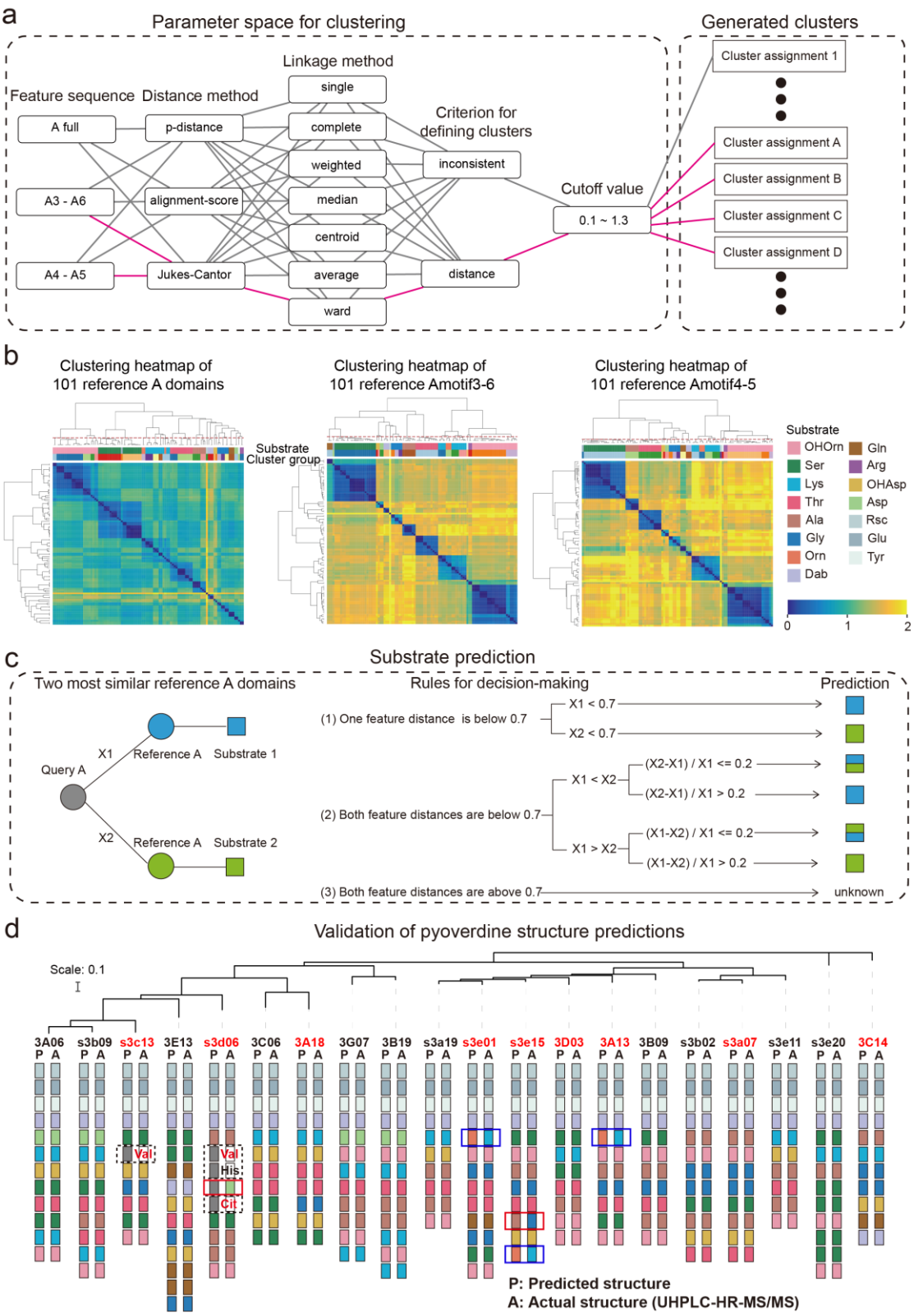


Figure 3 Phylogeny-focused substrate prediction for pyoverdine synthetase assembly lines. a. Information from 101 reference A domains with known amino acid substrates were used to

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. Color bar in the lower right shows the sequence distance. Left panel: complete A domain sequences. Middle panel: Amotif3-6 sequences. Right panel: Amotif4-5 sequences. The heatmaps show that hierarchical clustering, reliably associating sequence distances with substrates, worked best with the Amotif4-5. The experimentally validated substrates are shown on top of each heatmap and consist of 13 amino acids and 2 amino acid derivatives: Ser (Serine), Lys (Lysine), Thr (Threonine), Ala (Alanine), Gly (Glycine), Gln (Glutamine), Arg (Arginine), Asp (Aspartic acid), Glu (Glutamic acid), Tyr (Tyrosine), Orn (Ornithine), Rsc (succinic acid derivatives), Dab (2,4-diaminobutyric acid), OHAsp (Aspartic acid derivative), and OHOrn (a general term for three ornithine derivatives: FoOHOrn, AcOHOrn, cOHOrn). **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 sequence 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. Strains marked in red font indicate cases with novel (not yet characterized) pyoverdines structures. 151 out of the 160 substrates (94.4%) were correctly predicted. The nine inconsistencies are boxed in blue (Lysine and Ornithine are indistinguishable), in dashed black (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, for example, we cannot distinguish the three derivatives of ornithine (FoOHOrn, AcOHOrn, cOHOrn).

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 cases of valine and citrulline above). After considering the presence/absence of an E domain in each module, we derived the structures of 1664 pyoverdines according to our methods in section 2 (Figure 4).

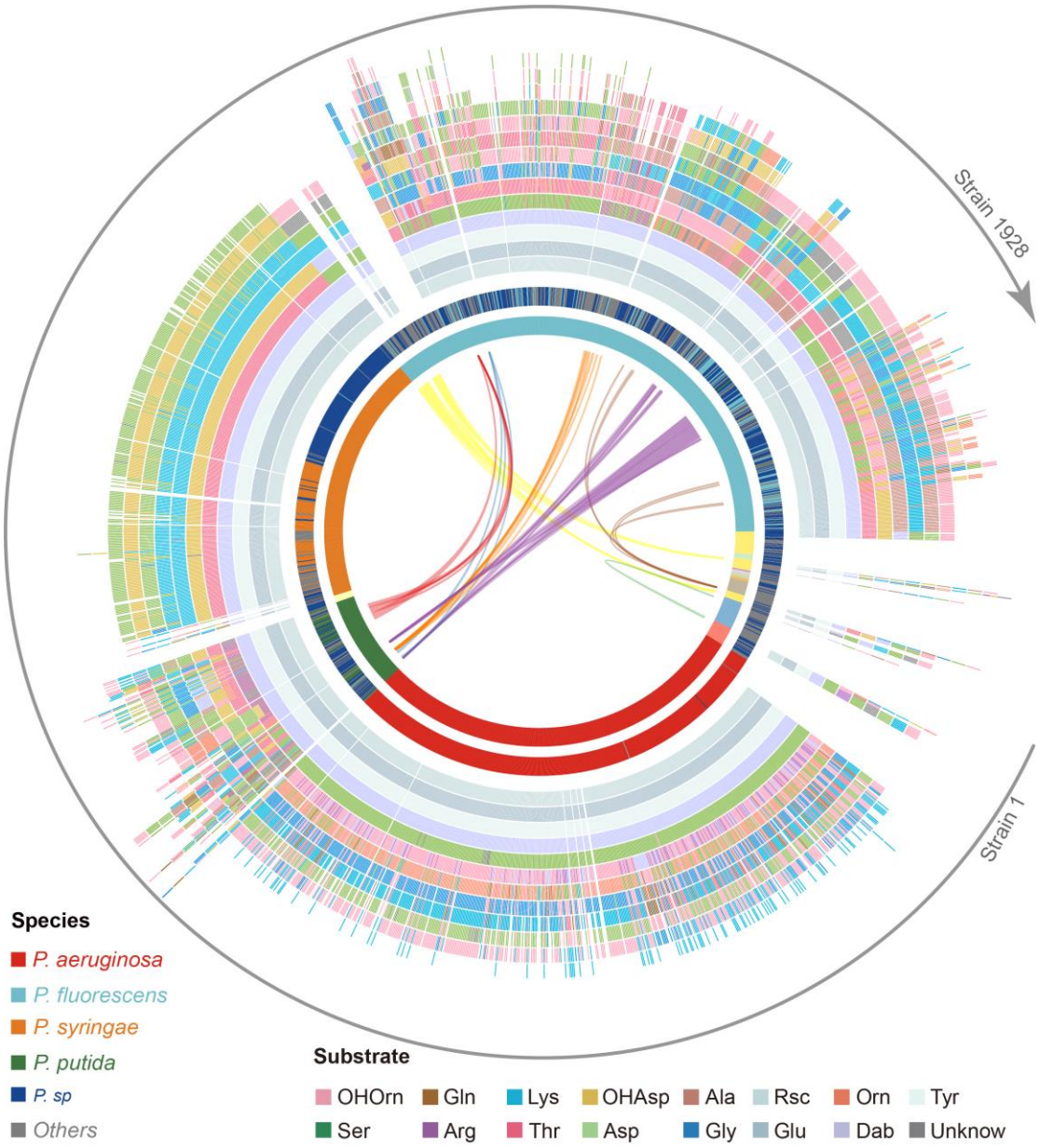


Figure 4 Predicted pyoverdine structural diversity based on our developed algorithm mapped onto the phylogenetic tree of all 1928 (non-redundant) *Pseudomonas* strains. In this figure, the strains are arranged based on the phylogenetic tree without showing the tree itself. The stacked boxes in the outermost 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 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 one 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.

372 Our prediction yielded 188 different pyoverdine molecules, out of which only 37 structures
373 had been previously reported. While these 37 reported structures were highly abundant across
374 strains (1103 out of 1664), our pipeline was powerful in identifying many of the rarer pyoverdine
375 variants. Agreeing with previous studies, we observed that the fluorophore is highly conserved
376 among the 188 predicted structures. Moreover, our analysis found that 13 amino acids form the
377 main structural substrates of all the 188 pyoverdine structures, with most of the variation being
378 attributable to different substrate combinations, peptide lengths, and substrate chirality (Figure
379 4).

380 Notably, pyoverdine structural diversity was not strongly linked to phylogeny because the
381 same pyoverdine structure could be found in completely unrelated species, while closely related
382 species often had different pyoverdine structures (Figure 4). These observations suggest that
383 there may be both frequent recombination and horizontal gene transfer of pyoverdine
384 synthetase clusters between species. Taken together, the bioinformatics methods developed in
385 our study can predict a suit of secondary metabolites (pyoverdines) from sequence data with
386 high accuracy, revealing an unprecedented richness and evolutionary history of siderophores
387 within pseudomonads and the discovery of 151 novel pyoverdine candidate variants.

388

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

391 In pseudomonads, iron-loaded pyoverdines are recognized by FpvA, a TonB-dependent
392 receptor, that transports the ferri-pyoverdine into the periplasm[19,36-38]. The protein structure of
393 characterized FpvA variants consists of three domains: The Secretin and TonB N-terminus short
394 domain (STN), the Plug domain (Plug), and the TonB-dependent receptor domain (TonB)[19].

395 While these domains are conserved across FpvA variants and other siderophore receptors,
396 there is substantial variation at the sequence level. This makes it challenging to reliably identify
397 FpvA receptors from sequence data by homologous search. As an example, we were unable
398 to find FpvA genes (with a 60% identity threshold) by homologous search in several genomes
399 although they had complete pyoverdine biosynthesis machineries. Moreover, there are many
400 other TonB-dependent receptors with fairly high sequence identity to FpvA but that transport
401 other siderophores than pyoverdine (e.g. FpvB, 55% identity, transporting pyoverdine,
402 ferrichrome and ferrioxamine B[39]). Therefore, it is imperative to develop a new comprehensive
403 method for identifying FpvA receptors in *Pseudomonas* genomes.

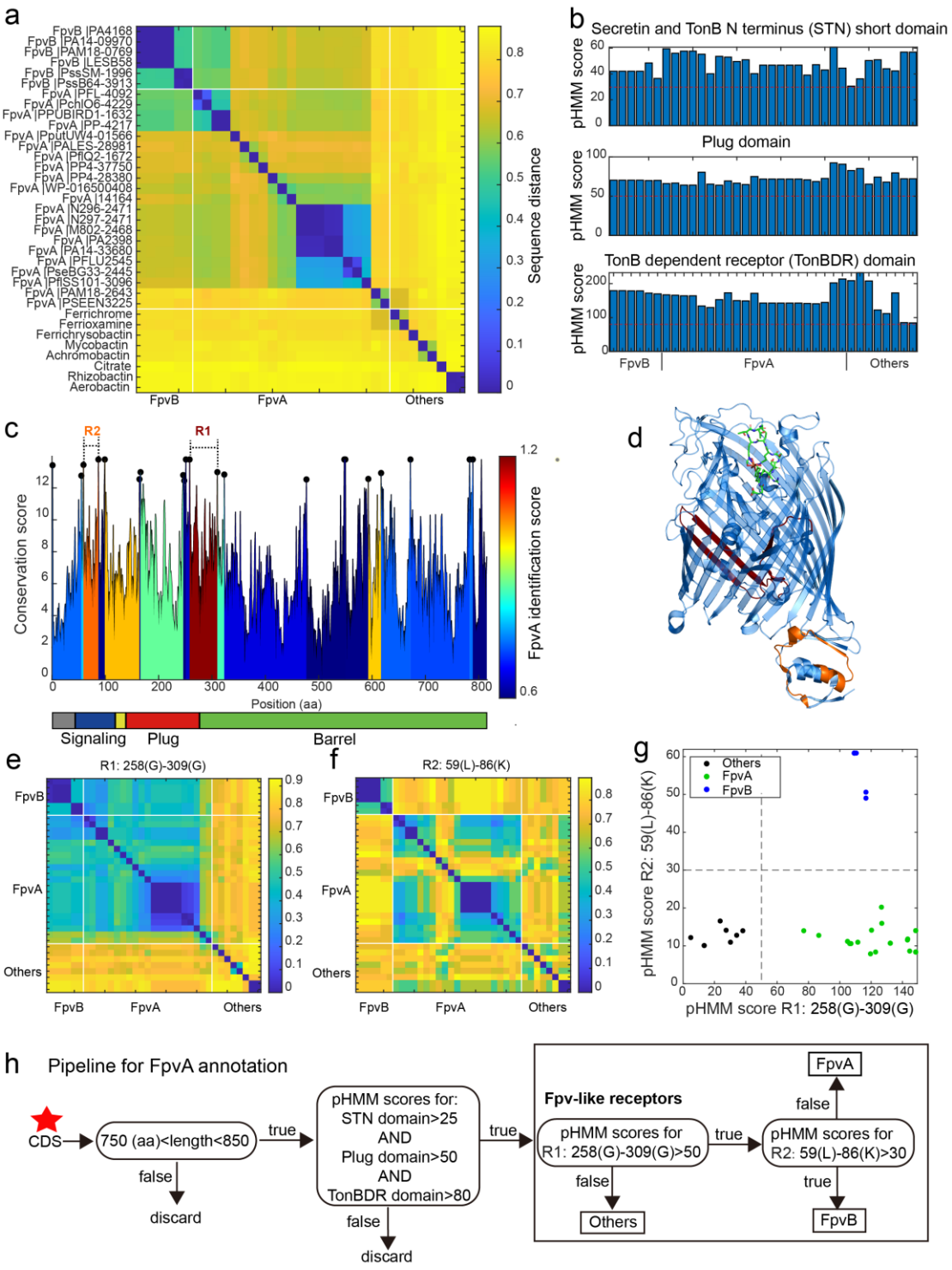
404 We started our approach by comparing the sequences of 35 reported siderophore
405 receptors, including 21 FpvA and 6 FpvB from The Pseudomonas Genome Database
406 (https://www.pseudomonas.com/), and 8 TonB-dependent siderophore receptor sequences
407 often found in *Pseudomonas* genomes from Uniprot (https://www.uniprot.org/), encoding
408 receptors for the uptake of heterologous siderophores (Supplementary_table10). We found that
409 all receptor sequences share a similar length of around 800 amino acids (FpvA and FpvB
410 sequences: 809 ± 10 amino acids). We then used the complete sequences to calculate the
411 pair-wise distances by global alignment before applying hierarchical clustering (Figure 5a). We
412 found substantial divergence between FpvA variants to an extent that was comparable to the
413 distance between FpvA and other siderophore receptors. Moreover, FpvB variants clustered
414 with FpvA variants, showing that FpvA identification based on full sequence distances is
415 unachievable. We hence focused on the three typical receptor domains (TonB, Plug, and STN,
416 retrieved from the Pfam database) and applied Profile Hidden Markov Models (pHMM) to each

417 reference sequence to calculate the pHMM probability scores for each domain. The probability

418 scores (calculated as the log-odd ratios for emission probabilities and log probabilities for state

419 transitions) had reasonably high scores but no distinction was apparent between the three

420 receptor classes (Figure 5b).



421

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 are 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. **g.** 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 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

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

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

457 Our analysis yielded two locations with noticeably high FpvA identification scores (Figure
458 5c). The region with the highest FpvA identification score (referred to as R1) locates at the
459 intersection of the Plug domain and the barrel structure of the TonB domain (Figure 5d, between
460 258 Gly and 309 Gly in the PAO1 FpvA). According to the sequence distance matrix, the R1
461 region allows to distinguish heterologous siderophore receptors from FpvA and FpvB receptors
462 (Figure 5e). The region with the second highest FpvA identification score (referred to as R2)
463 was located in the C-terminal signaling domain (Figure 5d, between 59 Leu and 86 Lys in the
464 PAO1 FpvA). The sequence distance matrix revealed that R2 allows to distinguish FpvB from
465 FpvA receptors (Figure 5f).

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

474 Based on these insights, we developed a decision flow chart for annotating FpvAs in
475 *Pseudomonas* genomes (Figure 5h). First, we considered sequences as Fpv-like receptors that
476 share similar properties to the ones identified in our reference database. Particularly, protein

477 coding sequence (CDS) length had to be between 750 and 850 amino acids and the phMM
478 scores for the three typical receptor domains STN, Plug, and TonB must be greater than 25, 50,
479 and 80, respectively (Figure 5b, red dashed lines). Second, we used the pHMM threshold
480 scores obtained for R1(FpvA) and R2(FpvB) (Figure 5g) to differentiate other siderophore
481 receptors (R1(FpvA) score < 50) from FpvB receptors (R1(FpvA) score > 50 and R2(FpvB)
482 score > 30) and FpvA receptors (R1(FpvA) score > 50 and R2(FpvB) score < 30). Our method
483 effectively identifies FpvA receptors from sequence data and can be readily applied to the entire
484 *Pseudomonas* dataset.
485
486 **Section 6: Application of the receptor annotation pipeline to the full dataset.**
487 The region-based receptor identification pipeline was applied to all 1928 *Pseudomonas*
488 genomes. The analysis identified 4547 FpvAs, 615 FpvBs, and 9139 other TonB-dependent
489 Fpv-like receptors across the dataset (Figure 6a). The 4547 FpvA sequences clustered
490 hierarchically into 114 groups, defined by an identity threshold of 60%. When comparing to the
491 21 reference FpvAs (Figure 6b), we found that 2293 FpvA sequences have close homologues
492 in the reference data base, while 2254 FpvA sequences lack such close homologues (sequence
493 identity < 50%). These latter sequences, termed as "dissimilar to reference", may represent
494 novel subtypes of FpvA receptors that could not be found by simple homology search. Our
495 analysis further shows that many strains have more than one FpvA receptor.
496 We then asked whether the 4547 FpvAs are found in proximity of pyoverdine Pep
497 synthetase genes as it is commonly the case for cognate FpvA receptors [40]. We thus calculated
498 the proximity between pyoverdine Pep genes and the Fpv-like receptor genes by counting the

499 number of base pairs between the two coding regions. All TonB-dependent receptors that have
500 not been classified as FpvAs were more than 20 kb away from the Pep genes (Figure 6c). In
501 contrast, 92% of the nearest FpvA genes were indeed located within 20 kb of their pyoverdine
502 Pep genes (Figure 6d, called proximate receptors). These proximate receptors encompassed
503 both those with close (66%) and more distant (34%) resemblance to the reference receptor
504 types. Overall, this proximity analysis confirmed that our region-based gene identification
505 method can reliably identify FpvA receptors.

506 Next, we explored the diversity among FpvA receptors in greater detail by focusing on the
507 1,534 strains with proximate receptors located within 20 kb of the pyoverdine Pep genes (Figure
508 6d). We utilized these high-confidence FpvAs for sequence feature extraction. When
509 considering the whole gene sequences, these receptors segregated into 44 groups according
510 to single-linkage clustering with an identity threshold of 60% (Figure S4a). To investigate which
511 sequence regions were the most informative for reliable clustering, we used a similar approach
512 as with FpvAs detection by quantifying the "group identification score" for variable regions
513 flanked by highly conserved sites. The higher the score, the stronger a region's capacity to
514 discriminate between FpvA groups. We found that the four regions with the top discrimination
515 capacities all located near the Plug domain surrounding the pyoverdine transmission channel
516 (Figure 6e). The plug domain is known to undergo conformational changes and is involved in
517 pyoverdine selectivity and import[41,42], suggesting that the four high-score regions are
518 responsible for pyoverdine specificity.

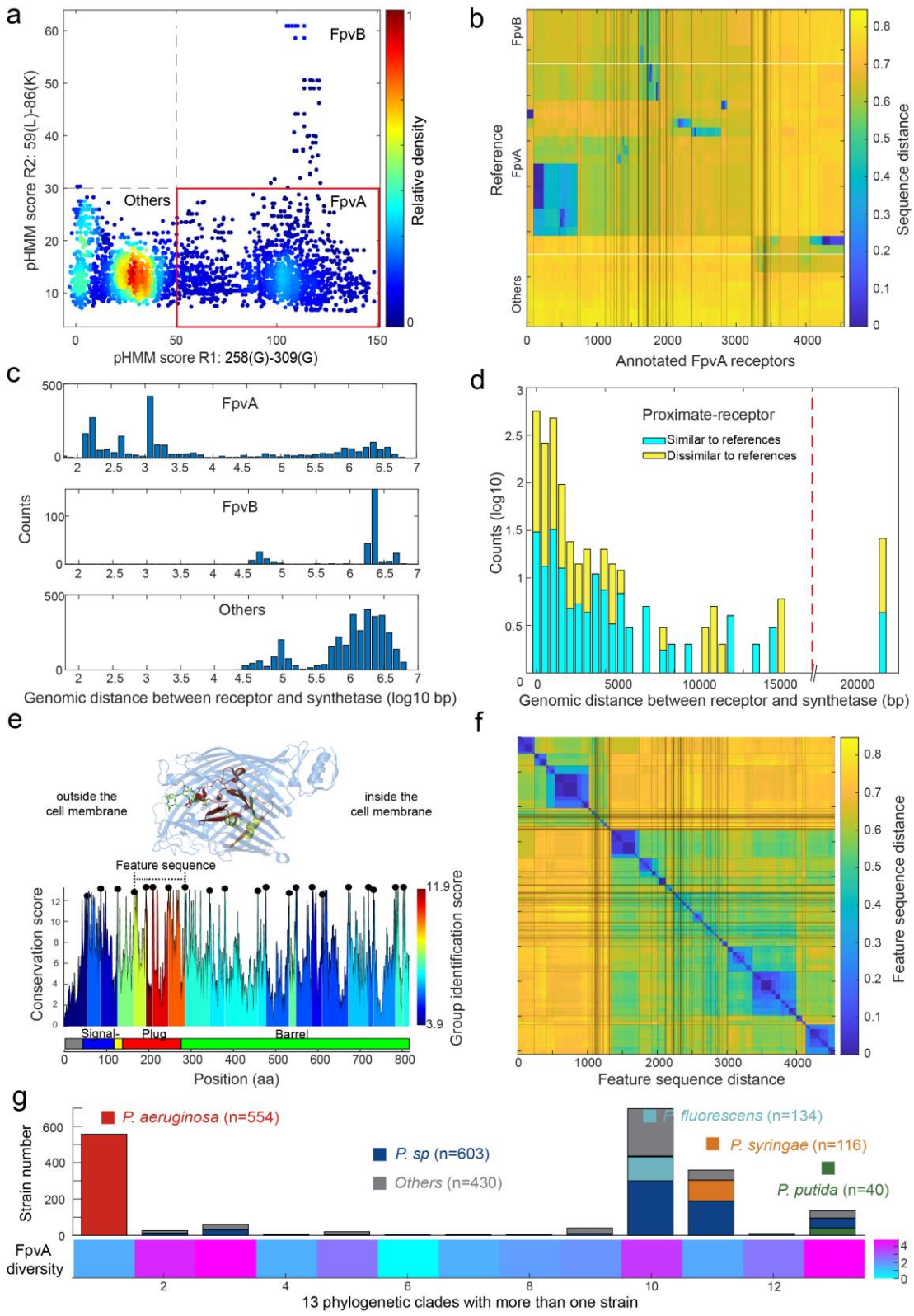


Figure 6 Application of the receptor annotation pipeline to the full database. a. Applying the receptor annotation pipeline to the genomes of the 1928 non-redundant *Pseudomonas* strains yields 14301 Fpv-like receptors, which segregate into 4547 FpvA receptors (red box), 615 FpvB receptors, and 9139 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 low sequence identity < 60% compared to the reference receptors, pointing at many 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 are 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, upper 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 S4a-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 pyoverdine receptors from genome data, revealing an undiscovered FpvA diversity that is unequally distributed across the different phylogenetic clades of pseudomonads.

We further extended the search for FpvA receptors, using our pipeline in Figure 5h, to the >30,000 complete bacterial genomes in NCBI. We found 31,936 receptors to pass the R1 and R2 thresholds (Figure S5), and surprisingly found that FpvA receptors seem to be widely dispersed across 12,944 strains spanning 13 phyla, 22 classes, 54 orders, 139 families, and 468 genera, and 2,598 species (Figure S6 and Supplementary_table11). While *Pseudomonas* remains the genus with the highest distribution of FpvAs, other genera with notable distributions comprise *Escherichia*, *Salmonella*, *Enterobacter*, *Acinetobacter*, *Xanthomonas*, *Bordetella*, and *Achromobacter*. While confirmatory experimental work is still pending, these analyses suggest that FpvA receptors for pyoverdine uptake may be widely distributed throughout the bacterial kingdom.

Discussion

The rapid expansion of sequencing data offers exciting opportunities for microbiology[43-45]. One key challenge of current research in the field is to infer biological functions of microbial communities from genome sequence data[46-48]. While this endeavor is increasingly successful for biological functions involving the primary metabolism and the associated complex metabolic

580 flux, reconstructing aspects of the secondary metabolism is much more challenging. The main
581 issue is that neither the function of a secondary metabolite enzyme nor the resulting metabolite
582 can be precisely predicted from gene sequence data. In our study, we tackled this challenge
583 and developed a bioinformatic pipeline to reconstruct the complete secondary metabolism
584 pathway of pyoverdines, a class of iron-scavenging siderophores produced by *Pseudomonas*
585 spp. These secondary metabolites are synthetized by a series of non-ribosomal peptide
586 synthetases and require a specific receptor (FpvA) for uptake. We combined knowledge-guided
587 learning with phylogeny-based methods to predict with high accuracy: (i) the full pyoverdine
588 assembly line, (ii) the substrate specificity for each enzyme within the assembly lines, (iii) the
589 complete chemical structure of pyoverdines, and (iv) the FpvA receptors from genome
590 sequences. After validation, we tested our pipeline with sequence data from 1664
591 phylogenetically distinct *Pseudomonas* strains and were able to determine 18,292 enzymatic A
592 domains involved in pyoverdine biosynthesis, to reliably predict 97.8% of their substrates, and
593 to identify 188 different pyoverdine molecule structures and 4547 FpvA receptor variants
594 belonging to 94 distinct groups. The uncovered diversity is stunning and goes far beyond
595 currently known levels of variation (73 pyoverdines and 3 FpvA groups).

596 We show that knowledge-guided learning is an extremely powerful tool to predict enzyme,
597 metabolite, and receptor properties. The establishment of our entire pipeline is based on only
598 101 previously known enzymatic A domains (from 13 known pyoverdine assembly lines) and
599 21 FpvA receptor sequences. Even with this limited amount of information, we were able to
600 predict the substrates of almost all the 18,292 enzymatic A domains and to identify 4547 FpvA
601 receptors from the sequence data. A key insight from our knowledge-guided learning is that

602 comparisons based on the full gene sequences (e.g., for pyoverdine synthetase or receptor)
603 are typically non-informative and unsuitable for obtaining functional information. This is because
604 overall diversity does not stand for functional diversity, meaning that A domains recognizing the
605 same substrate can diverge substantially in their full sequences. The same holds true for
606 receptor sequences: whole-sequence alignments can neither accurately identify FpvA
607 receptors nor reliably separate them into functional groups. Instead, it is imperative to extract
608 informative feature sequences that are defined as sequence stretches within a gene whose
609 diversity is tightly linked to variation in its functioning. We successfully extracted and applied
610 feature sequence comparisons for both A domain substrate prediction and FpvA identification.
611 It is important to note that a knowledge-guided pipeline does not have to be perfect right from
612 the start. For example, our pipeline for pyoverdine structure prediction returned unknowns for
613 several amino acid positions within the PEP. Our experimental verifications then revealed new
614 substrates such as valine and citrulline. This information can then be used to refine our
615 prediction algorithm in a feedback loop.

616 Another main advantage of our bioinformatic pipeline is that it can be applied to draft
617 genomes. This reflects a major improvement compared to existing annotation tools such as
618 antiSMASH[23], which typically has difficulties in recognizing NRPS structures in fragmented
619 genome assemblies. However, draft genomes are the most common data source in
620 microbiology. While our pipeline shows high performance, we need to acknowledge that we still
621 lose many genomes (6087 out of 9599, 63.4%). The reason for the loss is that the pyoverdine
622 biosynthesis machinery is large, which increases the probability that it is positioned at the end
623 of a contig. We decided to exclude those cases because the annotated biosynthesis machinery

624 might be truncated and thus incomplete. Thus, the high loss rate of draft genomes is rather due
625 to limitations in sequence quality (too many short contigs) and not due to a limitation of our
626 bioinformatic pipeline. We believe that this limitation may disappear in the future as long-read
627 sequencing technologies are quickly becoming cheaper and more reliable.

628 We further show that knowledge-guided learning combined with a phylogeny-focused
629 approach is a powerful tool for predicting the substrate specificity of A domains of synthetases.
630 It outperforms currently known bioinformatics prediction tools of NRPS substrates such as
631 antiSMASH[23]. Most current algorithms[49-53] perform poorly when applied to pyoverdines,
632 particularly when encountering non-proteogenic amino acids. The high accuracy of our
633 algorithm can largely be attributed to our reference set, composing only 13 pyoverdines from
634 *Pseudomonas spp*., yet capturing most of the substrate diversity. Similarly accurate
635 predictions based on a handful of known substrates among closely related species were
636 observed in several fungal NRPS systems[54]. It is worth noting that when the algorithm output
637 is "unknown," it signifies the presence of uncharacterized A domains that are not yet
638 incorporated into the reference data set. This should prompt researchers to pay attention to
639 these A domains, and like in our case, subject them to further experimental investigation. This
640 approach helped us discover new substrates (valine, citrulline) and link them to the
641 corresponding A domains. The novel substrates identified through our structural assessment
642 and mass spectrometry experiment can subsequently be used to enhance the precision of our
643 phylogeny-centered substrate prediction technique in the future, creating a progressive
644 feedback loop of expanding knowledge. Taken together, supervised learning based on a few
645 known compounds produced by species from the same genus can outperform generalized

646 prediction algorithms trained on many products from a diverse set of microbes for NRPS
647 substrate predictions.

648 Our results show that both pyoverdine and receptor diversity has been vastly
649 underestimated. While considerable pyoverdine diversity (n=73) has already been captured in
650 previous studies, here we predicted 151 putatively new variants. On the receptor side, the
651 uncovered novel diversity is more dramatic. One reason for this is that research on receptors
652 has mainly focused on the pathogen *P. aeruginosa*[19-22,55]. For this species, three different
653 pyoverdine types were described[21] together with three structurally different FpvA receptor types
654 that each recognize one of the pyoverdine types[22]. While our study confirmed that *P. aeruginosa*
655 strains (n = 554) indeed have only 3 pyoverdine-receptor types, we further discovered 91 new
656 FpvA groups among environmental *Pseudomonas* spp. Our findings raise the question why
657 there are so many different pyoverdine and receptor variants. One potential explanation is that
658 the benefit of specific siderophores could be context-dependent and locally adapted to the
659 multitude of different environmental conditions pseudomonads are exposed to. For example,
660 experimental work has revealed that pyoverdines can be cooperatively shared among strains
661 with matching receptors[37,56], or conversely, pyoverdines can serve as competitive agents by
662 locking away iron from species that have non-matching receptors[57]. Given that bioavailable iron
663 is limited in most natural and host-associated habitats[58-60], the unraveled functional diversity is
664 likely a direct evolutionary consequence of the struggle and competition of microbes for iron.
665 While experimental work is often restricted to a low number of strains, we propose that our
666 bioinformatic pipeline can be used to predict pyoverdine-mediated interaction networks across
667 thousands of strains and across different habitats. We have addressed this point in a

668 complementary study[61].

669 We are confident that our pipeline can be expanded to study iron competition in multi-
670 species communities and perhaps also in plant-microbe ecosystems, as siderophores exist
671 ubiquitously and are shared among microbes and certain hosts[62]. To move further, a key
672 question is whether our knowledge-guided approach can be applied to other microbes and
673 important secondary metabolites, such as antibiotics, toxins, biosurfactants and pigments? Our
674 preliminary investigation, focused on the curated dataset in *Burkholderiales* as a test case,
675 provides a highly promising response to this query. The outcomes from this evaluation
676 predominantly showcased the method's potential for extension to diverse microbial genera and
677 metabolites, given sufficiently accurate dataset. Meanwhile, we observed a notable decline in
678 accuracy (from 94% to 83%) when compared to pyoverdine. This indicates that our algorithm
679 possibly needs refinement, particularly concerning specific secondary metabolites or microbial
680 genera. As soon as sufficient case-by-case knowledge on a specific system is available, the
681 annotation strategies together with the feature sequence extraction and the phylogeny-focused
682 approach developed in our paper can be applied. It is worth noting that our current algorithm
683 does not predict tailoring reactions. While tailoring reactions are crucial for predicting the final
684 NRPS product structure, this challenge is not addressed by any existing pipelines and remains
685 an area for future research. In our study, it is important to note that the specificity of pyoverdines
686 is mainly determined by the backbone composition, with tailoring reactions playing a relatively
687 minor role[17]. For most of the secondary metabolites listed above, there are no receptors as the
688 compounds have purely extra-cellular functions, which substantially simplifies the development
689 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.

Method

Construction of phylogeny tree

The phylogenetic tree depicted in Figure 1e was constructed utilizing the PhyloPhlAn3 pipeline[30]. PhyloPhlAn is a comprehensive pipeline that encompasses the identification of phylogenetic markers, multiple sequence alignment, and the inference of phylogenetic trees. In this analysis, we employed 400 universal genes defined by PhyloPhlAn as our selected phylogeny markers. Subsequently, the taxonomic cladogram was generated using the iTOL web tool (http://huttenhower.sph.harvard.edu/galaxy/).

Application of our pipeline to identify errors in the substrate information associated with the A domain of *Pseudomonas* in the MIBiG database

We downloaded all A domain and substrate data for all NRPS secondary metabolism instances for *Pseudonomas* within MIBiG 3.0[33](Supplementary_table4). Following our pipeline, we initiated the process by utilizing NRPSMotifFinder to extract all Amotif4-5 sequences of the A domains. Subsequently, we calculated the Jukes-Cantor sequence distance between Amotif4.5 pairs. Finally, employing Ward linkage clustering, we conducted a cluster analysis on all Amotif4.5 sequences, resulting in the generation of Figure S2. We observed that within the same cluster, substrates often exhibit some subtle impurities. We speculate that such clusters might contain some errors in substrate information. Consequently, we focused on these potential errors, manually searched for the respective original literature, and verified that the A

713 domain and its corresponding substrate information entered in MIBiG indeed contain errors
714 (Supplementary_table5).
715
716 **Application of our pipeline to predict the structural composition of diverse NRPS**
717 **metabolites across various genera of *Burkholderiales***
718 To evaluate the extendibility of our bioinformatic pipeline to various metabolites and microbial
719 genera, we utilized a previously curated dataset of A domain sequences with experimentally
720 confirmed substrates for *Burkholderiales* [63] (Supplementary_table7). This manually curated
721 dataset, encompassing 7 genera, 34 secondary metabolite products, 203 A domains, and 25
722 substrates within *Burkholderiales*, served as a valuable resource and as a test set to validate
723 the extendibility of our bioinformatic approach. Simultaneously, we compiled a training set by
724 aggregating A domain and substrate data from all NRPS secondary metabolism instances
725 within *Burkholderiales*, utilizing experimentally validated biosynthetic gene cluster data from
726 MIBiG 3.0[33] (Supplementary_table8). This initial training set covered 6 genera, 19 secondary
727 products, 99 A domains, and 21 substrates. To ensure the inclusivity of all substrates from the
728 test set, we randomly selected an A domain from each substrate in the manually curated dataset,
729 incorporating it into our training set. Consequently, the finalized training set comprised 124 A
730 domains (Supplementary_table12), while the test set encompassed 178 A domains
731 (Supplementary_table9). Then, we concurrently employed our pipeline and antiSMASH[16] to
732 predict the corresponding substrate based on the sequence data of the A domain in the test set.
733 Subsequently, we compared these predictions with the actual substrate information, calculating
734 the accuracy rate for the two algorithms.

Application of our pipeline to annotate FpvA receptors across the bacterial domain

To explore the distribution of FpvA receptors across the bacterial domain, we acquired all complete bacterial genomes (33,207) from NCBI (https://www.ncbi.nlm.nih.gov). Employing the pipeline depicted in Figure 5h to examine the CDS of these genomes, we identified a total of 357,790 TonB-dependent receptors, of which 31,936 and 623 matched the criteria for being considered as FpvA and FpvB receptors, respectively. Subsequently, leveraging the taxonomy information from NCBI, we conducted a comparative analysis of the distribution differences between FpvA and FpvB among various species. The findings suggest that FpvA exhibit a widespread distribution across different species, whereas FpvB was exclusive to the genus *Pseudomonas*, with a maximum occurrence of one copy per genome.

Data Availability

The source code and parameters used are available in the supplementary material. The sequences of the 20 Pseudomonas strains are available on ENA accession PRJEB76792. One-click online notebook for the annotation and substrate prediction can be found here: https://drive.google.com/drive/folders/1JsfyPUGDTFo8BDDZk8JLSvKry8emzMhr?usp=drive_link , following a more detail explanation on the code.

Acknowledgements

We thank Richard Allen and Simon Maréchal for the handling of the genome sequencing of the

20 *Pseudomonas* strains used for the structure validation experiment. We further thank Vera Vollenweider for sample preparation for the experiment elucidating pyoverdine structures.

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, No. T2321001), National Postdoctoral Program for Innovative Talents (No. BX2021012). R.K. is supported 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 pyoverdine structure elucidation experiments. Di Zhang depicted the circular plot. Ruolin He standardized all NRPS structures to the motif-intermotif format. Jiqi Shao and Ruichen Xu assisted in revising the manuscript. Xiaoying Bian provided curated dataset of A domain sequences with experimentally confirmed substrates from *Burkholderiales*.

Xiaoying Bian, 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 Chaturvedi, K. S., Hung, C. S., Crowley, J. R., Stapleton, A. E. & Henderson, J. P. The siderophore yersiniabactin binds copper to protect pathogens during infection. *Nature Chemical Biology* **8**, 731-736, doi:10.1038/nchembio.1020 (2012).

10 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).

11 Yee, D. A. *et al.* Genome mining for unknown–unknown natural products. *Nature Chemical Biology* **19**, 633-640, doi:10.1038/s41589-022-01246-6 (2023).

12 Lautru, S., Deeth, R. J., Bailey, L. M. & Challis, G. L. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. *Nature Chemical Biology* **1**, 265-269, doi:10.1038/nchembio731 (2005).

821 13 Andryukov, B., Mikhailov, V. & Besednova, N. The Biotechnological Potential of
822 Secondary Metabolites from Marine Bacteria. *Journal of Marine Science and*
823 *Engineering* **7** (2019).
824 14 Kautsar, S. A., Blin, K., Shaw, S., Weber, T. & Medema, M. H. BiG-FAM: the biosynthetic
825 gene cluster families database. *Nucleic Acids Research* **49**, D490-D497,
826 doi:10.1093/nar/gkaa812 (2021).
827 15 He, R. *et al.* Knowledge-guided data mining on the standardized architecture of NRPS:
828 Subtypes, novel motifs, and sequence entanglements. *PLOS Computational Biology*
829 **19**, e1011100, doi:10.1371/journal.pcbi.1011100 (2023).
830 16 Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation,
831 chemical structures and visualisation. *Nucleic Acids Research* **51**, W46-W50,
832 doi:10.1093/nar/gkad344 (2023).
833 17 Ringel, M. T. & Brüser, T. The biosynthesis of pyoverdines. *Microb Cell* **5**, 424-437,
834 doi:10.15698/mic2018.10.649 (2018).
835 18 Hopkinson, B. M. & Morel, F. M. M. The role of siderophores in iron acquisition by
836 photosynthetic marine microorganisms. *BioMetals* **22**, 659-669, doi:10.1007/s10534-
837 009-9235-2 (2009).
838 19 Cobessi, D. *et al.* The Crystal Structure of the Pyoverdine Outer Membrane Receptor
839 FpvA from Pseudomonas aeruginosa at 3.6Å Resolution. *Journal of Molecular Biology*
840 **347**, 121-134, doi:10.1016/j.jmb.2005.01.021 (2005).
841 20 Diggle, S. P. & Whiteley, M. Microbe Profile: Pseudomonas aeruginosa: opportunistic
842 pathogen and lab rat. *Microbiology (Reading)* **166**, 30-33, doi:10.1099/mic.0.000860

843 (2020).

844 21 Meyer, J.-M. *et al.* Use of Siderophores to Type Pseudomonads: The Three
845 Pseudomonas Aeruginosa Pyoverdine Systems. *Microbiology* **143**, 35-43,
846 doi:10.1099/00221287-143-1-35 (1997).

847 22 Bodilis, J. *et al.* Distribution and evolution of ferripyoverdine receptors in Pseudomonas
848 aeruginosa. *Environmental Microbiology* **11**, 2123-2135, doi:10.1111/j.1462-
849 2920.2009.01932.x (2009).

850 23 Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining
851 pipeline. *Nucleic Acids Research* **47**, W81-W87, doi:10.1093/nar/gkz310 (2019).

852 24 Kümmerli, R. Iron acquisition strategies in pseudomonads: mechanisms, ecology, and
853 evolution. *BioMetals*, doi:10.1007/s10534-022-00480-8 (2022).

854 25 Meyer, J. M. Pyoverdines: pigments, siderophores and potential taxonomic markers of
855 fluorescent Pseudomonas species. *Arch Microbiol* **174**, 135-142,
856 doi:10.1007/s002030000188 (2000).

857 26 Xu, Z., Park, T. J. & Cao, H. Advances in mining and expressing microbial biosynthetic
858 gene clusters. *Crit Rev Microbiol*, 1-20, doi:10.1080/1040841x.2022.2036099 (2022).

859 27 Keller, N. P. Fungal secondary metabolism: regulation, function and drug discovery.
860 *Nature Reviews Microbiology* **17**, 167-180, doi:10.1038/s41579-018-0121-1 (2019).

861 28 Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Research* **32**,
862 D138-D141, doi:10.1093/nar/gkh121 (2004).

863 29 Winsor, G. L. *et al.* Enhanced annotations and features for comparing thousands of
864 Pseudomonas genomes in the Pseudomonas genome database. *Nucleic Acids*

865 *Research* **44**, D646-D653, doi:10.1093/nar/gkv1227 (2016).

866 30 Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and genomes from
867 metagenomes using PhyloPhlAn 3.0. *Nature Communications* **11**, 2500,
868 doi:10.1038/s41467-020-16366-7 (2020).

869 31 Felnagle, E. A. *et al.* Nonribosomal peptide synthetases involved in the production of
870 medically relevant natural products. *Mol Pharm* **5**, 191-211, doi:10.1021/mp700137g
871 (2008).

872 32 Süssmuth, R. D. & Mainz, A. Nonribosomal Peptide Synthesis—Principles and
873 Prospects. *Angewandte Chemie International Edition* **56**, 3770-3821,
874 doi:10.1002/anie.201609079 (2017).

875 33 Terlouw, B. R. *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally
876 validated biosynthetic gene clusters. *Nucleic Acids Research* **51**, D603-D610,
877 doi:10.1093/nar/gkac1049 (2023).

878 34 Butaitė, E., Kramer, J., Wyder, S. & Kümmerli, R. Environmental determinants of
879 pyoverdine production, exploitation and competition in natural Pseudomonas
880 communities. *Environmental Microbiology* **20**, 3629-3642, doi:10.1111/1462-
881 2920.14355 (2018).

882 35 Rehm, K., Vollenweider, V., Kümmerli, R. & Bigler, L. A comprehensive method to
883 elucidate pyoverdines produced by fluorescent Pseudomonas spp. by UHPLC-HR-
884 MS/MS. *Analytical and Bioanalytical Chemistry* **414**, 2671-2685, doi:10.1007/s00216-
885 022-03907-w (2022).

886 36 Butaitė, E., Baumgartner, M., Wyder, S. & Kümmerli, R. Siderophore cheating and

887 cheating resistance shape competition for iron in soil and freshwater Pseudomonas
888 communities. *Nature Communications* **8**, 414, doi:10.1038/s41467-017-00509-4
889 (2017).
890 37 Kramer, J., Özkaya, Ö. & Kümmerli, R. Bacterial siderophores in community and host
891 interactions. *Nature Reviews Microbiology* **18**, 152-163, doi:10.1038/s41579-019-
892 0284-4 (2020).
893 38 Jin, Z. *et al.* Conditional privatization of a public siderophore enables Pseudomonas
894 aeruginosa to resist cheater invasion. *Nature Communications* **9**, 1383,
895 doi:10.1038/s41467-018-03791-y (2018).
896 39 Chan, D. C. K. & Burrows, L. L. Pseudomonas aeruginosa FpvB
897 is a high-affinity transporter for xenosiderophores ferrichrome and ferrioxamine B.
898 *bioRxiv*, 2022.2009.2020.508722, doi:10.1101/2022.09.20.508722 (2022).
899 40 González, J. *et al.* Loss of a pyoverdine secondary receptor in Pseudomonas
900 aeruginosa results in a fitter strain suitable for population invasion. *The ISME Journal*
901 **15**, 1330-1343, doi:10.1038/s41396-020-00853-2 (2021).
902 41 Schalk, I. J., Mislin, G. L. & Brillet, K. Structure, function and binding selectivity and
903 stereoselectivity of siderophore-iron outer membrane transporters. *Curr Top Membr* **69**,
904 37-66, doi:10.1016/B978-0-12-394390-3.00002-1 (2012).
905 42 Greenwald, J. *et al.* FpvA bound to non-cognate pyoverdines: molecular basis of
906 siderophore recognition by an iron transporter. *Molecular Microbiology* **72**, 1246-1259,
907 doi:10.1111/j.1365-2958.2009.06721.x (2009).
908 43 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).

44 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).

45 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).

46 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).

47 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).

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

49 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).

50 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).

51 Prieto, C., García-Estrada, C., Lorenzana, D. & Martín, J. F. NRPSsp: non-ribosomal

931 peptide synthase substrate predictor. *Bioinformatics* **28**, 426-427,
932 doi:10.1093/bioinformatics/btr659 (2012).
933 52 Röttig, M. *et al.* NRPSpredictor2—a web server for predicting NRPS adenylation
934 domain specificity. *Nucleic Acids Research* **39**, W362-W367, doi:10.1093/nar/gkr323
935 (2011).
936 53 Zierep, P. F., Ceci, A. T., Dobrusin, I., Rockwell-Kollmann, S. C. & Günther, S. SeMPI
937 2.0-A Web Server for PKS and NRPS Predictions Combined with Metabolite Screening
938 in Natural Product Databases. *Metabolites* **11**, doi:10.3390/metabo11010013 (2020).
939 54 Fan, J. *et al.* Biosynthetic diversification of peptaibol mediates fungus-mycohost
940 interactions. *bioRxiv* (2022).
941 55 Smith, E. E., Sims, E. H., Spencer, D. H., Kaul, R. & Olson, M. V. Evidence for
942 diversifying selection at the pyoverdine locus of Pseudomonas aeruginosa. *J Bacteriol*
943 **187**, 2138-2147, doi:10.1128/jb.187.6.2138-2147.2005 (2005).
944 56 Gu, S. *et al.* Competition for iron drives phytopathogen control by natural rhizosphere
945 microbiomes. *Nature Microbiology* **5**, 1002-1010, doi:10.1038/s41564-020-0719-8
946 (2020).
947 57 Figueiredo, A. R. T., Özkaya, Ö., Kümmerli, R. & Kramer, J. Siderophores drive invasion
948 dynamics in bacterial communities through their dual role as public good versus public
949 bad. *Ecology Letters* **25**, 138-150, doi:10.1111/ele.13912 (2022).
950 58 Andrews, S. C., Robinson, A. K. & Rodríguez-Quiñones, F. Bacterial iron homeostasis.
951 *FEMS Microbiology Reviews* **27**, 215–237, doi:10.1016/s0168-6445(03)00055-x
952 (2003).

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

60 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).

61 Gu, S. *et al.* Siderophore synthetase-receptor gene coevolution reveals habitat- and pathogen-specific bacterial iron interaction networks. *bioRxiv*, 2023.2011.2005.565711, doi:10.1101/2023.11.05.565711 (2024).

62 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).

63 Chen, H. *et al.* Biosynthesis and engineering of the nonribosomal peptides with a C-terminal putrescine. *Nature Communications* **14**, 6619, doi:10.1038/s41467-023-42387-z (2023).

Supplementary figures

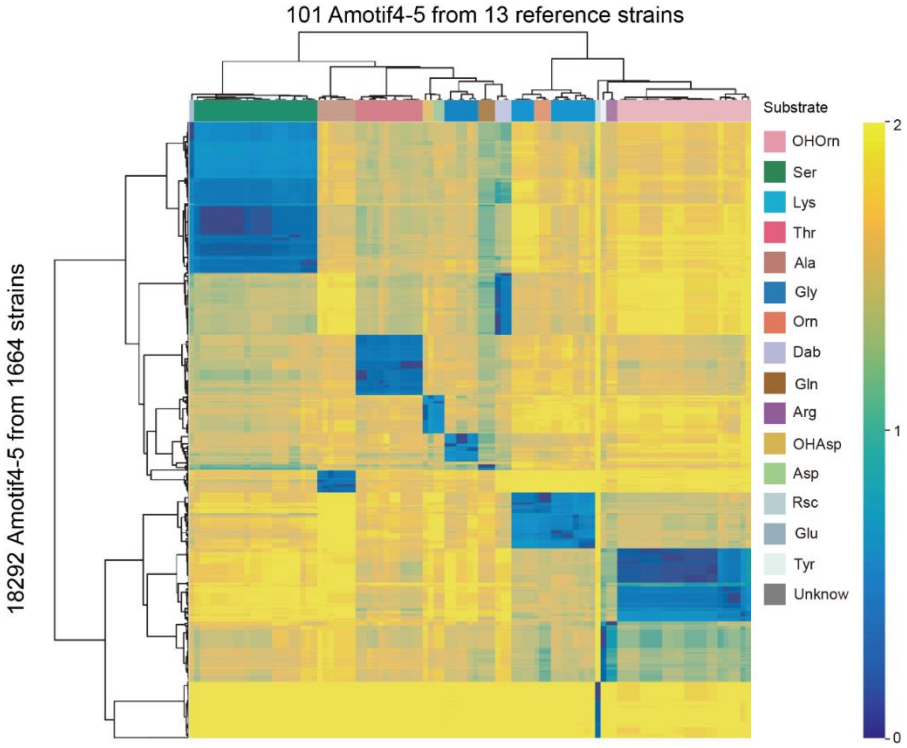


Figure S1 Clustering heatmap of sequence distance matrix between A domains from 1664 strains and 101 reference A domains. The color on top indicates the substrate of reference A domains.

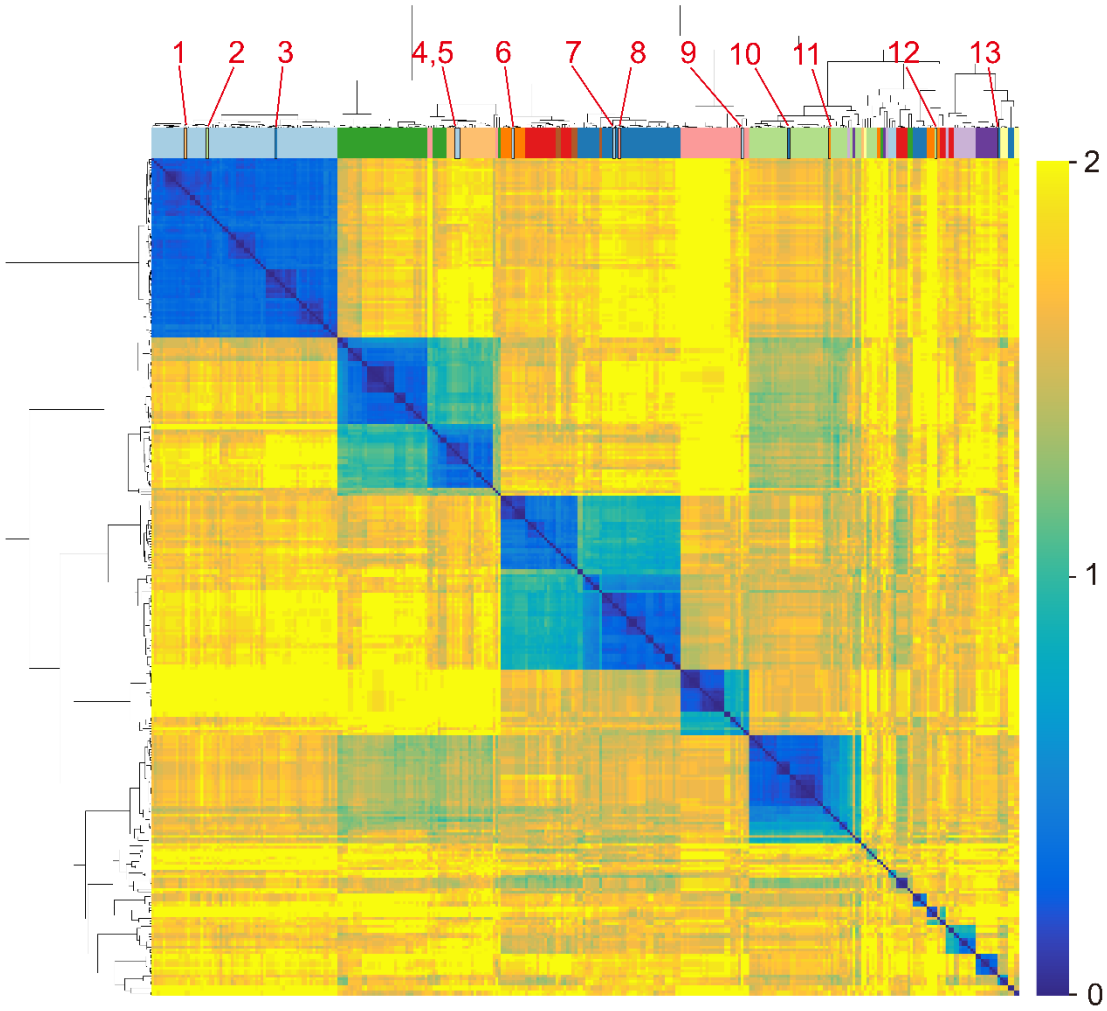


Figure S2 Clustering heatmap of sequence distance matrix between 318 Amotif4-5 from *Pseudomonas* strains in the MIBiG database (Supplementary_table4). MIBiG is a standard annotation database of biosynthetic gene clusters. We retrieved all secondary metabolites produced by members of the *Pseudomonas* genus from MIBiG. The colors on top of the heatmap indicate the substrate of the A domains. We focus on the domains with closed sequence distance but with distinct substrate specificity (marked with red text). After checking the original reference, theses domains were labeled incorrectly in MIBiG (Supplementary_table5).

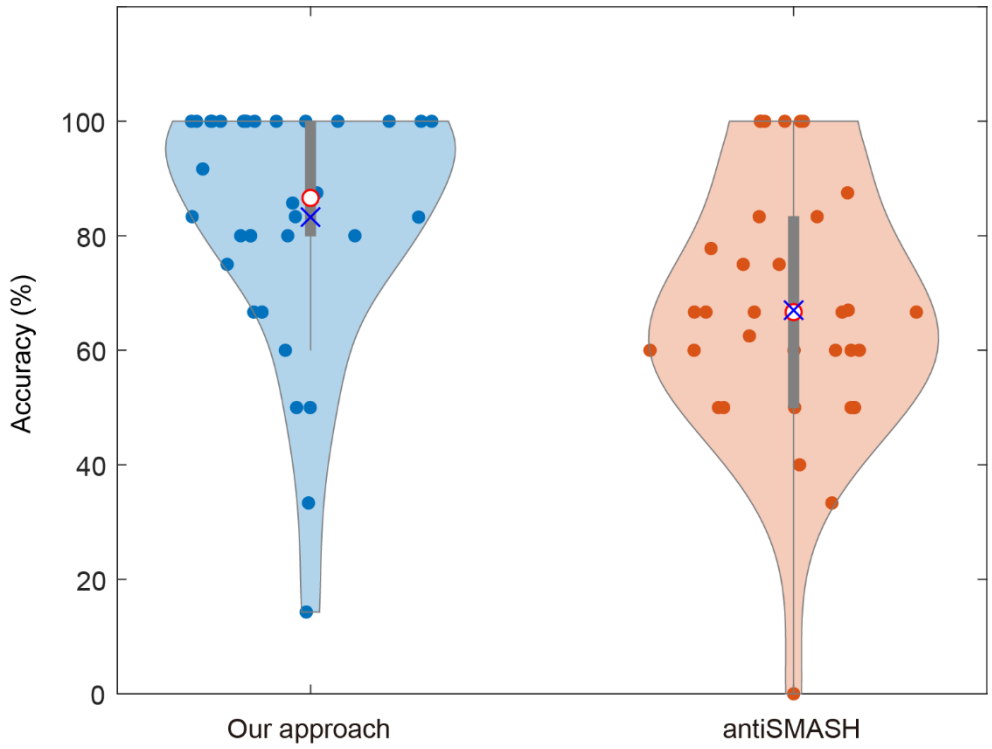


Figure S3 Comparison of prediction accuracy in the structural composition of secondary metabolites between our approach and antiSMASH in *Burkholderiales*. Each data point in the violin plot symbolizes a distinct secondary metabolite. The median is denoted by a red circle, while the mean is represented by a blue '×'.

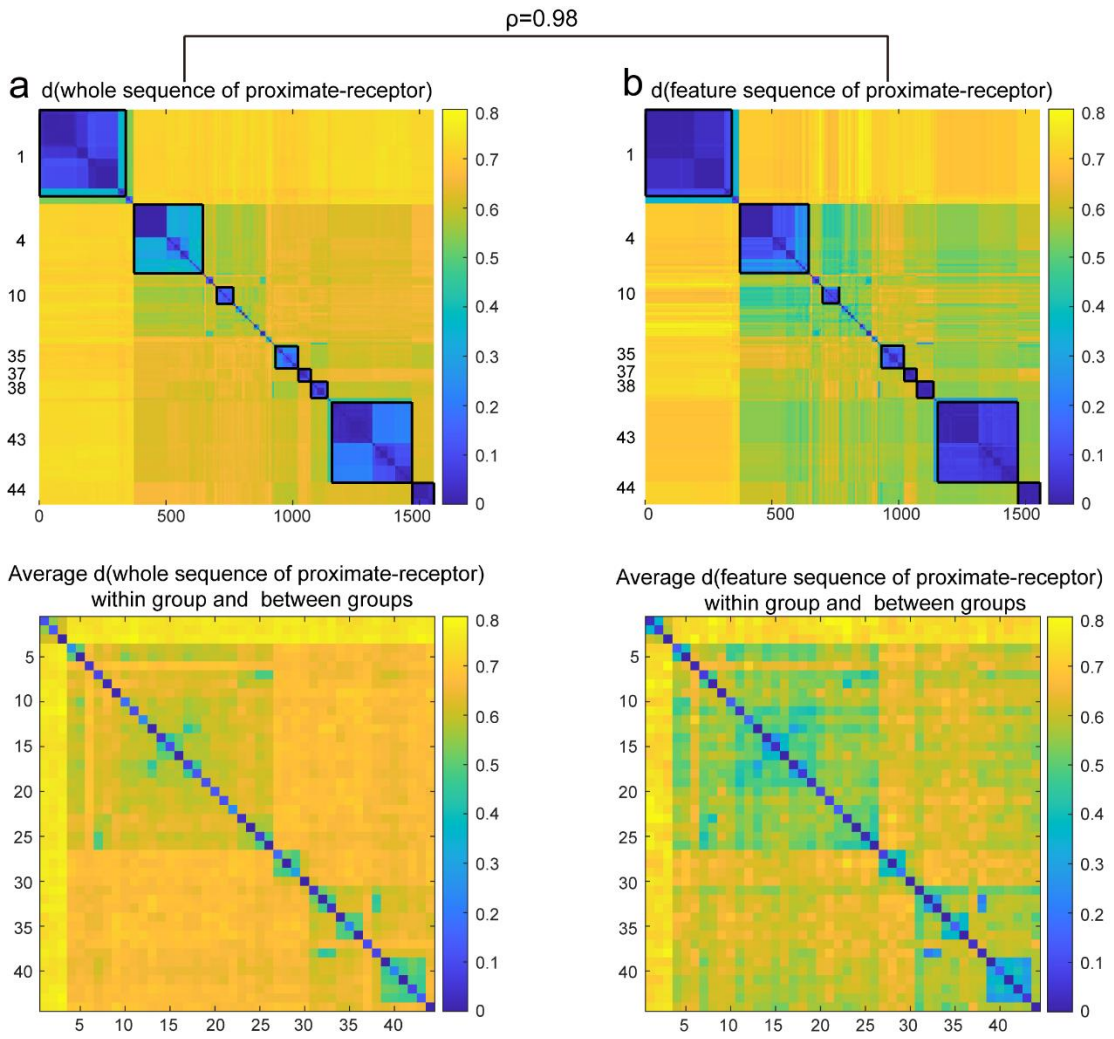


Figure S4 Comparison of clustering and grouping effects of complete sequences (a) and feature sequences (b) of all proximate-receptors.

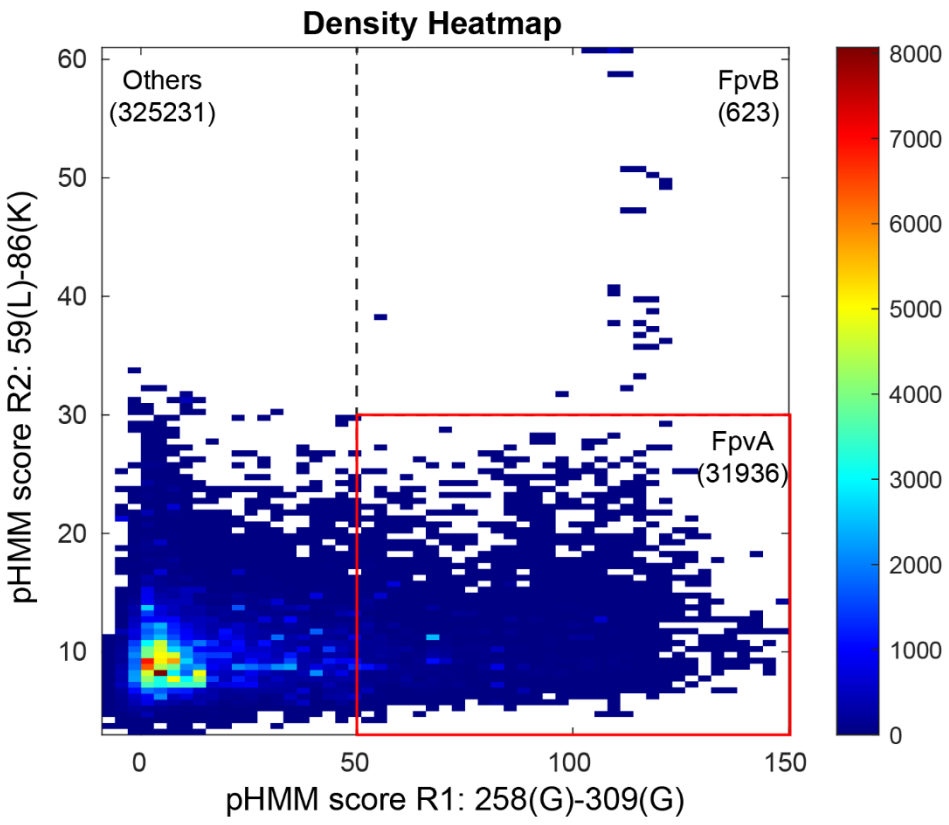


Figure S5 Applying the receptor annotation pipeline to >30,000 complete bacterial genomes in NCBI based on the pHMM score thresholds for regions R1 and R2. The heatmap indicates receptor density.

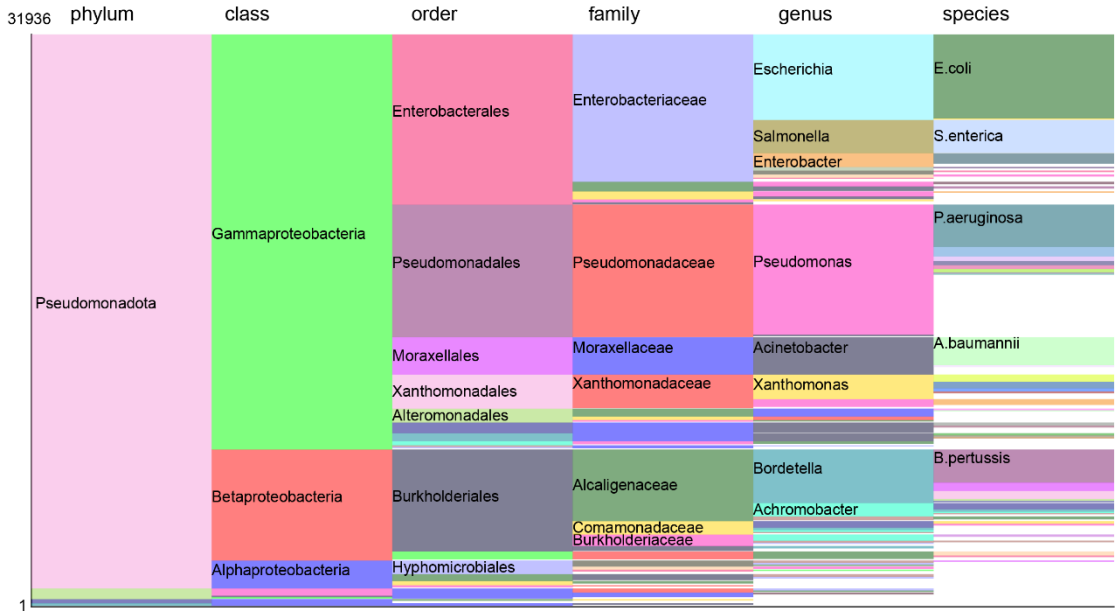


Figure S6 The distribution of 31,936 FpvAs spans various taxonomic categories. Taxa with an abundance exceeding 1/500 are highlighted with colors, while others remain white. Taxa surpassing an abundance of 1/50 will have their specific names marked.