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

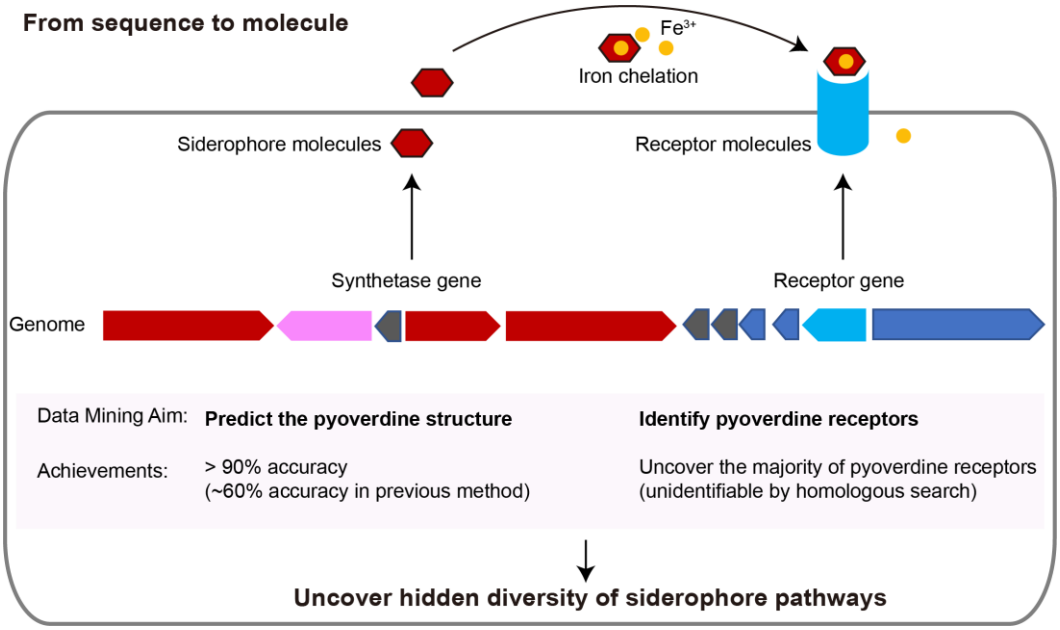
[8] University of Zurich, Department of Quantitative Biomedicine, Winterthurerstr. 190, 8057 Zurich, Switzerland

[*] 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

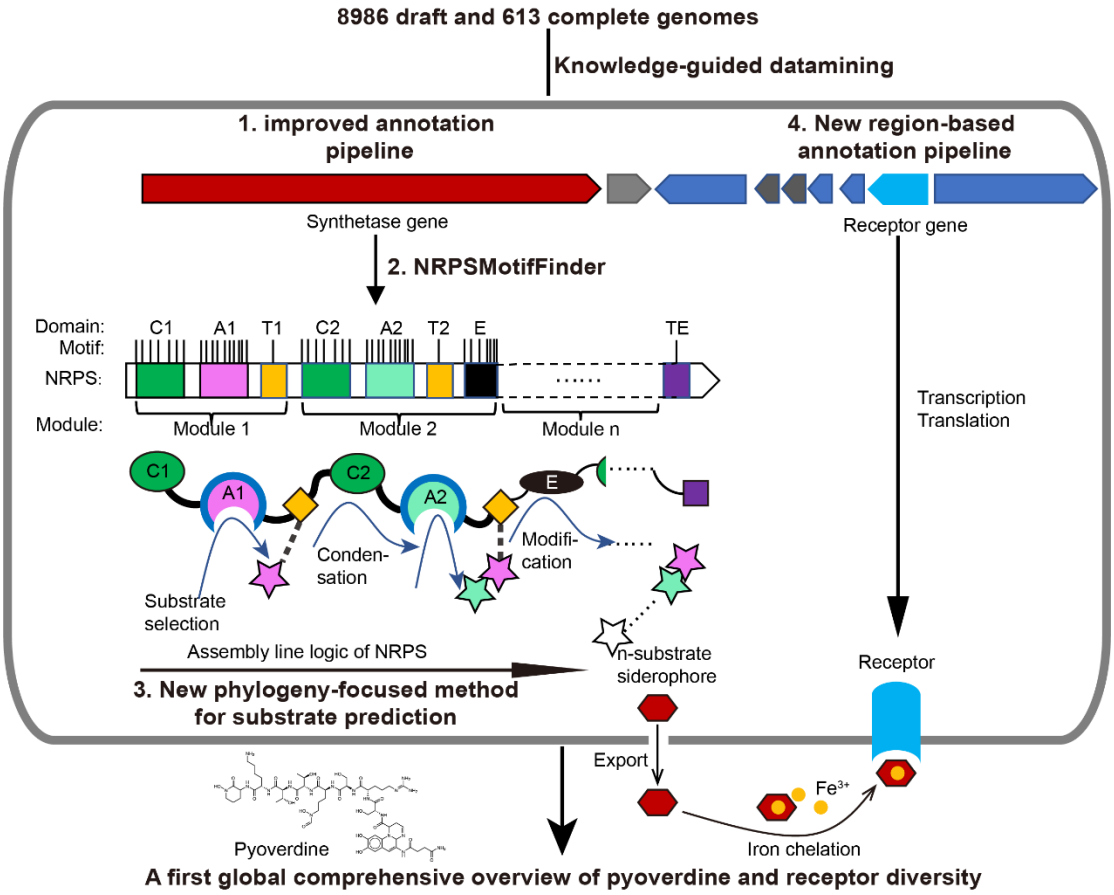
84 substrates[18]. Based on this knowledge, our pipeline entails the following steps (Figure 1): (i)

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

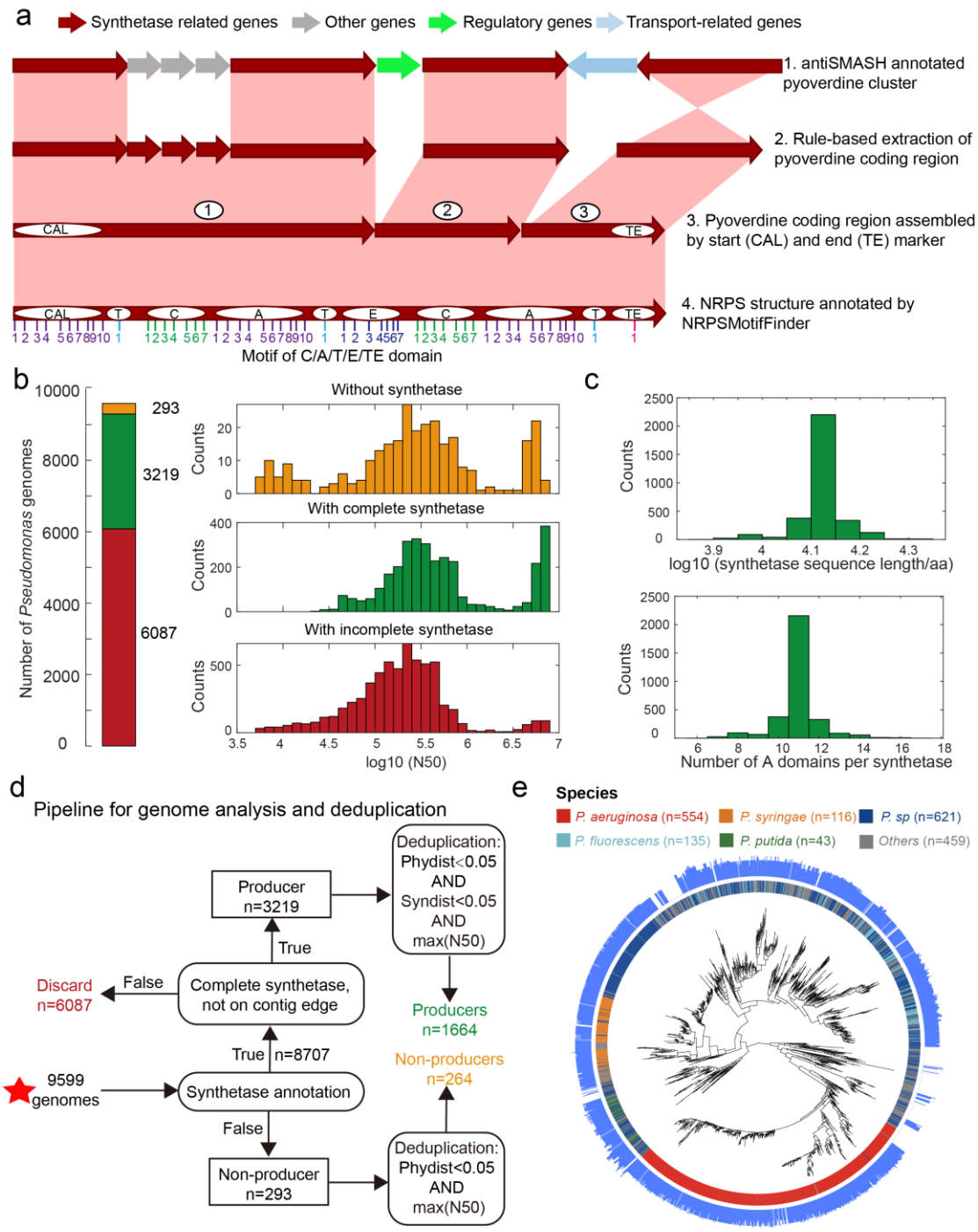


Figure 2 Improved annotation pipeline reveals a vast diversity of pyoverdine synthetase genes. a. Improved annotation pipeline based on the raw annotation from antiSMASH[23]. **b.** The annotation pipeline was applied to 9599 *Pseudomonas* genomes (94% draft genomes). Genomes could be separated into 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 highlight
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) and
218 completely sequenced synthetase genes were available (Supplementary_table2). In order to
219 make reliable molecule predictions, two challenges must be addressed: (i) the extraction of
220 relevant information from A domain sequences for which the substrate is known, and (ii) the
221 effective application of this information to predict specificity of A domain sequences for which
222 the substrate is unknown.

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

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

245 To address the second challenge, we developed a "phylogeny-focused method" to apply
246 the above-derived feature sequence distance approach to the 18,292 discovered A domains.
247 We realized that a direct construction of a phylogenetic tree including all 18,292 query A
248 domains and the 101 reference A domains would be computationally too demanding and
249 impossible to scale up. Furthermore, such an approach would result in phylogeny-interference
250 issues, where domains would cluster not only based on their substrate similarities but also

251 based on overall species relatedness[15]. To minimize the effect of phylogeny and speed up
252 calculations, we took each of the 18,292 query A domains and identified the two most similar A
253 domain clusters among our 101 reference A domains. We then compared the feature distance
254 between each query A domain and the two most similar reference A domains and assigned a
255 substrate to the query A domain using the following rules (Figure 3c). (1) If the feature sequence
256 distance is below the 0.7 threshold (corresponding to 50% identity) for only one of the two
257 reference A domains, then the substrate of the query A domain is matched to the substrate of
258 the more similar (lower distance) reference A domain. (2a) If the feature sequence distance is
259 below the 0.7 threshold for both reference A domains, then we considered the relative difference
260 of the query A domain towards the two reference A domains. If the relative difference is larger
261 than 0.2, the query A domain is matched to the substrate of the more similar reference A domain.
262 (2b) If the relative difference is smaller than 0.2, the substrate of the query A domain cannot
263 unambiguously be determined and is thus matched with both reference substrates. (3) If the
264 feature distance is above the 0.7 threshold (below 50% identity) for both reference A domains,
265 then the substrate of the query A domain is marked as “unknown”. For most query A domains,
266 rule (1) could be applied (17880 cases), whereas rules (2) and (3) had to be used rarely (133
267 and 279 cases, respectively). We applied our methodology termed “phylogeny-focused method”
268 to all following substrate and pyoverdine structure predictions.
269
270 **Section 3: Experimental validation of the annotation and prediction pipeline**
271 To test whether our bioinformatic pipeline can reliably predict pyoverdine structures, we
272 sequenced the genomes of 20 pyoverdine-producing *Pseudomonas* strains previously isolated

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

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

295 that our analysis pipeline can be used to highlight new substrates. Once experimentally verified,
296 the new A domains and their substrates can expand the reference dataset, allowing targeted
297 improvement of our phylogeny-focused prediction technique. Finally, there were only two cases
298 (0.8%) that represented true mismatches between observed and predicted amino acids.
299 Altogether, our phylogeny-focused method is highly accurate in predicting pyoverdine peptide
300 structures and in identifying unknown substrates in *Pseudomonas*.

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

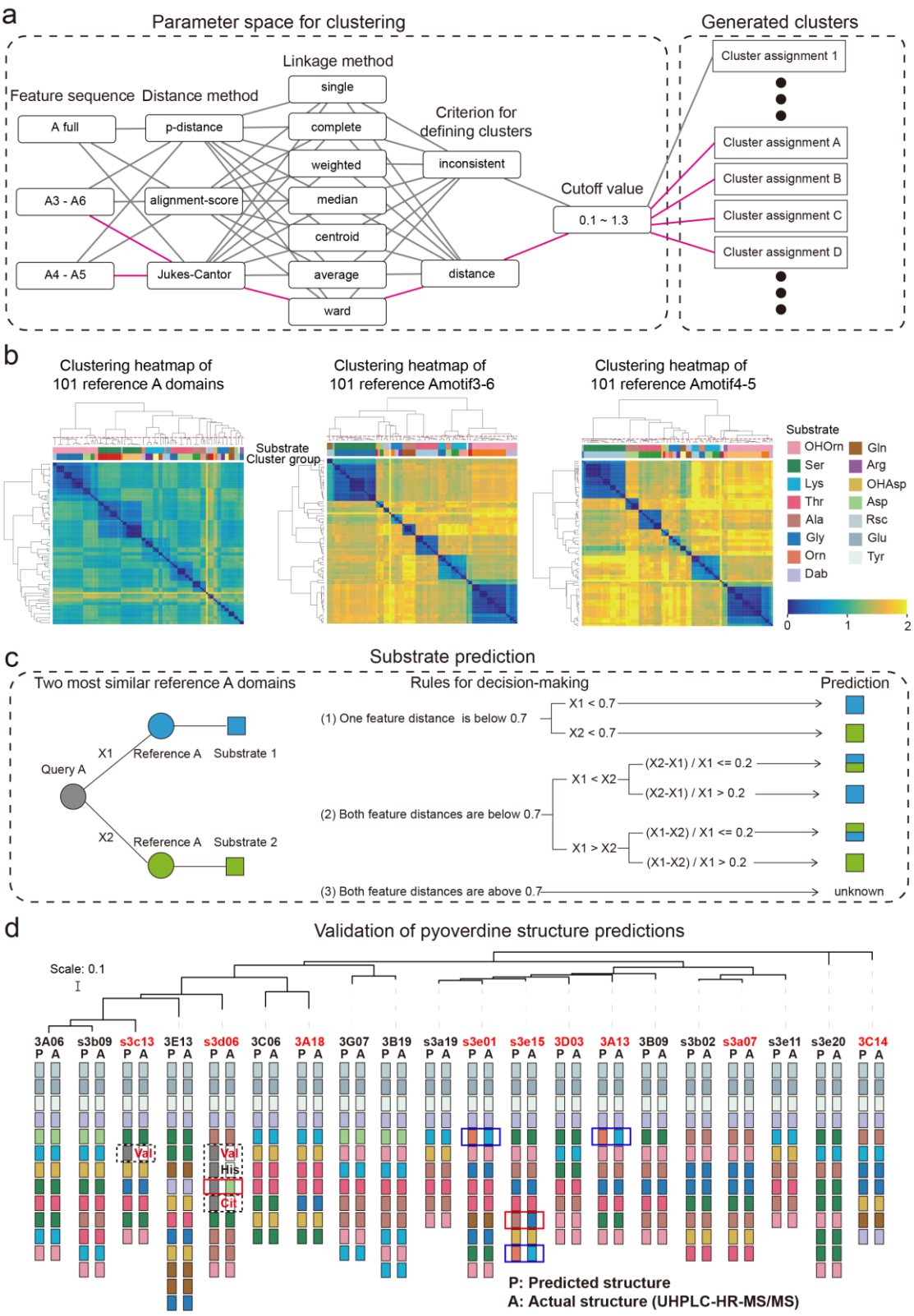


Figure 3 Phylogeny-focused substrate prediction for pyoverdine synthetase assembly lines. a. Information from 101 reference A domains with known amino acid substrates were used to develop an algorithm that predicts substrates from A domain sequence data with high accuracy. The challenge is to group the variable A domains into clusters that predict the same substrate (captured by the

322 silhouette index). To find the most distinctive algorithm, we combined different feature sequences of A
323 domains (Amotif) with different distance and linkage methods in our hierarchical clustering analyses. The
324 best performing path is shown in pink. **b.** Heatmap showing the hierarchically clustered distances of the
325 101 reference A domains as a function of the feature sequence used. Color bar in the lower right shows
326 the sequence distance. Left panel: complete A domain sequences. Middle panel: Amotif3-6 sequences.
327 Right panel: Amotif4-5 sequences. The heatmaps show that hierarchical clustering, reliably associating
328 sequence distances with substrates, worked best with the Amotif4-5. The experimentally validated
329 substrates are shown on top of each heatmap and consist of 13 amino acids and 2 amino acid derivatives:
330 Ser (Serine), Lys (Lysine), Thr (Threonine), Ala (Alanine), Gly (Glycine), Gln (Glutamine), Arg (Arginine),
331 Asp (Aspartic acid), Glu (Glutamic acid), Tyr (Tyrosine), Orn (Ornithine), Rsc (succinic acid derivatives),
332 Dab (2,4-diaminobutyric acid), OHAsp (Aspartic acid derivative), and OHOrn (a general term for three
333 ornithine derivatives: FoOHOrn, AcOHOrn, cOHOrn). **c.** Phylogeny-focused substrate prediction pipeline
334 for query A domains (grey circle) based on Amotif4-5 feature sequence comparisons. X1 and X2 represent
335 the feature distance between the query A domain and two closest reference A domains (blue and green
336 circles), respectively. Three rules are used, based on the feature sequence distances X1 and X2 and a
337 threshold value of 0.7 (50% similarity), to make substrate predictions for the query A domain. There are
338 three possible outcomes: unambiguous substrate prediction (blue or green squares), ambiguous
339 substrate prediction (dual-colored squares), and no prediction ("unknown). **d.** Phylogenetic tree of 20
340 *Pseudomonas* strains and visualization of their predicted and actual pyoverdine structures to validate our
341 phylogeny-focused substrate prediction pipeline. Strains marked in red font indicate cases with novel (not
342 yet characterized) pyoverdines structures. 151 out of the 160 substrates (94.4%) were correctly predicted.
343 The nine inconsistencies are boxed in blue (Lysine and Ornithine are indistinguishable), in dashed black
344 (detection of "unknown" substrates), and in red (true mismatches). Note that our prediction pipeline (as
345 any other pipeline) cannot distinguish between modified variants of the same amino acid, for example, we
346 cannot distinguish the three derivatives of ornithine (FoOHOrn, AcOHOrn, cOHOrn).

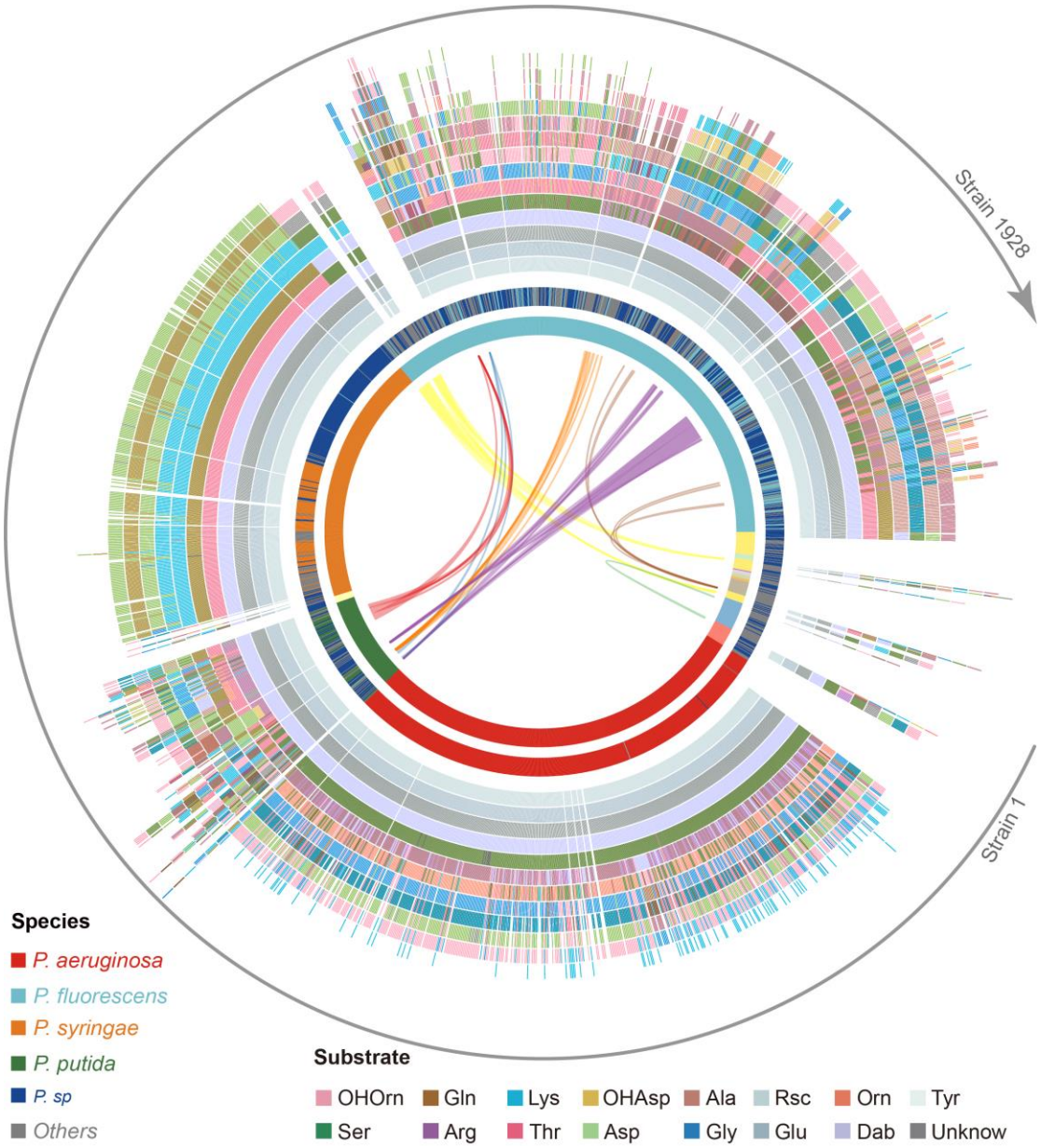
347

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

349 After successful validation, we applied our bioinformatic pipeline to the 1664 complete NRPS
350 assembly lines annotated in our genome analysis (Figure 2). Across all assembly lines, we
351 were able to predict the substrates of 17,880 A domains (97.75%) without ambiguity, whereas
352 133 A domains (0.73%) were associated with two different substrates, and 279 A domains
353 (1.52%) predicted an unknown substrate (similar to the cases of valine and citrulline above).
354 After considering the presence/absence of an E domain in each module, we derived the
355 structures of 1664 pyoverdines according to our methods in section 2 (Figure 4).

356 Our prediction yielded 188 different pyoverdine molecules, out of which only 37 structures

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



365 **Figure 4 Predicted pyoverdine structural diversity based on our developed algorithm**
366 **mapped onto the phylogenetic tree of all 1928 (non-redundant) *Pseudomonas* strains.** In
367 this figure, the strains are arranged based on the phylogenetic tree without showing the tree itself. The
368 stacked boxes in the outermost circle show the predicted structure of pyoverdines, whereby each color
369 represents a specific amino acid substrate. Strains without boxes represent non-producers (n = 264).
370 Boxes with two colors indicate cases of ambiguous (dual) substrate prediction. The inner circle shows the
371 taxonomic species classification following Figure 2e. Because the allocation of strains to species names
372 is often imprecise, we divided the 1928 strains by their phylogenetic distance into 18 clades (color
373 shadings in inner-most circle), out of which 13 contained more than one strain. Lines within the inner-most
374 circle link strains from different clades that share the same pyoverdine structures, whereby line colors
375 represent the shared unique pyoverdines. The bending of the lines represents the phylogenetic sequence
376 distances of the connected strain pairs.

377

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

386

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

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

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

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

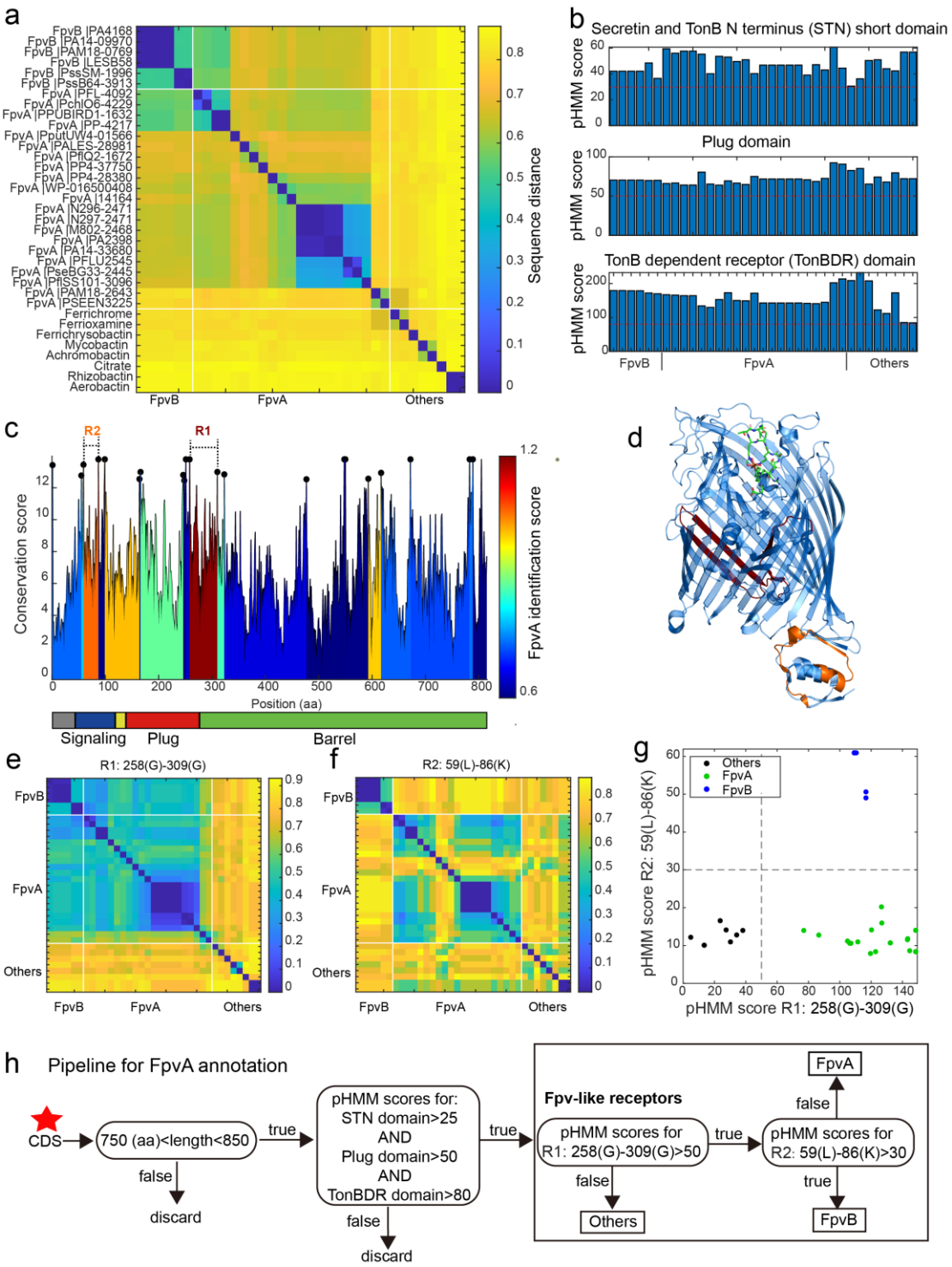


Figure 5 A sequence-region-based identification pipeline for annotating FpvA receptors. **a.** Heatmap displaying the hierarchically clustered sequence distances (p-distance calculation method,

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

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

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

454 distance between all 21 FpvA and the 14 non-FpvA sequences.

455 Our analysis yielded two locations with noticeably high FpvA identification scores (Figure
456 5c). The region with the highest FpvA identification score (referred to as R1) locates at the
457 intersection of the Plug domain and the barrel structure of the TonB domain (Figure 5d, between
458 258 Gly and 309 Gly in the PAO1 FpvA). According to the sequence distance matrix, the R1
459 region allows to distinguish heterologous siderophore receptors from FpvA and FpvB receptors
460 (Figure 5e). The region with the second highest FpvA identification score (referred to as R2)
461 was located in the C-terminal signaling domain (Figure 5d, between 59 Leu and 86 Lys in the
462 PAO1 FpvA). The sequence distance matrix revealed that R2 allows to distinguish FpvB from
463 FpvA receptors (Figure 5f).

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

472 Based on these insights, we developed a decision flow chart for annotating FpvAs in
473 *Pseudomonas* genomes (Figure 5h). First, we considered sequences as Fpv-like receptors that
474 share similar properties to the ones identified in our reference database. Particularly, protein
475 coding sequence (CDS) length had to be between 750 and 850 amino acids and the phMM

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

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

504 Next, we explored the diversity among FpvA receptors in greater detail by focusing on the
505 1,534 strains with proximate receptors located within 20 kb of the pyoverdine Pep genes (Figure
506 6d). We utilized these high-confidence FpvAs for sequence feature extraction. When
507 considering the whole gene sequences, these receptors segregated into 44 groups according
508 to single-linkage clustering with an identity threshold of 60% (Figure S4a). To investigate which
509 sequence regions were the most informative for reliable clustering, we used a similar approach
510 as with FpvAs detection by quantifying the "group identification score" for variable regions
511 flanked by highly conserved sites. The higher the score, the stronger a region's capacity to
512 discriminate between FpvA groups. We found that the four regions with the top discrimination
513 capacities all located near the Plug domain surrounding the pyoverdine transmission channel
514 (Figure 6e). The plug domain is known to undergo conformational changes and is involved in
515 pyoverdine selectivity and import[41,42], suggesting that the four high-score regions are
516 responsible for pyoverdine specificity.

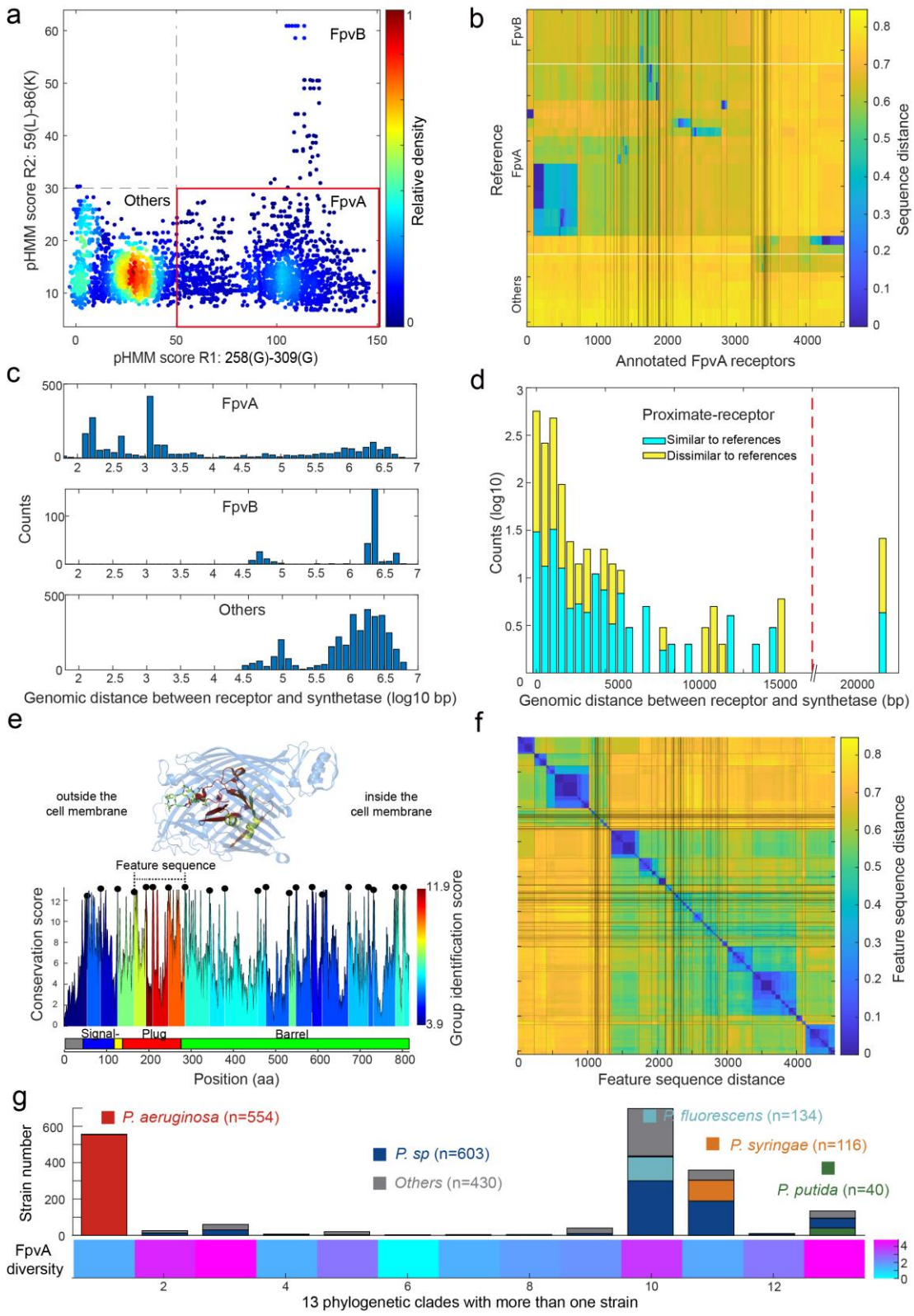


Figure 6 Application of the receptor annotation pipeline to the full database. a. Applying the receptor annotation pipeline to the genomes of the 1928 non-redundant *Pseudomonas* strains yields 14301 Fpv-like receptors, which segregate into 4547 FpvA receptors (red box), 615 FpvB receptors, and 9139 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

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

545

546 Based on the above insights, we concatenated the four high-score regions (from 168 Pro
547 to 295 Ala in PAO1) into a single "feature sequence". The feature sequence could characterize
548 98% of the distance matrix compared to the whole sequence (1534 FpvAs, r=0.98, Figure S4a-
549 b) and substantially reduced within-group distance. We applied the concatenated feature
550 sequence approach to all the 4547 annotated FpvAs to calculate the sequence distance matrix.
551 Single-linkage clustering with an identity threshold of 70% revealed a total of 94 groups, out of
552 which 43 groups contained more than 10 members (Figure 6f). The diversity of receptors is
553 hence much larger than currently anticipated as only 3 groups of FpvAs have previously been
554 reported. Finally, we calculated the diversity of receptor FpvAs for each of the 13 phylogenetic
555 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

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

594 We show that knowledge-guided learning is an extremely powerful tool to predict enzyme,
595 metabolite, and receptor properties. The establishment of our entire pipeline is based on only
596 101 previously known enzymatic A domains (from 13 known pyoverdine assembly lines) and
597 21 FpvA receptor sequences. Even with this limited amount of information, we were able to
598 predict the substrates of almost all the 18,292 enzymatic A domains and to identify 4547 FpvA
599 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 typically 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 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[23], 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 biosynthesis 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 biosynthesis machinery

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

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

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

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

complementary study[61].

We are confident that our pipeline can be expanded to study iron competition in multi-species communities and perhaps also in plant-microbe ecosystems, as siderophores exist ubiquitously and are shared among microbes and certain hosts[62]. To move further, a key question is whether our knowledge-guided approach can be applied to other microbes and important secondary metabolites, such as antibiotics, toxins, biosurfactants and pigments? Our preliminary investigation, focused on the curated dataset in *Burkholderiales* as a test case, provides a highly promising response to this query. The outcomes from this evaluation predominantly showcased the method's potential for extension to diverse microbial genera and metabolites, given sufficiently accurate dataset. Meanwhile, we observed a notable decline in accuracy (from 94% to 83%) when compared to pyoverdine. This indicates that our algorithm possibly needs refinement, particularly concerning specific secondary metabolites or microbial genera. 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. It is worth noting that our current algorithm does not predict tailoring reactions. While tailoring reactions are crucial for predicting the final NRPS product structure, this challenge is not addressed by any existing pipelines and remains an area for future research. In our study, it is important to note that the specificity of pyoverdines is mainly determined by the backbone composition, with tailoring reactions playing a relatively minor role. 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.

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

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

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

797 interactions: from community assembly to plant health. *Nature Reviews Microbiology*
798 **18**, 607-621, doi:10.1038/s41579-020-0412-1 (2020).
799 7 Price-Whelan, A., Dietrich, L. E. P. & Newman, D. K. Rethinking 'secondary' metabolism:
800 physiological roles for phenazine antibiotics. *Nature Chemical Biology* **2**, 71-78,
801 doi:10.1038/nchembio764 (2006).
802 8 Thirumurugan, D., Cholarajan, A., Raja, S. & Vijayakumar, R. An introductory chapter:
803 secondary metabolites. (2018).
804 9 Chaturvedi, K. S., Hung, C. S., Crowley, J. R., Stapleton, A. E. & Henderson, J. P. The
805 siderophore yersiniabactin binds copper to protect pathogens during infection. *Nature*
806 *Chemical Biology* **8**, 731-736, doi:10.1038/nchembio.1020 (2012).
807 10 Penn, K. *et al.* Genomic islands link secondary metabolism to functional adaptation in
808 marine Actinobacteria. *The ISME Journal* **3**, 1193-1203, doi:10.1038/ismej.2009.58
809 (2009).
810 11 Yee, D. A. *et al.* Genome mining for unknown–unknown natural products. *Nature*
811 *Chemical Biology* **19**, 633-640, doi:10.1038/s41589-022-01246-6 (2023).
812 12 Lautru, S., Deeth, R. J., Bailey, L. M. & Challis, G. L. Discovery of a new peptide natural
813 product by Streptomyces coelicolor genome mining. *Nature Chemical Biology* **1**, 265-
814 269, doi:10.1038/nchembio731 (2005).
815 13 Andryukov, B., Mikhailov, V. & Besednova, N. The Biotechnological Potential of
816 Secondary Metabolites from Marine Bacteria. *Journal of Marine Science and*
817 *Engineering* **7** (2019).
818 14 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).

15 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).

16 Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Research* **51**, W46-W50, doi:10.1093/nar/gkad344 (2023).

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).

841 22 Bodilis, J. *et al.* Distribution and evolution of ferripyoverdine receptors in Pseudomonas
842 aeruginosa. *Environmental Microbiology* **11**, 2123-2135, doi:10.1111/j.1462-
843 2920.2009.01932.x (2009).
844 23 Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining
845 pipeline. *Nucleic Acids Research* **47**, W81-W87, doi:10.1093/nar/gkz310 (2019).
846 24 Kümmerli, R. Iron acquisition strategies in pseudomonads: mechanisms, ecology, and
847 evolution. *BioMetals*, doi:10.1007/s10534-022-00480-8 (2022).
848 25 Meyer, J. M. Pyoverdines: pigments, siderophores and potential taxonomic markers of
849 fluorescent Pseudomonas species. *Arch Microbiol* **174**, 135-142,
850 doi:10.1007/s002030000188 (2000).
851 26 Xu, Z., Park, T. J. & Cao, H. Advances in mining and expressing microbial biosynthetic
852 gene clusters. *Crit Rev Microbiol*, 1-20, doi:10.1080/1040841x.2022.2036099 (2022).
853 27 Keller, N. P. Fungal secondary metabolism: regulation, function and drug discovery.
854 *Nature Reviews Microbiology* **17**, 167-180, doi:10.1038/s41579-018-0121-1 (2019).
855 28 Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Research* **32**,
856 D138-D141, doi:10.1093/nar/gkh121 (2004).
857 29 Winsor, G. L. *et al.* Enhanced annotations and features for comparing thousands of
858 Pseudomonas genomes in the Pseudomonas genome database. *Nucleic Acids*
859 *Research* **44**, D646-D653, doi:10.1093/nar/gkv1227 (2016).
860 30 Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and genomes from
861 metagenomes using PhyloPhlAn 3.0. *Nature Communications* **11**, 2500,
862 doi:10.1038/s41467-020-16366-7 (2020).

863 31 Felnagle, E. A. *et al.* Nonribosomal peptide synthetases involved in the production of
864 medically relevant natural products. *Mol Pharm* **5**, 191-211, doi:10.1021/mp700137g
865 (2008).
866 32 Süssmuth, R. D. & Mainz, A. Nonribosomal Peptide Synthesis—Principles and
867 Prospects. *Angewandte Chemie International Edition* **56**, 3770-3821,
868 doi:10.1002/anie.201609079 (2017).
869 33 Terlouw, B. R. *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally
870 validated biosynthetic gene clusters. *Nucleic Acids Research* **51**, D603-D610,
871 doi:10.1093/nar/gkac1049 (2023).
872 34 Butaitė, E., Kramer, J., Wyder, S. & Kümmerli, R. Environmental determinants of
873 pyoverdine production, exploitation and competition in natural Pseudomonas
874 communities. *Environmental Microbiology* **20**, 3629-3642, doi:10.1111/1462-
875 2920.14355 (2018).
876 35 Rehm, K., Vollenweider, V., Kümmerli, R. & Bigler, L. A comprehensive method to
877 elucidate pyoverdines produced by fluorescent Pseudomonas spp. by UHPLC-HR-
878 MS/MS. *Analytical and Bioanalytical Chemistry* **414**, 2671-2685, doi:10.1007/s00216-
879 022-03907-w (2022).
880 36 Butaitė, E., Baumgartner, M., Wyder, S. & Kümmerli, R. Siderophore cheating and
881 cheating resistance shape competition for iron in soil and freshwater Pseudomonas
882 communities. *Nature Communications* **8**, 414, doi:10.1038/s41467-017-00509-4
883 (2017).
884 37 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).

38 Jin, Z. *et al.* Conditional privatization of a public siderophore enables Pseudomonas aeruginosa to resist cheater invasion. *Nature Communications* **9**, 1383, doi:10.1038/s41467-018-03791-y (2018).

39 Chan, D. C. K. & Burrows, L. L. Pseudomonas aeruginosa 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).

40 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).

41 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).

42 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).

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).

907 45 Handelsman, J. Metagenomics: Application of Genomics to Uncultured
908 Microorganisms. *Microbiology and Molecular Biology Reviews* **68**, 669-685,
909 doi:10.1128/MMBR.68.4.669-685.2004 (2004).
910 46 Tsilimigras, M. C. B. & Fodor, A. A. Compositional data analysis of the microbiome:
911 fundamentals, tools, and challenges. *Annals of Epidemiology* **26**, 330-335,
912 doi:10.1016/j.annepidem.2016.03.002 (2016).
913 47 Weiss, S. *et al.* Correlation detection strategies in microbial data sets vary widely in
914 sensitivity and precision. *ISME J* **10**, 1669-1681, doi:10.1038/ismej.2015.235 (2016).
915 48 Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nature Reviews*
916 *Microbiology* **10**, 538-550, doi:10.1038/nrmicro2832 (2012).
917 49 Khayatt, B. I., Overmars, L., Siezen, R. J. & Francke, C. Classification of the
918 Adenylation and Acyl-Transferase Activity of NRPS and PKS Systems Using
919 Ensembles of Substrate Specific Hidden Markov Models. *PLOS ONE* **8**, e62136,
920 doi:10.1371/journal.pone.0062136 (2013).
921 50 Minowa, Y., Araki, M. & Kanehisa, M. Comprehensive Analysis of Distinctive Polyketide
922 and Nonribosomal Peptide Structural Motifs Encoded in Microbial Genomes. *Journal*
923 *of Molecular Biology* **368**, 1500-1517, doi:10.1016/j.jmb.2007.02.099 (2007).
924 51 Prieto, C., García-Estrada, C., Lorenzana, D. & Martín, J. F. NRPSsp: non-ribosomal
925 peptide synthase substrate predictor. *Bioinformatics* **28**, 426-427,
926 doi:10.1093/bioinformatics/btr659 (2012).
927 52 Röttig, M. *et al.* NRPSpredictor2—a web server for predicting NRPS adenylation
928 domain specificity. *Nucleic Acids Research* **39**, W362-W367, doi:10.1093/nar/gkr323

929 (2011).

930 53 Zierep, P. F., Ceci, A. T., Dobrusin, I., Rockwell-Kollmann, S. C. & Günther, S. SeMPI
931 2.0-A Web Server for PKS and NRPS Predictions Combined with Metabolite Screening
932 in Natural Product Databases. *Metabolites* **11**, doi:10.3390/metabo11010013 (2020).

933 54 Fan, J. *et al.* Biosynthetic diversification of peptaibol mediates fungus-mycohost
934 interactions. *bioRxiv* (2022).

935 55 Smith, E. E., Sims, E. H., Spencer, D. H., Kaul, R. & Olson, M. V. Evidence for
936 diversifying selection at the pyoverdine locus of Pseudomonas aeruginosa. *J Bacteriol*
937 **187**, 2138-2147, doi:10.1128/jb.187.6.2138-2147.2005 (2005).

938 56 Gu, S. *et al.* Competition for iron drives phytopathogen control by natural rhizosphere
939 microbiomes. *Nature Microbiology* **5**, 1002-1010, doi:10.1038/s41564-020-0719-8
940 (2020).

941 57 Figueiredo, A. R. T., Özkaya, Ö., Kümmerli, R. & Kramer, J. Siderophores drive invasion
942 dynamics in bacterial communities through their dual role as public good versus public
943 bad. *Ecology Letters* **25**, 138-150, doi:10.1111/ele.13912 (2022).

944 58 Andrews, S. C., Robinson, A. K. & Rodríguez-Quiñones, F. Bacterial iron homeostasis.
945 *FEMS Microbiology Reviews* **27**, 215–237, doi:10.1016/s0168-6445(03)00055-x
946 (2003).

947 59 Boyd, P. W. & Ellwood, M. J. The biogeochemical cycle of iron in the ocean. *Nat Geosci*
948 **3**, 675-682, doi:10.1038/ngeo964 (2010).

949 60 Emerson, D., Roden, E. & Twining, B. The microbial ferrous wheel: iron cycling in
950 terrestrial, freshwater, and marine environments. *Frontiers in Microbiology* **3**,

951 doi:10.3389/fmicb.2012.00383 (2012).

952 61 Gu, S. *et al.* Siderophore synthetase-receptor gene coevolution reveals habitat- and
953 pathogen-specific bacterial iron interaction networks. *bioRxiv*, 2023.2011.2005.565711,
954 doi:10.1101/2023.11.05.565711 (2024).

955 62 Ruolin, H. *et al.* SIDERITE: Unveiling Hidden Siderophore Diversity in the Chemical
956 Space Through Digital Exploration. *bioRxiv*, 2023.2008.2031.555687,
957 doi:10.1101/2023.08.31.555687 (2023).

958 63 Chen, H. *et al.* Biosynthesis and engineering of the nonribosomal peptides with a C-
959 terminal putrescine. *Nature Communications* **14**, 6619, doi:10.1038/s41467-023-
960 42387-z (2023).

961

962

Supplementary figures

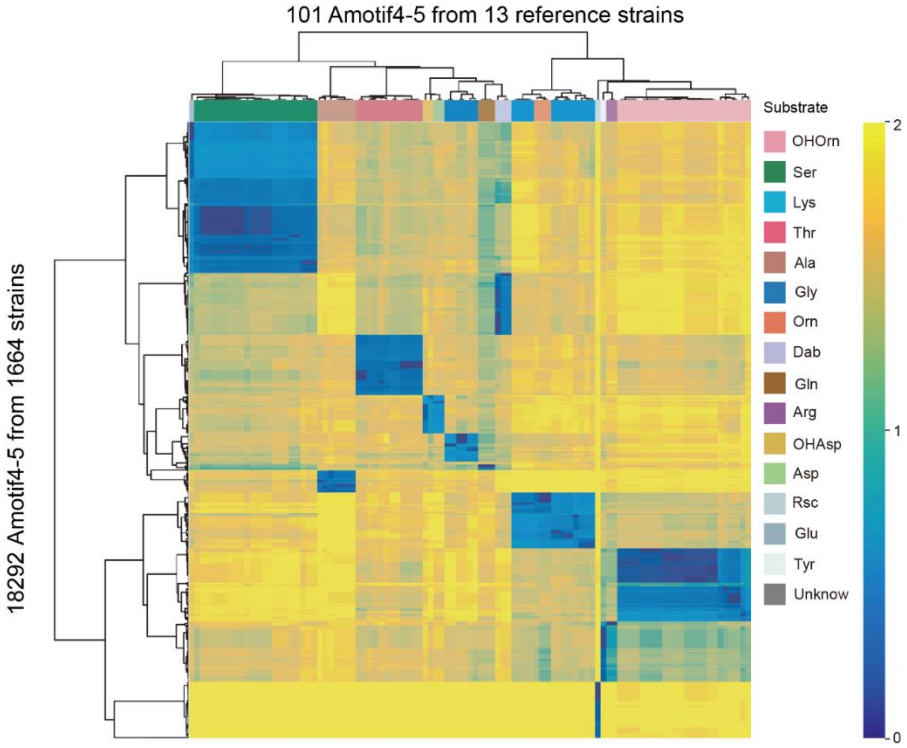


Figure S1 Clustering heatmap of sequence distance matrix between A domains from 1664 strains and 101 reference A domains. The color on top indicates the substrate of reference A domains.

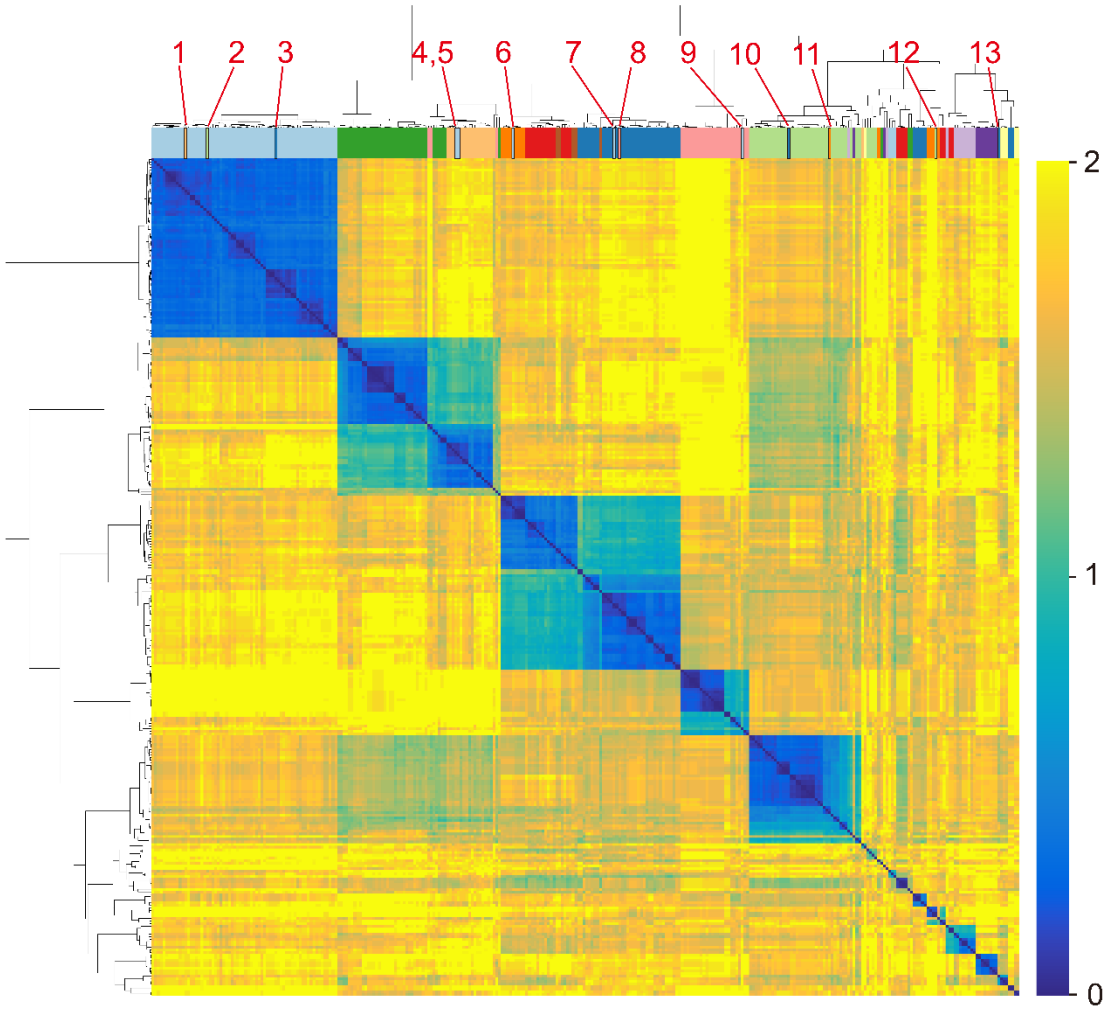


Figure S2 Clustering heatmap of sequence distance matrix between 318 Amotif4-5 from *Pseudomonas* strains in the MIBiG database (Supplementary_table4). MIBiG is a standard annotation database of biosynthetic gene clusters. We retrieved all secondary metabolites produced by members of the *Pseudomonas* 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).

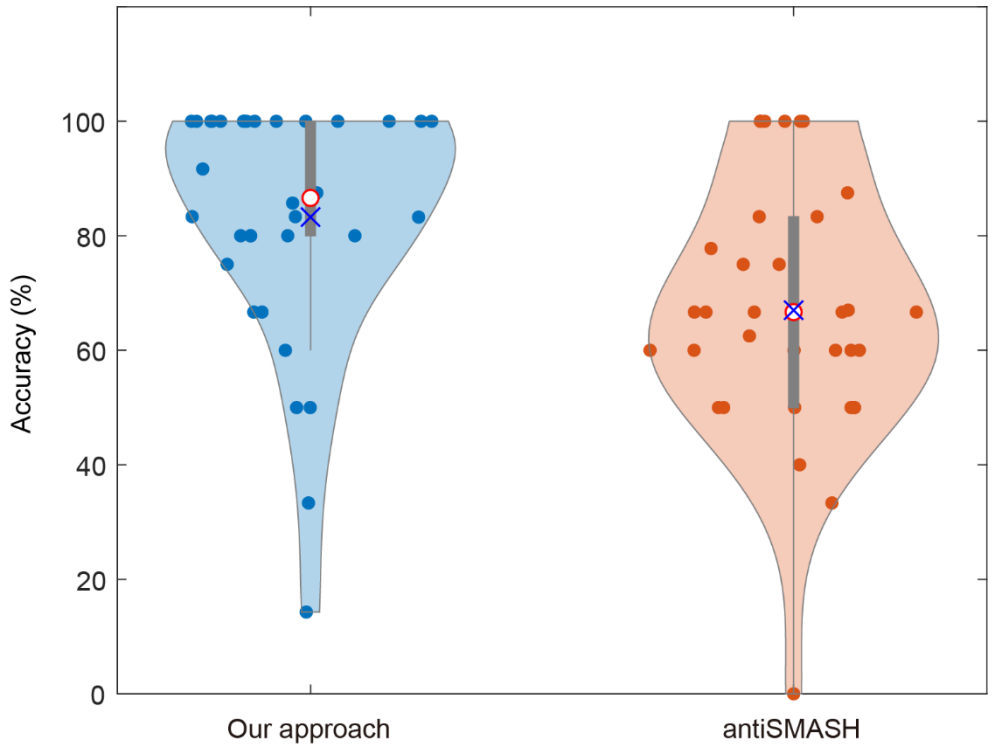


Figure S3 Comparison of prediction accuracy in the structural composition of secondary metabolites between our approach and antiSMASH in *Burkholderiales*. Each data point in the violin plot symbolizes a distinct secondary metabolite. The median is denoted by a red circle, while the mean is represented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

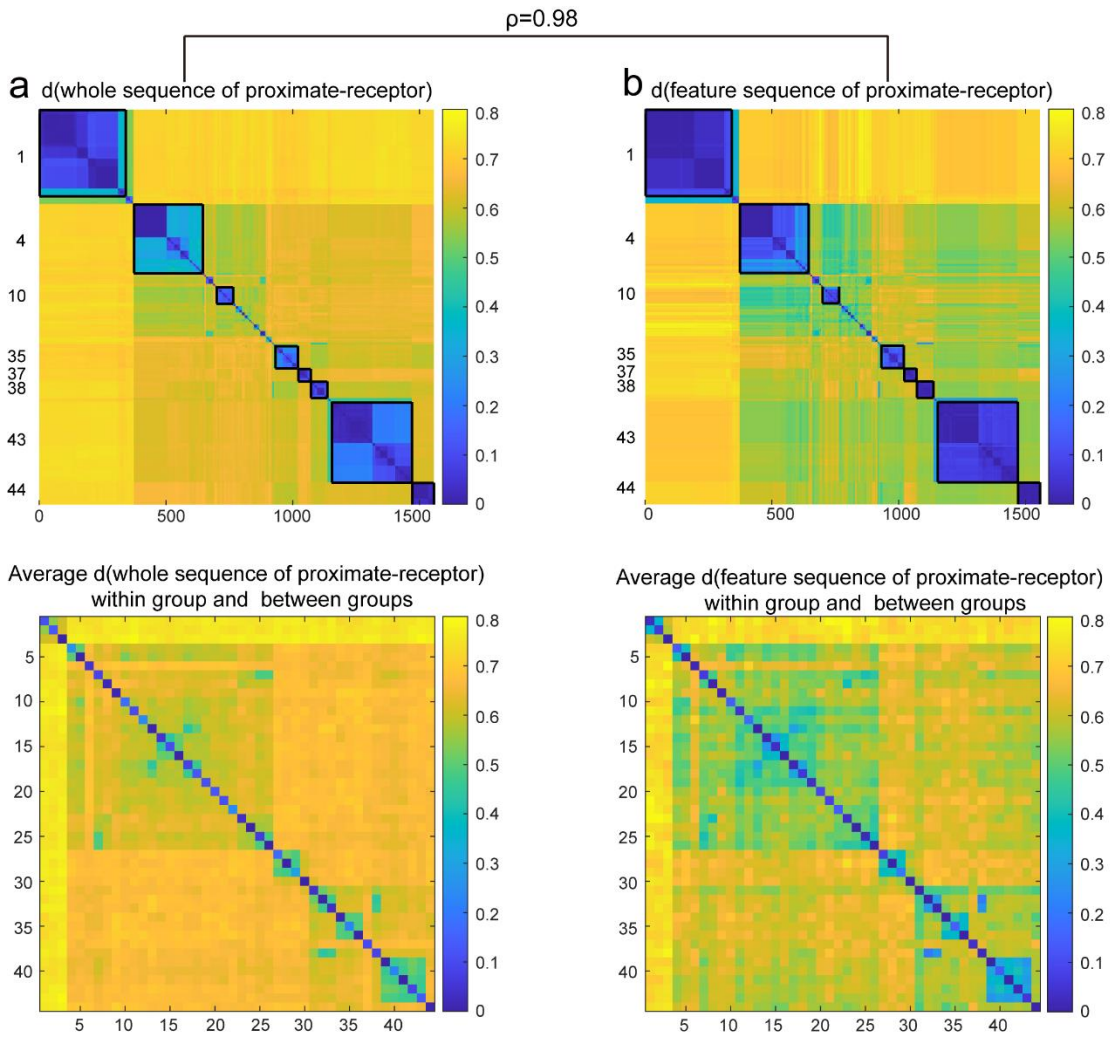


Figure S4 Comparison of clustering and grouping effects of complete sequences (a) and feature sequences (b) of all proximate-receptors.

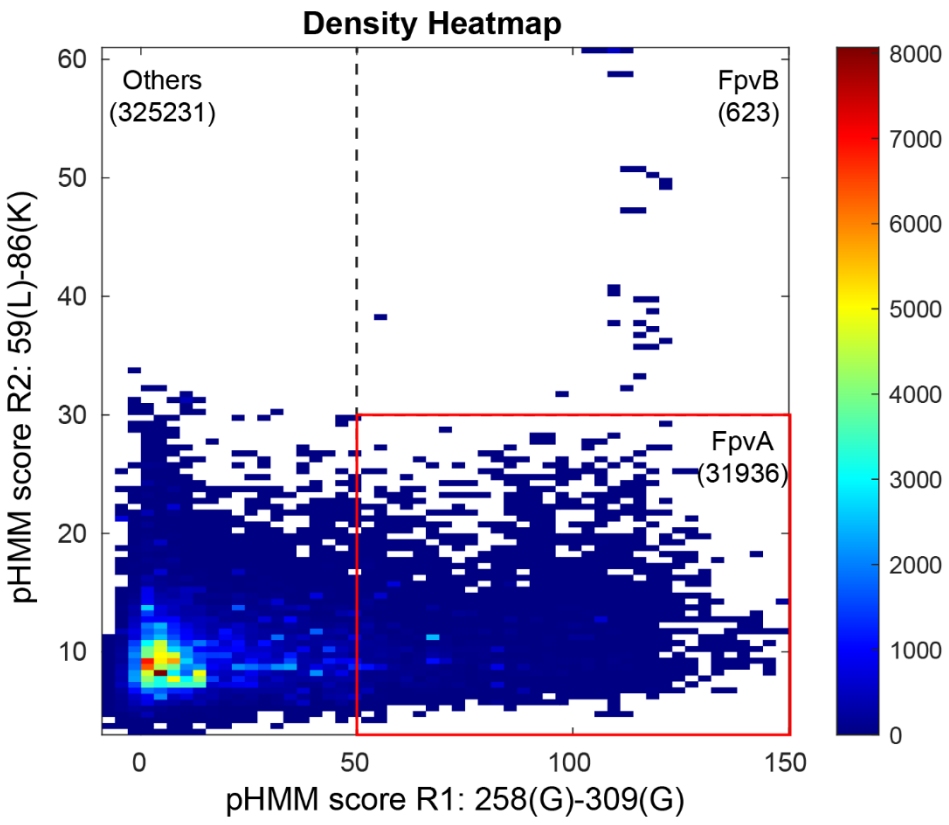


Figure S5 Applying the receptor annotation pipeline to >30,000 complete bacterial genomes in NCBI based on the pHMM score thresholds for regions R1 and R2. The heatmap indicates receptor density.

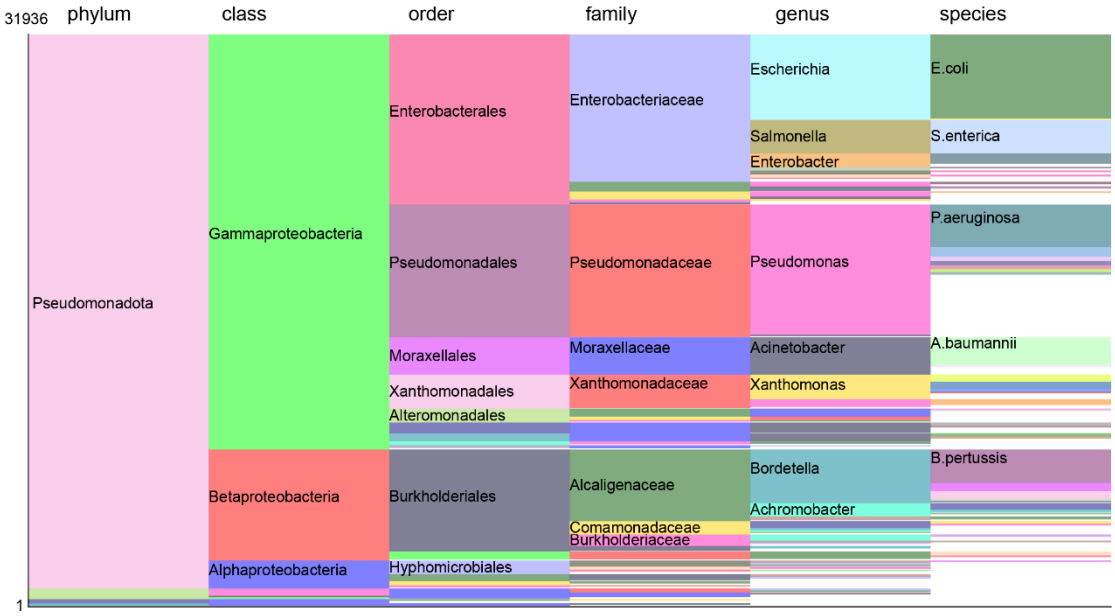


Figure S6 The distribution of 31,936 FpvAs spans various taxonomic categories. Taxa with an abundance exceeding 1/500 are highlighted with colors, while others remain white. Taxa surpassing an abundance of 1/50 will have their specific names marked.