

# **Prediction of gene cluster function based on transcriptional regulatory networks uncovers a novel locus required for desferrioxamine B biosynthesis**

Hannah E. Augustijn<sup>1,2,#</sup>, Zachary L. Reitz<sup>1,#,\*</sup>, Le Zhang<sup>2</sup>, Jeanine A. Boot<sup>1</sup>, Somayah S. Elsayed<sup>2</sup>, Gregory L. Challis<sup>3,4,5</sup>, Marnix H. Medema<sup>1,2,\*</sup>, Gilles P. van Wezel<sup>1,2,6,\*</sup>

1 Bioinformatics Group, Wageningen University, Wageningen, The Netherlands;

2 Institute of Biology, Leiden University, Leiden, The Netherlands;

3 Department of Chemistry, University of Warwick, Coventry, United Kingdom;

4 Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria, Australia;

5 ARC Centre of Excellence for Innovations in Peptide and Protein Science, Biomedicine Discovery Institute, Monash University, Clayton, Victoria, Australia

6 Department of Microbial Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands

\* Current address: Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, CA, USA

# Contributed equally

\* Co-corresponding authors

1 **ABSTRACT**

2 Bacteria produce a plethora of natural products that are in clinical, agricultural and  
3 biotechnological use. Genome mining revealed millions of biosynthetic gene clusters (BGCs)  
4 that encode their biosynthesis, and the major challenge is to predict the bioactivities of the  
5 molecules these BGCs specify, and how to elicit their expression. Here, we present an  
6 innovative strategy whereby we harness the power of regulatory networks combined with  
7 global gene expression patterns to predict BGC functions. Studying the regulon of iron master  
8 regulator DmdR1 in *Streptomyces coelicolor* combined with co-expression data and large-  
9 scale comparative genome analysis identified the novel *desJGH* gene cluster. Mutational and  
10 metabolomics analysis showed that *desJGH* is required for biosynthesis of the clinical drug  
11 desferrioxamine B. DesJGH thereby dictate the balance between the structurally distinct  
12 desferrioxamines B and E. We propose regulation-based genome mining as a promising  
13 approach to functionally prioritize BGCs to accelerate the discovery of novel bioactive  
14 molecules.

15

## 16 INTRODUCTION

17 Within the genetic blueprint of microorganisms lies an immense reservoir of chemical potential,  
18 which likely constitutes the mechanistic basis for numerous microbiome-associated  
19 phenotypes and offers a rich source of raw materials for discovery and development of among  
20 others antibiotics, anticancer agents, immunosuppressants, crop protection agents, and  
21 industrial ingredients <sup>1,2</sup>. Genome mining efforts have led to the identification of millions of  
22 biosynthetic gene clusters (BGCs) predicted to encode the biosynthesis of many thousands  
23 of natural product scaffolds <sup>3</sup>. However, only an estimated 3% of these specialized metabolites  
24 have undergone experimental characterization thus far, leaving a vast amount of untapped  
25 chemical diversity yet to be explored <sup>4</sup>.

26 Identifying the diverse roles of specialized metabolites in microbiome interactions is  
27 highly challenging, primarily due to the dynamic nature of the host environment and the  
28 difficulties in replicating such conditions in laboratory settings. Moreover, while these  
29 molecules exhibit a wide range of functions, only a small fraction of metabolites will directly  
30 contribute towards microbiome-associated phenotypes such as disease suppression or  
31 growth promotion, or have the necessary properties to yield the next generation of crop  
32 protection agents, antibiotics, or food additives <sup>5-7</sup>. As a result, there is a pressing need for  
33 generalized strategies to predict the functions of specialized metabolites, enabling us to  
34 understand their mechanistic roles in inter-organismal interactions and to gauge their  
35 usefulness for industrial and clinical applications.

36 A major aim in current natural product discovery is to identify ways to reduce the  
37 genetic space of sequenced BGCs to manageable numbers, to inform scientists on which  
38 BGCs to prioritize in the search for novel bioactivity. Historically, scientists have investigated  
39 two dimensions, namely the molecular space via high-throughput screening of compound and  
40 strain libraries, followed by the genomic space in the 21<sup>st</sup> century, by investigating BGCs in  
41 sequenced genomes, based on the identification of enzyme-coding genes <sup>8</sup>. Perhaps the most  
42 advanced strategy for the latter has thus far been target-based genome mining, which uses  
43 self-resistance genes inside BGCs as beacons for recognizing the macromolecular targets of

44 their products. However, the presence of recognizable self-resistance genes seems to be  
45 limited to a mere 5-10% of BGCs, necessitating complementary methods to predict the  
46 functions of the remaining specialized metabolic diversity<sup>9,10</sup>.

47 We anticipate that an attractive alternative would be regulation-guided approaches,  
48 given that the regulatory system plays a pivotal role in the transcription of BGCs.  
49 Overexpression or inactivation of cluster-situated regulatory genes have been used to activate  
50 their expression<sup>11-13</sup>. For example, targeting BGCs containing *Streptomyces* antibiotic  
51 regulatory protein (SARP) family regulators enabled the discovery of novel antibiotic BGCs  
52<sup>14,15</sup>. Also, the Identification of Natural compound Biosynthesis pathways by Exploiting  
53 Knowledge of Transcriptional regulation (INBEKT) strategy was able to unveil a previously  
54 undetectable BGC by identifying regulatory binding sites of the zinc-dependent regulator ZuR  
55<sup>16</sup>. These early successes at the single-gene or single-BGC level indicate that genome-wide  
56 analysis of regulatory networks may be even more successful at unveiling BGC functions.

57 Here, we introduce a computational omics strategy that leverages genome-wide gene  
58 regulation information to provide functional predictions of BGCs in microbes. This novel  
59 approach connects genome-wide regulatory information derived from transcription factor  
60 binding site (TFBS) prediction to gene co-expression networks, thereby associating genes to  
61 functions. Genome-wide regulatory analysis of BGCs of *Streptomyces coelicolor* M145 in  
62 combination with co-expression patterns unveiled a novel BGC that had escaped detection by  
63 current genome mining software tools. Subsequent mutational analysis and metabolic profiling  
64 experiments showed that this BGC plays an important role in the biosynthesis of the well-  
65 studied siderophore desferrioxamine B. These results illustrate the potential of our method to  
66 infer BGC function, facilitate the detection and prioritization of novel BGCs and ultimately pave  
67 the way for identifying genes responsible for the biosynthesis of novel bioactive molecules.

68

69 **RESULTS AND DISCUSSION**

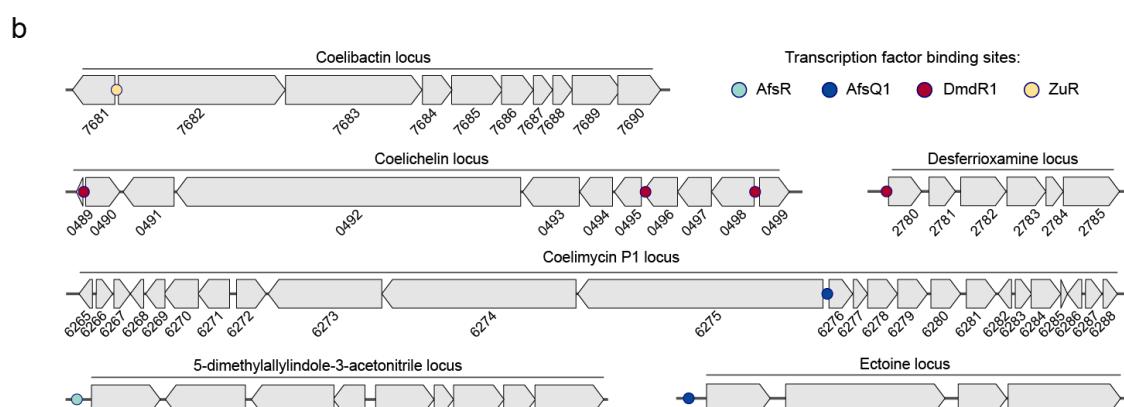
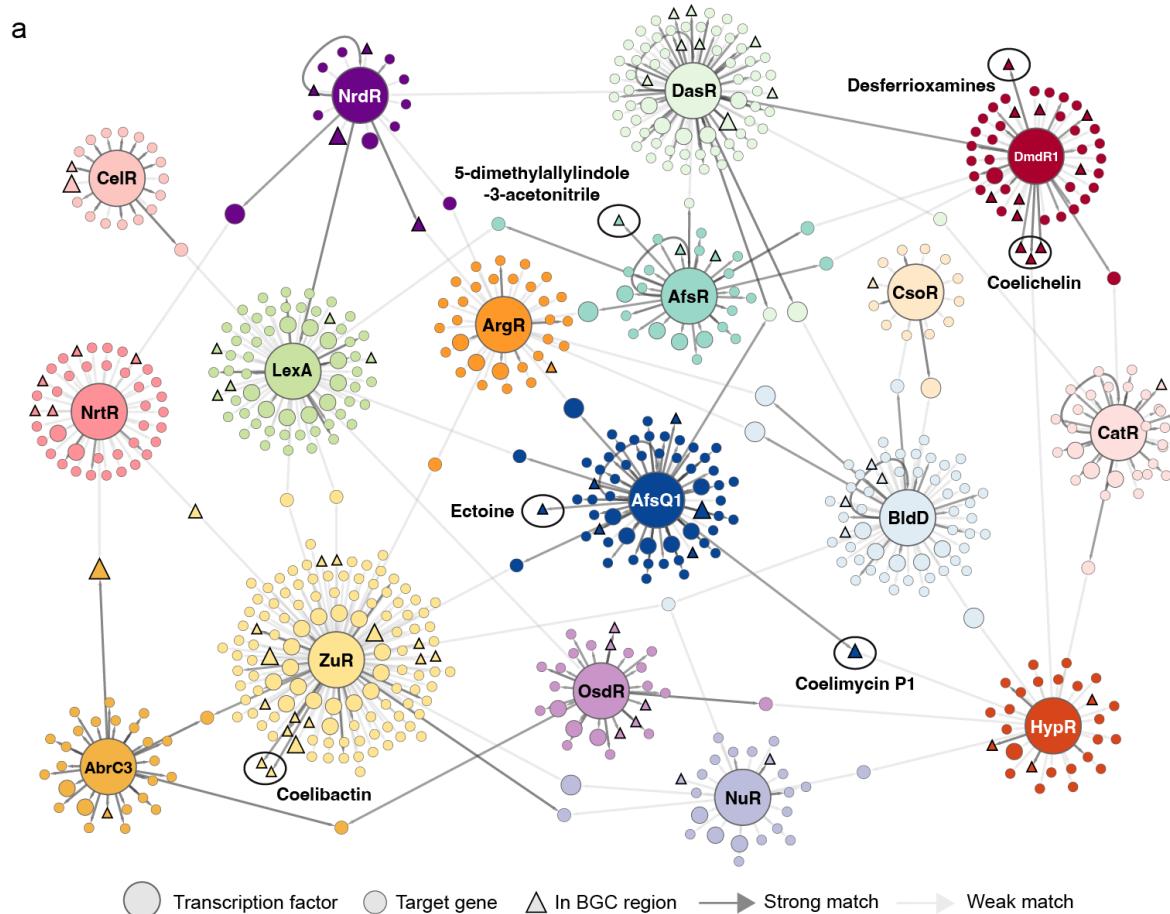
70

71 **Identifying functional associations through gene regulatory networks**

72 A major challenge in genome-mining-based drug discovery lies in prioritizing BGCs within the  
73 vast unexplored biosynthetic space, and in particular finding novel ways to predict their  
74 function. We hypothesized that regulatory networks that control BGC expression might form a  
75 new, third, dimension for screening for potential functions, complementing phenotypic and  
76 genomic screening. The concept we propose is that if an unknown BGC (or any cluster of  
77 genes) is predicted to be controlled by a transcriptional regulator that responds to a known  
78 signal and is connected to a specific physiological response, that BGC may functionally relate  
79 to known BGCs controlled in a similar manner.

80 To develop such a regulation-based genome mining strategy and assess its validity,  
81 we chose to focus on the model organism *Streptomyces coelicolor* M145. This microbe,  
82 belonging to the phylum Actinomycetota, is renowned for its exceptional ability to produce a  
83 wide array of bioactive compounds, making it an interesting target for natural product  
84 discovery <sup>17-20</sup>. Moreover, it is the bacterial species with currently the largest number of  
85 functionally characterized BGCs, with 17 out of its 27 BGCs having been connected to the  
86 production of a known metabolite, making it an ideal organism to assess how well regulation  
87 connects to function <sup>21</sup>. To investigate the functional relationships between this microbe's  
88 regulatory machinery and specialized metabolite biosynthesis, we investigated the binding of  
89 transcription factors (TFs) to their corresponding binding sites (TFBSs). For this purpose, we  
90 used the regulatory data of the LogoMotif database <sup>22</sup>. Seventeen precalculated and manually  
91 curated position weight matrices (PWMs) associated with TFs in this database were used for  
92 genome-wide predictions of 730 TFBSs, using automated computational matching. Based on  
93 these predictions, a gene regulatory network (GRN) was constructed in which TFBSs were  
94 identified within BGC regions predicted by antiSMASH (Fig. 1a). A total of 81 TFBSs were  
95 found within antiSMASH BGC regions; 55 of these were at the region peripheries and  
96 putatively unrelated to specialized metabolite biosynthesis. To identify which TFBSs were truly

97 linked to biosynthetic pathways, we then refined the boundaries of the BGCs (Table S1) using  
98 literature evidence and gene co-expression patterns (see below). This resulted in the  
99 identification of 17 low-confidence and 9 medium/high-confidence BGC-TFBS associations  
100 each matching the physiological or ecological functions associated with the corresponding  
101 regulon (Fig. 1a). These findings agree with existing experimental analyses, thus reinforcing  
102 the utility of our approach in accurately identifying BGC-TFBS connections (Fig. 1b). For  
103 example, there is a clear correlation between TFBSs of the zinc uptake regulator (Zur) and  
104 the zinc-regulated coelibactin locus<sup>23</sup>, as well as between the pleiotropic antibiotic biosynthesis  
105 regulator AfsQ1 and the antibiotic coelimycin P1<sup>24</sup>. Additionally, we observed a connection  
106 between the iron-dependent regulator DmdR1 and the biosynthesis of two iron-chelating  
107 compound families that function as siderophores: the desferrioxamines (DFOs) and  
108 coelichelin<sup>25,26</sup>.



109

110 **Figure 1. a**, Predicted gene regulatory network of *Streptomyces coelicolor* based on 17 well-  
111 known regulators. Each node in the network represents a (regulatory) gene, and every edge  
112 represents a regulatory interaction between two nodes. The edges colored in dark gray  
113 indicate strong PWM prediction scores, while the lighter gray shades represent weaker  
114 interactions. Matches within BGC regions are depicted as triangles. In six regions (black  
115 circled), the matches fall within a co-expressed region, highlighting their functional relation to  
116 these compounds. **b**, Representation of the four co-expressed regions, including the locations

117 of their detected TFBSs as colored dots. All predicted TFBSs have been experimentally  
118 validated in pre-existing work.

119

120 **Co-expression analysis and operon-level expansion of the predicted DmdR1 regulon**

121 Next, we aimed to go beyond antiSMASH-detectable BGCs and assess if we could infer the  
122 function of any uncharacterized operons and gene clusters using regulatory predictions.

123 Expectedly, the predicted DmdR1 regulon exhibited a clear functional association with

124 siderophores, as evidenced by the connection between its binding sites and known  
125 siderophore BGCs<sup>27</sup>. Therefore, we focused on exploring the functional connection between

126 DmdR1 binding sites (*iron boxes*) and iron metabolic genes. A critical issue when using PWMs

127 is the large number of false positive TFBS hits. To address this, we refined the general  
128 LogoMotif detection threshold for DmdR1 to be more accurate for *S. coelicolor* by applying the

129 principles previously described for the calibration of the PREDetector algorithm<sup>28</sup>. This

130 approach involves an analysis of the distribution of hits and the ratio of hits in non-coding  
131 versus coding regions (Fig. S1). The results demonstrated that higher PWM match scores

132 correlated with a greater frequency of hits detected in non-coding regions, where iron boxes

133 are typically found. By calculating the median score of the non-coding to coding ratio, we

134 established a refined threshold of 22.875, leading to the identification of a total of 39 predicted  
135 DmdR1 binding sites (Table S2). Among these 39 predicted binding sites, we identified 25

136 unique binding site locations, 22 of which corresponded to previously reported DmdR1 target

137 genes. Based on these predictions, we identified three novel putative DmdR1 target genes:

138 SCO2114, SCO2275, and SCO5998.

139 Bacterial regulons consist not only of genes with TFBSs in their regulatory region, but

140 also any downstream co-operonic genes. DmdR1-controlled operons were predicted using a

141 co-expression analysis of a previously published transcriptome. The RNA-Seq dataset of Lee

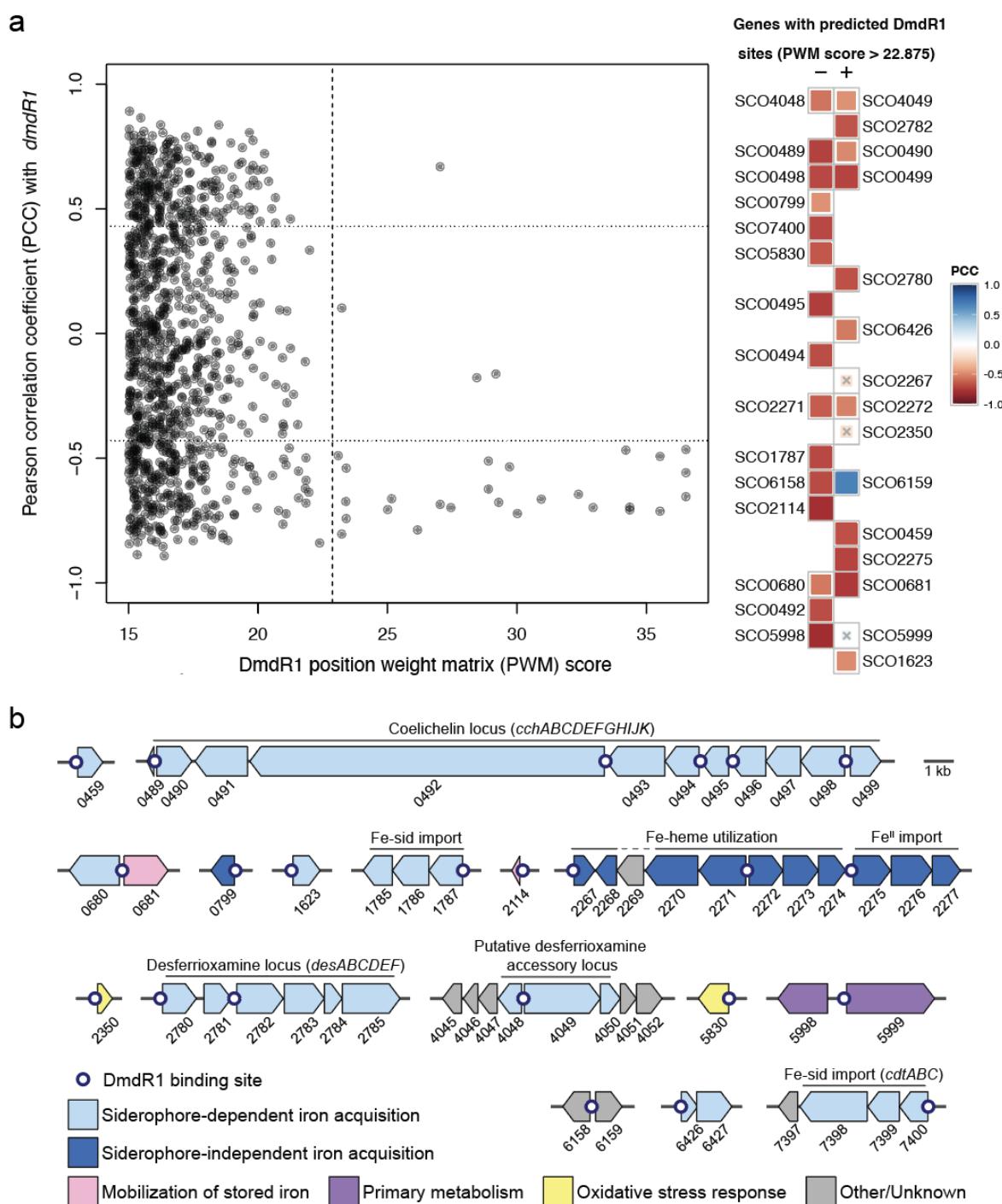
142 *et al.*<sup>29</sup> was chosen for its relatively high sample count (22 for *S. coelicolor*) and the study's

143 focus on iron restriction. Reads were retrieved from NCBI SRA and mapped to the *S. coelicolor*

144 M145 genome, and gene count data were processed using previously reported techniques to

145 generate a pairwise gene co-expression matrix<sup>30,31</sup>. Of the 30 predicted DmdR1 target genes  
146 with a significant PWM match score, 26 were anti-correlated with transcription of *dmdR1*  
147 (Pearson correlation coefficient [PCC] < -0.43,  $p < 0.05$ , Fig. 2a), including newly predicted  
148 target genes SCO2114, SCO2275, and SCO5998. The co-expression data support the  
149 minimum PWM match score of 22.875; below this threshold, no mean anti-correlated  
150 expression was identified. Only a single gene with a significant PWM score, the GntR-type  
151 regulator SCO6159, was positively co-expressed with *dmdR1* (PCC = 0.69), and the  
152 transcription pattern of three putative target genes did not correlate significantly with that of  
153 *dmdR1*, suggesting a more complicated regulation by multiple transcription factors. DmdR1  
154 target genes were placed into predicted operons using the gene co-expression matrix, as well  
155 as strand and intergenic distance, expanding the putative direct regulon of DmdR1 from 25 to  
156 58 genes, which are found across 16 genomic loci (Fig. 2b). A description of the predicted  
157 DmdR1 regulon, including functional predictions, is presented in SI Discussion 1. As expected,  
158 DmdR1 binding sites were recovered in the coelichelin and desferrioxamine BGCs but not the  
159 ZuR-controlled coelibactin BGC, supporting the use of regulatory analysis for linking  
160 metallophore BGCs to their corresponding metal. Other logical gene annotations present in  
161 the regulon include siderophore-independent iron acquisition, mobilization of stored iron, and  
162 oxidative stress response.

163



164

165 **Figure 2. a**, Anti-correlation of gene expression between *dmdR1* and its predicted regulon.

166 Left: Pearson correlation coefficients (PCCs) between *dmdR1* and all genes with a DmdR1  
 167 position weight matrix (PWM) score greater than 15 in their regulatory region. The vertical  
 168 dashed line marks the refined PWM score threshold of 22.875. The horizontal dotted lines  
 169 mark  $PCC = \pm 0.43$ , corresponding to an adjusted  $p$ -value of 0.05. Right: Target genes  
 170 immediately downstream of a predicted DmdR1 binding site, ordered by decreasing PWM

171 score. Plus and minus indicate the strand of the target gene. Genes marked with an x did not  
172 have significant co-expression with *dmdR1*. Binding site details are given in Table S2. **b**, The  
173 putative regulon of *DmdR1* in *S. coelicolor* M145. White dots indicate predicted *DmdR1*  
174 binding sites. Genes are labeled by SCO number and colored by putative function. Clusters  
175 are drawn to scale, and arrows represent the direction of transcription.

176

177 **Metabolic profiling of an unexplored *DmdR1*-controlled locus**

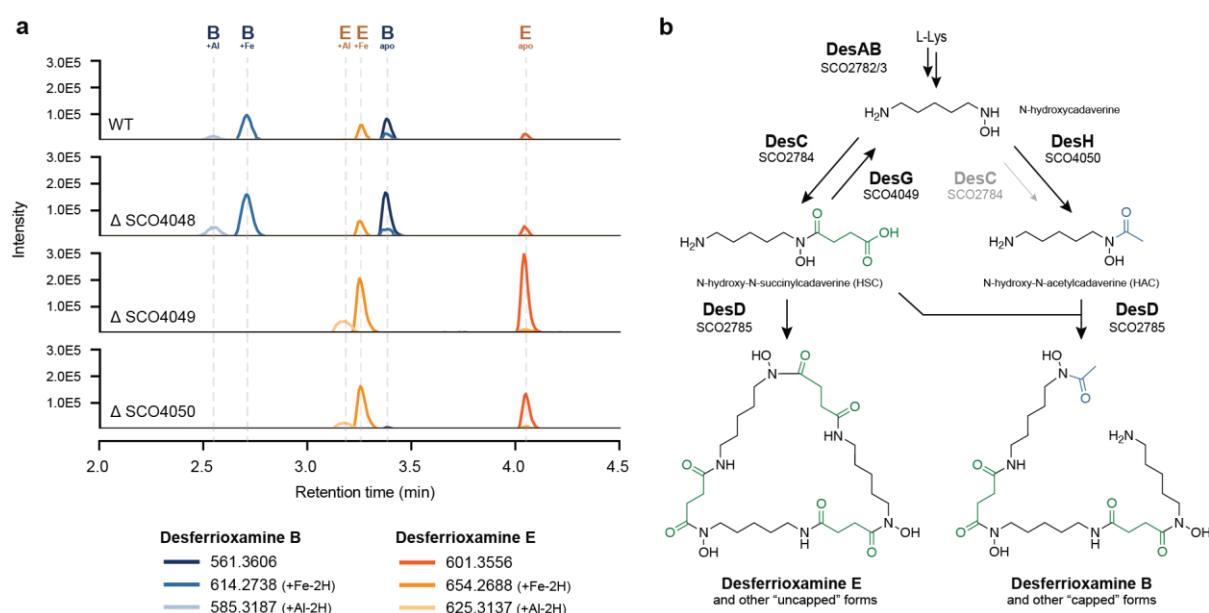
178 This systematic mapping of the *DmdR1* regulon then provided the opportunity to investigate  
179 whether new operons or gene clusters could be identified that would be predicted to function  
180 in iron acquisition. Upon close examination of all individual genes across the regulon, the  
181 uncharacterized region from SCO4045 to SCO4052 stood out due to sequence similarity to  
182 biosynthetic genes (Fig. 2b). Interestingly, SCO4050 encodes a protein similar to the *N*-  
183 acyltransferase *DesC* (encoded by SCO2784), which catalyzes the conversion of *N*-  
184 hydroxycadaverine to *N*-hydroxy-*N*-succinylcadaverine (HSC) and *N*-hydroxy-*N*-  
185 acetylcadaverine (HAC), the direct precursors of desferrioxamine B, *in vitro*<sup>32</sup>. SCO4048 is a  
186 paralog of *desF* (SCO2781), which encodes ferrioxamine reductase. Furthermore, SCO4049  
187 is homologous to genes designated as *desG* in other streptomycetes, and is predicted to  
188 encode a penicillin amidase family protein; phylogenetic analysis in Actinobacteria revealed  
189 that *desG*, if present, either colocalized with the DFO cluster, or with a separate *DmdR1*-  
190 controlled locus<sup>33</sup>. *DesG* was originally hypothesized to increase DFO structural diversity by  
191 producing phenylacetic acid-capped derivatives in some strains; however, no arylated DFOs  
192 have been identified in *S. coelicolor*. Together, SCO4048, SCO4049, and SCO4050 (further  
193 referred to as *desJ*, *desG*, and *desH*, respectively) appear to comprise a previously  
194 undetected locus putatively related to DFO biosynthesis<sup>33</sup>.

195 To analyze the role of the *DmdR1*-controlled locus in the production of DFOs, we  
196 applied the CRISPR-based editing system (CRISPR-BEST)<sup>34</sup> to construct three knock-out  
197 mutants in which either SCO4048 (*desJ*), SCO4049 (*desG*) or SCO4050 (*desH*) had been  
198 inactivated. The system allows introduction of a premature stop codon in the target ORF, thus

199 preventing the production of a functional protein. Using this method, we created null mutants  
200 of SCO4048 (*desJ*) with mutations W55\* or Q68\*, resulting in 186 aa or 173 aa truncation of  
201 the gene product, respectively. The introduction of a stop codon at W61 in SCO4049 (*desG*)  
202 led to a substantial 721 aa shortening, while mutations W43\* or Q91\* in SCO4050 (*desH*)  
203 resulted in truncations of 163 aa or 115 aa, respectively. PCR followed by DNA sequencing  
204 was used to verify the correctness of the knock-out mutants.

205 To obtain extracts for metabolomics, *S. coelicolor* M145 and its mutant derivatives  
206 were grown in a liquid iron-limited medium (ISP-2) for five days. The metabolites produced  
207 were adsorbed on Diaion® HP20 resin, which was subsequently extracted with methanol and  
208 analyzed using liquid chromatography-mass spectrometry (LC-MS), which revealed changes  
209 in the production of DFO-related metabolites in each of the mutants compared to the wild-type  
210 strain (Fig. 3a). The metabolites were annotated by matching the high-resolution mass  
211 spectrometry (HRMS) and tandem mass spectrometry (MS/MS) spectra to previously  
212 published ones (Fig. S2) <sup>35-37</sup>. Statistical analyses showed that only the levels of  
213 desferrioxamine B (DFOB) were significantly increased in extracts of the *desJ* mutant as  
214 compared to the parental strain (Fig. S3). Metabolomic analysis of  $\Delta$ *desG* and  $\Delta$ *desH* revealed  
215 an approximate 1000-fold and 16-fold decrease in DFOB production, respectively (Fig. 3 and  
216 Fig. S3). Conversely, the mutants exhibited a significant increase in desferrioxamine E (DFOE)  
217 and its metal complexes, most likely as a result of the nearly abolished DFOB production.

218



219

220 **Figure 3. New model for biosynthesis of desferrioxamines B and E. a,** Extracted ion  
221 chromatograms for  $m/z$  values corresponding to DFO-related metabolites in culture extracts  
222 of the knock-out mutants of *S. coelicolor* M145 strain. The *desG* mutant fails to produce DFOB,  
223 while a 16-fold decrease in DFOB biosynthesis was seen in *desH* mutants (cf. Fig. S3). **b,**  
224 Proposed biosynthetic pathway for assembly of desferrioxamines E and B. Main biosynthetic  
225 enzymes presented in bold face. DesG and DesH balance intracellular *N*-hydroxy-*N*-  
226 succinylcadaverine (HSC) and *N*-hydroxy-*N*-acetylcadaverine (HAC) concentrations by  
227 converting HSC to HAC. In the absence of DesG and/or DesH, the cells likely fail to produce  
228 sufficient levels of HAC, thereby strongly attenuating the production of DFOB. Although DesC  
229 has been shown to be able to catalyze the acetylation of *N*-hydroxycadaverine *in vitro*, the  
230 enzyme can only modestly compensate for the loss of DesH *in vivo*, underlining the important  
231 role played by DesG and DesH in DFOB production (Fig. S4).

232

233  
234 We genetically complemented the mutants to determine if the effects were due solely to the  
235 gene inactivation and not to second-site mutations. For this, constructs were introduced that  
236 expressed the respective wild-type genes *desJ*, *desG* or *desH* from the constitutive *gap*  
237 promoter. The complementation constructs were based on vector pSET152<sup>38</sup>, which

238 integrates at the bacteriophage  $\Phi$ C31 attachment site on the *S. coelicolor* genome. The  
239 complemented mutants showed recovery of DFOB production in the complemented strains  
240 (Fig. S5). Taken together, our mutational analysis shows that the attenuation of DFOB  
241 production in the mutants can be fully explained by the inactivation of *desG* and *desH*.

242 DFOB and other capped desferrioxamines have been isolated from many  
243 *Streptomyces* strains, as well as several other Actinomycetota. To see if the proposed  
244 biosynthetic role for DesGH applies more generally to DFO biosynthesis in other  
245 Actinomycetota, we performed a meta-analysis of published DFO producers. In total, we  
246 identified reports of DFO production in 46 sequenced strains, comprising mostly *Streptomyces*  
247 species (n=34), as well as other Actinomycetota (n=7), Pseudomonadota (n=4), and one  
248 member of Bacteroidota (Table S3). Homologues of *desG* and *desH* were found in 36 of the  
249 genomes, all Actinomycetota. One sequenced DFO producer, *Gordonia rupripertincta* CWB2,  
250 contained *desG* but not *desH*; however, the *G. rupripertincta* DFO locus is part of a larger  
251 BGC that putatively encodes the biosynthesis of the cryptic nocardichelins (see SI Discussion  
252 2), and one of the two other acyltransferase genes in the BGC has presumably replaced *desH*.  
253 In all other cases, *desG* and *desH* are putatively co-operonic, and the two genes are fused in  
254 *Streptomyces atratus* and *Micrococcus* spp. CH3 and CH7. Among collected reports of DFO  
255 production, DFOB (Fig. 3b) and other acetyl, fatty-acyl, or aryl “capped” DFOs were common,  
256 isolated from 34 of 47 sequenced strains. However, in line with our discovery, the nine strains  
257 lacking *desGH* exclusively produced DFOE (Fig. 3b) and other “uncapped” DFOs with  
258 succinylated monomers (Fig. S6).

259 Based on the combination of the above data, we propose the following pathway for  
260 desferrioxamine biosynthesis in *S. coelicolor* (Fig. 3b). The biosynthesis of DFOE is encoded  
261 by the canonical biosynthetic locus *desABCD* (SCO2782-85): DesA and DesB convert L-lysine  
262 to *N*-hydroxycadaverine, DesC succinylates *N*-hydroxycadaverine to form HSC<sup>32</sup>, and DesD  
263 cyclotrimerizes HSC to produce DFOE<sup>39</sup>. In contrast, DesG (SCO4049) and DesH (SCO4050)  
264 enable DFOB production (Fig. 3). A recent study of DesD concluded that the relative  
265 intracellular concentrations of HSC and HAC must be controlled for DFOB formation<sup>39</sup>.

266 Previous investigations of DesC *in vitro* have shown that it is able to catalyze the conversion  
267 of *N*-hydroxycadaverine to both HSC and HAC, using succinyl and acetyl-CoA, respectively  
268<sup>32</sup>. However, the relative catalytic efficiency of these two processes has yet to be elucidated.  
269 Our experiments strongly suggest that the main function of DesC *in vivo* is to catalyze the  
270 production of HSC, while HAC results primarily from the action of DesH. We propose that  
271 DesG, which shows sequence similarity to amidases, de-succinylates HSC to regenerate *N*-  
272 hydroxycadaverine, which is then acetylated by the putative acetyltransferase DesH to boost  
273 the levels of HAC relative to HSC in high level DFOB producers. Gene fusions of *desGH*  
274 observed in some strains are equipped to exploit the high local effective concentration of *N*-  
275 hydroxycadaverine generated by the DesG domain, enabling the DesH domain to acetylate  
276 *N*-hydroxycadaverine before it can be re-succinylated. The production of DFOB in the  $\Delta$ *desH*  
277 mutant is strongly attenuated but not abolished, consistent with the previously reported ability  
278 of DesC to catalyze acylation of *N*-hydroxycadaverine with acetyl-CoA in addition to succinyl-  
279 CoA (Fig. 3a). Taken together, these data indicate that DesC strongly prefers succinyl-CoA  
280 as a substrate over acetyl-CoA, and that DesG and DesH are required to ensure sufficient  
281 quantities of HAC are produced to support high level DFOB production *in vivo*. This  
282 biosynthetic model is in line with the available phylogenomic, metabolomic, and genetic  
283 evidence, as well as the canonical catalytic chemistry of DesG and DesH homologues.

284

## 285 CONCLUSION

286 In conclusion, we have developed a novel computational omics strategy for functional  
287 inference of BGCs in microbes, which uses regulatory information to provide clues regarding  
288 their functional roles in inter-organismal interactions and to gauge their usefulness for  
289 industrial and clinical applications. Uniquely, this method leverages genome-wide gene  
290 regulation information derived from TFBS detection combined with gene co-expression  
291 network analysis to link biosynthetic genes to their potential functions. A key application of this  
292 method is showcased in our study of *Streptomyces coelicolor* M145, a well-studied model  
293 organism, where we predict the regulons of 17 well-known regulators and 9 high-confidence

294 functional associations to known BGCs. Of these, we selected the iron-dependent repressor  
295 DmdR1 and its strong connection to the regulation of siderophore biosynthesis for showcasing  
296 the effectiveness of our approach. This analysis, which involved TFBS prediction of the  
297 DmdR1 regulon, alongside the detection of co-expression patterns under iron starvation  
298 conditions, allowed us to detect an uncharacterized gene cluster with a functional link to iron  
299 metabolism. Furthermore, we present evidence that the putative amidase and acyltransferase  
300 encoded by *desG* and *desH*, respectively, in this cluster collaborate in the efficient  
301 biosynthesis of desferrioxamine B by SCO4049 and SCO4050 CRISPR-cBEST knockout  
302 mutants and subsequent metabolic profiling experiments. These findings not only validated  
303 our hypothesis, but also enabled identification of a novel pathway within the complex  
304 biosynthetic route to desferrioxamines. Overall, our results demonstrate the effectiveness of  
305 our method in identifying and inferring the function of novel BGCs that escaped detection  
306 despite the availability of state-of-the-art genome mining tools. We anticipate that  
307 transcriptomics-guided regulatory genome mining, by combining function prediction with  
308 application of elicitors that may activate BGCs of interest, will provide pointers as to how to  
309 select and activate cryptic BGCs in the extant biosynthetic diversity. This will aid in the  
310 identification of their roles in microbiome interactions and guide the discovery of bioactive  
311 natural products that are of value for pharmaceutical, agricultural, and biotechnological  
312 applications.

313

## 314 **METHODS**

315

### 316 **General**

317 Default software parameters were used unless otherwise noted. Scripts are available at:

318 <https://github.com/zreitz/dmdR>.

319

320 **Construction of the position weight matrix and sequence motif**  
321 Ten previously reported DmdR1 binding sites from *Streptomyces coelicolor* were collected  
322 from literature<sup>26</sup>. Thereafter, the occurrences of each nucleotide across all positions of the  
323 sequences were counted to construct a position frequency matrix (PFM). This PFM was  
324 converted to a PWM by applying Bioconductor's seqLogo v5.29.8 algorithm<sup>40</sup>, which  
325 calculates the log-likelihood of each nucleotide in the matrix, while taking into account the  
326 background nucleotide distributions. Additionally, the information content (IC) of the resulting  
327 PWM was calculated using Shannon's entropy calculation methods. The IC was visualized as  
328 a sequence motif with the use of Logomaker v 0.8<sup>41</sup>.

329

### 330 **Identification of DmdR1 binding sites**

331 The genome assembly of *Streptomyces coelicolor* A3(2) was downloaded from NCBI using  
332 accession GCA\_000203835.1. The coding and non-coding regions, as well as the regions  
333 spanning from -350 bp to +50 bp relative to the start codons of each gene were extracted with  
334 MiniMotif<sup>22</sup> (<https://github.com/HAugustijn/MiniMotif>). We employed MOODS v1.9.4.1<sup>42</sup> to  
335 query these regions for DmdR1 PWM matches, using a p-value threshold of 0.01 and  
336 background distribution of 72% representing the GC percentage of *S. coelicolor*. The ratio of  
337 hits in non-coding versus coding regions was visualized using the R package ggplot2<sup>43</sup>.

338

### 339 **RNA-Seq data processing and co-expression analyses**

340 *Streptomyces coelicolor* A3(2) RNA-Seq data, collected by Lee *et al.*,<sup>29</sup> was retrieved  
341 from the European Nucleotide Archive (PRJEB25075).<sup>44</sup> Raw read quality was assessed with  
342 FastQC.<sup>45</sup> Reads were mapped to the reference genome NC\_003888.3 using STAR v2.7.6a:<sup>46</sup>  
343 Index files were generated with the parameters “--genomeSAindexNbases 10 --  
344 sjdbGTFfeatureExon CDS”, and reads were aligned with the parameter “--alignIntronMax 1”.  
345 Mapped reads were indexed using SAMtools v1.3.1<sup>47</sup> and visualized with the Integrative  
346 Genomics Viewer.<sup>48</sup> Per-gene read count tables were generated with featureCounts v2.0.1<sup>49</sup>  
347 using the parameters “-O -M -t CDS -s 2 --fraction”.

348 The per-gene RNA-Seq count data was further analyzed in R. A minimum gene  
349 expression cutoff was applied ( $\geq 5$  counts in 50% of samples), then counts were normalized  
350 by Trimmed Mean of M-values (TMM) and  $\log_2$  transformed using a hyperbolic arcsine  
351 pseudocount<sup>50</sup>. A co-expression bias associated with lowly- and highly-expressed genes (of  
352 unknown origin, but present in several other RNA-Seq datasets<sup>31</sup>) was mitigated by  
353 regressing out the first principal component using the *sva\_network* function from the *sva*  
354 package (Fig. S7)<sup>30</sup>. The resulting correlation matrix still had an expression-correlated  
355 broadening of correlation coefficients, which was corrected by spatial quantile normalization  
356 (Fig. S7)<sup>31</sup> and used for further analyses. An all-to-all Pearson Correlation Coefficient (PCC)  
357 matrix with corrected two-sided Student p-values was calculated using the *corAndPValue*  
358 function from the package *WGCNA*.<sup>51</sup> A p-value of 0.05 corresponded to a minimum absolute  
359 PCC value of 0.43. The correlation matrix was corrected for remaining expression-level-  
360 dependent PCC distribution broadening using spatial quantile normalization  
361 (*spqn::normalize\_correlation*) with the following parameters: ngrp = 20, size\_grp = 337,  
362 ref\_grp = 18.<sup>31</sup> Subsets of the resulting correlation matrix were used for all downstream  
363 analyses.

364

### 365 Comparative genomics

366 Desferrioxamine core loci (*desABCD*) and accessory loci (*desGH*) were found in  
367 *Streptomyces* genomes using a modified version of antiSMASH 7<sup>52</sup>  
368 (<https://github.com/zreitz/antismash/tree/desGH-7-1>). The “desABCD” rule requires matches  
369 to all of the following Pfam models with a maximum intergenic distance of 5 kbp: PF00282.22  
370 (*desA*), PF13434.9 (*desB*), PF13523.9 (*desC*), and PF04183.5 (*desD*). The “desGH” rule  
371 requires matches to PF01804.21 (*desG*) and PF13523.9 (*desH*) with a maximum intergenic  
372 distance of 1 kbp. Genome assemblies for previously reported DFO producers (Table S3)  
373 were downloaded from NCBI Genbank on 21 Nov, 2023, in Genbank format using ncbi-  
374 genome-download<sup>53</sup>. The multiSMASH pipeline<sup>54</sup> was used to scan the genomes with  
375 antiSMASH and tabulate the results<sup>52</sup>. A gene phylogeny of the resulting desABCD loci was

376 obtained from CORASON, run as part of BiG-SCAPE v1.1.5 using settings "--mix --no-classify  
377 --clans-off --cutoffs 1" <sup>55</sup>. The resulting phylogenetic tree was annotated using iTOL v5 <sup>56</sup>.

378

### 379 **Bacterial strains and media**

380 *E. coli* strains DH5α and ET12567/pUZ8002<sup>57</sup> were used for routine cloning and for  
381 interspecific conjugation, respectively. *E. coli* transformants were selected on Luria Bertani  
382 (LB) agar media containing the relevant antibiotics and grown O/N at 37 °C. *Streptomyces*  
383 *coelicolor* A3(2) M145 was used as parental strain to construct mutants. All media and routine  
384 *Streptomyces* techniques are described in the *Streptomyces* manual <sup>58</sup>. Soy flour mannitol  
385 (SFM) agar plates were used to grow *Streptomyces* strains for preparing spore suspensions.

386

### 387 **Growth conditions and extraction**

388 The cultures were grown in triplicate in 100 mL Erlenmeyer flasks with 1 g of Diaion® HP-20  
389 resin (Resindion, Mitsubishi) in 15 mL of International *Streptomyces* Project-2 medium (ISP-  
390 2; yeast extract 4 g/L, malt extract 10 g/L and dextrose 4 g/L at pH 7.2). The medium was  
391 inoculated using 1 µL of spore stock and incubated in a rotary shaker at 30 °C. After five days  
392 of growth, the resin was vacuum filtered, washed three times with Milli-Q water, and extracted  
393 with 3 x 5 mL of methanol. The crude extracts were then dried, weighed, and dissolved in  
394 methanol at a final concentration of 1 mg/mL. Media blanks were extracted and prepared in a  
395 similar way as negative controls.

396

### 397 **LC-MS based metabolic profiling**

398 Liquid chromatography-tandem mass spectrometry (LC-MS/MS) acquisition was performed  
399 using Shimadzu Nexera X2 ultra high-performance liquid chromatography (UPLC) system,  
400 with attached photodiode array detector (PDA), coupled to Shimadzu 9030 QTOF mass  
401 spectrometer, equipped with a standard electrospray ionization (ESI) source unit, in which a  
402 calibrant delivery system (CDS) is installed. A total of 2 µL of dissolved extracts were injected

403 into a Waters Acquity HSS C18 column (1.8 µm, 100 Å, 2.1 × 100 mm). The column was  
404 maintained at 30 °C, and run at a flow rate of 0.5 mL/min, using 0.1% formic acid in H<sub>2</sub>O as  
405 solvent A, and 0.1% formic acid in acetonitrile as solvent B. A gradient was employed for  
406 chromatographic separation starting at 5% B for 1 min, then 5–85% B for 9 min, 85–100% B  
407 for 1 min, and finally held at 100% B for 3 min. The column was re-equilibrated to 5% B for 3  
408 min before the next run was started. The LC flow was switched to the waste the first 0.5 min,  
409 then to the MS for 13.5 min, then back to the waste to the end of the run.

410 The MS system was tuned using standard NaI solution (Shimadzu). The same solution was  
411 used to calibrate the system before starting. Additionally, a calibrant solution made from ESI  
412 tuning mix (Sigma-Aldrich) was introduced through the CDS system, the first 0.5 min of each  
413 run, and the masses detected were used for post-run mass correction for the file, ensuring  
414 stable accurate mass measurements.

415 System suitability was checked by regularly measuring a standard sample made of the  
416 following compounds:

compound	concentration (µg/mL)	retention time (min)	expected m/z
paracetamol	25	2,375	152,0712
caffeine	5	3,246	195,0882
prednisolone	2,5	5,290	361,2015
reserpine	1,25	6,186	609,2812
clomipramine	1,25	6,379	315,1628

417  
418 All the samples were analyzed in positive polarity, using data dependent acquisition mode. In  
419 this regard, full scan MS spectra (*m/z* 100–1700, scan rate 10 Hz, ID enabled) were followed  
420 by two data dependent MS/MS spectra (*m/z* 100–1700, scan rate 10 Hz, ID disabled) for the  
421 two most intense ions per scan. The ions were selected when they reach an intensity threshold  
422 of 1500, isolated at the tuning file Q1 resolution, fragmented using collision induced  
423 dissociation (CID) with fixed collision energy (CE 20 eV), and excluded for 1 s before being  
424 re-selected for fragmentation. For the ESI source, the parameters were set to interface voltage  
425 4 kV, interface temperature 300 °C, nebulizing gas flow 3 L/min, and drying gas flow 10 L/min.

426 The parameters used for the CDS probe include an interface voltage 4.5 kV, and nebulizing  
427 gas flow 1 L/min.

428

## 429 Comparative metabolomics

430 Raw LC-MS data were converted to open source mzXML format using LabSolutions software  
431 (Shimadzu), and the converted files were imported into MZmine 3.3.0<sup>59</sup> for data processing.  
432 Unless specified otherwise, *m/z* tolerance was set to 0.002 *m/z* or 10.0 ppm, RT tolerance was  
433 set to 0.05 min, MS1 noise level was set to 1.0E3, MS2 noise level to 1.0E1 and the minimum  
434 absolute height was set to 5.0E2. The option to detect isotope signals below noise level was  
435 selected. For feature detection and chromatogram building, the ADAP chromatogram builder<sup>60</sup>  
436 was used with positive polarity, centroid mass detector, minimum group size of 5 in number of  
437 scans and a 2.0E3 group intensity threshold. The obtained peaks were smoothed (width: 9),  
438 and the chromatograms were deconvoluted using the local minimum search with a 90%  
439 chromatographic threshold, 1% minimum relative height, minimum ratio of peak top/edge of 2  
440 and peak duration of 0.03 to 3.00 min. The detected peaks were deisotoped (monotonic  
441 shape, maximum charge: 5; representative isotope: most intense). Peak lists from different  
442 extracts were aligned (weight for *m/z*: 20, weight for RT: 20, compare isotopic pattern with a  
443 minimum score of 50%). The gap filling algorithm was used to detect and fill missing peaks  
444 (intensity threshold 1%, RT tolerance: 0.1 minute). Duplicate peaks were filtered, and artifacts  
445 caused by detector ringing were removed (*m/z* tolerance: 1.0 *m/z* or 1,000.0 ppm). The aligned  
446 peaks were exported to a MetaboAnalyst. From here, peaks were additionally filtered to keep  
447 only peaks present in all 3 replicates and not in the media blanks, using in-house scripts. The  
448 resulting MetaboAnalyst peak list was uploaded to MetaboAnalyst<sup>61</sup>, log transformed, and  
449 normalized with Pareto scaling without prior filtering. Missing values were filled with half of the  
450 minimum positive value in the original data. Volcano plots were generated using default  
451 parameters. Additionally, extracted ion chromatograms have been obtained for the ions of the  
452 DFO-related metabolites (*m/z* tolerance 0.001 or 5 ppm, Table S4). An in-house python script  
453 was used to visualize these chromatograms with matplotlib v3.7.2 pyplot<sup>62</sup>.

454

#### 455 **Plasmids, constructs and oligonucleotides**

456 All plasmids and constructs described in this work are summarized in Table S5. The  
457 oligonucleotides are listed in Table S6.

458 Fragment containing *gapdh* promoter was digested from previously published plasmid  
459 pGWS1370<sup>63</sup> and cloned into pCRISPR-cBEST<sup>34</sup> via the same restriction sites to generated  
460 pGWS1384, where the expression of Cas9n (D10A), cytidine deaminase and uracil-DNA  
461 glycosylase inhibitor (UGI) were under the control of *gapdh* promoter instead of *tipA* promoter.  
462 Spacers of each targeted gene were selected on CRISPy-web<sup>64</sup> and cloned into Ncol-digested  
463 pGWS1384 via single strand DNA (ssDNA) oligo bridging method. Single strand DNA (ssDNA)  
464 oligos SCO4048\_W55 and SCO4048\_Q68b were used to generate SCO4048 knockout  
465 constructs pGWS1582 and pGWS1584, respectively. Similarly, SCO4049 knockout construct  
466 pGWS1585 was created using oligo SCO4049\_W61. SCO4050 knockout constructs  
467 pGWS1598 and pGWS1590 were created employing oligos SCO4050\_W43 and  
468 SCO4050\_Q91, respectively. All the generated knockout constructs were validated by Sanger  
469 sequencing using primer sg\_T7\_R\_SnaBI.

470 For the complementation of SCO4048 null mutant, pGWS1596 was used, an integrative vector  
471 based on pSET152 and harboring SCO4048 under the control of *gap* promoter. The *gap*  
472 promoter and the entire coding region (+1/+724) of SCO4048 were amplified from *S. coelicolor*  
473 M145 genomic DNA using primer pairs Pgap\_F and Pgap\_R, and SO4048\_F and  
474 SCO4048\_R, respectively. Fragments were cloned into EcoRI and XbaI digested pSET152  
475 via Gibson assembly to generate pGWS1596. Similarly, pGWS1597 and pGWS1598 were  
476 created for the complementation of SCO4049 and SCO4050 null mutants, respectively. The  
477 coding region (+1/+2347) of SCO4049 in pGWS1597 was amplified using primers  
478 SCO4049\_F and SCO4049\_R, while the coding region (+1/+619) of SCO4050 in pGWS1598  
479 was amplified using primer pair SCO4050\_F and SCO4050\_R.

#### 480 **ACKNOWLEDGEMENTS**

481 The work was supported by the European Union via ERC Advanced Grant 101055020-  
482 COMMUNITY to G.P.v.W. and ERC Starting Grant 948770-DECIPHER to M.H.M.

483 **COMPETING INTEREST STATEMENT**

484 M.H.M. is a member of the Scientific Advisory Board of Hexagon Bio.

485

486 **REFERENCES**

- 487 1. Wright, G. D. Unlocking the potential of natural products in drug discovery. *Microb. Biotechnol.* **12**, 55–57 (2019).
- 488 2. Atanasov, A. G., Zotchev, S. B., Dirsch, V. M. & Supuran, C. T. Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* **20**, 200–216 (2021).
- 489 3. Tran, P. N., Yen, M.-R., Chiang, C.-Y., Lin, H.-C. & Chen, P.-Y. Detecting and prioritizing biosynthetic gene clusters for bioactive compounds in bacteria and fungi. *Appl. Microbiol. Biotechnol.* **103**, 3277–3287 (2019).
- 490 4. Gavriilidou, A. *et al.* Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nat Microbiol* **7**, 726–735 (2022).
- 491 5. Beck, M. L., Song, S., Shuster, I. E., Miharia, A. & Walker, A. S. Diversity and taxonomic distribution of bacterial biosynthetic gene clusters predicted to produce compounds with therapeutically relevant bioactivities. *J. Ind. Microbiol. Biotechnol.* (2023) doi:10.1093/jimb/kuad024.
- 492 6. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition: surviving and thriving in the microbial jungle. *Nat. Rev. Microbiol.* **8**, 15–25 (2010).
- 493 7. van Bergeijk, D. A., Terlouw, B. R., Medema, M. H. & van Wezel, G. P. Ecology and genomics of Actinobacteria: new concepts for natural product discovery. *Nat. Rev. Microbiol.* **18**, 546–558 (2020).
- 494 8. Chen, R., Wong, H. & Burns, B. New approaches to detect biosynthetic gene clusters in the environment. *Medicines (Basel)* **6**, 32 (2019).
- 495 9. Yan, Y., Liu, N. & Tang, Y. Recent developments in self-resistance gene directed natural product discovery. *Nat. Prod. Rep.* **37**, 879–892 (2020).
- 496 10. Chen, X., Pan, H.-X. & Tang, G.-L. Newly Discovered Mechanisms of Antibiotic Self-Resistance with Multiple Enzymes Acting at Different Locations and Stages. *Antibiotics (Basel)* **12**, (2022).
- 497 11. Lu, F. *et al.* Regulatory genes and their roles for improvement of antibiotic biosynthesis in Streptomyces. *3 Biotech* **7**, (2017).

514 12. van der Heul, H. U., Bilyk, B. L., McDowall, K. J., Seipke, R. F. & van Wezel, G. P.

515 Regulation of antibiotic production in Actinobacteria: new perspectives from the post-

516 genomic era. *Nat. Prod. Rep.* **35**, 575–604 (2018).

517 13. Laureti, L. *et al.* Identification of a bioactive 51-membered macrolide complex by

518 activation of a silent polyketide synthase in *Streptomyces ambofaciens*. *Proc. Natl.*

519 *Acad. Sci. U. S. A.* **108**, 6258–6263 (2011).

520 14. Krause, J., Handayani, I., Blin, K., Kulik, A. & Mast, Y. Disclosing the Potential of the

521 SARP-Type Regulator PapR2 for the Activation of Antibiotic Gene Clusters in

522 Streptomycetes. *Front. Microbiol.* **11**, 225 (2020).

523 15. Ye, S. *et al.* Uncovering the Cryptic Gene Cluster for 3-amino-4-hydroxybenzoate

524 Derived Ahbamycins, by Searching SARP Regulator Encoding Genes in the Genome.

525 *Int. J. Mol. Sci.* **24**, (2023).

526 16. Spohn, M., Wohlleben, W. & Stegmann, E. Elucidation of the zinc-dependent regulation

527 in *Amycolatopsis japonicum* enabled the identification of the ethylenediamine-

528 disuccinate ([S,S]-EDDS) genes. *Environ. Microbiol.* **18**, 1249–1263 (2016).

529 17. Ward, A. C. & Allenby, N. E. Genome mining for the search and discovery of bioactive

530 compounds: the Streptomyces paradigm. *FEMS Microbiol. Lett.* **365**, (2018).

531 18. Belknap, K. C., Park, C. J., Barth, B. M. & Andam, C. P. Genome mining of biosynthetic

532 and chemotherapeutic gene clusters in Streptomyces bacteria. *Sci. Rep.* **10**, 1–9

533 (2020).

534 19. Nett, M., Ikeda, H. & Moore, B. S. Genomic basis for natural product biosynthetic

535 diversity in the actinomycetes. *Nat. Prod. Rep.* **26**, 1362–1384 (2009).

536 20. Hoskisson, P. A. & van Wezel, G. P. *Streptomyces coelicolor*. *Trends Microbiol.* **27**,

537 468–469 (2019).

538 21. Challis, G. L. Exploitation of the *Streptomyces coelicolor* A3(2) genome sequence for

539 discovery of new natural products and biosynthetic pathways. *J. Ind. Microbiol.*

540 *Biotechnol.* **41**, 219–232 (2014).

541 22. Augustijn, H. E. *et al.* LogoMotif: a comprehensive database of transcription factor

542 binding site profiles in Actinobacteria. *bioRxiv* (2024) doi:10.1101/2024.02.28.582527.

543 23. Kallifidas, D. *et al.* The zinc-responsive regulator Zur controls expression of the  
544 coelibactin gene cluster in *Streptomyces coelicolor*. *J. Bacteriol.* **192**, 608–611 (2010).

545 24. Wang, R. *et al.* Identification of two-component system AfsQ1/Q2 regulon and its cross-  
546 regulation with GlnR in *Streptomyces coelicolor*. *Mol. Microbiol.* **87**, 30–48 (2013).

547 25. Kim, Y., Roe, J.-H., Park, J.-H., Cho, Y.-J. & Lee, K.-L. Regulation of iron homeostasis  
548 by peroxide-sensitive CatR, a Fur-family regulator in *Streptomyces coelicolor*. *J.*  
549 *Microbiol.* **59**, 1083–1091 (2021).

550 26. Flores, F. J. & Martín, J. F. Iron-regulatory proteins DmdR1 and DmdR2 of  
551 *Streptomyces coelicolor* form two different DNA-protein complexes with iron boxes.  
552 *Biochem. J* **380**, 497–503 (2004).

553 27. Barona-Gómez, F. *et al.* Multiple biosynthetic and uptake systems mediate siderophore-  
554 dependent iron acquisition in *Streptomyces coelicolor* A3(2) and *Streptomyces*  
555 *ambofaciens* ATCC 23877. *Microbiology* **152**, 3355–3366 (2006).

556 28. Hiard, S. *et al.* PREDetector: a new tool to identify regulatory elements in bacterial  
557 genomes. *Biochem. Biophys. Res. Commun.* **357**, 861–864 (2007).

558 29. Lee, N. *et al.* Iron competition triggers antibiotic biosynthesis in *Streptomyces coelicolor*  
559 during coculture with *Myxococcus xanthus*. *ISME J.* **14**, 1111–1124 (2020).

560 30. Parsana, P. *et al.* Addressing confounding artifacts in reconstruction of gene co-  
561 expression networks. *Genome Biol.* **20**, 94 (2019).

562 31. Wang, Y., Hicks, S. C. & Hansen, K. D. Co-expression analysis is biased by a mean-  
563 correlation relationship. *bioRxiv* 2020.02.13.944777 (2020)  
564 doi:10.1101/2020.02.13.944777.

565 32. Ronan, J. L. *et al.* Desferrioxamine biosynthesis: diverse hydroxamate assembly by  
566 substrate-tolerant acyl transferase DesC. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**,  
567 (2018).

568 33. Cruz-Morales, P. *et al.* Actinobacteria phylogenomics, selective isolation from an iron  
569 oligotrophic environment and siderophore functional characterization, unveil new

570 desferrioxamine traits. *FEMS Microbiol. Ecol.* **93**, (2017).

571 34. Tong, Y. *et al.* Highly efficient DSB-free base editing for streptomycetes with CRISPR-  
572 BEST. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 20366–20375 (2019).

573 35. Han, E. J., Lee, S. R., Hoshino, S. & Seyedsayamdst, M. R. Targeted discovery of  
574 cryptic metabolites with antiproliferative activity. *ACS Chem. Biol.* **17**, 3121–3130  
575 (2022).

576 36. Groenewold, G. S. *et al.* Collision-induced dissociation tandem mass spectrometry of  
577 desferrioxamine siderophore complexes from electrospray ionization of  $\text{UO}_2^{2+}$ ,  $\text{Fe}^{3+}$  and  
578  $\text{Ca}^{2+}$  solutions. *J. Mass Spectrom.* **39**, 752–761 (2004).

579 37. Sidebottom, A. M., Karty, J. A. & Carlson, E. E. Accurate mass MS/MS/MS analysis of  
580 siderophores ferrioxamine B and E1 by collision-induced dissociation electrospray mass  
581 spectrometry. *J. Am. Soc. Mass Spectrom.* **26**, 1899–1902 (2015).

582 38. Bierman, M. *et al.* Plasmid cloning vectors for the conjugal transfer of DNA from  
583 *Escherichia coli* to *Streptomyces* spp. *Gene* **116**, 43–49 (1992).

584 39. Yang, J., Banas, V. S., Rivera, G. S. M. & Wencewicz, T. A. Siderophore Synthetase  
585 DesD Catalyzes N-to-C Condensation in Desferrioxamine Biosynthesis. *ACS Chem.*  
586 *Biol.* **18**, 1266–1270 (2023).

587 40. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo  
588 generator. *Genome Res.* **14**, 1188–1190 (2004).

589 41. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python.  
590 *Bioinformatics* **36**, 2272–2274 (2020).

591 42. Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search  
592 for position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181–3182  
593 (2009).

594 43. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer Science &  
595 Business Media, 2009).

596 44. Harrison, P. W. *et al.* The European Nucleotide Archive in 2020. *Nucleic Acids Res.* **49**,  
597 D82–D85 (2021).

598 45. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data.

599 *Babraham Bioinformatics* <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

600 (2010).

601 46. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21

602 (2013).

603 47. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,

604 2078–2079 (2009).

605 48. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV):

606 high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**,

607 178–192 (2013).

608 49. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for

609 assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

610 50. Johnson, K. A. & Krishnan, A. Robust normalization and transformation techniques for

611 constructing gene coexpression networks from RNA-seq data. *Genome Biol.* **23**, 1

612 (2022).

613 51. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network

614 analysis. *BMC Bioinformatics* **9**, 1–13 (2008).

615 52. Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation,

616 chemical structures and visualisation. *Nucleic Acids Res.* (2023)

617 doi:10.1093/nar/gkad344.

618 53. Blin, K. *Ncbi-Genome-Download*. (Zenodo, 2023). doi:10.5281/ZENODO.8192432.

619 54. Reitz, Z. L. *MultiSMASH*. (2023). doi:10.5281/zenodo.8276144.

620 55. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale

621 biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).

622 56. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic

623 tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

624 57. MacNeil, D. J. *et al.* Analysis of *Streptomyces avermitilis* genes required for avermectin

625 biosynthesis utilizing a novel integration vector. *Gene* **111**, 61–68 (1992).

626 58. Kieser, T. *Practical Streptomyces Genetics*. (2000).

627 59. Schmid, R. *et al.* Integrative analysis of multimodal mass spectrometry data in MZmine

628 3. *Nat. Biotechnol.* **41**, 447–449 (2023).

629 60. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. One Step Forward for Reducing

630 False Positive and False Negative Compound Identifications from Mass Spectrometry

631 Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and

632 Detecting Chromatographic Peaks. *Anal. Chem.* **89**, 8696–8703 (2017).

633 61. Lu, Y., Pang, Z. & Xia, J. Comprehensive investigation of pathway enrichment methods

634 for functional interpretation of LC-MS global metabolomics data. *Brief. Bioinform.* **24**,

635 (2023).

636 62. Caswell, T. A. *et al.* *Matplotlib/Matplotlib: REL: V3.7.2.* (Zenodo, 2023).

637 doi:10.5281/ZENODO.8118151.

638 63. Zhang, L. *et al.* An Alternative and Conserved Cell Wall Enzyme That Can Substitute for

639 the Lipid II Synthase MurG. *MBio* **12**, (2021).

640 64. Blin, K., Pedersen, L. E., Weber, T. & Lee, S. Y. CRISPy-web: An online resource to

641 design sgRNAs for CRISPR applications. *Synth Syst Biotechnol* **1**, 118–121 (2016).

642