Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity

Alexander Crits-Christoph[1,2], Nicholas Bhattacharya[3], Matthew R. Olm[4], Yun S. Song[5,6,9], and Jillian F. Banfield[2,4,7,8,9+]

[+]Corresponding author: jbanfield@berkeley.edu
[1]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA
[2]Innovative Genomics Institute, Berkeley, CA, 94704, USA
[3]Department of Mathematics, University of California, Berkeley, CA, USA
[4]Department of Microbiology and Immunology, Stanford University, CA, USA
[5]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA
[6]Department of Statistics, University of California, Berkeley, CA, USA
[7]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA
[8]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
[9]Chan Zuckerberg Biohub, San Francisco, CA, USA

Running title: Transporter genes in biosynthetic gene clusters
Keywords: Transporters, specialized metabolism, siderophore

Abstract

Biosynthetic gene clusters (BGCs) are operonic sets of microbial genes that synthesize specialized metabolites with diverse functions, including siderophores and antibiotics, which often require export to the extracellular environment. For this reason, genes for transport across cellular membranes are essential for the production of specialized metabolites, and are often genomically co-localized with BGCs. Here we conducted a comprehensive computational analysis of transporters associated with characterized BGCs. In addition to known exporters, in BGCs we found many importer-specific transmembrane domains that co-occur with substrate binding proteins possibly for uptake of siderophores or metabolic precursors. Machine learning models using transporter gene frequencies were predictive of known siderophore activity, molecular weights, and a measure of lipophilicity (log *P*) for corresponding BGC-synthesized metabolites. Transporter genes associated with BGCs were often equally or more predictive of metabolite features than biosynthetic genes. Given the importance of siderophores as pathogenicity factors, we used transporters specific for siderophore BGCs to identify both known and uncharacterized siderophore-like BGCs in genomes from metagenomes from the infant and adult gut microbiome. We find that 23% of microbial genomes from the infant gut have siderophore-like BGCs, but only 3% of those assembled from adult gut microbiomes do. While siderophore-like BGCs from the infant gut are predominantly associated with *Enterobactericaee* and *Staphylococcus*, siderophore-like BGCs can be identified from taxa in the adult gut microbiome that have rarely been recognized for siderophore production. Taken together, these results show that consideration of BGC-associated transporter genes can inform predictions of specialized metabolite structure and function.

Introduction

Microbes produce specialized metabolites with diverse functions, including siderophores, ionophores, antibiotics, antifungals, and signalling molecules (Osbourn 2010). Specialized metabolites therefore often underlie both cooperative and competitive interactions between microbes, and microbial interactions with the physiochemical environment (Davies 2013; Sharon et al. 2014; Tyc et al. 2017). The vast majority of specialized metabolites in bacteria are produced by biosynthetic gene clusters (BGCs), which are sets of genomically colocalized genes. Colocalization of genes into BGCs is thought to occur because of selection for co-inheritance and co-regulation (Fischbach, Walsh, and Clardy 2008). While thousands of microbial natural products have been characterized, genomic BGC predictions made using programs such as antiSMASH (Blin, Shaw, et al. 2019) and ClusterFinder (Cimermancic et al. 2014) suggest that characterized molecules represent just a small fraction of all existing microbial natural products (Medema and Fischbach 2015; Kim et al. 2017). Many of these unknown metabolites may be highly novel due to enzymatic and combinatorial diversity of genes in BGCs (Jenke-Kodama, Börner, and Dittmann 2006; Chevrette et al. 2019).

Because of the sheer number of sequenced but otherwise uncharacterized BGCs and the time and costs required for chemical characterization, there is a pressing need for predictions of BGC metabolite structures or functions to enable prioritization of targets for laboratory study (Tran et al. 2019). Prediction of metabolite structure or function for a novel BGC from gene content alone is challenging. For many biosynthetic nonribosomal peptide synthetases (NRPS) and polyketide synthases (PKS), there is a ‘co-linear’ assembly-line regulation in which the order of genes relates to the order of enzymatic modifications on the metabolite during synthesis (Fischbach and Walsh 2006). Using this co-linearity rule can help predict some degree of structural detail in NRPSs and PKSs, as is done by by antiSMASH and PRISM (Skinnider et al.

2017), but there are many known exceptions to this rule (Wenzel and Müller 2005), and the accuracies of these software predictions have not been formally assessed using a large training dataset.

Prediction of BGC metabolite function generally relies on contextual genes associated with BGCs. The observation that genes conferring resistance to the produced metabolite are also colocalized with the BGC motivates investigation of putative resistance genes (self-resistance gene mining (Yan, Liu, and Tang 2020)) for functional prediction. For siderophore activity prediction, AntiSMASH assigns a functional 'siderophore' label for BGCs that contain the IucA / IucC gene family, but this gene is only specific for siderophores with biosynthetic pathways similar to aerobactin (Hider and Kong 2010). More recently, Hannigan et al. (2019) trained neural networks to both identify BGCs in genomes and classify BGCs by known metabolite functions. These networks used Protein Families from the PFAM database ("PFAM Database," n.d.) (PFAMS) found in each BGC as features. They predicted activity labels of antibacterial, antifungal, cytotoxic, and inhibitor, with precisions of 36%, 47%, 61%, and 69% on each class respectively.

Many specialized metabolites perform their ecological roles extracellularly, and thus require transport across cellular membranes. Transporter genes often colocalize in BGCs and have been shown to be compound specific and necessary for export of the product in many cases (Severi and Thomas 2019; Martín, Casqueiro, and Liras 2005; Méndez and Salas 2001). Therefore, transporters may also inform predictions of BGC metabolite structure and function. The distribution of transporters associated with biosynthetic gene clusters has so far been assessed only in characterized BGCs with experimental validation, a small fraction of the total number BGCs sequenced. At least 40 BGC-associated exporters have been characterized, mostly in the *Actinomycetes*, with varying degrees of experimental validation (Severi and Thomas 2019).

89 Transporters associated with BGCs are commonly either ATP-dependent active
90 transporters or ion-gradient dependent transporters (Martín, Casqueiro, and Liras 2005). ATP-
91 dependent transporters include the ATP-binding cassette (ABC) superfamily of both importers
92 and exporters (Rees, Johnson, and Lewinson 2009), and the MacB tripartite efflux pump (Greene
93 et al. 2018b). Examples of characterized structures of each transporter class and their substrates
94 are shown in **Fig 1a**. In brief, Type I ABC importers are characterized by the *BPD_transp_1*
95 transmembrane (TM) protein family, and include MalFGK and MetNI for malate and methionine
96 import in *E. coli* (Beek et al. 2014). Type II ABC importers are characterized by the *FecCD* TM
97 protein family, and examples include BtuCD and HmuUV (Beek et al. 2014), and the FecBCDE
98 system for Iron(III) dicitrate import in *Escherichia coli* (Staudenmaier et al. 1989). Both types
99 of ABC importers often associate with Substrate Binding Proteins (SBPs), small membrane or
100 periplasmic proteins for substrate uptake (Beek et al. 2014; Berntsson et al. 2010). Periplasmic
101 binding proteins, Type II ABC importers, and TonB-dependent receptors are also known to play
102 key roles in siderophore uptake in multiple bacterial species (Ellermann and Arthur 2017).
103 Meanwhile, examples of ABC exporters include McjD for lasso peptide microcin J25
104 export (Romano et al. 2018) and the *Staphylococcus aureus* multidrug exporter Sav1866
105 (Dawson and Locher 2007), composed of the *ABC_mem* TM protein family, while the O-antigen
106 polysaccharide exporter is composed of the *ABC2_membrane* TM protein family (Bi et al. 2018).
107 Export of Nystatin, Doxorubicin, and Mccj25 was found to be dependent upon ATP-dependent
108 transporters (Severi and Thomas 2019). The vast majority of ribosomally synthesized and post-
109 translationally modified peptides (RiPPs) and a number of antibiotics from *Actinomycetes* also
110 rely on characterized ATP-dependent ABC transporters (Gebhard 2012) (Méndez and Salas
111 2001).
112 Ion-gradient dependent transporters (also known as secondary active transport systems)
113 do not require ATP and facilitate transport of small molecules in response to chemiosmotic

114 gradients (Quistgaard et al. 2016). Those found in BGCs are often examples of the major
115 facilitator superfamily (MFS) and occasionally, the resistance nodulation division (RND) family
116 or the multidrug and toxic compound extrusion (MATE) family. Examples of characterized
117 secondary active transporters for antibiotics include an RND transporter for pyoluteorin, and
118 MFS transporters specific for mitomycin C, virginiamycin S, and landomycin (Severi and
119 Thomas 2019).

120 Thousands of BGCs with chemically characterized metabolites open up the possibility for
121 a broad genomic and computational analysis of phylogenetically and functionally diverse BGCs.
122 Here, we used a curated version of the Minimum Information about a Biosynthetic Gene cluster
123 database (MIBiG 2.0) of BGCs (Kautsar et al. 2020) and selected transporter-specific protein
124 domain hidden markov models (HMMs) to perform a wide genomic assessment of the
125 distribution of transporters in BGCs. We found clear correlations between transporter domains
126 and corresponding metabolite features, especially siderophore activity, that indicate underlying
127 logical structure to transporter associations and can inform functional and structural prediction of
128 specialized metabolites from genomics alone.

129 **Results**

130 Genome mining of transporters associated with biosynthetic gene clusters

131 Using two compiled sets of transporter-specific HMMs (PFAM and CATHDB), we
132 cataloged all classes of transporters across the MIBiG 2.0 database of characterized and
133 experimentally validated biosynthetic gene clusters. We found that 56% of the bacterial BGCs in
134 MIBiG contained at least one PFAM transporter hit and an additional 6% contained a CATHDB
135 transporter hit without a PFAM domain (**Fig 1c**). These percentages increased among BCGs that
136 produce antibiotics (71%) and siderophores (78%), indicating that BGCs with these activities are
137 more likely to contain at least one transporter. BGCs with transporters contained 2.5 transporter-

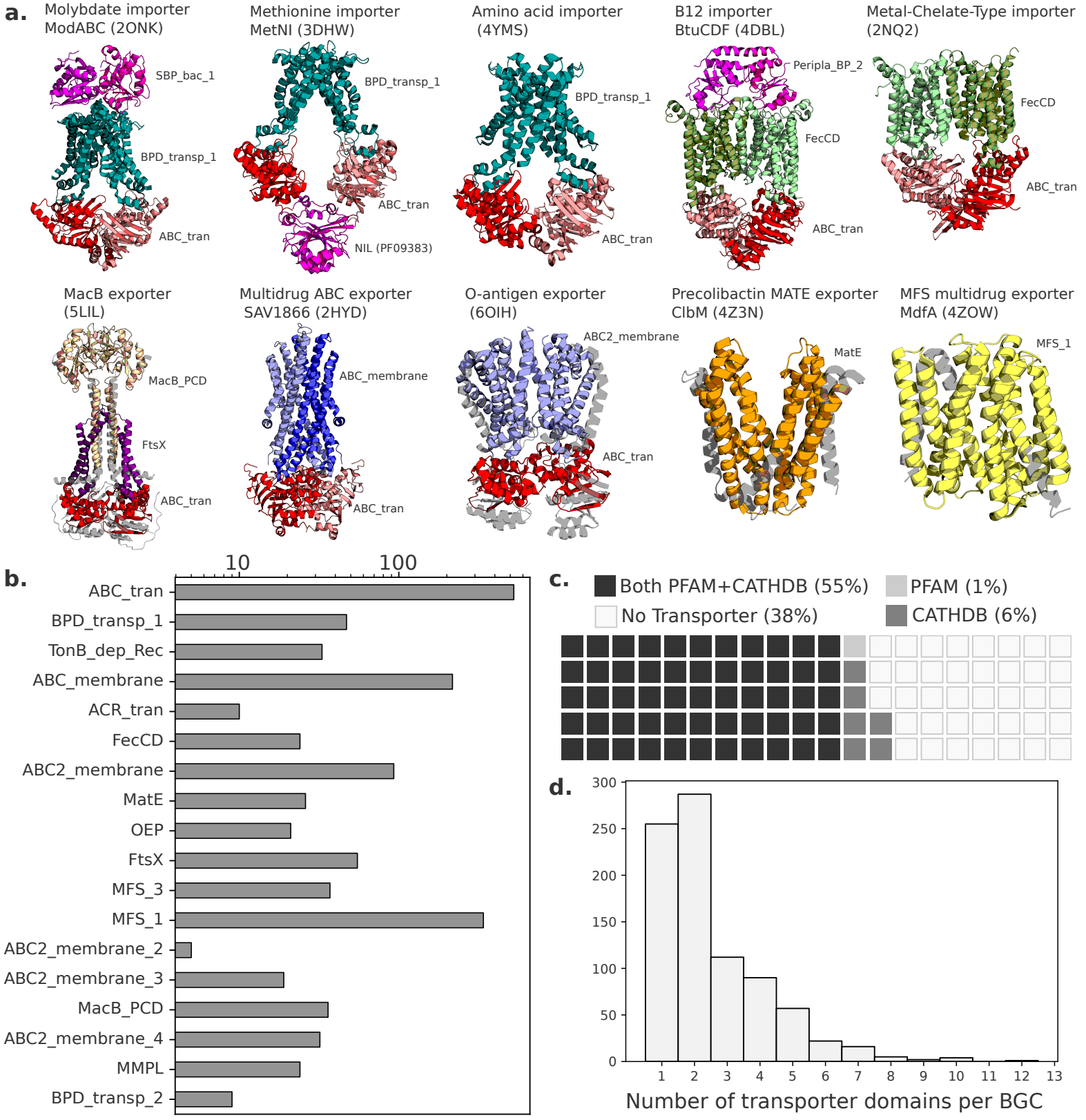

Figure 1 Distributions of transporter classes in biosynthetic gene clusters. (a) Structures of characterized examples of major transporter classes often found in BGCs, colored and labeled by PFAM domains. (b) The frequencies of common PFAM transporter domains across the bacterial BGCs in the MIBiG database. (c) The percentages of bacterial BGCs in MiBiG that do and do not contain transporter domains. Each square represents 1% of BGCs. (d) The counts of transporter domains per each bacterial BGC that contains at least 1 transporter gene across MiBiG.

138 associated domains across transport-annotated genes on average (**Fig 1d**), which is expected as
139 many ATP-dependent transporter systems have at least 2 domain complexes. However, some
140 BGCs contained considerably more transporter ORFs and domains, indicating that sometimes
141 multiple transport systems can be associated with one BGC (**Supplementary Tables S2-S3**).
142 The number of transporters in a BGC had no association with the number of metabolite
143 structures reported for that BGC. The ATP-binding ABC transporter (*ABC_tran*) domain and the
144 Major Facilitator Superfamily 1 (*MFS_1*) domain were the two most common transporter
145 domains found in BGCs (**Fig 1b**). A variety of proteins had nucleotide-binding domains along
146 with several different transmembrane domains – *ABC_membrane* and *ABC2_membrane* domains
147 were most common but *ABC2_membrane_2*, -*_3*, and -*_4* domains were also represented.
148 Examining domains specific for export, the *ABC_membrane* domain is often
149 characteristic of exporters (e.g., Sav1866 (Velamakanni et al. 2008)), but recently has been
150 reported in the genes for siderophore uptake (YbtPQ) in *Yersinia (Wang, Hu, and Zheng 2020)*,
151 and is therefore not necessarily indicative of export or import alone. The second most common
152 transmembrane domain, *ABC2_membrane*, has been observed in the O-antigen polysaccharide
153 exporter (Bi et al. 2018). The *MacB-FtsX* tripartite efflux pump was found in 60 BGCs, while the
154 RND (*ACR_tran*) efflux pump was less common, and found in only 10 BGCs. Other known
155 efflux systems, such as SMR, MatE, and the MFS families 2 through 5 were comparatively rare
156 across BGCs.
157 We next calculated co-occurrence correlations between all transporter protein families
158 across BGCs and observed a strong negative correlation between *MFS* transporters and ATP-
159 dependent transporters relying on the nucleotide binding domain, and a weaker negative
160 correlation of *MatE* domains from the ATP-dependent NBD (**Fig 2a**). This points towards a
161 dichotomous choice between ATP-dependent and ATP-independent transport associated with a
162 BGC.

163 Multiple lines of annotation evidence indicated that many of the transporter genes
164 associated with BGCs were likely to be importers. Importers can be involved in the uptake or re-
165 uptake of molecules like siderophores, and may also play roles in importing precursor
166 metabolites for a BGC. The membrane domains specific to Type I importers (*BPD_transp_1*)
167 and Type II importers (*FecCD*) were the most often observed ATP-dependent transmembrane
168 domains besides *ABC_membrane* and *ABC2_membrane* (**Fig 1b**). CATH Protein Structure
169 Classification database HMMs (CATHDBs) that were specific for importer families in the
170 Transporter Classification Database and found in BGCs included permeases for sugars,
171 oligopeptides, and iron siderophores (**Fig 2c**). 11% of BGCs with a transporter also contained a
172 substrate binding protein. Among the substrate binding proteins we searched for, the most
173 common contained domain was *Peripla_BP_2*, also found in the *E. coli* B12 importer complex
174 BtuCDF, and variants of this protein are specific for siderophores and cobalamin (Berntsson et
175 al. 2010). We also observed many substrate binding proteins with specificities predicted to
176 include carbohydrates, oligopeptides, and peptide uptake (Berntsson et al. 2010) (**Fig 2b**).
177 In the co-occurrence data, we observed pairing of different substrate binding proteins
178 with different transmembrane domains. *Peripla_BP_2* positively correlated strongly with *FecCD*
179 and *TonB_dep_Rec*, genes known to be involved in siderophore uptake. *BPD_trans_1* co-
180 occurred with either *SBP_bac_5* or *SBP_bac_1*, while *BPD_transp_2* co-occurred with
181 *Peripla_BP_4*. The *ABC2_Membrane_2* domain loosely correlated with *SBP_bac_3*, while the
182 largely uncharacterized *MMPL* transmembrane domain correlated with *SBP_bac_8*. Taken
183 together, these results show a logical organization of importer-specific transporter domains
184 within BGCs that may be involved in either siderophore uptake, precursor uptake, or other roles.
185 Regardless of the substrate specificity of these proteins, care must be taken when assuming that a
186 transporter in a BGC is definitively for export of the matured product.

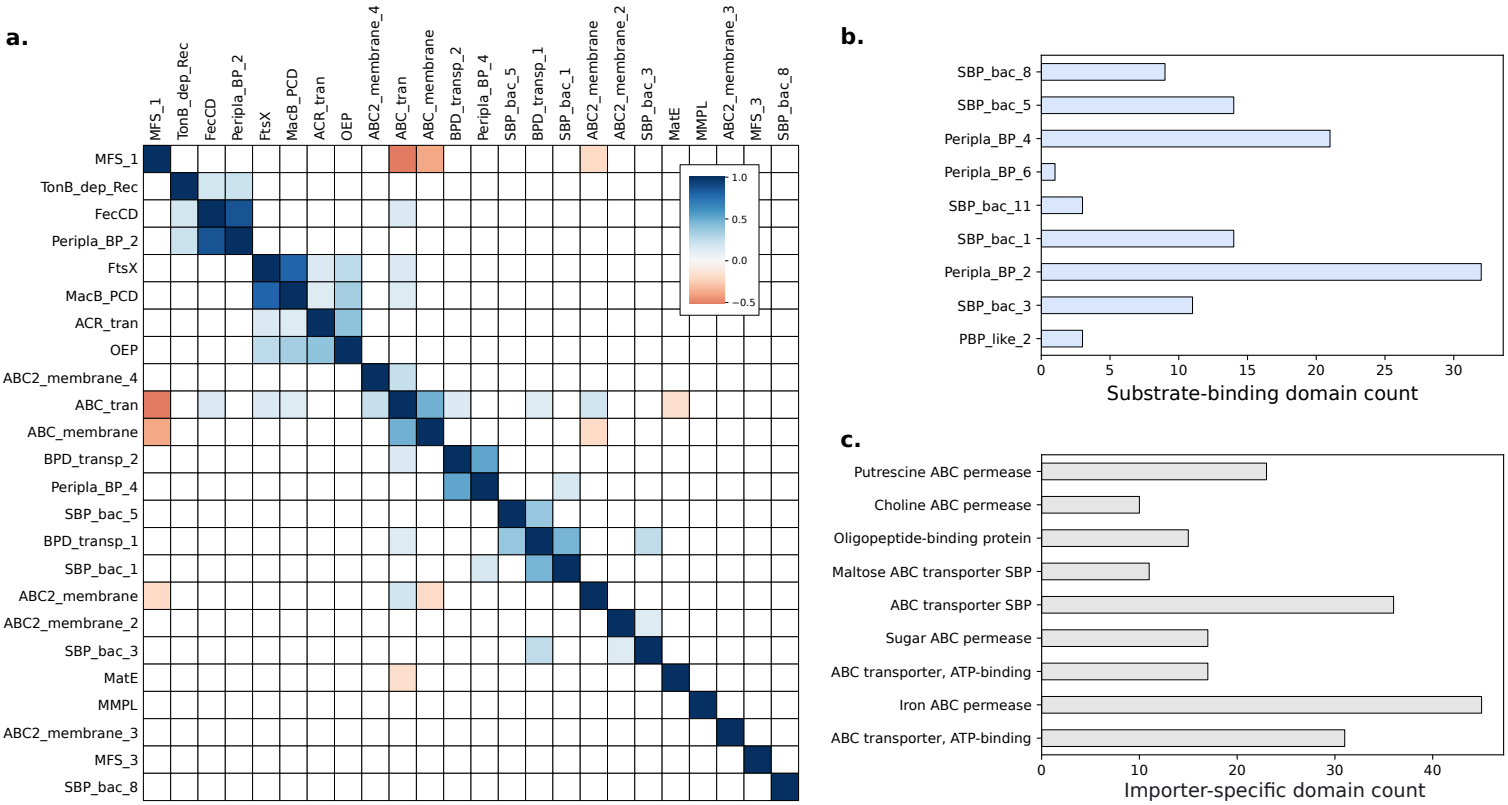

Figure 2: Presence of importer-specific domains and co-occurrence between transporters across BGCs. (a) Spearman correlations between commonly occurring PFAM transporter domains across MiBiG BGCs - only correlations with $p < 0.001$ are shown. (b) Counts of PFAM transporter substrate binding domain families, and (c) and counts of importer-specific CATHDB domains across MiBiG BGCs. The CATHDB functional family “Iron ABC permease” is essentially synonymous with the FecCD PFAM.

187 Prediction of siderophore and antibacterial activity from biosynthetic transporters

188 Because transporters are required for the ecological functions of, and are in close contact
189 with, biosynthesized specialized metabolites, we used machine learning to test if transporter
190 classes were predictive of BGC-synthesized metabolite structures and functions. We noticed that
191 metabolite activity labels in MIBiG were strongly associated with phylogeny: 83% of antibiotic
192 BGCs were from gram-positive bacteria while only 40% of siderophore BGCs were from gram-
193 positive bacteria in the dataset of curated MIBiG BGCs. Although this pattern may indicate that
194 the incidence of these functions varies substantially with phylogeny, it is also possible that it
195 reflects a sampling or culturability bias. To reduce the impact of this potential bias, we created
196 separate training and testing datasets for activity prediction for gram-positive and gram-negative
197 organisms.

198 Using our curated set of BGCs with transporters from the MIBiG 2.0 database, we tested
199 for associations between BGC transport genes and metabolite functions. We generated two
200 activity classification tasks: (a) distinguishing siderophores (including known ionophores) (n=16
201 gram-positive, n=24 gram-negative), from non-siderophores (n=142 gram-positive, n=42 gram-
202 negative), and (b) distinguishing antibiotics and antifungals (n=131 gram-positive, n=27 gram-
203 negative) from non-antibiotics (n=57 gram-positive, n=37 gram-negative). We observed several
204 statistically significant (Fisher's exact test; $q<0.05$) associations in the distribution of transporter
205 types between both the siderophore and other activity classes. Among gram-positive bacteria,
206 60% of siderophore BCGs contained the SBP *Peripla_BP_2* and 55% contained the *FecCD*
207 importer, while no BGCs with other activities had either (**Fig 3a; Fig S1**). The situation was
208 similar for gram-negative siderophore BGCs. The TonB-dependent receptor (completely absent
209 from gram-positive bacteria) was the strongest signal, found in almost 80% of gram-negative
210 siderophore BGCs with a transporter and never in BGCs with other activities.

211 To assess siderophore predictability from BGC gene content, we used simple decision
212 trees with only two layers applied to different feature sets of protein domain annotations-
213 transport-affiliated PFAMs, transport-affiliated CATHDB HMMs, and biosynthetic PFAMs. To
214 avoid issues with class imbalance, we report precision and recall on the siderophore class, as
215 siderophore prediction requires searching for a minority class (siderophores) within a
216 background of mostly non-siderophores. With the transport-only features, we found that just with
217 two gene decisions, it is possible to achieve 100% precision with over 80% recall using either
218 PFAM or CATHDB transporter annotations for gram-negative siderophores, and 100% precision
219 with over 80% recall using CATHDB transporter annotations for gram-positive siderophores
220 (**Fig 3b; Supplementary Table S5**) On the other hand, when using all biosynthetic annotations
221 within BGCs we found that 2-layer decision trees trained on biosynthetic genes performed
222 substantially worse at predicting siderophore activity than those trained on transporter genes (**Fig**
223 **3b; Supplementary Table S5**). We further validated our results by training LASSO linearized
224 regression models, which do not model interactions between features. These models obtained a
225 slightly improved area under the precision-recall curve, indicating that very simple transporter
226 patterns are highly predictive of whether a BGC is siderophore producing or not in our dataset
227 (**Fig 3b**). Transporter features predictive of siderophores were consistently selected by LASSO
228 across stratified cross-validation repeats, giving evidence that these patterns are robust (**Fig S2**).
229 The top predictive biosynthetic features of siderophores were the lucA/lucC protein family (used
230 by antiSMASH to label siderophores and which is known to be involved in aerobactin
231 biosynthesis) and condensation domains (likely to capture non-ribosomal peptide siderophores),
232 but predictive effect sizes were smaller than those for transporters (**Fig S2**).
233 Using our models for siderophores, we attempted to predict siderophore classes for all
234 gene clusters that have no annotated function in MIBiG2 (and that we had not already hand
235 curated). We searched for gene clusters containing the siderophore-predictive genes *FecCD* and

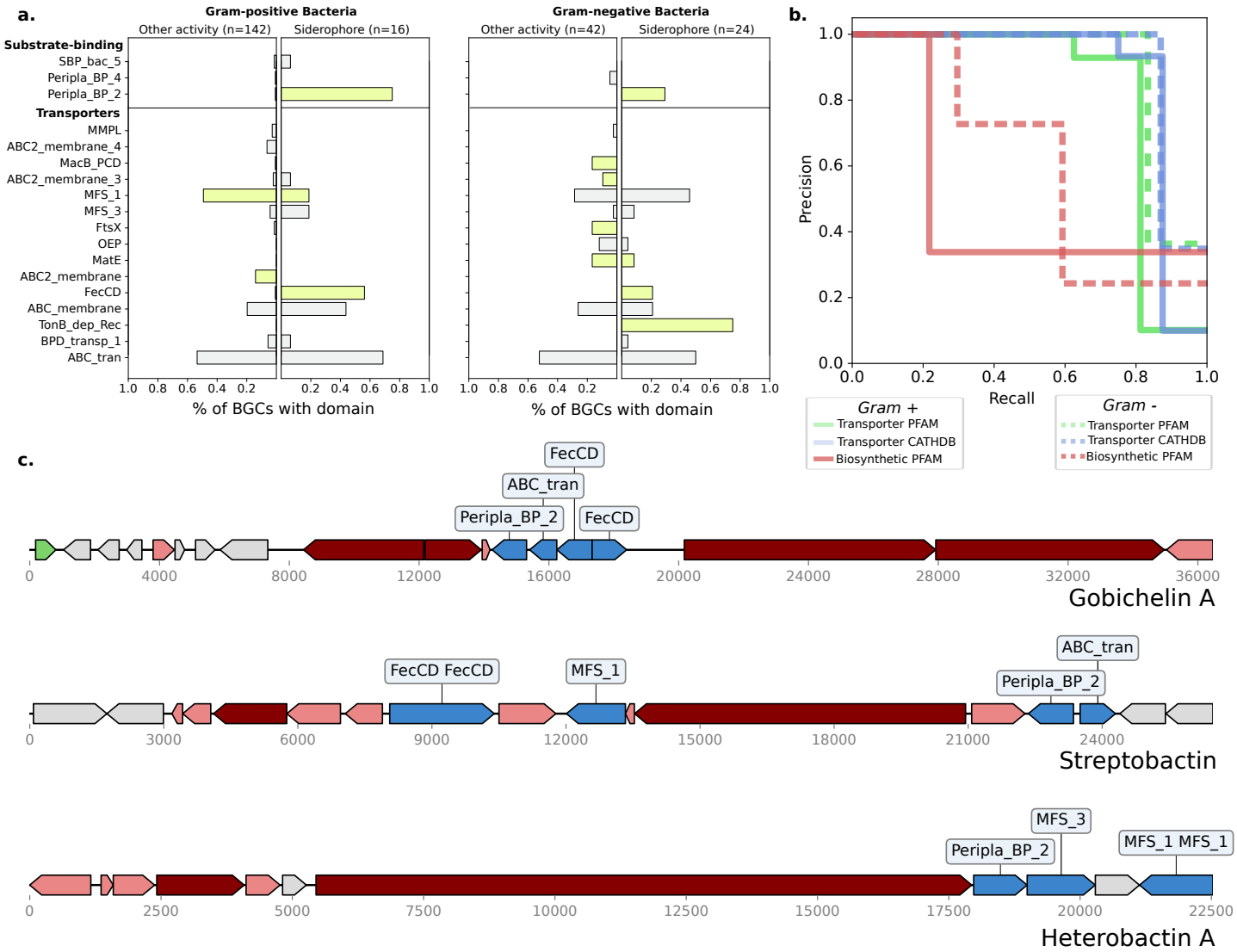

Figure 3: Transporter domains are predictive of siderophore BGCs. (a) The frequencies of common transporter PFAM domains across siderophore BGCs and BGCs of other known activities in gram-positive and gram-negative bacteria. Bars in green were significantly different in frequency between the two classes (Fisher's exact test; $q<0.05$) (b) Precision-Recall curves for 2-layer decision trees classifying siderophore BGCs using PFAM transporter, CATHDB transporter, and PFAM biosynthetic gene features in gram-negative and gram-positive bacteria. (c) Examples of three siderophore BGCs without activity labels in MiBiG 2.0, which could be identified using transporter frequencies. Transporter genes are colored in blue, core biosynthetic genes (NRPS and PKS) are colored in dark red, accessory biosynthetic genes in light red, and regulatory genes in green.

236 *Peripla_BP_2* and found 6 additional BGCs with no annotated activity, three of which were
237 experimentally validated by the literature to be siderophores (Matsuo et al. 2011; Y. Chen et al.
238 2013; Carran et al. 2001) (**Fig 3c**). The remaining three identified BGCs were false positives.
239 One of them, the BGC for the antibiotic Ficellomycin, only contained these transport genes in
240 flanking regions not known to be involved in biosynthesis (Liu et al. 2017) while the herbimycin
241 A BGC contains the transporters genes in the reverse reading frame from the BGC, separated by
242 an unusual 16 Kbp intergenic region with only 7% coding density, and the genes appear to be
243 fragmented (Rascher et al. 2005). The other false positive was Lividomycin, an antibiotic that
244 does seem to be a rare non-siderophore BGC with *FecCD* and *Peripla_BP_2* transporters in the
245 MIBiG2.0 database.

246 Classifying antibiotics and antifungals from either transporters or biosynthetic genes
247 proved more challenging than classifying siderophores (**Supplementary Table S5; Fig S3**). In
248 gram-negative bacteria, we observed positive associations between the MacB-FtsX tripartite
249 efflux pump with antibacterial or antifungal activity (Fisher's exact test; $q<0.05$). MacB and
250 associated components (FtsX and OEP) were positively associated with antibiotic activity by
251 LASSO logistic regression. This result was stable across both cross-validation folds and repeats
252 of cross-validation (**Fig S3**). We found that 7 out of 27 gram-negative antibacterial BGCs
253 contained a MacB, while no BGCs in our classes of other activities contained MacB. Although
254 MacB is involved in export of the siderophore Pyoverdine (for which there is not an accurate
255 BGC in MiBiG) in *Pseudomonas* (Greene et al. 2018a), in general MacB may be a strong
256 indicator of antibacterial activity for a BGC. Previously, we identified a number of MacB-FtsX
257 exporters in BGCs from novel Acidobacteria (Crits-Christoph et al. 2018), possibly indicating a
258 role in antibacterial activity for these BGCs. LASSO effect sizes for individual biosynthetic
259 genes were substantially lower (**Fig S3**).

260 Association of biosynthetic transporter classes with the molecular weights and lipophilicity of
261 their putative substrates

262 We next hypothesized that transporter classes could be predictive of other molecular
263 features beyond functional activity. There was no strong correlation between the molecular size
264 of the metabolite produced and Gram status of the bacteria encoding each corresponding BCG in
265 the MiBiG dataset. Thus, we tested for differences in transporter classes in BGCs producing
266 metabolites that were (a) less than and (b) greater than 1000 Da in size across all bacteria. There
267 was a striking difference in the frequencies of some transporters between BGCs with different
268 metabolite molecular weights (**Fig 4a**). The strongest difference was in the distribution of MFS
269 transporters - found in 57% of BGCs with products under 1000 Da, but only 14% of those over
270 1000 Da (Fisher's exact test; $q<0.05$). The 95th percentile of metabolite molecular weights for
271 clusters with *MFS_1* and without *ABC_tran* was 1082 Da, which may be approaching a
272 biological limit for the molecular weights of substrates for these transporters. Conversely, the
273 ATP-dependent *ABC_tran* domain was found in 89% of BGCs producing high molecular weight
274 compounds but only 42% of those producing low molecular weight compounds (Fisher's exact
275 test; $q<0.05$). *MacB/FtsX* and the rarer transmembrane domains were also associated with higher
276 molecular weight compounds. In particular, *ABC2_membrane_4* was most associated with high
277 molecular weight compounds, with almost all of the BGCs in which it is found in producing
278 compounds over 1500 Da in size (**Fig 4c; Fig S4**). After training both LASSO logistic regression
279 and 2-layer decision tree models to classify whether produced molecules are greater than 1000
280 Da, we found that transporter genes were able to distinguish large from small metabolites with
281 moderate precision and recall (**Supplementary Table S5**) and an AUPRC of up to 42%, while
282 biosynthetic genes performed similarly (**Fig 4b**). Top transporter features consistently had larger
283 effect sizes than top biosynthetic features, indicating that transporter-based features provided
284 clearer signals. This result was stable across both cross-validation folds and repeats. (**Fig S4**).

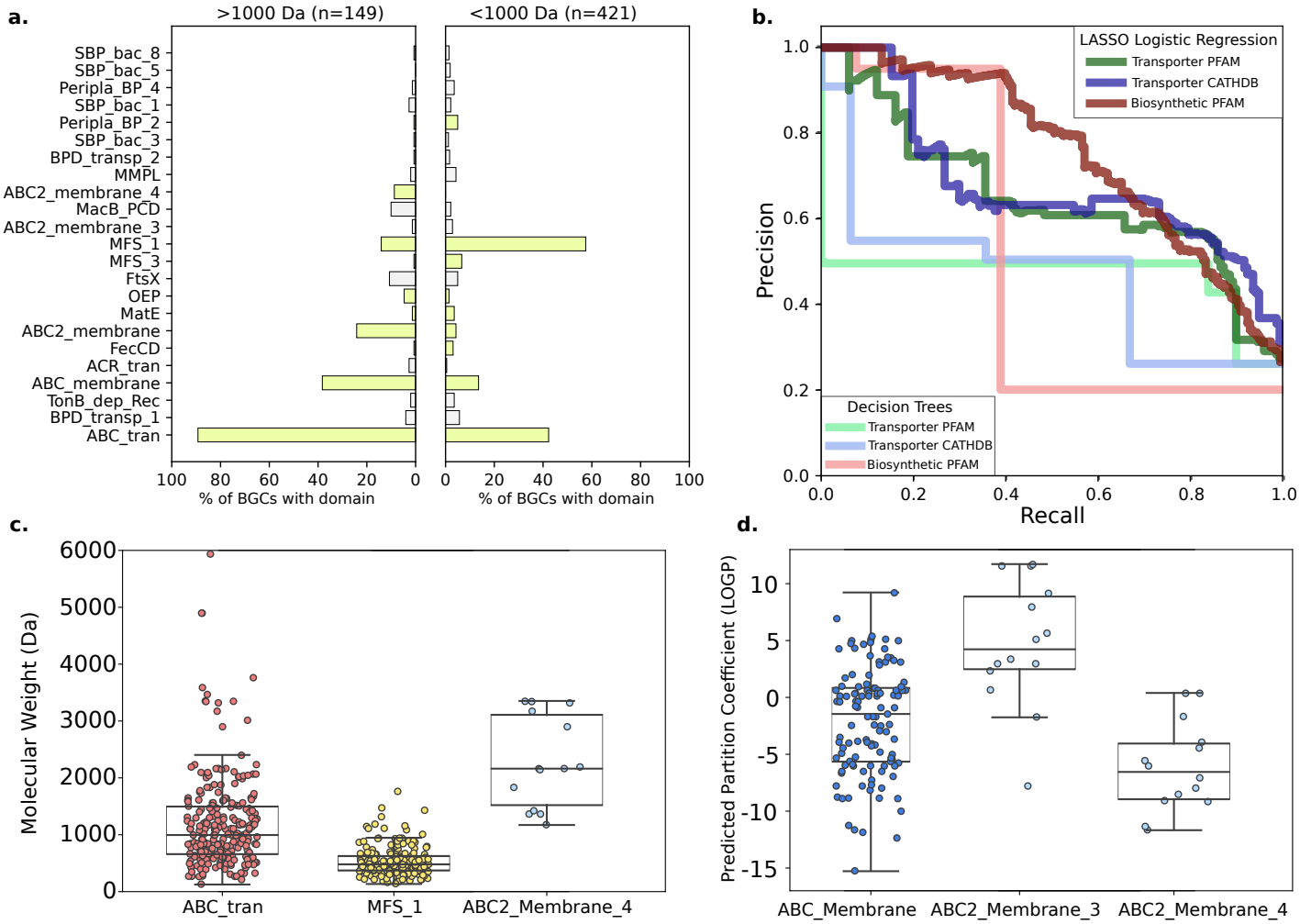

Figure 4: Transporter domains associated with molecular size and partition coefficient. (a) The frequencies of common transporter PFAM domains in BGCs that synthesize metabolites larger than 1000 Da (left) and smaller than 1000 Da (right). Bars in green were significantly different in frequency between the two classes (Fisher's exact test; $q<0.05$) (b) Precision-Recall curves for 2-layer decision trees and LASSO logistic regression models classifying BGCs producing metabolites >1000 Da using PFAM transporter, CATHDB transporter, and PFAM biosynthetic gene features. (c) The distribution of metabolite molecular weights synthesized by BGCs with at least one NBD-binding ABC transporter domain, at least one MFS domain, and the ABC2_Membrane_3 transmembrane domain. (d) Predicted partition coefficients (log P) for metabolites synthesized by BGCs that contain at least one variant of three different ABC transporter transmembrane domains.

284 The biosynthetic protein family most associated with high molecular weight metabolites was
285 *Glycos_transf_2*, likely due to the addition of sugar groups to metabolites by these enzymes in
286 BGCs.

287 It has previously been reported that transporters can be specific for compounds with
288 similar hydrophilicity (Rempel et al. 2020). The lipophilicity of a metabolite is often considered
289 critical for its success in clinical development for human therapeutics (Arnott and Planey 2012).
290 With LASSO logistic regression, we predicted partition coefficients (log *P*), a measure of
291 lipophilicity, for all of the metabolites in MiBIG, and tested how well metabolite partition
292 coefficients could be predicted by gene content. We found that the presence of 5 transporter
293 classes was significantly associated with increased lipophilicity (Fisher's exact test; $q<0.05$). In
294 particular, we observed an association between varying ATP-dependent transmembrane domains
295 and log *P*, with *ABC2_Membrane_3* domain co-occurring with BGC-metabolites with a high log
296 *P* (median 4.4) and *ABC2_Membrane_4* domain co-occurring with BGC-metabolites with a low
297 log *P* (median -6.2) (**Fig 4d**). LASSO logistic regression distinguished log P > 0 from log P < 0
298 with 77% AUPRC (**Supplementary Table S5; Fig S5**). Interestingly, on this task biosynthetic
299 genes were distinctly superior over transporters at prediction of log P, obtaining a 83% AUPRC
300 with a LASSO logistic regression trained on biosynthetic genes (**Supplementary Table S5**).

301 <u>Identifying novel siderophore-like biosynthetic gene clusters in the human microbiome</u>

302 To demonstrate the predictive utility of BGC-associated transporters, we used
303 antiSMASH to annotate all BGCs in sets of genomes assembled from metagenomes from (a) the
304 infant gut microbiome in a study of premature infants from a neonatal intensive care unit (Olm et
305 al. 2019) and (b) a cross-study collation of genomes assembled from multiple human gut studies
306 (Nayfach et al. 2019). We searched these BGCs for the two sets of transporter classes that can
307 achieve near 100% siderophore specificity in the MiBiG database - (a) *Peripla_BP_2* and

307 *FecCD* in gram-positive bacteria, and (b) *Peripla_BP_2*, *FecCD*, and *TonB_dep_Rec* in gram-
308 negative bacteria. We annotated BGCs with either all of the first two genes or all of the latter
309 three genes as "siderophore-like", i.e., BGCs that putatively produce siderophores. We identified
310 1442 BGCs with siderophore-like transporter classes (**Fig 5a; Supplementary Table S6**) and
311 then grouped them into novel gene cluster families using BiG-SCAPE, resulting in 75
312 siderophore-like gene cluster families (**Fig 5b**). 23% of microbial genomes from the neonatal
313 infant gut microbiomes had siderophore-like BGCs, but only 3% of those assembled from adult
314 gut microbiomes did. Siderophore-like BGCs were identified across a range of bacterial genera,
315 but the vast majority of which were from the *Staphylococcus* or *Enterobacteriales*, which are
316 known to be in high abundance in the neonatal gut microbiome. The genera with the most
317 siderophore-like BGCs were *Staphylococcus* and *Klebsiella*, common hospital-acquired
318 pathogens of neonates. Five of the 75 identified siderophore-like BGC families included a known
319 representative gene cluster in MiBIG, and four of the known representatives were siderophores -
320 again pointing to the specificity of these transporter classes.
321 The rest of the siderophore-like gene cluster families that were identified had no closely
322 characterized representative, indicating that there is likely capacity for production of multiple
323 novel siderophores in the human gut microbiome (**Fig 5c**). In the adult gut microbiome samples,
324 these novel siderophore-like gene clusters were identified in phylogenetically diverse genomes,
325 particularly including members of the *Lachnospiraceae*, often considered to be important
326 commensals in the human gut (Duvallet et al. 2017; L. Chen et al. 2017). Most siderophore-like
327 BGCs were in large gene cluster families with other BGCs which also contained the same set of
328 transporter hits.

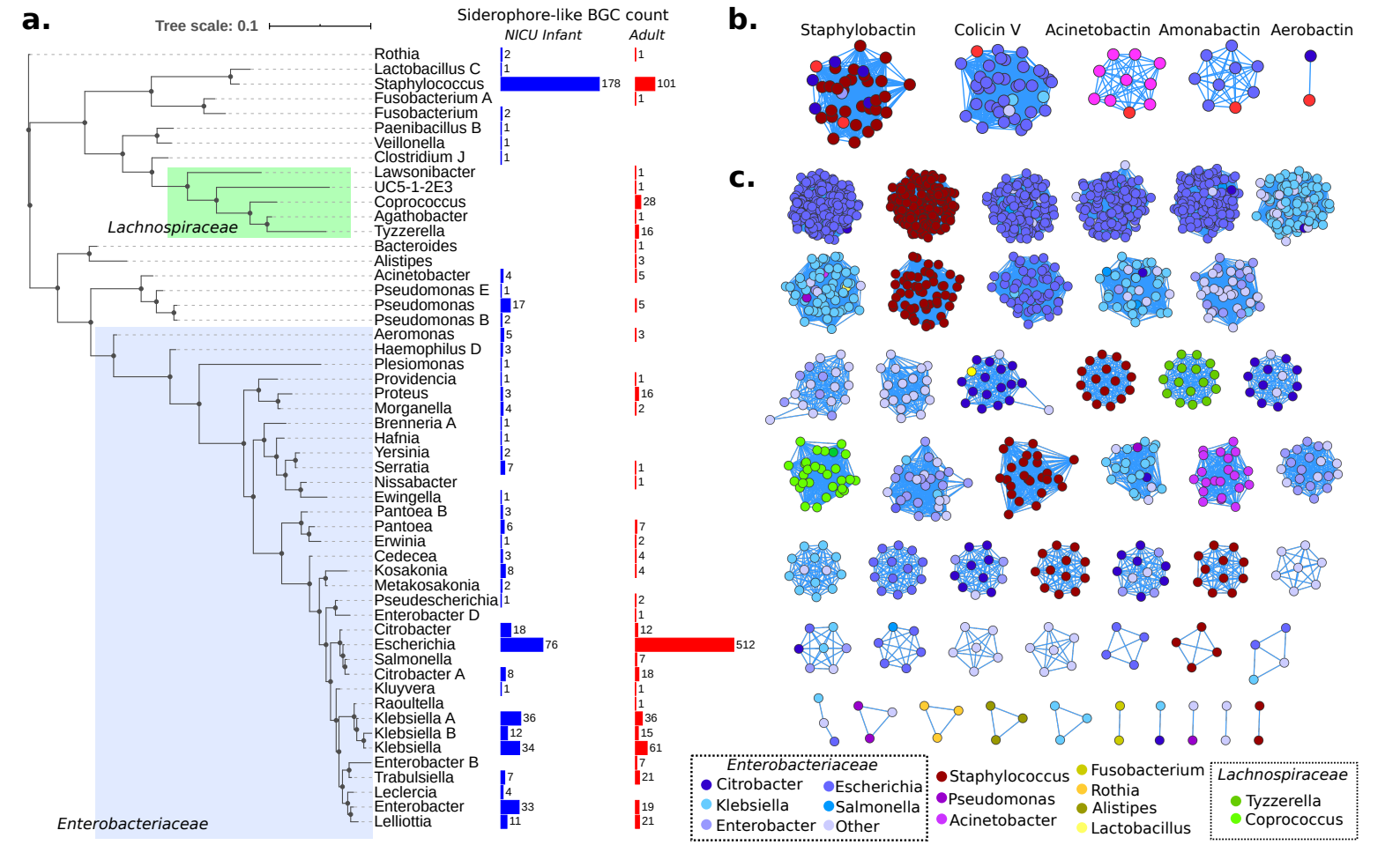

Figure 5: BGCs with siderophore-like transporters from human gut microbiomes. (a) Concatenated ribosomal protein tree (collapsed to the genus level) for high-quality genomes from the infant and adult gut microbiomes that encode siderophore-like BGCs. On the right are counts of siderophore BGCs from infant gut genomes (blue) and adult gut genomes (red). (b) Gene Cluster Families of BGCs containing known siderophores (bright red) and human-microbiome derived BGCs with siderophore-like transporters. BGCs are connected by similarity to other BGCs in the same gene cluster family, calculated using BiG-SCAPE. (c) Families of siderophore-like BGCs without any similarity to existing known BGCs. BGCs in the network are colored by the taxonomy of the genome of origin.

329 **Discussion**

330 We uncovered several strong associations between transporters within characterized
331 BGCs and features of the corresponding BGC-synthesized metabolites. With regards to
332 prediction of metabolite activity, we quantified the specificity of TonB-dependent receptors,
333 *FecCD*, and Periplasmic-binding protein 2 for siderophore-producing BGCs. This complements
334 existing literature indicating that genes in these families are specific for siderophore import in
335 both gram-positive and gram-negative bacteria (Chu et al. 2010). We also identified a putative
336 association between the MacB tripartite efflux pump and antibacterial/antifungal activity. Based
337 on these findings, a strategy of targeting novel BGCs containing MacB for characterization may
338 be useful for antibiotic prospecting. In addition to activity prediction, we used metabolite
339 structural information in MiBiG in order to predict metabolite molecular weight and lipophilicity
340 from BGC gene content. We discovered a strong relationship between the transporters in
341 characterized BGCs and the molecular weight of their synthesized metabolites. The strong
342 dichotomy between ATP-dependent transporters (utilizing the *ABC_tran* nucleotide binding
343 domain) and *MFS* family transporters points towards required ATP-dependence for transporting
344 metabolites larger than 1000 Da. We also identified relationships between two understudied
345 membrane components (*ABC2_Membrane_3* and *ABC2_Membrane_4*) and substrate log *P*,
346 possibly indicating tradeoff in membrane domains for molecules of different chemical properties.
347 Future phylogeny-based subdivision of these families may improve upon general protein family
348 annotations to increase the predictive power of transporter substrate characteristics.

349 There are multiple caveats to our work. Molecular activity of specialized metabolites
350 based on functions proven in the laboratory may be very different from the ecological roles that
351 metabolites play in natural settings (van der Meij et al. 2017; Kramer, Özkaya, and Kümmerli
352 2020; Behnsen and Raffatellu 2016). Further, many BGCs may produce multiple variants of a
353 metabolite (Fischbach and Clardy 2007), only some of which may be reported. There may also

355 be reference-database biases in our gene searches - while they are sensitive, it is possible that
356 phylogenetically divergent microbes use transporter genes that are not hit by our sequence
357 models. Finally, as reported, a significant proportion of BGCs contain no transporter at all or a
358 transporter gene genomically adjacent to a BGC may not be functionally linked. Despite these
359 caveats, it appears as though transporter genes provide simple and strong signals for inferring
360 both activity and chemical properties of metabolites produced by BGCs.

361 Siderophores are both considered critical pathogenicity factors for many human-
362 associated microbes (Weakland et al. 2020) and are also known to facilitate interactions with
363 other microbes and the innate immune system in the human gut microbiome (Behnsen and
364 Raffatellu 2016; Holden et al. 2016; Zhu et al. 2020; Lam et al. 2018). Therefore, being able to
365 annotate genes for the production of siderophores across diverse bacterial species may be critical
366 for understanding the distribution of virulence factors, yet it is difficult to do using traditional
367 annotation pipelines alone. We observed a high prevalence of siderophore-like BGCs in bacterial
368 genomes from NICU premature infant guts, suggesting that the premature infant gut could be
369 more prone to invasion by pathogens with siderophore virulence factors. Potentially novel
370 siderophore-like BGCs were most consistently found to be encoded in the genomes of members
371 of the *Enterobacteriaceae* and *Staphylococcus* in the premature infant microbiome. Only in the
372 adult microbiome datasets did we identify siderophore-like BGCs in the *Lachnospiraceae*, that
373 are often considered important commensals, indicating that commensal siderophore production
374 may be more common in developed gut microbiomes. Importantly, we identified siderophore-
375 like BGCs in these taxa that are not homologous to known siderophore clusters, indicating that
376 there is still substantial unknown chemical diversity of siderophores, even within well-studied
377 lineages.

378 In general, here we demonstrated that consideration of transporter genes can aid holistic
379 functional prediction of BGC products. A transporter guided approach could be especially useful

380 for identification of siderophore targets for medical (Nagoba and Vedpathak 2011) and
381 biotechnological applications(Ahmed and Holmström 2014). Given the large diversity of BGCs
382 and that chemical characterization of their products can be time and resource intensive, better
383 functional prediction of BGCs for targeted study can improve selection of targets for
384 antimicrobial discovery and downstream activity tests.

385 **Methods**

386 Curation and selection of BGCs and transporter annotations

387 We parsed the MIBiG 2.0 database of biosynthetic gene clusters metadata and extracted
388 information including host genus, compound count, chemical structures of the metabolite
389 product, known metabolite activities, and the number of open reading frames for each BGC.
390 Using Entrez and NCBI, we assigned the expected Gram status for each BGC based on phylum,
391 coded 0=Gram-negative, 1=Gram-positive, 2=Fungal, 3=other. For the purpose of this
392 manuscript, we only analyzed BGCs from Gram-positive and Gram-negative bacteria. We
393 noticed that the activity labels in MIBiG 2.0 were often incomplete, and manually added a set of
394 antibacterial, siderophore, and antifungal labels derived from the literature (**Supplementary**
395 **Table S1**). We found that 28 BGCs (1.8%) in MIBiG were unusually large in length, and
396 comparisons to published papers on these BGCs showed that their MIBiG counterparts were
397 overextended in comparison to the validated BGC. For this reason, we eliminated the 28 BGCs
398 over 60 ORFs in length. Using Python and RDKit, we calculated molecular weights and partition
399 coefficients (log *P*) using the algorithm described in (Wildman and Crippen 1999) for all 1,042
400 MIBiG BGCs with a single associated compound structure. 238 BGCs have more than 1
401 associated metabolite structure, and these multi-structure BGCs were not used in our structural
402 association analyses. We annotated biosynthetic genes in MIBiG with HMMER *hmmsearch*

404 *(Eddy 1998)* and *cath-resolve-hits (Lewis, Sillitoe, and Lees 2019)* on a set of the 99 most
405 commonly represented biosynthetic PFAMs in antiSMASH BGCs obtained from (Cimermancic
406 et al. 2014), to generate a counts table of biosynthetic protein families for each BGC
407 (**Supplementary Table S4**).
408 To obtain a comprehensive overview of the distribution of transport-associated protein
409 domains in biosynthetic gene clusters, we generated two separate feature tables: (a) using
410 CATHDB (Sillitoe et al. 2019) HMMs and (b) using PFAM ("PFAM Database," n.d.) HMMs.
411 For the first, we downloaded all proteins in the Transporter Classification Database (TCDB)
412 (Saier et al. 2016) and annotated them with *hmmsearch* and *cath-resolve-hits*, using all
413 CATHDB Functional Family HMMs. We then selected all CATHDB HMMs which were
414 represented at least 5 times. We then manually curated this list down to 180 final CATHDB
415 HMMs that were transport-specific. We then calculated the specificities of each CATHDB
416 HMM for TCDB families; 80 were specific for exactly one TCDB family.
417 For the second set of features, we took all protein sequences with an annotation including
418 "Transport" in the antiSMASH database v2 (Blin, Pascal Andreu, et al. 2019) and annotated
419 these proteins with the Pfam-A set of HMMs using *hmmsearch* and the option --cut_ga. We then
420 selected highly represented HMMs and manually curated this list to be transporter-specific and
421 representative of the major transporter classes in TCDB. We then also selected the PFAM
422 Substrate Binding Protein and Periplasmic Binding Protein HMMs that were represented more
423 than 5 times in MIBiG. When comparing both the PFAM and CATHDB set of HMMs, we found
424 substantial overlap, but the CATHDB set is composed of 166 domain features while the PFAM
425 set only contains 18.
426 <u>Machine Learning to predict metabolite structural and functional characteristics</u>
427 To identify associations between metabolite functional classes and structural properties
428 with BGC gene content, we used traditional statistical tests and different machine learning

429 models. Our classification tasks were (a) siderophore (n=16 gram-positive; n=24 gram-negative)
430 vs other activity (n=142 gram-positive; n=42 gram-negative), (b) antibiotic and antifungal
431 (n=131 gram-positive; n=27 gram-negative) vs other activity (n=57 gram-positive; n=37 gram-
432 negative), (c) metabolite molecular weight > 1000 Daltons (n=149) vs metabolite molecular
433 weight < 1000 Daltons (n=421), and (d) predicted partition coefficient log P <0 (n=220) vs log P
434 >= 0 (n=350). For the functional classification tasks, we noticed a strong class imbalance with
435 gram-status, so we performed functional classification separately for BGCs from gram-positive
436 and gram-negative bacteria. We first tested for univariate differences in proportional
437 representation of each transporter BGC between classes for classification tests using Fisher's
438 Exact Test in Python (q<0.05, Benjamini-Hochberg (Hochberg and Benjamini 1990) correction).
439 We then assessed the predictive power of the three sets of features for each BGC:
440 transporter PFAM HMMs, transporter CATHDB HMMs, and biosynthetic PFAM HMMs.
441 Features were counts of protein families, which were standardized using the StandardScaler
442 function in the scikit-learn package. Given the nature of our study, we used simple models to
443 ensure reliability and interpretability of our results. We fit two classes of machine learning
444 models (a) LASSO-penalized logistic regression, which fits a linear model with a sparsity
445 penalty on weights, and (b) shallow decision trees (of depth one or two), which can classify
446 based on splitting at most two features (Franklin 2005). All models were trained using the
447 Python package scikit-learn. Due to data size and class imbalance, we fit models using repeated,
448 stratified k-fold cross-validation ("RepeatedStratifiedKFold" in scikit-learn) with five repeats
449 and five folds. On each cross-validation split of our data, we computed the area under the
450 precision-recall curve (AUPRC) to evaluate performance for our class-imbalanced tasks. Thus
451 for the final output of this procedure we reported the mean of accuracy, precision, recall, and
452 AUPRC each generated from five repeats of different random 5-fold partitions of the data. We

452 further use these repeated cross-validation splits to see which features are consistently used for
453 classification across repeats.

454 Annotating metagenomic siderophore-like BGCs from the human microbiome

455 To assess the distribution of siderophore-like BGCs in the human microbiome, we
456 downloaded two sets of genomes assembled from metagenomes obtained from the human gut
457 microbiome: 2,425 genomes from a neonatal intensive care unit (NICU) premature infant
458 microbiome (Olm et al. 2019) and 24,345 genomes from a diverse set of mostly adult human
459 cohorts (Nayfach et al. 2019). We ran antiSMASH 5.0 on these genomes and then scanned
460 predicted BGCs for at least two of the PFAMs that were found to be specific for siderophore
461 BGCs (*FecCD*, *Peripla_BP_2*, and *Ton_dep_Rec*). We dereplicated these BGCs and compared
462 them to known MIBiG siderophore BGCs using the software BiG-SCAPE (Navarro-Muñoz et al.
463 2020) run with default settings. We then considered BGCs with either set of hits and reported
464 their genomic taxonomic distribution based on the closest BLAST hit representatives of genomic
465 ribosomal proteins to taxonomic genera defined by GTDB (Parks et al. 2018) (minimum percent
466 identity of hits >80%).

467 Acknowledgements

468 This research was supported, in part, by the NIH under awards RAI092531A, R01-GM109454,
469 and R35-GM134922. J.F.B. and Y.S.S. are Chan Zuckerberg Biohub Investigators.

470 Data Access

471 All Python code used in this paper, along with the data analyzed, antiSMASH BGCs, and
472 additional data tables, is available at https://github.com/nickbhat/bgc_tran.

473 **Supplementary Figure Legends**

474 **S1. Heatmap of PFAM transporter domain abundance in all annotated siderophores in our**
475 **modified version of the MIBiG database.**

476 **S2. Distributions of and LASSO logistic regression feature importances of transporter and**
477 **biosynthetic gene features for prediction of siderophore activity.**

478 **S3. Distributions of and LASSO logistic regression feature importances of transporter and**
479 **biosynthetic gene features for prediction of antibacterial activity.**
480
481 **S4. LASSO logistic regression feature importances of transporter and biosynthetic gene**
482 **features for prediction of molecular weights >1000 Da.**

483 **S5. Precision-recall curves and LASSO logistic regression feature importances of**
484 **transporter and biosynthetic gene features for prediction of log P > 0.**

485 **Supplementary Table Legends**
486 **S1. Metadata for all MIBiG BGCs included in this study, including additional hand**
487 **curated labels for data.**

488 **S2. Abundances of PFAM transporter domains across all MIBiG BGCs.**

489 **S3. Abundances of CATHDB transporter domains across all MIBiG BGCs.**

490 **S4. Abundances of PFAM biosynthetic domains across all MIBiG BGCs.**

491 **S5. Classification metrics (precision, recall, accuracy, and AUPRC) for all classification**
492 **tasks in this study.**

493 **S6. Siderophore-like Biosynthetic Gene Clusters identified in the human microbiome by**
494 **this study.**

References

Ahmed, E., and S. J. M. Holmström. 2014. "Siderophores in Environmental Research: Roles and Applications." *Microbial Biotechnology* 7 (3): 196–208.

Arnott, John A., and Sonia Lobo Planey. 2012. "The Influence of Lipophilicity in Drug Discovery and Design." *Expert Opinion on Drug Discovery* 7 (10): 863–75.

Beek, Josy Ter, Josy ter Beek, Albert Guskov, and Dirk Jan Slotboom. 2014. "Structural Diversity of ABC Transporters." *The Journal of General Physiology*. https://doi.org/10.1085/jgp.201411164.

Behnsen, Judith, and Manuela Raffatellu. 2016. "Siderophores: More than Stealing Iron." *mBio*. https://doi.org/10.1128/mBio.01906-16.

Berntsson, Ronnie P-A, Ronnie P. -A. Berntsson, Sander H. J. Smits, Lutz Schmitt, Dirk-Jan Slotboom, and Bert Poolman. 2010. "A Structural Classification of Substrate-Binding Proteins." *FEBS Letters*. https://doi.org/10.1016/j.febslet.2010.04.043.

Bi, Yunchen, Evan Mann, Chris Whitfield, and Jochen Zimmer. 2018. "Architecture of a Channel-Forming O-Antigen Polysaccharide ABC Transporter." *Nature* 553 (7688): 361–65.

Blin, Kai, Victòria Pascal Andreu, Emmanuel L. C. de Los Santos, Francesco Del Carratore, Sang Yup Lee, Marnix H. Medema, and Tilmann Weber. 2019. "The antiSMASH Database Version 2: A Comprehensive Resource on Secondary Metabolite Biosynthetic Gene Clusters." *Nucleic Acids Research* 47 (D1): D625–30.

Blin, Kai, Simon Shaw, Katharina Steinke, Rasmus Villebro, Nadine Ziemert, Sang Yup Lee, Marnix H. Medema, and Tilmann Weber. 2019. "antiSMASH 5.0: Updates to the Secondary Metabolite Genome Mining Pipeline." *Nucleic Acids Research* 47 (W1): W81–87.

Carran, C. J., M. Jordan, H. Drechsel, D. G. Schmid, and G. Winkelmann. 2001. "Heterobactins: A New Class of Siderophores from Rhodococcus Erythropolis IGTS8 Containing Both Hydroxamate and Catecholate Donor Groups." *Biometals: An International Journal on the Role of Metal Ions in Biology, Biochemistry, and Medicine* 14 (2): 119–25.

Chen, Liang, Justin E. Wilson, Mark J. Koenigsknecht, Wei-Chun Chou, Stephanie A. Montgomery, Agnieszka D. Truax, W. June Brickey, et al. 2017. "NLRP12 Attenuates Colon Inflammation by Maintaining Colonic Microbial Diversity and Promoting Protective Commensal Bacterial Growth." *Nature Immunology* 18 (5): 541–51.

Chen, Yunqiu, Michelle Unger, Ioanna Ntai, Ryan A. McClure, Jessica C. Albright, Regan J. Thomson, and Neil L. Kelleher. 2013. "Gobichelin A and B: Mixed-Ligandsiderophores Discovered Using Proteomics." *MedChemComm*. https://doi.org/10.1039/c2md20232h.

Chevrette, Marc G., Karina Gutiérrez-García, Nelly Selem-Mojica, César Aguilar-Martínez, Alan Yañez-Olvera, Hilda E. Ramos-Aboites, Paul A. Hoskisson, and Francisco Barona-Gómez. 2019. "Evolutionary Dynamics of Natural Product Biosynthesis in Bacteria." *Natural Product Reports*, December. https://doi.org/10.1039/c9np00048h.

Chu, Byron C., Alicia Garcia-Herrero, Ted H. Johanson, Karla D. Krewulak, Cheryl K. Lau, R. Sean Peacock, Zoya Slavinskaya, and Hans J. Vogel. 2010. "Siderophore Uptake in Bacteria and the Battle for Iron with the Host; a Bird's Eye View." *Biometals: An International Journal on the Role of Metal Ions in Biology, Biochemistry, and Medicine* 23 (4): 601–11.

Cimermancic, Peter, Marnix H. Medema, Jan Claesen, Kenji Kurita, Laura C. Wieland Brown, Konstantinos Mavrommatis, Amrita Pati, et al. 2014. "Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters." *Cell* 158 (2): 412–21.

22

542 Crits-Christoph, Alexander, Spencer Diamond, Cristina N. Butterfield, Brian C. Thomas, and
543 Jillian F. Banfield. 2018. "Novel Soil Bacteria Possess Diverse Genes for Secondary
544 Metabolite Biosynthesis." *Nature* 558 (7710): 440–44.
545 Davies, Julian. 2013. "Specialized Microbial Metabolites: Functions and Origins." *The Journal*
546 *of Antibiotics* 66 (7): 361–64.
547 Dawson, Roger J. P., and Kaspar P. Locher. 2007. "Structure of the Multidrug ABC Transporter
548 Sav1866 fromStaphylococcus Aureusin Complex with AMP-PNP." *FEBS Letters*.
549 https://doi.org/10.1016/j.febslet.2007.01.073.
550 Duvallet, Claire, Sean M. Gibbons, Thomas Gurry, Rafael A. Irizarry, and Eric J. Alm. 2017.
551 "Meta-Analysis of Gut Microbiome Studies Identifies Disease-Specific and Shared
552 Responses." *Nature Communications* 8 (1): 1–10.
553 Eddy, S. R. 1998. "Profile Hidden Markov Models." *Bioinformatics*.
554 https://doi.org/10.1093/bioinformatics/14.9.755.
555 Ellermann, Melissa, and Janelle C. Arthur. 2017. "Siderophore-Mediated Iron Acquisition and
556 Modulation of Host-Bacterial Interactions." *Free Radical Biology & Medicine* 105 (April):
557 68–78.
558 Fischbach, Michael A., and Jon Clardy. 2007. "One Pathway, Many Products." *Nature Chemical*
559 *Biology*. https://doi.org/10.1038/nchembio0707-353.
560 Fischbach, Michael A., and Christopher T. Walsh. 2006. "Assembly-Line Enzymology for
561 Polyketide and Nonribosomal Peptide Antibiotics: Logic, Machinery, and Mechanisms."
562 *Chemical Reviews* 106 (8): 3468–96.
563 Fischbach, Michael A., Christopher T. Walsh, and Jon Clardy. 2008. "The Evolution of Gene
564 Collectives: How Natural Selection Drives Chemical Innovation." *Proceedings of the*
565 *National Academy of Sciences of the United States of America* 105 (12): 4601–8.
566 Franklin, James. 2005. "The Elements of Statistical Learning: Data Mining, Inference and
567 Prediction." *The Mathematical Intelligencer*. https://doi.org/10.1007/bf02985802.
568 Gebhard, Susanne. 2012. "ABC Transporters of Antimicrobial Peptides in Firmicutes Bacteria -
569 Phylogeny, Function and Regulation." *Molecular Microbiology* 86 (6): 1295–1317.
570 Greene, Nicholas P., Elise Kaplan, Allister Crow, and Vassilis Koronakis. 2018a. "Antibiotic
571 Resistance Mediated by the MacB ABC Transporter Family: A Structural and Functional
572 Perspective." *Frontiers in Microbiology* 9. https://doi.org/10.3389/fmicb.2018.00950.
573 ———. 2018b. "Corrigendum: Antibiotic Resistance Mediated by the MacB ABC Transporter
574 Family: A Structural and Functional Perspective." *Frontiers in Microbiology* 9
575 (September): 2318.
576 Hannigan, Geoffrey D., David Prihoda, Andrej Palicka, Jindrich Soukup, Ondrej Klempir, Lena
577 Rampula, Jindrich Durcak, et al. 2019. "A Deep Learning Genome-Mining Strategy for
578 Biosynthetic Gene Cluster Prediction." *Nucleic Acids Research* 47 (18): e110.
579 Hider, Robert C., and Xiaole Kong. 2010. "Chemistry and Biology of Siderophores." *Natural*
580 *Product Reports*. https://doi.org/10.1039/b906679a.
581 Hochberg, Yosef, and Yoav Benjamini. 1990. "More Powerful Procedures for Multiple
582 Significance Testing." *Statistics in Medicine*. https://doi.org/10.1002/sim.4780090710.
583 Holden, Victoria I., Paul Breen, Sébastien Houle, Charles M. Dozois, and Michael A. Bachman.
584 2016. "Klebsiella Pneumoniae Siderophores Induce Inflammation, Bacterial Dissemination,
585 and HIF-1α Stabilization during Pneumonia." *mBio* 7 (5).
586 https://doi.org/10.1128/mBio.01397-16.
587 Jenke-Kodama, Holger, Thomas Börner, and Elke Dittmann. 2006. "Natural Biocombinatorics in
588 the Polyketide Synthase Genes of the Actinobacterium Streptomyces Avermitilis." *PLoS*
589 *Computational Biology* 2 (10): e132.
590 Kautsar, Satria A., Kai Blin, Simon Shaw, Jorge C. Navarro-Muñoz, Barbara R. Terlouw, Justin

J. J. van der Hooft, Jeffrey A. van Santen, et al. 2020. "MIBiG 2.0: A Repository for Biosynthetic Gene Clusters of Known Function." *Nucleic Acids Research* 48 (D1): D454–58.

Kim, Hyun Uk, Kai Blin, Sang Yup Lee, and Tilmann Weber. 2017. "Recent Development of Computational Resources for New Antibiotics Discovery." *Current Opinion in Microbiology* 39 (October): 113–20.

Kramer, Jos, Özhan Özkaya, and Rolf Kümmerli. 2020. "Bacterial Siderophores in Community and Host Interactions." *Nature Reviews Microbiology*. https://doi.org/10.1038/s41579-019-0284-4.

Lam, Margaret M. C., Kelly L. Wyres, Louise M. Judd, Ryan R. Wick, Adam Jenney, Sylvain Brisse, and Kathryn E. Holt. 2018. "Tracking Key Virulence Loci Encoding Aerobactin and Salmochelin Siderophore Synthesis in Klebsiella Pneumoniae." *Genome Medicine* 10 (1): 77.

Lewis, T. E., I. Sillitoe, and J. G. Lees. 2019. "Cath-Resolve-Hits: A New Tool That Resolves Domain Matches Suspiciously Quickly." *Bioinformatics* 35 (10): 1766–67.

Liu, Yang, Meng Li, Huiyan Mu, Shuting Song, Ying Zhang, Kun Chen, Xihong He, et al. 2017. "Identification and Characterization of the Ficellomycin Biosynthesis Gene Cluster from Streptomyces Ficellus." *Applied Microbiology and Biotechnology* 101 (20): 7589–7602.

Martín, Juan F., Javier Casqueiro, and Paloma Liras. 2005. "Secretion Systems for Secondary Metabolites: How Producer Cells Send out Messages of Intercellular Communication." *Current Opinion in Microbiology* 8 (3): 282–93.

Matsuo, Yoshihide, Kaneo Kanoh, Jae-Hyuk Jang, Kyoko Adachi, Satoru Matsuda, Osamu Miki, Toshiaki Kato, and Yoshikazu Shizuri. 2011. "Streptobactin, a Tricatechol-Type Siderophore from Marine-Derived Streptomyces Sp. YM5-799." *Journal of Natural Products* 74 (11): 2371–76.

Medema, Marnix H., and Michael A. Fischbach. 2015. "Computational Approaches to Natural Product Discovery." *Nature Chemical Biology*. https://doi.org/10.1038/nchembio.1884.

Meij, Anne van der, Sarah F. Worsley, Matthew I. Hutchings, and Gilles P. van Wezel. 2017. "Chemical Ecology of Antibiotic Production by Actinomycetes." *FEMS Microbiology Reviews* 41 (3): 392–416.

Méndez, C., and J. A. Salas. 2001. "The Role of ABC Transporters in Antibiotic-Producing Organisms: Drug Secretion and Resistance Mechanisms." *Research in Microbiology* 152 (3-4): 341–50.

Nagoba, Basavraj, and Deepak Vedpathak. 2011. "Medical Applications of Siderophores." *Electronic Journal of General Medicine*. https://doi.org/10.29333/ejgm/82743.

Navarro-Muñoz, Jorge C., Nelly Selem-Mojica, Michael W. Mullowney, Satria A. Kautsar, James H. Tryon, Elizabeth I. Parkinson, Emmanuel L. C. De Los Santos, et al. 2020. "A Computational Framework to Explore Large-Scale Biosynthetic Diversity." *Nature Chemical Biology* 16 (1): 60–68.

Nayfach, Stephen, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard, and Nikos C. Kyrpides. 2019. "New Insights from Uncultivated Genomes of the Global Human Gut Microbiome." *Nature* 568 (7753): 505–10.

Olm, Matthew R., Nicholas Bhattacharya, Alexander Crits-Christoph, Brian A. Firek, Robyn Baker, Yun S. Song, Michael J. Morowitz, and Jillian F. Banfield. 2019. "Necrotizing Enterocolitis Is Preceded by Increased Gut Bacterial Replication, Klebsiella, and Fimbriae-Encoding Bacteria." *Science Advances*. https://doi.org/10.1126/sciadv.aax5727.

Osbourn, Anne. 2010. "Secondary Metabolic Gene Clusters: Evolutionary Toolkits for Chemical Innovation." *Trends in Genetics: TIG* 26 (10): 449–57.

Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski,

Pierre-Alain Chaumeil, and Philip Hugenholtz. 2018. "A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life." *Nature Biotechnology* 36 (10): 996–1004.

"PFAM Database." n.d. *SpringerReference*. https://doi.org/10.1007/springerreference_36286.

Quistgaard, Esben M., Christian Löw, Fatma Guettou, and Pär Nordlund. 2016. "Understanding Transport by the Major Facilitator Superfamily (MFS): Structures Pave the Way." *Nature Reviews. Molecular Cell Biology* 17 (2): 123–32.

Rascher, Andreas, Zhihao Hu, Greg O. Buchanan, Ralph Reid, and C. Richard Hutchinson. 2005. "Insights into the Biosynthesis of the Benzoquinone Ansamycins Geldanamycin and Herbimycin, Obtained by Gene Sequencing and Disruption." *Applied and Environmental Microbiology* 71 (8): 4862–71.

Rees, Douglas C., Eric Johnson, and Oded Lewinson. 2009. "ABC Transporters: The Power to Change." *Nature Reviews. Molecular Cell Biology* 10 (3): 218–27.

Rempel, S., C. Gati, M. Nijland, C. Thangaratnarajah, A. Karyolaimos, J. W. de Gier, A. Guskov, and D. J. Slotboom. 2020. "A Mycobacterial ABC Transporter Mediates the Uptake of Hydrophilic Compounds." *Nature* 580 (7803): 409–12.

Romano, Maria, Giuliana Fusco, Hassanul G. Choudhury, Shahid Mehmood, Carol V. Robinson, Séverine Zirah, Julian D. Hegemann, et al. 2018. "Structural Basis for Natural Product Selection and Export by Bacterial ABC Transporters." *ACS Chemical Biology* 13 (6): 1598–1609.

Saier, Milton H., Vamsee S. Reddy, Brian V. Tsu, Muhammad Saad Ahmed, Chun Li, and Gabriel Moreno-Hagelsieb. 2016. "The Transporter Classification Database (TCDB): Recent Advances." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv1103.

Severi, Emmanuele, and Gavin H. Thomas. 2019. "Antibiotic Export: Transporters Involved in the Final Step of Natural Product Production." *Microbiology* 165 (8): 805–18.

Sharon, Gil, Neha Garg, Justine Debelius, Rob Knight, Pieter C. Dorrestein, and Sarkis K. Mazmanian. 2014. "Specialized Metabolites from the Microbiome in Health and Disease." *Cell Metabolism* 20 (5): 719–30.

Sillitoe, Ian, Natalie Dawson, Tony E. Lewis, Sayoni Das, Jonathan G. Lees, Paul Ashford, Adeyelu Tolulope, et al. 2019. "CATH: Expanding the Horizons of Structure-Based Functional Annotations for Genome Sequences." *Nucleic Acids Research* 47 (D1): D280–84.

Skinnider, Michael A., Nishanth J. Merwin, Chad W. Johnston, and Nathan A. Magarvey. 2017. "PRISM 3: Expanded Prediction of Natural Product Chemical Structures from Microbial Genomes." *Nucleic Acids Research* 45 (W1): W49–54.

Staudenmaier, H., B. Van Hove, Z. Yaraghi, and V. Braun. 1989. "Nucleotide Sequences of the fecBCDE Genes and Locations of the Proteins Suggest a Periplasmic-Binding-Protein-Dependent Transport Mechanism for iron(III) Dicitrate in Escherichia Coli." *Journal of Bacteriology* 171 (5): 2626–33.

Tran, Phuong Nguyen, Ming-Ren Yen, Chen-Yu Chiang, Hsiao-Ching Lin, and Pao-Yang Chen. 2019. "Detecting and Prioritizing Biosynthetic Gene Clusters for Bioactive Compounds in Bacteria and Fungi." *Applied Microbiology and Biotechnology* 103 (8): 3277–87.

Tyc, Olaf, Chunxu Song, Jeroen S. Dickschat, Michiel Vos, and Paolina Garbeva. 2017. "The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria." *Trends in Microbiology* 25 (4): 280–92.

Velamakanni, Saroj, Yao Yao, Daniel A. P. Gutmann, and Hendrik W. van Veen. 2008. "Multidrug Transport by the ABC Transporter Sav1866 from Staphylococcus Aureus." *Biochemistry* 47 (35): 9300–9308.

Wang, Zhiming, Wenxin Hu, and Hongjin Zheng. 2020. "Pathogenic Siderophore ABC Importer

690 YbtPQ Adopts a Surprising Fold of Exporter." *Science Advances* 6 (6): eaay7997.
691 Weakland, Danelle R., Sara N. Smith, Bailey Bell, Ashootosh Tripathi, and Harry L. T. Mobley.
692 2020. "The Serratia Marcescens Siderophore, Serratiochelin, Is Necessary for Full
693 Virulence during Bloodstream Infection." *Infection and Immunity*.
694 https://doi.org/10.1128/iai.00117-20.
695 Wenzel, Silke C., and Rolf Müller. 2005. "Formation of Novel Secondary Metabolites by
696 Bacterial Multimodular Assembly Lines: Deviations from Textbook Biosynthetic Logic."
697 *Current Opinion in Chemical Biology* 9 (5): 447–58.
698 Wildman, Scott A., and Gordon M. Crippen. 1999. "Prediction of Physicochemical Parameters
699 by Atomic Contributions." *Journal of Chemical Information and Computer Sciences*.
700 https://doi.org/10.1021/ci990307l.
701 Yan, Yan, Nicholas Liu, and Yi Tang. 2020. "Recent Developments in Self-Resistance Gene
702 Directed Natural Product Discovery." *Natural Product Reports*, January.
703 https://doi.org/10.1039/c9np00050j.
704 Zhu, Wenhan, Maria G. Winter, Luisella Spiga, Elizabeth R. Hughes, Rachael Chanin, Aditi
705 Mulgaonkar, Jenelle Pennington, et al. 2020. "Xenosiderophore Utilization Promotes
706 Bacteroides Thetaiotaomicron Resilience during Colitis." *Cell Host & Microbe* 27 (3): 376–
707 88.e8.